The reliability and validity of adverse-event measures of the quality of healthcare

by

Kieran Walshe

A thesis submitted to
the Faculty of Commerce and Social Science
of the University of Birmingham
for the degree of
Doctor of Philosophy

Health Services Management Centre
School of Public Policy
Faculty of Commerce and Social Science
The University of Birmingham
May 1998.

Abstract

The quality of healthcare is increasingly the subject of scrutiny by a range of stakeholders, including healthcare provider organisations, health professionals and their representative bodies, healthcare purchasers and funders, policy makers and national governments, patients and users of health services. The use of a variety of quality measures has become widespread in the healthcare systems of many developed countries, including the United Kingdom. The twin tasks of measuring and improving the quality of care - often termed quality assurance - have been addressed by new arrangements for professional accountability, new approaches to managing and comparing organisational performance, and new statutory and legal mechanisms.

Adverse events in healthcare, which can be loosely defined as instances which indicate or may indicate that a patient has received poor quality care, offer an important opportunity for quality measurement and improvement. There is extensive evidence that adverse events are relatively common, that they can have serious and lasting impacts on patients, and that they represent a considerable cost to healthcare organisations. Equally importantly, evidence in healthcare and experience in other sectors suggests that adverse events offer an important insight into the strengths and weaknesses of healthcare processes, and an invaluable opportunity to bring about improvements in the quality of care.

Adverse events have been used quite widely, particularly in the United States of America, as the basis of a number of measures of the quality of healthcare. However, these measures have rarely been developed and tested rigorously before they have entered widespread usage, and there has been considerable debate about their advantages and disadvantages.

A series of empirical studies were undertaken, using data collected through the use of adverse-event measures of quality in a British acute hospital, aimed at investigating the validity and reliability of those measures. The results showed that the adverse-event measures being tested had moderate to good face, content and construct validity. Although their validity was capable of improvement, it was still clear that they were measuring meaningful and important dimensions of the quality of healthcare. However, the reliability of the measures being tested was more mixed. While experimental studies of interrater and intrarater reliability indicated that they had moderate to good reliability (though, again, it was capable of improvement) observational studies suggested that the reliability in actual use might be lower than that found during testing.

This research concludes that adverse-event measures of quality are important measures of the quality of healthcare, which should be used in healthcare quality assurance with two main provisos. Firstly, the development of measures should be more rigorous, and should pay more attention to both validity and reliability issues. Secondly, the routine use of such measures should incorporate some element of ongoing reliability testing, in order to ensure that good reliability is maintained.

My thanks to all those who helped me in the research reported in this thesis, particularly to Jenny Bennett, James Coles and David Ingram whose support for and involvement in the occurrence screening project at the Royal Sussex County Hospital was invaluable; to Cathy Callaghan, Daphne Fox, Becky Reynolds and Tracy Strickland who worked on the project; and to all the clinicians who took part.

I would not have completed this thesis without the patient advice and continuing support of my supervisor, Penelope Mullen, for which I am very grateful. I am also indebted to Catherine, Siobhan, Ruth and Michael for putting up with the invasion of family life it has entailed.

Outline Table of Contents

Complete Table of Contents

Appendices

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Adverse events can be loosely defined as "instances which indicate or may indicate that a patient has received poor quality care" (Walshe, Bennett and Ingram 1995). They are events which exhibit three key characteristics - negativity, patient impact, and healthcare process causation. Firstly, they are circumstances or happenings which are, by their very nature, undesirable, and which we would prefer not to occur if that were possible. Secondly, they are occurrences which have, or may have, some impact on the care that patients receive and so on their health status, and which could result in additional morbidity or mortality. Thirdly, they are events which result not from the diseases or conditions of patients, nor from patients' actions or behaviour, but from the healthcare process - the way that healthcare is organised, managed and delivered, the decisions that are taken by clinicians and others, and the acts of omission or commission of both the healthcare organisation and the individuals who make up that organisation.

There is a wealth of evidence to suggest that adverse events are not infrequent occurrences in the modern healthcare systems of the developed world. The incidence of adverse events found in empirical studies depends on the definition of the term which they adopt and the data sources and methods they use to identify adverse events. However, even the most conservative studies, which limit themselves to the consideration of events in which patients have clearly suffered some significant harm as a direct result of the healthcare process, have estimated that about 1 in 25 inpatients suffers such an adverse event during their stay in hospital (Mills 1978; Harvard Medical Practice Study 1990; and others). Other evidence suggests that up to 5% of hospital admissions may actually be the result of previous adverse events, either during earlier admissions or in outpatient or community care, which would make adverse events perhaps the single most common reason for admission to hospital (Lakshmanan, Hershey and Breslau 1980).

Adverse events represent a substantial burden, in both financial and non-financial terms, for healthcare organisations and for patients. The only published study which has followed up patients who have suffered adverse events found that while 76% of patients had returned to normal health status within 6 months and suffered little or no economic loss, a substantial minority of patients experienced longer term morbidity or reduced life expectancy with the usual economic consequences for them and their families (Harvard Medical Practice Study 1990). Although it is difficult to unpick the effects of adverse events from the underlying disease processes which had cause patients to enter the healthcare system in the first place, it seems clear that adverse events resulted in substantial pain, suffering and financial loss for patients.

The cost of adverse events for healthcare organisations is equally large but difficult to estimate. If, as has been suggested above, 1 in 25 inpatients has an adverse event and up to 5% of inpatient admissions result from some form of earlier adverse event, then the costs of such events to healthcare organisations are huge, and might amount to billions of pounds across the British National Health Service, but such an estimate should be treated with considerable caution. More specifically, the costs to the NHS of clinical negligence litigation can be argued to be a direct consequence of some adverse events. In the NHS, these costs have been estimated at about £125 million pa, though they are rising rapidly (Dingwall and Fenn 1995), while in the USA it has been suggested that negligence litigation costs the healthcare system about $5 billion pa (Relman 1989). Moreover, it has been demonstrated that since only a tiny proportion of adverse events result in clinical negligence litigation, the potential costs of such litigation are much higher than these actual costs (Localio et al 1991).

But perhaps the most persuasive evidence that adverse events in healthcare are worthy of study comes from a range of sources, both inside and outside healthcare, which suggest that many adverse events are avoidable and so could be prevented (Craddick 1979; Goldman and Walder 1992; and others), and that adverse events offer an invaluable insight into the strengths and weaknesses of healthcare processes (Deming 1986; Reason 1995; and others). In short, adverse events both indicate that the quality of care is capable of improvement, and provide information and understanding crucial to making such improvements happen.

Improving quality has become a central concern in healthcare systems throughout the developed world over recent years, through changes in patient and user expectations, shifts in attitudes to professionals and public services, pressure on the mechanisms and levels of funding for healthcare, and the pace of technological advance in medicine. As many healthcare systems have developed systems for attempting to measure, monitor and improve the quality of care, adverse events have often been used in those approaches to measurement. But while the use of adverse-event measures of the quality of healthcare has become increasingly widespread, their worth as measures has not been widely researched, and some commentators have expressed fundamental concerns about their validity, reliability and utility in quality measurement and improvement (Goldman 1989; Sanazaro and Mills 1991).

1.2 **Objectives of this research**

Although there is an extensive literature on quality measurement in healthcare, and a substantial body of published work related to the characteristics of adverse events in healthcare, it was noted above that the scientific worth of measures of quality which are based on adverse events has not been much researched. While a whole range of aspects or characteristics of such measures clearly need to be studied (as the evaluation framework set out in chapter 3 makes evident), it was felt from the review of the literature reported in chapter 4 that the most fundamental and urgent concerns related to the validity and reliability of adverse-event measures of the quality of care. There were, therefore, two main research questions or hypotheses which the research was intended to address:

a) Can information about adverse events in healthcare be used to provide valid measures of the quality of care? The meaning of validity in this context is discussed in more detail in chapter 2, but in broad terms the concern was to establish whether such measures had meaning for those who might use them, offered insight into and understanding of the healthcare processes they measured, and provided a basis for subsequent quality improvement activities.

b) Can adverse-event measures of the quality of healthcare be developed which are sufficiently reliable for wider use? Again, the meaning of reliability is discussed further in chapter 2, but here the main issue was the extent to which these measures could be used in "real-world" settings with the usual constraints of time and resources, and provide data which was appropriately consistent across different users, applications and subjects of the measures.

With these objectives in mind, a number of separate but related studies were designed and undertaken, each focused on different aspects of validity and reliability:

a) The face and content validity of an adverse-event measure were assessed through a questionnaire study of clinician opinion, drawing on both practising clinicians and public health physicians, and through an interview study involving a small group of clinicians involved in directing clinical audit activities in healthcare organisations.

b) The construct validity of some adverse-event measures of quality was explored by testing whether a number of constructs or theories about the incidence and characteristics of adverse events (some of which had been previously tested in studies published in the literature) were supported by data drawn from the use of these measures in a British acute hospital.

c) The reliability of some adverse-event measures of quality was assessed through a series of experimental studies in which the measures were applied repeatedly to a sample of patient admissions in order to measure interrater and intrarater reliability, and through a further analysis of the data drawn from a British acute hospital which used a number of adverse-event measures, in which the extent to which variations in the data could be attributed to the staff applying the measures was assessed.

The overall purpose of this research was to improve our understanding of the validity and reliability of adverse-event measures of quality, in ways that would facilitate the development of such measures in ways that maximised validity and reliability and that would aid the interpretation of the results from such measures.

1.3 Structure of thesis

This thesis first explores the science of quality measurement, with a particular focus on the evaluation of quality measures, and then reviews the development and testing of adverse-event measures of quality. It then reports on a series of research studies designed to test out the validity and reliability of some adverse-event measures of quality. It concludes by drawing together findings from the literature with the results from these studies, and identifying a number of areas for further research.

The thesis falls into three main parts. The first part (chapters 2 and 3) is intended to set the context for the research, through a review of the relevant literature on three main themes - quality measurement, adverse events in healthcare, and testing the validity and reliability of measures. The second part (chapters 4, 5 and 6) report on the research studies which were undertaken to address the objectives of the research set out above. Each chapter presents a separate study or a related set of studies, and so each separately details the aims of the research, the methods used, the results, and the conclusions for the study or studies it describes. The third and final part of the thesis (chapter 7) brings together the key findings and conclusions from the research (reported in chapters 4, 5 and 6) with the findings from the wider literature (reported in chapters 2 and 3). It suggests how these findings might be interpreted and applied, and identifies areas for further research.

In order to make the thesis more accessible to the reader, each chapter commences with its own introduction, which sets out a short overview of the areas and issues covered by the chapter itself. In addition, two tables of contents are provided - an outline table, intended as a general guide to the structure of the thesis, and a detailed table of contents, which lists the content of each chapter down to section, subsection and unnumbered headings levels. Both tables of contents contain page numbers. In addition, a brief overview of the contents of the thesis is given below.

Chapter 2 presents an overview of the science of quality measurement and its application in managing and improving quality in healthcare. It explores the definition of quality and some common models of quality, and concludes that deconstructing the concept of quality is essential to meaningful measurement. With this in mind, the construction of measures of quality is discussed,

using a framework for defining and assessing measures of quality, and the use of quality measurement in healthcare is briefly reviewed. Finally, the chapter turns to the business of evaluating quality measures, and uses a framework for such evaluations to set out a number of dimensions on which the performance of such measures could be assessed. It focuses particularly on the issues involved in assessing the validity and reliability of quality measures.

Chapter 3 presents a comprehensive review of the place of adverse events in healthcare, and the use of adverse events in quality measurement. It begins by examining the development of interest in adverse events in healthcare from a number of perspectives, and goes on to explore the definition and classification of adverse events in measures of healthcare quality. It reviews the use of such measures, with a particular focus on the epidemiology of adverse events and the experience of using such measures in healthcare quality assurance. Finally, it presents an analysis of the rather limited published literature on the validity and reliability of such measures, which highlights the need for further research in this area.

The next three chapters present the empirical research which was undertaken to address the research questions set out in section 1.2 above. As was noted earlier, each chapter reports on the findings from a separate study or series of studies, and contains details of the aims, methods, results and conclusions from that study or studies.

Chapter 4 describes two studies undertaken to assess the face and content validity of one adverse-event measure. It reports on a questionnaire study and interview study of clinician opinion, which explored both quantitatively and qualitatively clinicians' assessments of the validity of a generic adverse-event measure of quality.

Chapter 5 explores the construct validity of adverse-event measures through an analysis of data collected from a British acute hospital at which adverse-event measures were used. That extensive database is used to test out a number of constructs about the behaviour of such measures, some of which have been tested elsewhere and reported in the literature.

Chapter 6 examines the interrater and intrarater reliability of some adverse-event measures of quality. It reports on a series of experimental studies which were undertaken to assess both the interrater and intrarater reliability of several measures, in different specialties. It also uses a further analysis of part of the large database referred to above to explore the variations in that data, and the extent to which they are associated with differences in the raters or screeners who collected the data.

Finally, *chapter 7* draws together the findings from the studies reported in chapters 4, 5 and 6, and the results of the review of the literature in chapters 2 and 3. It presents an overview of the research findings, suggests how they might be interpreted and applied, identifies some limitations which ought to be noted, and suggests some areas for further research.

Chapter 2

Measuring quality in healthcare

2.1 Introduction

This chapter presents an overview of the science of quality measurement and its application in managing and improving quality in healthcare. It begins by exploring the definition of quality itself, through some example definitions and a discussion of their common characteristics, and a brief critique of two commonly cited models of quality. It is concluded that no single definition is generally accepted, and that the concept of quality is more likely to be measurable when it is deconstructed into a number of constituent dimensions or parts concerned with particular features or characteristics of the healthcare process.

With this in mind, the construction of quality measures is then discussed. A distinction is drawn between the data used in measurement and the valuations or meanings which are attached to that data. Some sources of data are reviewed, and the development of the valuations or meanings which underlie all quality measures (either implicitly or explicitly) is explored. The chapter then turns from theory to practice, with a brief overview of some practical examples of measures of healthcare quality. A selection of systems or instruments are described and compared, using a framework based on the earlier work in the chapter. The potential for confusion inherent in the lack of any agreed definitions or standardised terminology is highlighted, and a number of areas of overlap and duplication between approaches to quality measurement are noted. It is argued that a framework for developing and assessing quality measures is an important tool for those involved in using such measures in practice.

Next, the application of quality measurement in healthcare quality assurance is reviewed. A brief historical account of the development of quality assurance in healthcare is presented, and the role of quality measurement in the wider context of health policy and healthcare systems is discussed.

Finally, the chapter turns to the business of evaluating quality measures. After considering the potential objectives of quality measurement and quality assurance, a framework developed for the purpose of evaluating quality measures is used to explore a series of dimensions on which the performance of quality measures could be assessed. The task of assessing the worth of quality measures is set in its wider context, as part of the process of evaluating the quality assurance or improvement activities in which quality measures are used. The assessment of the validity and reliability of quality measures is discussed in some detail.

2.2 Concepts and models of quality

2.2.1 Defining quality in healthcare

In order to measure something, we must define it. If we are to measure the quality of healthcare, we first have to decide what is meant by quality in healthcare. Our definition needs to be objective and robust enough to support the measurement process, and should be shared by those involved in measurement and in using the results of measurement in clinical practice, healthcare management, health services research or whatever.

The British Standards Institute, in defining the terminology of quality assurance and quality management for industry and commerce, sets out three alternative senses or meanings for the concept of quality (British Standards Institute 1979, p3):

a) *Comparative* sense - degree of excellence, in which products may be ranked relative to each other.

b) *Quantitative* sense - the degree of conformity of a product with its quantitative specification.

c) *Fitness for purpose* sense - relating the ability of a product or service to satisfy a given need.

In common with most industrial quality theorists, the BSI defines quality in the third of these senses, as "the totality of features and characteristics of a product or service which bear on its ability to

satisfy a given need" (British Standards Institute 1979, p3). More succinctly, but to the same effect, Juran (1979) defines quality as "fitness for purpose or use".

The central theme in these definitions of quality, and the common thread throughout industrial quality assurance and quality management, is a definition of quality which is squarely founded on "meeting the true requirements of the customer" (Oakland 1989, p7). In this milieu, the process of measuring, assessing and improving quality is predicated on knowing who the customers are; knowing what their true requirements are (or how to find out what they are); knowing how to measure one's ability to meet those requirements; knowing whether one has the capability to meet the requirements (or how to change to acquire the capability); knowing whether one actually and continually meets those requirements; and knowing when the customers' requirements change.

In healthcare, however, things seem less clear-cut. To begin with, despite four decades of extensive research and development activity in healthcare quality assurance in which many possible definitions of quality in the context of healthcare have been advanced, none has become universally accepted (Steffen 1988). Walter Deming, one of the foremost writers on industrial quality management, observed:

> "A suitable definition for quality of medical care is a perennial problem among administrators of medical care and people doing research in the subject. It seems simple to anyone that has not tried it." (Deming 1986, p171)

Donabedian, who has laid much of the conceptual and terminological framework for healthcare quality assurance, expressed similar reservations:

> "The quality of care is a remarkably difficult notion to define ... The criteria of quality are nothing more than value judgements that are applied to several aspects ... of a process called medical care. As such, the definition of quality may be almost anything anyone wishes it to be, although it is, ordinarily, a reflection of values and goals current in the medical care system and in the larger society of which it is a part." (Donabedian 1966)

Defining quality in the context of healthcare (as opposed to in the context of an industrial or commercial setting) may be more problematic for four reasons: the complexity of healthcare itself; the way that healthcare organisations work; the problems involved in identifying customers and defining customer requirements; and the existence of additional non-commercial dimensions in healthcare which are not present in most commercial contexts (Pollitt 1996).

Firstly, it can be argued with some justification that healthcare is a uniquely complex business. The products and services provided can be both highly technical and highly heterogeneous, they are often more personal and intimate than those provided in other service or manufacturing sectors, and they interact with a multitude of external factors (such as patient behaviour and socioeconomic conditions) in intricate and dimly understood ways. The complexity of healthcare activity makes defining and understanding quality more difficult for all these reasons (Øvretveit 1994).

Secondly, healthcare organisations are, in some ways, less straightforward in their structure and management than equivalent industrial organisations. They contain large numbers of professionals whose culture and attitudes are not sympathetic to managerialism; they are structured into many separate departments which have their own aims and microcultures and often have considerable autonomy; and they are often functionally very diverse. This means that establishing a shared view of what is meant by quality in healthcare - across these different professional groups and services, through negotiation and persuasion rather than by directive, is a challenging task (Norman and Redfern 1995).

Thirdly, it is harder to identify customers' requirements in the healthcare arena than in industry or commerce. There are many different and competing groups which can claim to be customers - such as individual patients, patient groups, society, government, purchasers, and so on - with potentially conflicting requirements and frequently different perspectives on the quality of healthcare services. In many ways, healthcare professionals or healthcare organisations act as customers in their own right or on behalf of patients as well. It can be argued that there is far more potential for dissonance between these various stakeholders in the healthcare process than there is in the more straightforward relationships found in a commercial or industrial setting. Moreover, even when the

customers of the healthcare process are identified, some will be unable, because of the nature of healthcare, to specify, comprehend or monitor the important technical elements of their requirements for healthcare. This makes the industrial definition of quality outlined above less suitable, or less workable, in healthcare (Morgan and Murgatroyd 1994, p168).

Finally, the social, moral, ethical and political dimensions of healthcare (concepts such as the right to healthcare, differences between need and demand, equity of access, and so on) have no parallel in industry and commerce, but may be judged to be essential components of a definition of quality. Healthcare is more than simply another economic product, and the definition of quality of healthcare has to give these other considerations due weight (Pollitt 1996).

However, the difficulties involved in defining quality in the healthcare setting have not discouraged people from trying. Perhaps the best known definition of the quality of medical care is that offered by Donabedian himself:

> "The extent to which the care provided is expected to achieve the most favourable balance of risks and benefits." (Donabedian 1980, p5)

Donabedian calls this an *absolutist* definition - other factors such as monetary costs and patient expectations and valuations are not a part of it. When "the judgement of quality takes into account the patient's wishes, expectations, valuations and means" he terms the definition *individualised*. If "the aggregate net benefit for an entire population [and] the social distribution of that benefit" are added, he terms the result a *social definition* of quality.

Echoing Donabedian, but placing greater emphasis on the process of healthcare, Brook and Kosecoff (1988) offer a definition which is modelled around the medical processes of diagnosis and treatment:

> "Quality of care consists of two components: the selection of the right activity or task or combination of activities, and the performance of those activities in a manner that produces the best outcome." (Brook and Kosecoff 1988)

Alternatively, the American Medical Association (1986) defines quality healthcare much more widely, in terms of a list of its desired characteristics - it should produce optimal improvement in the patient's health, emphasise the promotion of health and the prevention of disease, be provided in a timely manner, seek to achieve the patient's informed cooperation and participation, be based on accepted principles of medical science, be provided with sensitivity and concern for the patient's welfare, make efficient use of technology, and be sufficiently documented to allow for continuity of care and peer evaluation.

A fourth (and, yet again, different) definition is that offered by the Institute of Medicine:

> "Quality of care is the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge." (Institute of Medicine 1990, p21).

The Institute of Medicine reviewed over a hundred definitions of quality from the literature before producing this definition. In doing so, it found there were eighteen common elements or themes each of which appeared in some of the definitions examined (table 2.1).

Scale of quality stated.	Interpersonal skills of provider.
Nature of entity being evaluated specified.	Acceptability of provider.
Goal-orientation of definition.	Statements about use of service.
Aspects of outcome specified.	Constraint of resources.
Acceptability of service.	Constraint of consumer and patient circumstances.
Type of recipient of healthcare.	Constraint of technology and state of scientific knowledge.
Role and responsibility of recipient.	Risk versus benefit tradeoff.
Continuity, management and coordination of service.	Documentation requirements.
Professional standards of service.	
Technical competency of provider.	

Table 2.1. Elements in definitions of quality.
Institute of Medicine (1990).

It seems from these quite different definitions that there are four properties or characteristics of definitions of quality which might be used to understand the similarities and differences between them.

a) *Specificity*.

Some definitions of quality are highly specific, listing long series of elements which go to make up good quality care. Others are highly non-specific, simply stating that quality consists of meeting the goals or objectives of care without predefining those objectives in any way (Steffen 1988).

b) *Scope*.

Some definitions of quality are of very limited scope - focused solely on the individual patient-practitioner interaction - whereas others are much broader in scope, encompassing wider issues such as resource usage, equity, access, etc. It is possible, though perhaps of questionable value, to draw the boundaries of quality so widely that they include almost all properties and characteristics of healthcare. The breadth and inclusivity of the preferred definition may reflect the cultural or social values and the structure of the healthcare system (Black 1990).

c) *Perspective*.

All definitions of quality might be viewed as subjective, in that they implicitly or explicitly contain value judgements about what constitutes the quality of care. They differ in the perspective from which these subjective judgements are made (provider, patient, society, or a combination of these) and the extent to which they acknowledge the subjectivity of their perspective.

d) *Derivation*.

Most definitions are the fruit of intellectual thought or of some kind of professional consensus rather than practical fieldwork, but some are based on (or draw on) empirical evidence from studies of what clients, clinicians, managers and policymakers think constitutes quality (Donabedian 1980, p35).

Much research into the quality of healthcare proceeds with little reference to or apparent need for a definition of quality, and it might be suggested that the process of definition outlined above is of

theoretical interest rather than of real practical benefit. However, a definition and its theoretical infrastructure are essential guides for both researchers and clinicians (Reerink 1990). Firstly, a definition provides a common frame of reference and a common language which researchers and clinicians concerned with healthcare quality can use to communicate. Secondly, a definition provides the basis for a theoretical framework within which hypotheses can be framed and tested, and knowledge accumulated and linked. Thirdly, a definition is essential as the foundation for the development and validation of measures of the concept. Lastly, without a theoretical structure it is difficult to generalise the results of empirical research. It is true that much research into the quality of healthcare seems to proceed without much reference to definitions, theories or models of quality - but this focus on the study of practice, to the detriment of attention to theory, may be a factor in the perceived inability of such research to date to produce measures of the quality of healthcare which gain widespread acceptance, whose validity and reliability are universally acknowledged, and which are well described and used (Berwick 1988).

2.2.2 **Models of quality in healthcare**

Deming (1986, p279) wrote that "meaning starts with the concept, which is in someone's mind and only there: it is ineffable ... An operational definition puts communicable meaning into a concept." Having developed or selected a definition of the concept of quality in the context of healthcare, the next step is to operationalise it - commonly through developing a model of quality which can be used to structure and guide research.

Inherent in all definitions of quality is an acknowledgement of the multidimensionality of the concept - that it consists of a "heterogeneous assortment of attributes, gathered into a bundle" (Donabedian 1980, p3). Models of quality work by organising these attributes into logically coherent or conceptually meaningful groupings or dimensions. In so doing, they can act as the interface between the abstract definition of quality and the real world of measurement.

The most widely used and influential quality model is that which was first articulated by Donabedian (1966), though it owes much to preceding work both within and outside healthcare. It

divides the attributes of quality into three dimensions - structure, process and outcome. Donabedian defines these three dimensions as follows (Donabedian 1990, p14):

a) "Structure denotes the attributes of the settings in which care occurs. This includes the attributes of material resources, of human resources, and of organisational structure."

b) "Process denotes what is actually done in giving and receiving care. It includes the patient's activities in seeking care and carrying it out as well as the practitioner's activities in making a diagnosis and recommending or implementing treatment."

c) "Outcome denotes the effects of care on the health status of patients and populations. Improvements in the patient's knowledge and salutary changes in the patient's behaviour are included under a broad definition of health status, and so is the degree of the patient's satisfaction with care."

Donabedian's model has been extensively used by researchers and clinicians as a frame of reference for their work on quality measurement, assurance and improvement, almost to the exclusion of other theoretical viewpoints. The predominance of the model may reflect its conceptual strength, though it may also indicate a preference amongst researchers for applied research rather than theory and model building. Donabedian's model has also engendered longrunning debates over two issues - the relative merits and demerits of each dimension, and the existence of links or correlations between measures of the different dimensions.

It is sometimes assumed that structure, process and outcome are, in that order, increasingly difficult to measure but increasingly important and valid as measures of quality - what Berwick (1988) has critically termed a "hierarchy of validity". Indeed, as Berwick and Knapp (1990) report, some researchers have treated structure and process as of value only to the extent that they act as proxies for outcome, arguing that since healthcare exists to improve or maintain people's health, outcome measures offer the most direct and inherently valid measure of their success (Frater 1992; Frater and Costain 1992; Shanks and Frater 1993). Other researchers have suggested that outcome measurement is really the province of clinical research rather than quality measurement and quality

assurance, and that process measures are more feasible in practice, and provide information which is more directly applicable in making changes to systems or processes to improve quality (Sanazaro 1974; Mant and Hicks 1995). Donabedian (1980, p119) finds there are significant advantages and disadvantages to measures in each dimension, and argues that we should distrust generalities about methods of assessment, preferring to judge each situation on the setting and particular objectives concerned. However, to gain a full picture of the quality of care, he advocates the use of multidimensional assessment methods, which include elements from each dimension of quality.

The relationships between the three dimensions of Donabedian's model have also been examined - often in the course of the debate of their relative validity, since if structure and process measures are valid only as proxies for outcomes, the establishment of a relationship between measures of structure and process and measures of outcome becomes vitally important. In practice, it has proved difficult to detect meaningful relationships between measures in different dimensions and the literature abounds with studies which have demonstrated little or no significant correlation (Komaroff 1978; Murphy and Jacobson 1984), a fact which some outcomes advocates have taken as evidence of the weakness of process and structure measures (Vuori 1989), while others have criticised the rigour and conceptual oversimplicity of the process-outcome studies themselves (McAuliffe 1978). Berwick (1988) contends that the search for relationships amongst the different dimensions betokens a mistakenly unidimensional view of quality, and a misplaced conviction that there is a single latent property called quality. In fact, he asserts, while the various aspects of quality may cluster in certain providers, "it is very unlikely that all of these and other valued attributes of care march strictly to a single underlying drummer we would call quality" (Berwick 1988).

An alternative model of quality, based on a much broader definition of the concept of quality and structured much more specifically around healthcare, was propounded by Maxwell (1984), who suggested that there were six dimensions each of which should be recognised separately, and each of which required different measures and different assessment skills:

a) Access to services - issues such as waiting times and referral patterns, geographic variations in service availability, or physical proximity and availability of services.

b) Relevance to need of the community - issues such as the pertinence of services provided to the actual needs of people, assessed through health needs studies or surveys.

c) Effectiveness for individual patients - aspects such as the technical effectiveness of treatment and the outcome for the patient.

d) Equity or fairness - the equitability of the distribution or provision of services across different social groups or sections of society.

e) Social acceptability - aspects of care relating to whether care is provided in the way people want, and meets their expectations.

f) Efficiency and economy - considerations such as the cost of services provided, both in absolute terms and relative to other providers.

This model of quality might be fairly accused of encompassing almost every conceivable attribute of healthcare, and so expanding the definition of quality as to make it ultimately less meaningful. Maxwell's model has not been widely used in practice outside the UK to develop or support approaches to quality measurement, though it is often cited, particularly by researchers who approach quality assurance with an epidemiological or public health perspective, to demonstrate the dangers of adopting too narrow a definition of quality (Black 1990).

Many other researchers have developed and published models of quality or structures which might be employed as models (Freeborn and Greenlick 1973; Williamson 1978; Merry 1987; Kitson 1989; and others). Often, these models are largely derived from Donabedian's seminal work, though sometimes they categorise or group the attributes of quality slightly differently. However, none has been anywhere near as widely used or applied.

These models of quality seem to serve one important purpose. They break down or deconstruct what is a complex, subjective concept which seems to defy adequate definition into a number of constituent dimensions or parts, each of which is rather more capable of being defined, understood,

measured and analysed. There is no generally accepted single definition of quality in healthcare, perhaps because the concept defies sensible definition. Models of quality, such as the structure-process-outcome paradigm first advanced by Donabedian three decades ago, serve a useful purpose in breaking down the complex portmanteau concept of quality into an ordered hierarchy or framework of simpler and more measurable concepts.

2.3 The measurement of quality

2.3.1 **The theory of quality measurement**

The problems of measurement in healthcare are almost self-evident. Kind (1988) contrasts measurement in the physical sciences which "conveys the impression of a precise operation based on well-established procedures, carried out in controlled laboratory settings and producing results which are expressed in terms of standardised units of measure" with the measurement of aspects of healthcare "where not only is the phenomenon under investigation defined in many different ways, but there are varying opinions as to how it might be represented, and on whether it could or should be quantified". The latter description neatly sets out the problems which face those who would design and develop measures of the quality of healthcare.

To measure quality, the concepts and constructs used to define and model quality must be translated into unambiguous and concrete representations. These will always consist of:

a) Actual items of data, representing characteristics of the healthcare system, processes, or results. These are often termed elements or parameters.

b) Valuations, giving meaning to the data elements by providing some general rule about what represents goodness (or good quality) in respect of individual data items or groups of data items. These are often termed criteria or standards.

Every measure of healthcare quality can be reduced to these component parts - data elements and valuations - though the manner in which they are selected, derived, defined, collected and used in individual measures varies tremendously.

Nature and source of data elements

The nature of the data elements used in measures of quality is so heterogeneous it is almost impossible to generalise about them. They represent characteristics or properties of almost every conceivable part, process and effect of the healthcare system. They may be quantitative or qualitative; categorical, nominal, ordinal or ratio scaled; parametric or non-parametric; dimensioned or dimensionless.

The sources of the data elements used in measures of quality are easier to categorise. First and foremost, data elements are commonly drawn from the clinical record of care. The clinical record is used as the primary data source for the many quality measures, sometimes supplemented by other data sources. The importance of the clinical record to quality measurement makes its completeness and accuracy (which are often questioned) fundamental issues for the developers and users of quality measures. In using clinical records, measures should take account of the acknowledged problems of completeness and veracity, and the risk that the development of measures is shaped by the availability (or nonavailability) of data in the clinical record. Secondly, some measures require the completion of some additional record above and beyond that normally kept in the clinical record, with data elements collected or recorded specifically for the purpose of quality measurement. While this frees the developer from the restrictions of the clinical record content, it imposes a new set of problems in ensuring that complete and accurate data is collected. A third source of data, less frequently used because of its cost, is direct observation of practice or the situations or circumstances in which care is provided. The fourth main source of data elements is the participants in the healthcare process (providers and consumers) who, through questionnaires, surveys or interviews, provide some data elements which might also be obtained from other sources and others which could only be collected from the participants themselves.

Nature and source of valuations or criteria

The valuations which are attached to data elements are used to interpret them, identifying what constitutes goodness (or good quality) and what does not. In many quality measures it can be difficult to separate these valuations from the data elements themselves, and the derivation of the valuations or criteria can be far from clear.

Donabedian (1982) developed a useful classification of the nature and source of the valuations or criteria used in the measurement of healthcare quality. He defines the nature of the criteria in terms of:

a) *Specification*.

In some measures, the criteria are explicitly stated - usually in written form - in unambiguous terminology which leaves little room for differences in interpretation. However in others, the criteria are implicit, often unstated or stated in terms which allow ample latitude for professional differences of opinion or interpretation. These latter measures often rely on individual professional judgements made by the users of the measure.

b) *Referent*.

The topic, subject, patient group or whatever to which the criteria are to be applied is termed the referent. Again, some measures make their intended referent very clear, while others do not. It is not unknown for quality measures designed for one referent to be used in a quite different setting with another referent.

c) *Monotony versus inflection*.

Some criteria are what Donabedian has termed monotonic - it can always be said that "the more the better" or "the less the better" (an example might be postoperative mortality rates). Other criteria are inflected, which means that there is an optimum value or range, below or above which quality decreases. In practice, few criteria are likely to be wholly monotonic. The certainty with which the optimum value or range can be specified for inflected criteria varies considerably.

d) *Stringency*.

The level of quality envisaged in a criterion or set of criteria is called its stringency. The more stringent the criteria used in a measure, the harder it might be for providers to achieve those criteria.

e) *Importance*.

Clearly, all criteria are not of the same importance, since some will relate to aspects of healthcare that are more crucial than others. There may, therefore, need to be some weighting process, which assigns greater weight to those criteria judged to be more important.

The source of the criteria used in quality measures is clearly crucial to their validity. If a measure were based on criteria which were not supported by available evidence from relevant clinical research, were not accepted by key stakeholders such as clinicians and service users, and did not match the context or setting in which the measure was to be used, then its validity would be very much open to question. Donabedian categorises the source of the criteria used in measures of quality on these three axes:

a) *Normative versus empirical derivation*.

Criteria may be derived from the opinions of participants or developers about what constitutes good quality care (normative criteria), or they may result from actual research into the healthcare process and its outcomes (empirical criteria). In the past, many measures made extensive use of normatively derived criteria, drawing on expert opinion either informally through consultation or discussion, or more formally through consensus building approaches like the Delphi technique. However, it has been increasingly recognised that such criteria may be misleading, representing a skewed, partial or simply misplaced professional consensus which is not supported by available empirical evidence. For that reason, there is now a much greater focus on grounding criteria more firmly in the evidence from well designed and rigorously conducted empirical research (using qualitative, quantitative, experimental and observational methods).

b) *Exogenous, endogenous and autogenous criteria*.

When the criteria are developed by one group of practitioners or developers for use on or by others, they are said to be exogenous. Endogenous criteria are those developed by the group of practitioners whose performance they will be used to measure. Autogenous criteria are those that an individual practitioner develops and uses to assess his or her individual performance.

c) *Representative versus elitist criteria*.

Criteria which are derived from or based on the performance of practitioners or providers who are perceived to be of above average performance or who practice in atypical circumstances (such as leading academic clinicians, tertiary service specialists or professional opinion-formers) are termed elitist, while those which are based on the experience and practice of all providers or practitioners are termed representative.

While all measures of quality are made up of data elements and valuations or criteria, as described above, there are important differences in the way measures use data elements or criteria. Some measures give primacy to the data elements - focusing on defining and collecting those data elements for analysis, with little attention to the valuations which will be placed on those data elements. These approaches can be said to be *data-driven*. Other measures give primacy to the valuations or criteria - focusing on elucidating and defining those valuations very carefully, and only then using them to determine what data elements are required. These approaches can be said to be *criterion driven*. The distinction is not simply bipolar - there is a continuum of variation from one extreme to the other.

The process through which measures of quality are developed, applied, tested and evaluated is far from well defined. In theory, at least, seven distinct stages in the process can be distinguished - conceptualisation, definition, development, testing, evaluation, application and revision. In practice, many measures are inadequately developed, with scant attention being given to aspects such as conceptualisation (the theoretical background or framework within which the measure fits, the theoretical concept which the measure is intended to measure) and evaluation (testing the reliability and validity of the measure, and formally assessing other aspects of its utility). Such

inadequately developed and tested instruments have often then been widely and improperly used in quality assessment (Sanazaro and Mills 1991).

There is no accepted classification of the bewildering variety of quality measures which have been developed and used in healthcare over the last four decades. However, Donabedian's model and the preceding discussion of the theory of quality measurement offer one approach to grouping and ordering actual measures. We can categorise measures according to two principal characteristics:

a) *Dimension measured - structure, process or outcome*. Whether the measure is primarily focused on issues of structure, process or outcome quality.

b) *Data-driven versus criterion-driven* - whether the measure gives primacy to the data elements on which it is based or the valuations placed on those data elements.

In the following sections, this approach to classification will be used to present an overview and brief critique of some of the measures of quality which have been developed and used in healthcare. For the purposes of clarity, the methods and measures to be reviewed are summarised in table 2.2, which also uses the same classification matrix.

Of course, the boundaries between measures and dimensions are not clearcut, and so the classification must be to some degree imperfect. Some measures address both structure and process, or combine some data-driven elements with some criterion-driven elements. In classifying them below, their primary or principal orientations have been used.

	Criterion-driven measures	Data-driven measures
Structure quality measures	Accreditation Organisational audit	Performance indicators
Process quality measures	Criterion audit Standard setting Practice guidelines Adverse events	Large clinical databases Clinical indicators Patient satisfaction
Outcome quality measures	Avoidable deaths	Patient satisfaction Mortality indicators Health status measures

Table 2.2. Summary of measures of quality reviewed.

2.3.2 **Measures of structure quality**

Criterion-driven measures

Criterion-driven measures of structural quality are largely based on sets of valuations or criteria which specify the desirable or preferable elements in the structure of a healthcare organisation such as a hospital. These criteria may simply specify the presence or absence of some elements in the structure (such as the availability of certain equipment), or may outline preferred arrangements or procedures (such as the preferred management structure).

Criterion-driven measures of structural quality have been widely used, particularly in the United States, Canada, Australia and (latterly) the United Kingdom in a process which is frequently termed accreditation (because the measures have been used to accredit or licence institutions). In several countries, these measures of quality have been an important component of the process of monitoring and reviewing institutions' adherence to commonly accepted standards of service provision (Scrivens 1995).

Accreditation has been used in the USA since the establishment of the Hospital Standardisation Programme by the American College of Surgeons in 1918, which set out to assess hospitals with more than 100 beds against a simple one-page statement of minimum requirements, through on-site visits. The independent, nongovernmental Joint Commission on Accreditation of Healthcare Organisations (JCAHO) was set up in 1951 to take forward the programme, which by then was

accrediting over 3,000 hospitals. It currently assesses over 80% of US hospitals. Although accreditation is voluntary, many public and private payors impose an accreditation requirement which gives the JCAHO considerable authority over individual provider institutions (Couch 1989).

The JCAHO standards are developed through a series of expert panels and committees, and are published annually in the Accreditation Manual for Hospitals (Joint Commission on Accreditation of Healthcare Organisations 1988). They consist of a series of statements of structural and procedural attributes of healthcare organisations which are deemed to be associated with high quality care. Compliance with the standards does not demonstrate that good quality care is being provided, but is claimed to show the potential for the provision of good quality care exists. The measures are applied through site visits by survey teams, whose assessment of compliance with the standards is the basis for the decision whether or not to accredit the institution.

In the UK, the King's Fund has sponsored the development of an accreditation programme for hospitals, modelled on the work of accreditation agencies in the USA, Canada and Australia (King's Fund 1996). Accreditation has also long been used by the Royal Colleges to assess the suitability of hospitals for medical staff training (Hopkins 1990, p13). In 1979, the British Standards Institute (BSI) developed a set of quality standards against which the quality management systems of industrial and commercial organisations could be assessed (British Standards Institute 1987). The BS5750 accreditation system was subsequently adopted by the International Standards Organisation and, as ISO9000, is now widely used throughout the world (Department of Trade and Industry 1995). These standards have been adapted for use within the health service (Rooney 1988) and a number of NHS organisations have sought and achieved BS5750 or ISO9000 accreditation (McDonald 1991).

Data-driven measures

By gathering a wide range of information about the resources and facilities for healthcare across a large number of institutions, comparisons can highlight those providers with substantially different arrangements from their peers, and may be the starting point for more detailed analyses of the effects of those differences in structure.

Perhaps the best known large-scale application of this approach in the UK is the development of performance indicators led by Yates and colleagues which led to the longstanding Department of Health sponsored system of Health Service Indicators (Sanderson 1987). In the 1970s, Yates and colleagues demonstrated that the comparative analysis of readily available structural information about long-stay mental handicap hospitals could be used to identify those institutions where the quality of care was unacceptably low (Yates and Vickerstaff 1982). Yates and others went on to use the same approach in the acute hospital sector to considerable effect (Yates and Davidge 1984; Ham 1985), and it formed the foundation of the current national system of Health Service Indicators (Department of Health 1989a), which present a range of comparative information from providers throughout the NHS, much of it relating to structure (though some indicators are focused on process).

Merits and demerits of structure quality measures

The advantages of structural measures of quality are that they are relatively easy to define and use, and the costs of collecting the data they require are relatively low. Because structures tend to change quite slowly, measurement can be relatively infrequent and thus less costly. In addition, it has been claimed that their clear factual base in the presence or absence of facilities or procedures makes them relatively objective (Rosenberg 1990). However, measures of the quality of structure have often been criticised for being "a rather blunt instrument in quality assessment" (Donabedian 1990, p21) with little demonstrable relationship to process and outcome measures. Their validity is also frequently questioned by practising clinicians, who see little relation between such measures and the realities of providing good quality care. Indeed, most of the structural quality measures described above are normatively rather than empirically derived, and so doubts about their validity may be quite justifiable (Longo, Wilt and Laubenthal 1986). On the other hand,, it may be argued that many elements of high structural quality - such as adequate staffing, equipment and training - should be an end in themselves, regardless of any hypothetical relationship to other dimensions of quality. While the presence of good structural quality does not guarantee good process and outcome quality, it seems probable that poor structural quality will make the provision of good process and outcome quality much less likely - so good structural quality is a necessary but not sufficient condition for good quality on other dimensions.

2.3.3 **Measures of process quality**

Criterion-driven measures

The development and application of sets of criteria which define the constituent parts, elements or stages in a high quality process of care has played a large and continuing role in quality measurement. Usually, these criteria map processes such as the diagnosis, investigation and treatment of conditions, defining the appropriate decisions, actions and recordkeeping of the health professionals involved.

Lembcke (1956) first described the development of process criteria, which he termed "medical auditing by scientific methods", and he set out a method for developing the criteria and for measuring performance against them. He also demonstrated the effectiveness of the method in changing physicians' practice patterns. The approach was adopted and promulgated by the JCAHO, and became the basis for mandatory quality assurance activities throughout the USA in the early 1970s (Sanazaro 1974). A tremendous investment was made in developing criterion-based measures of quality for a large number of conditions and patient groups and using them in medical audits (Brook 1977).

However, a number of major studies demonstrated that criterion-based process measures were neither as valid nor as reliable as Lembcke (1967) had suggested. Brook and Appel (1973) found that the results of such measures depended crucially on how the criteria used were derived and how the measures were actually applied. Hulka (1978) found that nonadherence to criteria often resulted from patient heterogeneity rather than poor quality practice. Sanazaro and Worth (1978) showed that the use of criterion-based measures had minimal effects on actual practice, and Nobrega et al (1977) and others found no significant relationship between criterion-based process measures and the quality of clinical outcomes. Komaroff (1978) concluded that many criterion-based measures were too complex, overambitious and poorly tested and should be simplified. Nevertheless, criterion-based measures continue to be widely used in a variety of settings (Mayer, Clinton and Newhall 1988) and form an important part of current quality measurement efforts in the USA (Agency for Health Care Policy and Research 1995).

Criterion-based measures of process quality have also been widely used in the UK, particularly over the last decade. Clinical guidelines for a range of conditions and treatments, from asthma to electroconvulsive therapy, have been promulgated and used as a basis for measurement (Department of Health 1993). However, a similar set of concerns has been voiced about the validity of these measures (Grimshaw et al 1995) and their utility in both quality measurement and improvement (Hopkins 1995).

Traditional criterion-based measures of process quality have been developed by attempting to define what constitutes a good process of care. The heterogeneity of patients and the complexity of the disease process and treatment modalities make this very difficult, and the resulting criterion-based measures are often complex and unwieldy. An alternative approach has been to develop criterion-based measures focused on poor quality care - in which the valuations or criteria define instances or events which are indicative of poor quality care rather than defining the constituents of good quality care. These instances of poor quality care (or "disquality") are often called adverse events. A range of quality measures and quality assurance systems using this approach have been developed and widely used in the USA (Craddick 1979; Craddick and Bader 1983; Goldman 1989), and the concept has also been applied in Australia (Wolff 1992) and the UK (Bennett and Walshe 1990; Walshe, Bennett and Ingram 1995). However, such adverse-event measures of quality have been criticised, on grounds of their relative inefficiency, high error rates (especially false positives), emphasis on patient safety, and poor validity as quality measures (Sanazaro and Mills 1991). The theoretical background, development and testing, practical application, and evaluation of these measurement techniques is discussed in more detail in chapter 3.

Data-driven measures

Data-driven approaches to measuring process quality largely centre around the collection and analysis of large databases of information about the process of care. The nature of modern healthcare organisations is such that relatively detailed computerised databases of information are often readily available. Though the data sets they contain - defined by operational need rather than the requirements of quality measurement - may be restrictive, the opportunity offered by these databases has been widely used in quality measurement.

The earliest developer of such a database was the Commission on Professional Hospital Activities (CPHA), which first began to gather a uniform data set on each patient from hospitals in the United States in the 1950s (Commission on Professional Hospital Activities 1990). CPHA established a database which at one point gathered data from over 2,300 acute hospitals in the USA, and now contains data on over 200 million patient admissions. Many other organisations and researchers, including federal agencies such as the Health Care Financing Administration, have followed suit, using data from payor databases (Dubois et al 1987), billing systems (Mendenhall 1987a; 1987b), and other sources to develop measures of quality. In 1989, the JCAHO, recognising the weaknesses of its largely structural quality measures, established an extensive programme to develop and collect clinical indicator information from accredited hospitals to provide new measures of process quality (Robinson 1988; O'Leary et al 1989).

In the UK, routine data about the healthcare process has been available for some years for some areas of care (for example, acute inpatient services) but very little data has been routinely collected in other areas (such as mental health, or community services). Some information from these sources has been used in the Health Service Indicators discussed earlier to provide process measures such as length of stay, admission and readmission rates, and so on. More recently, an independent UK organisation linked to CPHA has established a similar comparative database providing a range of process quality measures largely but not exclusively oriented around resource usage (CHKS 1995).

Another major data-driven approach to both process and outcome quality measurement is the monitoring of patient satisfaction and patient opinions about the care they receive - either proactively through patient surveys and other mechanisms for understanding patients' views (Kelson 1996) or reactively through patient complaints (Allsop and Mulcahy 1996). Patients' opinions can be viewed solely as an outcome (of their interaction with the healthcare process or healthcare system) but it is perhaps more meaningful to view patients' opinions as a data source for information on all three dimensions of quality - structure, process and outcome. After all, patients have valid views on the quality of the physical environment (structure), the quality of communication from staff (process), and the improvement in their health status (outcome) and it would seem perverse to label all these views as outcomes.

Though the orthodox professional view is that patients lack the knowledge to make valid judgements of the quality of process and outcome, research has shown that patients can distinguish between interpersonal and technical aspects of process and outcome quality, and can make valid assessments in both domains (Kaplan and Ware 1989). Patient satisfaction measures have been widely used both in the USA and the UK (Fitzpatrick 1990). Patient satisfaction may be termed either a process measure or an outcome measure, depending on whether satisfaction with the processes or the results of care are being assessed.

Merits and demerits of process quality measures

Process measures of quality have many advantages. In particular, they provide information which is focused on process performance and so is readily used in changing systems or practices, since the need for and direction of change are often clearly indicated by the process measures. Indeed, advocates of continuous quality improvement (CQI) would argue that without a comprehensive understanding of healthcare systems and processes (which can only be gained through process quality measurement) opportunities for quality improvement cannot be identified and realised (Plsek 1997). In addition, process quality measures are often relatively inexpensive to apply (certainly compared to many measures of outcome) since all measurement takes place within the healthcare organisation, at or around the time of treatment. Their primary weakness is the frequent lack of evidence that what is deemed to represent a good process of care is related to the achievement of good outcomes. In some cases, this represents a failure on the part of those developing the measure, who have not made proper use of available research findings, but more commonly it indicates that the evidence to link processes and outcomes is not available for a variety of reasons. This results at least in part from the nature of clinical science, in which there is limited evidence for the efficacy of much of accepted custom and practice (Smith 1991) and it is, perhaps, unrealistic to expect the developers of quality measures to establish process-outcome correlations where clinical research has not done so.

2.3.4 **Measures of outcome quality**

Criterion-driven measures

Most outcome quality assessment tools are data rather than criterion-driven - they rely on the collection of data about the outcomes of healthcare which is then interpreted in analysis, when valuations are placed on it. However, there are some examples of criterion-driven measures of outcome which have been developed and used, in a process akin to that described above for process quality criteria. Rutstein and colleagues, for example, established a set of criteria defining causes of mortality which they considered avoidable and therefore indicators of poor quality (Rutstein et al 1976). Charlton and colleagues applied a variant of these outcome criteria in the UK and internationally (Charlton et al 1983; Charlton and Velez 1986).

Data-driven measures

The most fundamental data-driven measures of outcome quality are those concerned with mortality, which assess quality through comparisons of various general or disease-specific mortality rates. Mortality measures are of limited validity, for while the aim of healthcare is usually to postpone death, this objective is often not the primary one, and almost never the only one. Because mortality is rare, especially in some services and specialties, measures based on mortality are founded on the care provided to a very small proportion of patients. In addition, even for those patients there are a number of causes and contributory factors involved in mortality which make the interpretation of general mortality rates difficult at best (Kahn et al 1988). Nevertheless, mortality statistics have been widely used in the USA to compare quality at different institutions, and similar mortality indicators have been proposed by the Department of Health for use in the UK (Walshe 1997). The validity of mortality rates as indicators of outcome quality has been widely questioned, especially by healthcare providers (Berwick and Wald 1990), but Dubois et al (1987) have demonstrated that hospitals with high outlier mortality statistics had a greater proportion of preventable deaths, suggesting that there is some relationship with quality of care.

The weakness of mortality as an outcome measure has led some researchers to attempt to develop more sophisticated measures which are risk-adjusted - in other words, take account of differences in patient populations such as severity (DesHarnais et al 1990). It has also led researchers to broaden

their interest to other relatively readily available outcome data, such as the incidence of hospital readmissions and post-discharge complications (Roos et al 1985; Roos, et al 1988; DesHarnais et al 1990) though the validity of these data as measures of quality has also been challenged (Milne and Clarke 1990).

The paucity of routinely available data on healthcare outcomes has led some to establish new data-driven outcome measures drawing on information which is specifically collected for the purpose, often using the growing library of health status measures developed both for specific diagnoses or patient groups and for global or generic use (McDowell and Newell 1991; Hopkins 1990, p44). The practicality of gathering such information (given that it involves data collection before and/or after the healthcare process) has been demonstrated (Tarlov et al 1989; Coles 1990; Bardsley and Coles 1992), and they have been increasingly widely used in the UK (UK Clearing House on Health Outcomes 1996) despite some criticism of their validity and utility (Mant and Hicks 1995; Sheldon 1994)

Merits and demerits of outcome quality measures

The single greatest advantage of outcome quality measures is the almost universal acceptance of their validity - they can be said to be intrinsically "self-validating" (Rosenberg 1990). Proponents of outcome quality measures argue that, since healthcare services exist to improve peoples' health, direct measures of the outcomes of healthcare (in terms of peoples' health) are the best way to assess healthcare quality. However, there are both theoretical and practical disadvantages to take into consideration. Firstly, outcome measurement often reveals little about *why* good or poor quality outcomes occur (in other words, the processes that lead to those good or poor outcomes). This can make the data of little practical value in making quality improvements, since the causes of poor quality and the workings of the healthcare processes involved have not been explored and are not understood. Secondly, the attribution of health outcomes to healthcare interventions (as opposed to other factors or life events) is problematic and methodologically complex, and becomes more so, the more distant those outcomes are from the healthcare processes being studied. Even with quite short term outcomes, the confounding effect of variables outside the control of the healthcare process (such as community support, patient behaviour, housing, and other economic and social factors) can be considerable. Thirdly, since outcome measurement reaches out beyond the period of the

healthcare intervention (both before it and after it) data collection is almost inevitably expensive and involved. Fourthly, the low prevalence of some outcomes means that unfeasibly large numbers of cases or amounts of data may need to be collected to support a sound statistical analysis (Mant and Hicks 1995). Arguing cogently against basing quality measurement around outcomes, Berwick and Knapp (1990) describe it as "a formula for paralysis". They point out that it burdens quality measurement with the agenda of almost all clinical and health services research, and suggest that we simply know too little about what in healthcare produces health to make outcome central to the definition of quality.

2.4 Applying quality measurement in healthcare

2.4.1 From quality measurement to quality assurance

Quality assurance is an umbrella term, widely used to describe all sorts of programmes, activities and systems designed to monitor and improve the quality of healthcare. It is commonly defined as "the measurement of the actual quality of care against preestablished standards, followed by the implementation of corrective actions to achieve those standards" (Vuori 1989). Essentially, quality assurance consists of two interconnected activities (Walshe et al 1991):

a) *Quality measurement*, in which a variety of tools and techniques (some of which are described above) are used to measure the quality of healthcare.

b) *Quality improvement*, in which information about the quality of healthcare and about the workings of the healthcare system is used to make changes to individual or organisational practices which are designed to improve the quality of healthcare.

These two activities are often visualised as part of a cycle of activity - which has been termed the *audit cycle* - in which measurement and changes to practice alternate to produce continuing improvements in quality (Fowkes 1982). This rather simplistic cycle of activity can be (and has been) further subdivided to break down the processes of measurement and improvement into a

number of separate but interlinked steps, and to illustrate the processes involved, such as objective setting, planning, data collection, reporting, change management and ongoing monitoring (Lang and Clinton 1984; Crombie et al 1993, p27; and others).

There are a number of other terms commonly (and confusingly) used to describe this process - medical audit, clinical audit, continuous quality improvement, total quality management, and so on. Shaw (1980a) suggests 96 possible combinations of words, and while there are certainly some important distinctions to be drawn between them, there are far more similarities than differences in the processes they describe. What terminology is current seems to depend more on fashion than any more secure foundation, and the absence of clear demarcations between these terms means they are often used interchangeably. In this thesis, for the sake of simplicity and consistency, the terms quality assurance, quality measurement and quality improvement will be used as described above.

Quality measurement is an essential part of the quality assurance process, and it is almost impossible to conceive of a successful quality assurance programme being established without good measurement tools to serve it. By the same token, quality measurement by itself would be a sterile occupation of little value, since it would involve data collection and analysis to no actual purpose. This recognition of the importance of quality measurement to quality assurance and vice versa must contribute to our ideas about what constitutes a good measure of quality. Not only must measures be valid, reliable, and all the things measurement systems are meant to be, they must satisfy the needs of the quality assurance process, which means they must be practical, workable, affordable, applicable, and meaningful for the business of quality improvement.

Berwick, Godfrey, and Roessner (1990) argue that research in the area of healthcare quality assurance has focused too much on the issues surrounding quality assessment, and has given too little attention to the problems of quality improvement - treating the subject as a science rather than an applied technology, and being directed towards "unveiling the fact of flaw, not its cause" (Berwick, Godfrey, and Roessner 1990, p11). Two decades ago Jessee (1977) and Brook (1977) made the same point, noting that while quality measurement techniques were far from perfect, they had been able to identify important deficiencies in the quality of care. However, remedying those deficiencies had proved to be a much more complex task than simply identifying them.

2.4.2 **Quality measurement in historical perspective**

Professional and public concern over the quality of care is not new. In 1518, when the Royal College of Physicians was founded, its charter explicitly committed it to uphold the standards of medicine "both for their own honour and the public benefit" (Shaw 1989). However, healthcare quality assurance has only become a part of everyday practice in recent years, as a result of modern social, financial and technological pressures.

Perhaps the first documented quality assessment studies were undertaken by Florence Nightingale, who used data such as mortality rates for diagnostic categories to highlight unsafe conditions in Crimean British Army hospitals, by showing that soldiers in the Crimean hospitals were much more likely to die of common medical conditions like pneumonia than civilians in the UK. She also developed a system of hospital statistics designed to monitor the outcomes of surgical operations and the efficiency and effectiveness of bed usage (Wilkin and McColl 1987). Her interest in surgical mortality was pursued by Groves (1908) who surveyed operative mortality rates in 50 UK hospitals and published the results. Codman, a Bostonian surgeon of the same era, undertook a systematic assessment of every patient he treated which included a classification of the outcome of all surgery - an approach he named his end results system (Codman 1914). These pioneers were all unusual individuals whose interest in healthcare quality was not shared by most of their colleagues, and their innovations were not widely adopted or applied.

The first systematic quality assurance programme was established in the USA 1918, by the then recently formed American College of Surgeons. The Hospital Standardisation Programme, as it was called, set out to regularly review hospital facilities and practices against a set of minimum standards. While the first annual inspections revealed that only 89 of 692 hospitals visited met the very basic standards, the programme grew and expanded rapidly (Couch 1989). In 1951, responsibility for the programme was passed to the new Joint Commission on Accreditation of Hospitals (JCAH), which was founded to provide an independent, nongovernmental centre for quality assurance in US healthcare.

In the 1950s and 1960s, interest in the quality of healthcare flourished in the USA. Research into the development of quality assessment techniques grew, and the academic foundations of quality assurance were laid (Donabedian 1966). The growing costs of medical care and the problems of controlling the highly fragmented US healthcare system led to the establishment in 1974 of a mandatory, externally managed quality assurance programme throughout the USA - the Professional Standards Review Organisations (Sanazaro 1974). The PSRO programme made extensive use of process criteria to assess the quality of care, but its effectiveness was widely questioned (Komaroff 1978). During the 1970s, an explosion in clinical negligence litigation brought a further focus on quality improvement, and the development of risk management in US hospitals (Mills and Von Bolschwing 1995). In the 1980s, new legislation replaced the PSRO programme with a system of state-based Peer Review Organisations (PROs), with a similar mandate but a different methodology, based largely on generic occurrence screening (Jost 1989). In the late 1980s, interest in industrial approaches to quality measurement and improvement grew, and a strong and influential continuous quality improvement (CQI) movement developed (Berwick Godfrey and Roessner 1990; Berwick 1996). With the growth of managed care in the 1990s came a number of new approaches to quality measurement, including the growing use of so-called balanced scorecards of performance measures, and the development of a new organisation, the National Committee on Quality Assurance, to monitor and accredit managed care plans (National Committee for Quality Assurance 1996).

In the USA today, the sheer extent and depth of quality assessment and monitoring is striking. All hospitals have substantial internal quality assurance programmes, with quality review staff who use a variety of techniques to monitor the quality of care and to intervene when problems are identified. These programmes are managed and directed by a complex arrangement of committees, with high level medical and managerial participation, and are often linked to physician credentialling and risk management (King and Jones 1989; Chassin 1996).

Almost all hospitals have to satisfy the Joint Commission's quality standards to gain accreditation, which means undergoing periodic inspections by external survey assessors and making changes to practice where required (Couch 1989). All hospitals receiving payments from the federal Medicare programme (which pays for care for all elderly people) have to submit a sample of cases to their

local Peer Review Organisation for detailed case reviews and quality assessments, and many other healthcare funding agencies (such as large health insurers) impose a similar requirement (Jost 1989). Detailed comparative information on hospital performance, based on a range of clinical data, is widely used by both providers and funders of healthcare to address quality issues, and such information is even available to the general public (Dubois et al 1987).

Almost all care provided in the US healthcare system is subject to some form of quality review - and much is often reviewed more than once by different bodies. Despite these extensive (and expensive) programmes, there is widespread dissatisfaction with the effectiveness of quality assurance in US healthcare, and a common perception that much current activity has little or no worth in terms of its effect in improving quality or controlling costs (Berwick 1989; Welch and Grover 1991; Angell and Kassirer 1996).

In the UK, quality assurance in healthcare has developed much more slowly than in the USA, and has tended to consist of separate initiatives addressing individual issues rather than integrated quality assurance programmes. For example, a confidential enquiry into maternal deaths was established in the 1930s (and still exists, in modified form, today) to undertake a confidential investigation of all instances of maternal mortality in childbirth (Department of Health 1991a). A national quality control scheme for pathology laboratories was established in 1969, and continues to monitor the quality of laboratory test results through a programme of sampled retesting (Whitehead and Woodford 1981). The Health Advisory Service was established in 1969, to deal with the quality problems of many long-stay hospitals for the mentally handicapped and mentally ill (Lancet 1984). The National Confidential Enquiry into Perioperative Deaths (NCEPOD) was set up in 1986 to examine instances of postoperative mortality and to identify avoidable causes and factors in these cases (NCEPOD 1989).

In the 1970s and 1980s there were a number of calls for the development of more comprehensive mechanisms for quality assurance than the isolated initiatives described above. In 1979, the Royal Commission on the NHS reported:

> "...We are not convinced that the professions generally regard the introduction of audit or peer review of standards of care and treatment with a proper sense of urgency. We recommend that a planned programme for the introduction of such procedures should be set up for the health professions by their professional bodies and progress monitored by the health departments." (Merrison 1979)

A few years later, Griffiths levelled a similar criticism in his highly influential Management Inquiry report:

> "Clinical evaluation of particular practices is by no means common, and economic evaluation of those practices extremely rare. Nor can the NHS display a ready assessment of the effectiveness with which it is meeting the needs and expectations of the people it serves ... Whether the NHS is meeting the needs of the patient and the community, and can prove that it is doing so, is open to question." (Griffiths 1983)

The professions - especially medicine - showed remarkably little interest in or enthusiasm for quality assurance. Editorials in the British Medical Journal asserted that "motives for audit in other countries had little relevance to the NHS", and that if audit had to be introduced it must be voluntary and totally under the control of the profession (British Medical Journal 1974). They also asked, rhetorically, "could we not rely on the innate sensitivity of the profession to the need to maintain standards rather than constantly thrust the minutiae of performance under the noses of those who are at the sharp end of clinical practice" (British Medical Journal 1976). Others offered more direct opposition, writing of the "chopping block of audit", asking why medicine should be singled out from other professions for this unwelcome attention, claiming that the best qualities of medical practice could not be audited, and asserting that "medical audit, if it set one doctor to assess or judge another, would be impracticable and even distasteful" (Practitioner 1980). Understandably, Maxwell (1984) wrote that the medical profession seemed "collectively allergic to rational examination of the case for medical audit in any form".

Interest in quality assurance in the UK grew during the 1980s, as the series of changes in managerial arrangements initiated by Griffiths took effect, and a number of local quality assurance programmes developed (National Audit Office 1988; Gruer, Gunn, Gordon et al 1986; Walshe, Lyons, Coles et al 1991). Indeed, in many public services - not just healthcare - there was an increasing recognition of the importance of service quality and client satisfaction (Pollitt 1990). In the health service, the climate of professional opinion was gradually transformed, so that when a major NHS reform programme was launched in 1989 with comprehensive medical audit as a key constituent (Department of Health 1989b), that aspect of the proposed reforms was generally accepted and welcomed by the medical profession. Medical audit was described as "an important professional obligation" which met the "need for a more systematic evaluation of the quality and effectiveness of doctor's work" (Royal College of Physicians 1989) and participation in medical audit became a requirement for the accreditation of training posts (Royal College of Surgeons 1989). Far from being voluntary, it was made clear by the leaders of the medical profession that all doctors should participate, and that nonparticipation would not be accepted (Standing Medical Advisory Committee 1990). Editorial comment in the British Medical Journal, in tones distinctly different from those of a decade earlier, declared that "the whole profession needs to claim ownership of audit and see a constant search for improvement as a central part of being a doctor" (Moss and Smith 1991).

The Department of Health issued guidance on the requirements for medical audit (Department of Health 1991b) and allotted substantial ringfenced financing to the programme - a total of £221 millions between 1989/90 and 1994/95 - with the result that the level of activity in quality assessment and quality assurance in the NHS expanded rapidly (Department of Health 1993b; Buttery et al 1994). At the same time a Patient's Charter, setting out some national quality standards was established, and national league tables of performance were produced and published (Bagust 1996). Some UK healthcare providers experimented with total quality management (Joss and Kogan 1995), and many professional organisations like the Royal Colleges became involved in leading or coordinating audit or quality assurance activities in their areas (Amess et al 1995)

Currently, almost all UK healthcare providers have well established medical audit, clinical audit or quality assurance departments, with some staff and other resources allocated to supporting quality measurement and improvement. While the nature of those activities varies very widely, and there is

evidence to suggest that the effectiveness of those activities is rather mixed the scale and substance of healthcare quality assurance activities in the UK today is very different from the position just a decade ago (Walshe 1995).

In other European countries, there has been a similar upsurge of interest in quality in healthcare over recent years, with mandatory quality assurance introduced in Germany, Belgium, France, Sweden and the Netherlands (Jost 1990; Øvretveit 1997). All European members of the World Health Organisation signed a declaration in 1985 that they would establish effective mechanisms for quality assurance in their healthcare systems by 1990 (World Health Organisation 1985).

2.4.3 **Objectives of quality assurance**

From this brief review of approaches to quality assurance in healthcare it is clear that the objectives of quality measurement and quality assurance are framed by their environment: the structure, financing and organisation of the healthcare system; the social and economic characteristics of the society it serves; and the culture and attitudes of the people who provide and use healthcare services. It is also evident that the objectives of these quality measurement and quality improvement programmes are often not well defined, which makes their evaluation rather difficult (Walshe and Coles 1993b). However, in almost any setting, some common principal objectives of quality measurement and assurance programmes can be identified:

a) *Improvement of quality*.

It is almost axiomatic that quality assurance programmes are designed to improve the quality of healthcare, though this is not necessarily their sole or even their primary objective. The aim of quality improvement may be expressed through goals such as the identification of substandard providers; the correction of substandard practices; the improvement of average performance (sometimes referred to as shifting the curve); and the identification and reward of superior performance (Institute of Medicine 1990, p46).

b) *Control of costs*.

Rising costs have been a perennial feature of most healthcare systems (especially over the last two decades). In the USA, healthcare costs amounted to 11% of GNP, or $600 billion in 1989 (Berwick, Godfrey and Roessner 1990, p5) and have consistently grown in real terms despite a range of measures designed to cap spending. Fuchs (1979) argues that quality assurance in the USA has been principally driven by the need to control costs, through more efficient, appropriate and consistent healthcare.

c) *Regulation of practice*.

Though self-regulation is traditionally part of the definition of a profession, growing demands for professional accountability in many spheres and accumulating evidence of confirmed poor professional practices which have not been rectified by the profession itself have increased the need for objective mechanisms for monitoring and regulating professional practice (Graham 1990). Quality assurance has been used to provide those mechanisms.

d) *Management of innovation and new technology*

The uncontrolled adoption of new technology in healthcare has been both a cause of spiralling healthcare costs and a source of controversy over the real benefits and harms associated with new services, treatments and procedures. Equally, the tardy uptake by clinicians and healthcare organisations of new practices of demonstrated effectiveness has also been a cause for concern (Appleby, Walshe and Ham 1995). Quality measurement and assurance programmes have a part to play in managing the dissemination of innovation.

e) *Control of clinical negligence litigation*.

The costs of litigation from patients over the quality of the healthcare services they have received have risen consistently over recent years, both in the USA and the UK. In 1989 the costs of malpractice insurance in the USA were estimated at $5 billion a year (Relman 1989), while in the UK litigation costs rose at 17% per annum over the 1980s (Tribe and Korgaonkar 1989; Dingwall and Fenn 1995). The presumed capability of quality assurance

to detect and prevent potential quality problems, reducing litigation costs, has often been an important factor in the development and direction of quality assurance programmes.

f) *Assuring and informing consumers*.

Consumers or users of healthcare have become increasingly willing to question the quality of service they encounter, and to make discriminating choices where possible between different providers. This trend has been evident both in the USA and the UK (where it can be seen as part of a larger trend towards greater consumer sovereignty in both public and private services) (Hopkins, Gabbay and Neuberger 1994). Quality assurance programmes can provide information to assure consumers and to allow them to make informed choices about the healthcare services they use.

g) *Assuring and informing purchasers*.

In the highly fragmented and decentralised healthcare system of the USA, quality assurance has been an important source of information on which purchasers of healthcare - state and federal agencies, insurers, health maintenance organisations, and so on - have made purchasing decisions. Indeed, it may be contended that this demand for information has been the most important single cause of the development of the extensive quality assurance industry in US healthcare. By contrast, the absence of a healthcare marketplace in the UK's centrally planned NHS until 1989 was accompanied by a marked dearth of significant quality assurance programmes. The development of the internal market in healthcare within the NHS over recent years was accompanied by a growth in quality assurance activities, at least in part to provide information to healthcare purchasers (Gill 1993).

In practice, the objectives of quality assurance programmes are rarely spelt out, and expressed in terms which make their evaluation against those aims possible. This lack of clarity about the objectives of quality assurance is an important problem. It hampers programme implementation, makes monitoring and progress assessment more difficult, and hamstrings any proper programme evaluation.

2.4.4 **Effectiveness of quality assurance**

There is extensive evidence, through reports from participants and observers of quality assurance programmes, that they can be effective in achieving at least some of the objectives set out above, in some circumstances. However, the available evidence from publications is really only able to demonstrate their *efficacy* (that they can work in some circumstances), and not their *effectiveness* (that they do work, generally) - an important distinction, which Brook and Kosecoff (1988) defined in relation to healthcare interventions in general. A brief review of this literature demonstrates the striking absence of consensus about the effectiveness of quality assurance.

For example, Shaw (1989, p11) cites several examples of successful programmes improving resource usage and producing better outcomes for patients. Gruer et al (1986) report statistically significant improvements in reoperation and operative mortality rates resulting from a five year surgical audit programme. Sellu (1986) describes significant improvements in postoperative infection rates and cost savings resulting from a similar audit programme. On a larger scale, the Confidential Enquiry into Maternal Deaths has been claimed to have improved maternity care in the UK (Department of Health 1991a). Craddick (1979) cites a series of examples of important individual quality improvements resulting from an occurrence screening programme. Kleefield, Churchill and Laffel (1991) report real improvements in pharmacy services resulting from a continuous quality improvement (CQI) programme. Gabbay et al (1990) found significant improvements in recordkeeping resulting from a year-long criterion audit programme, and Reynolds (1995) showed that a CQI programme in obstetrics reduced episiotomy rates significantly. O'Connor et al (1996) demonstrate the effectiveness of multi-institutional comparisons of practice in reducing mortality following cardiac surgery. Walshe and Buttery (1995) highlight the complexities of measuring the impact of quality assurance programmes and Palmer et al (1996) report the results of an experiment intended to identify the factors which affect the impact of quality assurance.

At the same time, there are many reports in the literature of quality assurance programmes which have not met the expectations of participants, observers or funding agencies. Twenty years ago, Nelson (1976) argued that criterion-based process audits were ineffective, and simply produced

"vast amounts of unusable data". Escovitz et al (1978), reviewing the effects of quality assurance in 17 US hospitals, found that extensive criterion-based audits had resulted in little or no action to address quality problems. Sanazaro and Worth (1978) studied the effects of concurrent quality assurance across 50 hospitals in the United States, and found virtually no evidence that it was effective in making any improvements. Sanazaro and Mills (1991), reviewing the widespread use of generic occurrence screening in the USA, conclude that it provides information which is of limited use and minimal relevance to quality assessment. Lohr (1990), reviewing the results of an extensive evaluation of the Medicare Peer Review Organisations' work in quality assurance, observes that the $300 million pa programme is "in general, not very effective, and may have serious unintended consequences". Within the UK, Krukowski and Matheson (1988), relating the experience of a decade of surgical audit, question the value of the "onerous and obsessional collection of data .. together with its time-consuming analysis" which their study involved, and with some understatement observe that "in terms of improved patient care, the benefit of this audit lacks tangibility". Walshe (1995b) suggests that while some quality assurance activities in the UK have been effective and worthwhile, others have not, indicating that effectiveness is very variable. Hopkins (1996) criticises the absence of tangible benefits from the investment in clinical audit in the UK, and suggests that much quality measurement and improvement activity lacks methodological rigour, while others argue that it is not possible to tell what either the benefits or the costs of clinical audit in the UK have been (Lord and Littlejohn 1997).

Reviews of the literature produce a similarly mixed message. Mitchell and Fowkes (1985) reviewed the use of a variety of approaches to using information to change performance, and concluded that the evidence for their effectiveness was poor. Mugford, Banfield and O'Hanlon (1991) examined 36 studies of the effect of audit in a wide range of settings, using a variety of feedback mechanisms, and found that in 24 (67%) there was evidence of lasting practice changes and quality improvements. A systematic review of experimental, quantitative studies of the use of audit and feedback to change clinical practice is being prepared for the Cochrane Library by the Cochrane Collaboration on Effective Professional Practice (Thompson et al 1997) and is expected to conclude, like the two reviews that have preceded it, that the evidence for the effectiveness of quality assurance is rather limited and inconclusive.

Bearing in mind the acknowledged bias of published literature in favour of positive results, it is notable that the literature yields perhaps as many accounts of unsuccessful quality assurance programmes as of demonstrably successful ones - and it is unfortunate that in most reports, the impact of quality assurance activities on the objectives set out above is not assessed.

It seems, however, that the main obstacles for quality assurance lie more in the domain of quality improvement than in quality measurement. Brook (1977) and Jessee (1977) both highlight this issue, arguing that quality measurement techniques have been able to identify the need for change, but quality improvement techniques have not been able to ensure that change then takes place. Brook observed:

> "The central failing of quality assessment is that it has rarely been used to change behaviour and hence has not contributed much to the goal of improving the health of the American people. The literature on quality of care is replete with studies showing deficiencies in medical care no matter what standard or method is used." (Brook 1977)

Shaw (1980b), surveying the available evidence for the effectiveness of audit, admits it is "not overwhelming" and acknowledges that there is conflicting evidence on the ability of audit to bring about change.

The absence of conclusive evidence of the effectiveness (or, indeed, the ineffectiveness) of healthcare quality assurance is partly a result of the complexity of the mechanisms and behaviours involved. Quality measurement is complicated by a range of definitional and methodological hurdles; quality improvement faces similar organisational and behavioural obstacles. There have been many calls for clearer objective setting and better evaluation in the field of quality measurement and quality assurance. For example, Phelps wrote candidly:

> "The evidence used to justify these various regulations of quality has been, to be generous, sparse. The primary justification for many of these programs is that the

> present state of affairs is scandalous, so that change must lead to improvement. While not necessarily quarrelling with the premise, the conclusion is unwarranted. It therefore seems appropriate to begin to evaluate the evaluators, to develop a framework with which one could assess the gains … from undertaking a quality assurance program of one type or another." (Phelps 1976).

More recently, Carr-Hill and Dalley (1992), reporting on a survey of quality assurance activity in the UK, wrote:

> "In the NHS our knowledge about our own quality assurance activities is abysmal. The vast majority of [quality assurance] procedures are not costed; more than half are not evaluated, and where they are, the procedures used do not appear to be a model of rigour. ... What is missing is accurate description and monitoring of what goes on … Without this, management and quality assurance is a nonsense." (Carr-Hill and Dalley 1992)

The need for quality assurance activities - including both quality measurement and quality improvement - to be properly evaluated is clear.

2.5 Evaluating quality measurement

2.5.1 **Theory, concepts and dimensions in evaluating quality measures**

St Leger, Schneiden and Walsworth-Bell (1992, p1) define evaluation as "the critical assessment, on as objective a basis as possible, of the degree to which entire services or their component parts fulfil stated goals". This places the goals or objectives of the programme to be evaluated at the centre of its evaluation - which presents problems when those goals are unstated or implicit, or are inadequately defined, as is often the case in quality measurement and quality assurance.

However, an alternative approach to the evaluation of quality measurement or quality assurance programmes is to consider the core objectives that most or all programmes would (or should) share (Walshe and Coles 1993). While each quality measurement technique or quality assurance programme should have its own, specific goals, to which evaluation must be tailored to some extent, it is possible to draw up a list of desired attributes, components or characteristics of most or all techniques or programmes. Although few researchers have addressed the issues of evaluation in any detail (Walshe and Coles 1993), the Institute of Medicine (1990), in its extensive review of the working of the Medicare quality assurance programme, established generic frameworks for evaluation which are certainly of wider relevance.

Quality measurement techniques evaluation

The Institute of Medicine (1990, p312) suggests that quality measurement techniques should be assessed against a set of structural attributes or dimensions - to do with the inherent characteristics of the tool itself - and a set of process attributes - to do with the way in which the tool is applied. The somewhat daunting list of proposed structural attributes is shown in table 2.3.

Four main clusters or groupings of attributes can be determined. Certain attributes concern the *scientific grounding* of the instrument (issues such as its reliability and validity, and the quality of its documentation). Others concern the *latitude* the instrument provides for variations in patient characteristics and clinical circumstances - such as its flexibility, appealability, patient responsiveness, clinical adaptability and inclusiveness. A third group of attributes concerns the instruments *general design* - issues such as its clarity, the concordance of professionals, its acceptability and its appropriateness. The last group of attributes concern its *efficiency* - specifically, its sensitivity and specificity.

Sensitivity	High true positive rate in detecting deficient or inappropriate care.
Specificity	High true negative rate in passing over cases of adequate care.
Reliability	Known to produce the same decisions or evaluations when applied by the user groups for which the tool is intended.
Validity	Based on outcome studies or other scientific evidence of validity.
Documentation	Documents methods of development and cites literature.
Patient responsiveness	Allows for eliciting or taking account of patient preferences.
Flexibility	Respects the role of clinical judgement, with clinical judgement explicable
Clinical adaptability	Allows for or takes into consideration clinically relevant differences among different classes of patients; population to which tool applies is specified.
Inclusiveness	Covers all major foreseeable clinical situations and full range of clinical problems.
Concordance from	Reflects consensus of professionals with extensive experience in field, with input academic and nonacademic practitioners, specialists and generalists.
Acceptability	Acceptable to the majority of professionals.
Clarity	Written in unambiguous language; terms, populations, data elements, and collection approach clearly defined.
Appropriateness	Specifies appropriate, inappropriate, and equivocal indications (procedure and technology appropriateness guidelines).

Table 2.3. Structural attributes for evaluation of quality measurement techniques. Institute of Medicine (1990).

The slightly shorter list of proposed process attributes of quality measurement techniques is shown in table 2.4. Again, four major groups of attributes can be discerned. The *ease of use* of the instrument is addressed by the attributes of comprehendability, manageability and nonintrusiveness. The *ease of implementation* of the instrument is the focus of issues such as its feasibility, computerisation and executability. The *latitude* for discretion offered by the instrument is addressed by issues such as flexibility and appealability. Finally the *progressive refinement* of the instrument is addressed by its pretesting, its dynamism and its evaluation.

Pretesting	Tool is tested before implementation.
Dynamism	Mechanism and commitment exists for reviewing and updating tool to incorporate new information and cover new situations.
Evaluation	Mechanism exists to review and evaluate outcome or impact of tool.
Comprehendability	Format of tool easily understood by nonphysician reviewers, by practitioners, and by patients and consumers.
Manageability	Not unduly burdensome for nonphysician reviewers to apply, for physician reviewers to apply, or for professionals to follow.
Nonintrusiveness	Minimises inappropriate direct interaction with treating physicians.
Appealability	Allows for appeals process for professionals and patients.
Feasibility	Ease of obtaining information.
Computerisation	Has been or could easily be computerised.
Executability	Includes instructions for implementation.

Table 2.4. Process attributes for evaluation of quality measurement techniques. Institute of Medicine (1990).

While there is clearly some overlap between different attributes, these checklists of desirable attributes of quality measurement techniques are undoubtedly potentially useful. The Institute of Medicine made some efforts to weight the attributes (or at least to identify the more important ones) through the use of an expert panel, which rated each attribute in terms of its relative importance. Clarity, validity and sensitivity emerged as the most important structural characteristics, while appealability was the most critical process attribute. These weighting perhaps reflect the healthcare environment of the US, and its emphasis on inspection.

Quality assurance programme evaluation

Similarly, the Institute of Medicine (1990, p50) suggests an ambitious list of fifteen attributes against which quality assurance programmes (as opposed to quality measurement techniques) might be evaluated. These attributes are listed in table 2.5.

Range	Ability to address a full range of quality problems, including poor technical quality, overuse and underuse.
Intrusion	Intrudes minimally into the patient-provider relationship, and does not jeopardise trust, clinician best judgement or patient autonomy.
Acceptability	Acceptable to professionals and providers - seen as supportive to practitioner's goals.
Improvement focus	Fosters improvement through the healthcare system - is focused on improving processes of delivery of care.
Outliers	Able to identify and ameliorate outlier practice - individual practitioners or institutions with problematic patterns of care.
Incentives performance.	Can invoke positive and negative incentives for change and improvement in
Information	Provides practitioners and providers with timely and relevant information which they can use to compare and to improve their practice.
Face validity reasonable to both.	Has face validity to the public and professionals - must be understandable and
Scientific worth	Individual elements meet requirements for reliability, validity and generalisability - programme must demonstrate scientific rigour.
Outcome focus	Improves patient outcomes. Ultimately, a quality assurance programme should affect patient outcomes. When outcomes cannot be measured directly, there should be a demonstrated link between healthcare processes and expected patient outcomes.
Individual/population	Can address both individual patient and population based outcomes.
Documentation	Documents improvements in quality and progress towards excellence.
Ease of use	Easily implemented and administered. Excessively costly, complex and labour intensive programmes may detract from quality themselves.
Cost	Is affordable and cost-effective.
Public participation	Includes patients and the public.

Table 2.5. Attributes for the evaluation of quality assurance programmes. Institute of Medicine (1990).

Again, while the length of the list of attributes is perhaps a little offputting, its comprehensiveness and conceptual clarity make the Institute of Medicine's framework potentially valuable in evaluating quality assurance programmes. Like almost any other checklist, it is possible to identify areas of overlap (such as between ease of implementation and use and affordability/cost effectiveness) and

potential conflicts (such as between minimal intrusion into practitioners' best judgement and the ability to change outlier practice and invoke sanctions for change). Nevertheless, these frameworks developed by the Institute of Medicine are of worth, if only because few (if any) other researchers have developed similar tools. Some of the dimensions of evaluation identified and discussed above are more amenable to assessment than others, but their inclusion is still important, if only to ensure that we do not fall into the trap of measuring only what is measurable and failing to take account of important but less assessable dimensions of performance.

Since the focus of this thesis is the reliability and validity of some quality measures based on adverse events in healthcare, it is appropriate to examine the meaning of these concepts in more detail. They are important components in the structural attributes set out for evaluation by the Institute of Medicine in table 2.3, but it should be recognised that the other attributes described there also deserve consideration.

2.5.2 **Validity of quality measurement**

In all measurement, there exists both random and non-random error. A simple mathematical definition of the validity of a measure is that it is inversely related to the extent of non-random error. By the same token, the reliability of a measure is inversely related to the extent of random error (Carmines and Zeller 1979, p13).

More usefully, validity is commonly defined as "the extent to which a measuring instrument measures what it is intended to measure" (Carmines and Zeller 1979, p17). However, this definition has been criticised because it places too much emphasis on the measure itself, and too little emphasis on the environment in which it is used or the phenomena it is used to measure. "One validates not the measuring instrument itself, but the measuring instrument in relation to the purpose for which it is being used" (Carmines and Zeller 1979, p17). In the context of healthcare quality measurement, it is important to validate an instrument in the healthcare environment it which it is intended for use.

There are a number of different approaches to assessing validity, each of which measures different aspects of the extent to which a measure is appropriate for use in a given set of circumstances. It is often recommended that validation studies use a variety of approaches rather than depending on a single type of validation procedure (McDowell and Newell 1991, p31).

Criterion-related validity

Criterion-related validity is concerned with the relationship between a measure and some external variable (the criterion) with which it is expected to correlate. The degree of correlation between the measure and the external variable or criterion (which is also often termed the *gold standard*) indicates how well the measure represents the criterion, and is frequently referred to as the validity coefficient.

For example, one might assess the criterion-related validity of a physical health status measure by correlating it with objective measurements of patients' physical health such as certain laboratory tests. The scientific tests would constitute the gold standard for physical health, and the degree of correlation would indicate the validity of the measure.

Criterion-related validity is also often termed *concurrent validity* or *predictive validity* (where the criterion is measured at some future point). It has been observed that, of all the types of validity, criterion-related validity has the greatest intuitive meaning and closest relationship with the everyday usage of the term validity (Carmines and Zeller 1991, p17). Its main disadvantage is that in many circumstances, no obvious criterion variable exists. For example, in measuring certain dimensions of healthcare quality, such as patient satisfaction, it may be difficult to identify any objective criterion. There may be a temptation to use as a criterion some measure of another dimension of quality, such as clinician compliance with treatment protocols, but this involves making unfounded and precarious assumptions about the relationship between different dimensions or attributes of quality.

Criterion-related validity is measured empirically by correlating the results of the measure with the criterion variable. The selection of appropriate statistical tests to measure this correlation depends on the nature of the two variables (Dick and Hegarty 1971). However, where true or forced

dichotomous data is available, the concepts of sensitivity and specificity are frequently used in assessing validity. The sensitivity of the measure is the proportion of cases positive on the criterion variable which are assessed as positive on the test variable while the specificity is the proportion of cases negative on the criterion variable which are assessed as negative on the test variable. The sensitivity can be construed as the positive accuracy of the measure - the higher it is, the more successful the measure is at identifying accurately those cases where the criterion variable is positive. Similarly, the specificity can be construed as a measure of the negative accuracy of the test. The acceptability of the sensitivity and specificity statistics for a measure depend crucially on the use to which the measure is to be put. For example, a high sensitivity but low specificity measure will successfully identify almost all positive cases, but will also throw up a considerable number of false positives. This may be quite acceptable in a measure designed to identify possible drug prescribing problems for subsequent review by a pharmacist, since the false positive cases will be discarded at that review. By contrast, it would be quite unacceptable in a measure used in a population screening programme to assess the risk of bowel cancer, since it would entail large numbers of healthy people being subjected to expensive, unpleasant and unnecessary further diagnostic tests, not to mention considerable temporary anxiety.

Content validity

Content validity refers to how adequately the items within a measure reflect the conceptual definition of its scope. (McDowell and Newell 1991, p27). For example, in a patient satisfaction measure, the content validity would be assessed by considering whether the individual items appeared relevant to the concept of satisfaction, and whether all aspects of the concept of satisfaction were covered.

Content validity is assessed by studying the measure itself, rather than data produced by using the measure. It is generally assessed through the use of expert panels or through a review of the measure against available literature or other evidence as to what makes up the concept. It has been pointed out that content validity is harder to assess for more abstract concepts (such as quality of healthcare) than for simpler, more concrete concepts (such as driving proficiency), since it is generally easier with more concrete concepts to define their scope.

Construct validity

Construct validity is concerned with the extent to which a particular measure supports or conforms with a given theory or construct, and begins with the theoretical definition of the construct, its relationship to other constructs, and its relationships with observable behaviours or phenomena (Ghiselli, Campbell and Zedeck 1981, p284). It is particularly important when the measurement of criterion validity is difficult or impossible, because of the lack of an appropriate criterion variable - a situation which arises more frequently when the concept or construct being measured is complex or abstract.

For example, if a theory suggests that patients' satisfaction with their care is related to a number of cultural and demographic factors - such as their ethnic background, social class, and age - then the construct validity of a patient satisfaction measure might be tested by examining whether it did indeed demonstrate differences in satisfaction as the theory predicts. Construct validation has been called "part science, but to a large extent an art form" (McDowell and Newell 1991, p31), perhaps because it sits at the interface between theory testing and theory building, but it is generally regarded as a valuable addition to the developer's armoury of testing mechanisms.

Construct validity is usually measured empirically, by examining the results of the measure being studied alongside other data which can be used to test whether the relationships predicted by theory actually exist.

Face validity

The face validity of a measure refers to whether users or potential users of the instrument report that, on the face of it, the measure seems reasonable and produces reasonable data (Horn and Horn 1986). As its definition makes clear, face validity is an inherently subjective concept, and it is not difficult to imagine circumstances where the production of unexpected but valid results from a study might lead to an unwarranted low estimate of the face validity of the measures used. Conversely, it is possible for a measure to seem reasonable and to produce apparently reasonable data because of our incomplete understanding of the phenomena it measures, while in fact the measure is not valid.

Assessments of face validity, especially in areas of measurement which are incompletely understood or explored, deserve to be treated with some caution.

Like content validity, face validity is not measured empirically. Instead, surveys of users of the measure are employed to assess its face validity.

Other concepts in validity

The discriminant validity of a measure is its ability to differentiate between different categories or groups of respondents or cases. It is often investigated when such discriminatory ability is the crux of the measure's worth. For example, the validity of a measure of psychiatric well-being, which is intended to identify people at risk of psychiatric illness, might be tested by assessing its ability to distinguish between samples of people who have (or have not) been users of psychiatric services.

Some researchers have used factor analysis to investigate the factorial validity of measures - a concept which is closely related to that of content validity. Essentially, factor analysis can show how far the various items in a test accord in measuring one or more common themes. When a measure deliberately sets out to yield separate scores on several dimensions (such as a measure of general health status such as the Nottingham Health Profile) factorial validity can be assessed by identifying whether the items belonging to each dimension actually measure a common theme.

Finally, Dick and Hegarty (1971) coined the appealing (though perhaps not very scientific) term *cash validity* to describe the validity a measure acquires simply from being widely used and applied. Though they offer no empirical measure of the concept, it is undoubtedly true that measures do acquire some validity simply from widespread use.

2.5.3 **Reliability of quality measurement**

Reliability is concerned with the random errors which occur in all forms of measurement (non-random errors constitute bias, and are the province of validity assessment). The reliability of a

measure is an indicator of the degree to which it can be replicated (McDowell and Newell 1991, p31).

There are three forms of reliability which are commonly assessed: interrater reliability (also termed observer variation); intrarater reliability (also termed stability or repeatability); and internal consistency. All are important to the developers of measures of healthcare quality.

Interrater reliability

Interrater reliability measures whether, when the same test is applied to the same respondent or subject by different raters or observers, the same results are produced. In other words, it assesses the extent of random variation which results from the use of different raters to apply the measure. There may be logistical problems associated with having two applications of the measure, depending on its nature (for example, reinterviewing a subject) which distort the results.

For example, the interrater reliability of a measure based on a set of treatment standards might be tested by using different raters to apply the measure to the same sets of casenotes, and comparing their results. Both interrater and intrarater reliability are fundamentally affected by the arrangements which are made for rater training and guidance - few measures are wholly self-explanatory, and a rater with no training or experience in using an instrument will certainly be less reliable than one who has been properly trained in how to apply the measure. While the developers of instruments usually cite results of interrater and intrarater reliability tests, it must be borne in mind that instrument developers may be able to train raters more thoroughly than ordinary users of the instrument, and may also be able to use raters of different levels of ability and commitment.

Interrater reliability is measured empirically by measuring the correlation between the scores of the two (or more) raters. The nature of the correlation statistic used depends on the type of data yielded by the measure, and the number of raters.

Intrarater reliability

Intrarater reliability (also known as test-retest reliability, stability or consistency) measures whether, when the same test is applied twice by the same rater or observer to the same respondent or subject,

the same results are produced. It assesses the extent of random variation which results from reapplication of the measure. It suffers from two main logistical problems - that the time interval between test must be brief enough to prevent changes in the characteristic being measured affecting the results, and that carryover effects from the first to the second application, such as practice or learned responses, may distort the results.

For example, the intrarater reliability of an adverse-event measure of healthcare quality might be tested by arranging for a rater to apply the instrument to the same sets of casenotes on two occasions, sufficiently far apart to minimise the effects of memory.

Intrarater reliability is measured empirically by measuring the correlation between the scores of the two applications of the measure. As for interrater reliability, the nature of the correlation statistic used depends on the type of data yielded by the measure, and the number of raters.

Internal consistency

Internal consistency measures how well the items within a measure which is designed to measure a particular characteristic correlate with each other. If the measure is measuring a single trait, then a reliable measure should show high inter-item correlations. If inter-item correlations are low, then it suggests that the measure may actually be measuring different traits or characteristics, and factor analysis might then be used to identify these separate themes.

For example, internal consistency tests might be used to examine the performance of a functional disability scoring system used in assessing the improvements in functional disability resulting from orthopaedic surgery. By testing the inter-item correlations within the measure, using data from its application to sets of cases, the ability of the measure to focus on the trait of functional disability could be assessed.

Internal consistency is usually measured through the use of Cronbach's alpha statistic, a generalised form of the split-halves approach to reliability testing which effectively calculates the mean correlation of each item in the instrument with every other item (Ghiselli, Campbell and Zedeck 1981, p256).

Other issues in reliability measurement

There are two alternative theoretical models of reliability, which offer different conceptual approaches to assessing the reliability of measures. The classical theory school postulates that observed scores are made up of two components - the true score and the error score. It assumes that the subject of the test possesses stable characteristics or traits which can be measured and remeasured, and that all error scores are wholly random. Approaches to measuring reliability are founded on assessing correlations between parallel tests of various forms. The assumptions of the classical school are clearly open to challenge. In particular, the error components of scores may not be randomly distributed, since different raters or cases may have different distributions of error scores. However, it can be demonstrated statistically that this stringent assumption can be relaxed without invalidating the mathematical logic which leads to classical school formulae for reliability coefficients (Ghiselli, Campbell and Zedeck 1981, p208).

By contrast, the domain sampling (or generalisability theory) school of thought suggests that the true score/error score model on which classical theory is based is too simplistic, especially for behavioural and social sciences where traits may frequently not be stable and where error scores may not always be wholly random. It suggests instead that, through some complex statistics, the reliability of a measure can be assessed by correlating it with hypothetical parallel tests in the same domain (Ghiselli, Campbell and Zedeck 1981, p195). While the statistical formulae for reliability produced by the two theoretical approaches are identical, their differing assumptions produce different assessments of the value of the empirical approaches to reliability testing discussed above. The concepts of interrater and intrarater reliability testing are founded in the ideas of the classical school, and are in many ways incompatible with domain sampling theory, which places greater weight on tests of internal consistency.

Finally, the reliability of a measure is also related to its validity. It can be demonstrated mathematically that the reliability of a measure places an effective limit on its criterion-related validity, and that the lower the reliability coefficient is, the lower the maximum achievable validity coefficient can be. However, adjustment formulae allow validity coefficients to be corrected for this attenuation, so that the hypothetical validity (assuming perfect reliability) can be calculated (Carmines and Zeller 1979, p48).

2.6 Conclusions

The quality of healthcare is a difficult concept to define in simple terms. There is no single generally accepted definition, and the content of the concept varies considerably depending on the perspectives of those using it and the setting in which it is used. When drawn widely, the concept of quality seems to embrace almost everything which is germane to performance and evaluation in healthcare. It is, perhaps, more usefully deconstructed into a number of constituent parts or dimensions which are more capable of being conceptualised and defined and therefore more amenable to measurement.

Quality measurement involves both the collection of data describing the structure, process or outcome of the healthcare system and the development and application of valuations which give meaning to that data by defining what represents good quality. A wide variety of quality measures have been developed and are in use, and they form an important part of quality assurance activities in healthcare in the UK and abroad. The need for quality assurance in healthcare, long established in some countries like the USA, has been increasingly recognised in other countries in Europe and the developed world, and investment in quality measurement and improvement has grown apace. However, there are important unanswered questions about the worth of these quality assurance systems. There is a need to evaluate them more rigorously, both during their development and when they are applied in practice.

Chapter 3

Adverse events and the quality of healthcare

3.1 Introduction

This chapter presents a comprehensive review of the place of adverse events in healthcare. It begins by examining the development of interest in adverse events from a number of different perspectives, concerned with epidemiology, health services research, and quality improvement. In each, the need to be able to identify, measure, analyse and understand adverse events, and to have some form of definition and classification or typology as a basis for measurement, is established.

Next, the definition and classification of adverse events in healthcare is explored. A number of alternative definitions from the literature are compared and contrasted, and a single definition is proposed. Then, some of the frameworks proposed in chapter 2 are used to develop some criteria on which adverse events could be classified. Some issues involved in measurement - such as the sources of data, arrangements for its collection, and construction of quantitative measures - are considered.

The chapter then reviews the use of adverse event measures of quality in healthcare. Firstly, a number of different measures which have been proposed or used are presented and discussed. Then, the epidemiology of adverse events in healthcare is explored, through a review of studies which have examined empirically the incidence, characteristics and causation of adverse events. The diversity of definitions and approaches evident from the studies highlights the need for rigour in defining the approach to measurement discussed earlier. Nevertheless, some common themes and issues raised by these empirical studies are identified and discussed.

Next, the chapter turns to explore the experience of using adverse-event measures of quality in healthcare quality assurance activities. Experience from healthcare organisations in a range of

countries and settings is presented and discussed, and the importance of adverse-event measures of quality (in terms of their widespread application) is highlighted.

Finally, the chapter explores the relatively limited literature on the validity and reliability of adverse-event measures. A number of studies which have attempted to assess validity, reliability or both are presented and critiqued. It is noted that there are methodological problems associated with measuring some dimensions of validity and reliability, and that the absence of published literature on some aspects of validity and reliability is of some concern.

3.2 The concept of adverse events in healthcare

3.2.1 The origins of the adverse event

Adverse events in healthcare can be loosely defined as "instances which indicate or may indicate that a patient has received poor quality care" (Walshe, Bennett and Ingram 1995). The idea that it would be useful or important to study the incidence, circumstances or causes of adverse events in healthcare arises from a number of different but related schools of thought. For example, researchers concerned with various forms of iatrogenic disease, those interested in organisational performance, others investigating medical malpractice, clinical negligence and litigation, those concerned with quality measurement and improvement - all have developed approaches to studying adverse events in healthcare which, though they reflect the different interests of their developers have much in common. Below, each is briefly reviewed and discussed in turn, before some common themes and concerns are identified.

Iatrogenesis

While the term *iatrogenesis* was not coined by Ivan Illich, his influential writings are probably responsible for its widespread use today. Illich defines clinical iatrogenesis as:

> "the adverse or undesirable effects of healthcare on patients... Clinical iatrogenic disease comprises all clinical conditions for which remedies, physicians or hospitals are the sickening agents" (Illich 1975, p22).

In one of the first comprehensive reviews of iatrogenic disease, Moser (1956) drew on over 140 research studies to highlight the causation of a wide range of clinical conditions which he labelled "diseases of medical progress". He argued that "the history of medicine is replete with examples of illness resulting from sound therapeutic endeavour... However, in recent years the development of potent new therapeutic agents, improved surgical procedure and more efficient equipment has forced this facet of medicine into unprecedented prominence". Writing around the same time, Barr (1955) was one of the first researchers to attempt to quantify the impact of iatrogenic disease on patients and healthcare organisations. He estimated that about 5% of hospital admissions resulted in or were caused by some form of iatrogenesis which qualified "iatrogenic disease [as] one of the commonest conditions encountered".

Since Moser's review, an enormous and ever-growing literature on the iatrogenic effects of individual therapies and healthcare interventions has developed. Sartwell (1974) surveyed iatrogenesis from an epidemiologist's perspective and catalogued an alarming series of "iatrogenic epidemics". Kane (1980) classified iatrogenic diseases in four categories - those resulting from known risks of therapy; those arising from unknown or unexpected risks; instances of inept care (lack of skill, errors of judgement, inadequate knowledge, process failures, etc); and those resulting from unnecessary therapies or what he termed overzealous care. There are 6,138 studies catalogued on Index Medicus under the MeSH term "iatrogenic disease" (from 1966 to 1997), with 241 reports in 1996 alone, covering issues as diverse as iatrogenic damage to the facial nerve during surgery (Prass 1996), iatrogenic cardiac tamponade (Yim, Lam and Haines 1996) and iatrogenic congestive heart failure (Rich et al 1996). But while most of these reports concern the iatrogenic effects of an individual drug, or a particular surgical intervention, a limited number have focused on the incidence, causation and prevention of iatrogenesis in operational settings, such as ordinary acute hospitals serving a general population. It is these investigations which provide the greatest insight into the importance of iatrogenesis for healthcare organisations. As long ago as the 1960s, Schimmel (1964) found that 16% of inpatients in general medicine suffered some form of iatrogenic

disease during their stay, and McLamb and Huntley (1967) found similarly that about 20% of inpatients in their general medical service had some form of iatrogenic disease while in hospital. These and a number of other studies of the epidemiology of iatrogenic disease are reviewed later in this chapter. Research into iatrogenesis seems to suggest that the rapid pace of technological advance in healthcare has its costs, in the growing risk of iatrogenic effects from increasingly complex and potentially toxic therapies (Friedman 1982).

Using adverse events as measures of performance

David Rutstein, an American public health physician, was one of the first people to recognise the importance of adverse events to healthcare quality measurement. He designed, with colleagues, a system for measuring the incidence of *sentinel health events* which he defined as "unnecessary disease, disability or untimely death" (Rutstein et al 1976). Writing about the theoretical background to this methodology, Rutstein and colleagues observed:

> "Most previous efforts to measure quality have failed because of the almost insurmountable difficulty of establishing objective criteria for the measurement of increasing gradations of positive health. Our proposed system overcomes this difficulty by establishing quantitative negative indexes of health. Cases of unnecessary disease, and unnecessary disability, and unnecessary untimely deaths can be counted. Their occurrence is a warning signal, a sentinel health event, that the quality of care may need to be improved." (Rutstein et al 1976).

Rutstein argued that quality control systems based on negative indexes were able, through the study of undesirable health events, to yield crucial information on the causes of those events which could be used to make improvements, and that to focus on such events was a much more productive way of using resources than to attempt to measure global characteristics of all patients.

Rutstein's method was based around an inventory of conditions, tabulated in the standard International Classification of Diseases, which a working group of physicians and epidemiologists had deemed partially or wholly preventable or treatable. With his public health perspective, Rutstein included a whole range of diseases resulting from non-healthcare preventable causes (such

as cancers resulting from unnecessary exposure to carcinogens, and diseases resulting from poor diet or housing). Rutstein applied his own techniques in studying occupational diseases (Rutstein et al 1983), while others have used it to examine geographical variations in avoidable mortality (Charlton et al 1983) and to explore levels of unmet need for healthcare in the population (Carr et al 1989).

Measures of the quality of healthcare based on the detection and classification of adverse events were first developed and used in the USA in the 1970s. In the light of four years experience of mandatory quality assurance in the USA which had largely been based on the development of standards based audits of the process of care, Komaroff (1978) wrote of the growing awareness that these efforts had made little or no progress in establishing meaningful measures or in improving the quality of care. He argued for the simplification of programmes, and for a focus on areas in which experience suggested there were "deficiencies in care". Other commentators also urged less emphasis on complex, standards based measures of process and a greater attention to adverse events, critical incidents, or instances of what Fifer termed *disquality* (Craddick and Bader 1983, p3). During the 1980s, a growing number of adverse-event measures of quality were developed and applied in the USA (Craddick and Bader 1983; Goldman 1989; Sanazaro and Mills 1991; and others), Australia (Burr 1990; Wilson et al 1995; Wolff 1995), Canada (Carlow 1988; Nordal and Ang 1988), the UK (Bennett and Walshe 1990; Walshe, Bennett and Ingram 1995; and others) and Holland (Bomhof, Arends and van der Beek 1993; Bomhof, Nieman and Reerink 1993). These and other studies are reviewed in some detail later in this chapter.

The critical incident technique

Flanagan (1954) described the critical incident technique as "a set of procedures for collecting direct observations of human behaviour in such a way as to facilitate their potential usefulness in solving practical problems and developing broad psychological principles". It is a tool for "collecting observed incidents having special significance and meeting systematically defined criteria". Flanagan cites as examples of the early application of the critical incident technique its use in understanding near-miss accidents in flying, errors in reading and interpreting aircraft instruments, army personnel failures in highly stressful emergency or combat situations, and the study of motivation and leadership in the armed forces. In each case the approach involves identifying

critical incidents (which might often be thought of as adverse events) and exploring their causes and impacts in order to develop a wider understanding of the area being studied.

The critical incident technique has been used widely in many areas of research related to healthcare. For example, it has been employed to explore what causes clinicians to change their practice (Allery, Owen and Robling 1997), and what factors influence general practitioners' prescribing decisions (Bradley 1992). Norman et al (1992) applied it to develop indicators of quality in nursing care from the critical incidents identified by patients and their nurses.

The focus of most critical incident studies is a set of events which are particularly significant to participants because they represent either instances of good or poor performance from which lessons can be learned. Sometimes, the definition of those events is left to the participants in the study, but broad indicative areas of interest are often suggested. There is clearly a link to other approaches concerned with adverse events. Though critical incidents are not necessarily adverse in nature, and can represent exemplars of good practice or performance, they are more usually instances of some concern.

The critical incident technique has been used to study adverse events in healthcare, where it has sometimes been called "significant event auditing" (Robinson, Stacy and Spencer 1995; Pringle et al 1995). For example, Berlin, Spencer and Bhopal (1992) employed it to audit deaths in general practice, and identified a series of avoidable factors in the cases studied which led to changes in clinical practices and organisational arrangements. Diamond, Kamien and Sim (1995) used it to study the learning experiences of trainees in general practice and found that many incidents concerned participants' communication and interpersonal skills, and knowing when and how to obtain support in difficult or complex clinical decisions. Some studies of the quality of care, such as the various confidential enquiries into perioperative deaths, maternal deaths and neonatal deaths (NCEPOD 1989; Department of Health 1991a) can be seen as using the critical incident technique implicitly, both in their approach to identifying significant events and in the way in which those events are analysed and used to develop an understanding of causation and prevention.

It is notable that studies of adverse events in healthcare which use the critical incident technique place considerable emphasis on the importance of a deep understanding of the circumstances surrounding the event, often based on a detailed discussion with participants. Although these studies may involve the analysis of large numbers of adverse events, they are usually qualitative in nature, focused on the common themes identified in the investigation of events rather than on the use of these data in developing quantitative measures.

Negligence and malpractice in healthcare

Medical malpractice is "negligent care by a health services provider that causes injury to a patient" (Morlock, Lindgren and Mills 1989). Malpractice has attracted the attention of researchers because of its considerable and growing importance to healthcare organisations. It is shown below that there is strong evidence that malpractice is a relatively frequent occurrence; that it is costly to both providers and consumers of healthcare; that the costs of malpractice have risen consistently for two decades and are continuing to rise; and that recourse to the law provides patients with very little remedy for malpractice.

Researchers studying the incidence of malpractice have examined both the frequency with which claims arise and the incidence of instances of negligent care which could result in a successful claim if they were pursued. For the latter, adverse events have been an important research tool. Three major studies have established that negligent care is alarming common, and that in comparison, the numbers of claims made by patients are very low (California Medical Association 1977; Mills 1978; Harvard Medical Practice Study 1990; Wilson et al 1995). Overall, these studies suggest that about 1% of inpatients suffer a negligent adverse event, and that only 1.5% of such negligent adverse event actually result in a malpractice claim. These studies are reviewed in detail later in this chapter.

While no epidemiological studies of malpractice performed in the UK were found, there is some research on the causation and nature of malpractice. For example, Woodyard (1990) surveyed British orthopaedic consultant surgeons and found that 377 surgeons (58% of respondents, as the survey had a 50% response rate) had been or were at the time of the survey the subject of a malpractice action, and 185 (27%) had had one or more malpractice claims awarded against them.

Of the 236 successful claims that the respondents reported, 67 (28%) resulted from technical surgical errors on the part of the consultant surgeon, and 33 (14%) from the actions of junior medical staff. Asked to name the causes of the errors which led to litigation, the commonest factors cited were communication failures, clinical inexperience, and inadequate help or supervision. Drawing conclusions about the prevalence of malpractice in the UK is difficult in the absence of epidemiological evidence. However, the similarities in practice between the UK and the USA, and the strong evidence for widespread unrecognised malpractice in the USA support the tentative conclusion that there may be substantial unrecognised malpractice in the UK.

In the USA, the UK and most of Europe, medical malpractice is dealt with through the civil law. Patients who suffer malpractice seek redress by taking their doctors and hospitals to court, where the decision on whether to award damages hinges primarily on whether or not the doctor and/or hospital have been negligent, not on the nature of the harm suffered by the patient or its effects. Doctors and hospitals generally insure themselves against the risk of an action for negligence - but the costs of malpractice insurance have risen explosively in recent decades (Morlock, Lindgren and Mills 1989). In 1955, the average New York physician paid annual insurance premiums of $123. Three decades later, in 1985, the average premium was $38,000, a rise of 31,000%. Even allowing for inflation, the 1985 costs of malpractice insurance were about 35 times the costs in 1955. The causes are clear. In 1955, there were 1.9 successful claims per 100 doctors per year, but in 1985 there were 8.4 successful claims per 100 doctors per year. The cost in damages of each successful claim was at least 8 times as much, in real terms, in 1985 as it was in 1955 (Harvard Medical Practice Study 1990). In the UK, by comparison, there were just 1.3 successful claims per 100 doctors in 1988 - though this was ten times the level of a decade earlier (Woodyard 1990). The value of awards to plaintiffs has also been much lower in the UK (Jost 1990) though these too are rising - at about 17% per annum through the 1980s, according to some commentators (Tribe and Korgaonkar 1990). Dingwall and Fenn (1995) report that in 1992 there were about 6,000 malpractice claims a year against the NHS in England, and the costs of medical negligence litigation to the NHS were estimated at around £125 million pa.

Malpractice litigation has been less common in the UK than the USA for a number of reasons: contingency fee arrangements (in which lawyers take a percentage of the award instead of charging

a fee) were not permitted in the UK until recently, thus restricting access to the law to the very rich who could afford the costs involved and the very poor and children who receive Legal Aid; many costs which in the USA must be recovered from the doctor or hospital responsible through legal action are met in the UK through the social insurance and welfare system; some aspects of the law make it harder to win cases for negligence in the UK; and it may be argued that Britain is simply a less litigious society (Jost 1990, p52).

However, in both the UK and the USA it seems that malpractice litigation provides a poor and ineffective remedy for most of the problems of malpractice and its effects (Tancredi and Bovbjerg 1992). When the results of the Harvard Medical Practice Study (1990) review of 30,121 patient admissions were compared with the New York malpractice claims records, it was found that 47 patients had made malpractice claims relating to their admissions. Of these 47 patients, 18 (38%) had been judged by the study team to have suffered an adverse event, and only 8 (17%) had been rated as a negligent adverse event. Thus, only 1.5% of negligent adverse events actually led to a malpractice claim, and 83% of malpractice claims filed related to care that had been judged to be adequate (Localio, Lawthers, Brennan et al 1991). In view of these results, it is unsurprising that the study concluded that there was no evidence that the threat of malpractice litigation had any effect on the incidence of adverse events or negligent adverse events. It seems that most negligent injuries are never brought to court, and so those patients are never compensated. Many patients who have not suffered malpractice do bring actions, which are sometimes successful. The legal process is laborious and expensive to administer, and rarely compensates people in a timely fashion. Court decisions often seem haphazard, both in the way liability is assessed and in the size of damages awarded. Finally, the legal process encourages an adversarial battle between plaintiff and defendant rather than a careful, joint examination of the facts.

Human factors in medical accidents

In a number of other industries, particularly those where accidents or errors might be dangerous or even disastrous (like aviation and nuclear power), there is an established record of research into the factors involved in accidents and the way that future accidents can be prevented. Reason (1995) sets out a widely used classification of the human causes of accidents, which highlights the need to look beyond the immediate circumstances of the error in order to understand its causes. He argues

that the way organisations are structured and make decisions, and the way that processes are designed create latent failures - circumstances which make eventual accidents more likely or even certain when certain conditions occur or circumstances coincide. He suggests that studies of medical accidents have been too much concerned with classifying their consequences, and too little concerned with their causes.

To date, the human factors approach has not been widely employed in investigating adverse events in healthcare, though there are some examples of its use in anaesthetics (Gaba 1989; Runciman et al 1993; and others). Leape (1994) argues that it requires a change of paradigm in healthcare, from regarding adverse events as the responsibility of individuals to be corrected and even punished after the event. Instead, he believes we should learn from other industries and adopt an approach which is more focused on error prevention, on designing systems and processes of care to be able to absorb errors, and by creating a culture which is more tolerant of individual slips and lapses but less tolerant of organisational failures.

Conclusions from a diversity of approaches

The research reported above may describe itself as being focused on iatrogenic disease, critical incidents, medical malpractice, quality measurement, human factors in medical accidents or whatever, but it can all be seen as studying different facets of the same phenomenon - what might be called the epidemiology of adverse events in healthcare. As the brief review above demonstrates, researchers in these different areas have defined the concept of an adverse event quite differently. They have also focused their attention on different aspects of the epidemiology of adverse events - their consequences for patients, the costs for healthcare organisations, the perceptions of clinicians and others involved in these events, the causes and factors which contribute to their occurrence, their preventability, their use in performance measurement, and so on.

However, some common themes can be identified. Researchers concur that adverse events are important and worthy of study and investigation because they are more prevalent than might be expected, have important impacts on healthcare organisations and patients, and are often preventable. The case was neatly summarised by McIntyre and Popper, who wrote:

> "Mutual criticism is not personal and pejorative, but ... springs from a mutual respect and a desire to improve the lot of patients. It then becomes important not only to acknowledge mistakes but to search for them, in order to correct them as quickly as possible. When errors are due to lack of skill we will, we hope, try to improve our skill; and when, as is sometimes the case, our errors are due to carelessness, or our failure to do what we know we ought to do, then we will look for ways of improving our behaviour." (McIntyre and Popper 1983).

They also seem to agree that the study of adverse events should look beyond the performance of the individual clinician, and recognise the importance of the wider process of care and the organisational context in which it takes place.

Looking outside healthcare, other industries use adverse events extensively in quality measurement and quality improvement. For example, Oakland (1989, p165) places error/defect detection and prevention at the centre of the process of quality monitoring, though he emphasises that constructive investigations focused on error prevention should not become destructive inquisitions aimed at placing blame. Deming, whose influence on modern industrial quality assurance has been profound, reviewed the needs of quality measurement in healthcare from an outsider's perspective, and recommended a series of indicators of which two thirds were based on the incidence of errors or defects (Deming 1986, p203).

In short, the proven importance of adverse events in healthcare, and their confirmed worth in quality measurement and quality assurance outside healthcare both suggest that adverse-event measures of quality may have an important role to play in healthcare quality assurance. The rest of this chapter explores the definition, development and use of adverse events in healthcare quality measurement and quality improvement.

3.2.2 **Concepts and theory of adverse-event measures of quality**

Those involved in developing adverse-event measures of quality have tended to place greater emphasis on developing and applying those instruments than on either defining the concepts they use or on evaluating the measures they develop and apply. Goldman (1989), having reviewed the literature, remarked that "given the widespread use of occurrence screening programmes in hospital QA and external peer review programmes, it was striking how little scientific literature existed on the subject". As a result, a welter of different and sometimes conflicting definitions of adverse events, sets of screening criteria, and classifications of causation and standards of care have evolved. It is important, in order to understand how different studies or instruments relate and how they differ, that some of these concepts are explored and clearly delineated. To develop this understanding, there are six principal areas which need to be considered: the definition of adverse events; the classification of actual instances of adverse events; the various sources of information on adverse events; the use of sampling in identifying adverse events; the timing of measurement; and the construction of quantitative measures based on information about adverse events.

Defining adverse events

A number of different researchers have developed definitions for the term *adverse event*, and some of the principal ones are listed in table 3.1. Craddick and Bader's (1983) definition of an adverse event (which they termed an adverse patient occurrence) has probably been most widely cited and often adapted or modified by others.

McLamb and Huntley (1967)	"Any response to medical care in the hospital that is unintended, undesirable and harmful to the patient."
Mills (1978)	"A potentially compensable event is a disability caused by health care management: • disability - is a temporary or permanent impairment of physical or mental function (including disfigurement) or economic loss • causation - is established when the disability is more probably than not attributable to health care management • health care management - includes both actions and inactions of any healthcare provider or attendant."
Craddick and Bader (1983, p23).	"Adverse patient occurrences ... refer to untoward patient events which, under optimal conditions, are not a natural consequence of the patient's disease or treatment. The common thread of all APOs is that they are events which health professionals agree are not desirable outcomes of medical management."
Harvard Medical Practice Study (1990)	"An unintended injury caused by medical management rather than by the disease process. The injury is sufficiently serious to lead to prolongation of hospitalisation or temporary or permanent impairment or disability in the patient."
Wilson RM, Runciman WB, Gibberd RW (1995)	"An unintended injury or complication which results in disability, death or prolonged hospital stay and is caused by health care management."

Table 3.1. Some definitions of adverse events drawn from the literature.

Reviewing the definitions offered by different researchers, it is clear that they largely agree that an *adverse event* is a happening, incident or set of circumstances which exhibits three key characteristics to some degree:

a) *Negativity*.

It must be an event which is, by its very nature, undesirable, untoward, or detrimental to the healthcare process or to the patient. This is a theme which is common to all definitions.

b) *Patient involvement/impact.*

It must in some way involve or have some negative impact or potential impact on a patient or patients. The wider definitions of adverse events include occurrences in which there is no actual effect on any patient, though there is the potential for harm. More restrictive

definitions often only include events where the patient has suffered some definable and identifiable ill effect from the event.

c) *Causation.*

There must be some indication that the event is a result of some part of the healthcare process (either through commission or omission), rather than a result of events outside the healthcare process, such as the patient's own actions or the natural progression of the disease. Again, definitions vary, with some accepting events as adverse events with little or no evidence of causation, while others insist on strong and direct evidence of causation.

Combining these three key characteristics gives the following definition, which is used throughout this thesis:

> An *adverse event* is an untoward or undesirable occurrence in the healthcare process which has or potentially has some negative impact on a patient or patients and results or may result from some part of the healthcare process.

Of course, having defined the term *adverse event* does not in itself provide a basis for identifying or measuring such events in practice. Since adverse events can take so many different forms, from patient falls to drug errors, most researchers have developed some form of list or classification of adverse events which sets out the kind of occurrences, incidents or sets of circumstances which make up an adverse event. A number of examples of these lists or classifications are described and compared later in this chapter, but one example, drawn from Craddick and Bader (1983) is given in table 3.2.

1.	Admission for complications or adverse results of outpatient management.
2.	Admission for complications or incomplete management of problems on previous hospitalisation.
3.	Operative consent incomplete, missing, or otherwise incorrect.
4.	Unplanned removal, injury or repair of organ or structure during surgery, invasive procedure or vaginal delivery.
5.	Unplanned return to operating room on this admission.
6.	Invasive procedure with tissue removed where pathology report does not match preoperative diagnosis, or non-diagnostic or no tissue removed.
7.	Transfusion required for bleeding/anaemia/other iatrogenic reason, not clinically indicated, or resulting in reaction.
8.	Nosocomial (hospital acquired) infection.
9.	Antibiotic or drug utilisation problems.
10.	Cardiac or respiratory arrest, or low Apgar score.
11.	Transfer from general care unit to special care unit.
12.	Other patient complications.
13.	Hospital-incurred patient incident, such as fall, IV problem, medication error, or skin problem.
14.	Abnormal laboratory, X-ray or other test results not addressed by physicians.
15.	Neurological deficit present at discharge which was not present on admission.
16.	Transfer to another acute care facility.
17.	Death.
18.	Subsequent visit to ER or OPD for complication or adverse results related to this hospitalisation.
19.	Length of stay above certain percentile or allotted days.
20.	Medical record review.
21.	Nursing record review.
22.	Departmental or other problems.
23.	Patient or family dissatisfaction.

Table 3.2. Generic screening criteria used to identify adverse events.
Craddick and Bader (1983).

The list of circumstances or occurrences which are deemed to be adverse events is sometimes called a set of *screening criteria* (since it is used to screen patients or admissions to find those who have suffered an adverse event). Almost all adverse-event measures of quality consist of a set of screening criteria like those set out above in table 3.2, but they vary greatly both in their content (the type of occurrences which are deemed to be adverse events) and in their nature (the way in which those occurrences are defined in the measure). The classification of the criteria used to give value or meaning to data in quality measurement which was developed by Donabedian (1982) and was discussed in chapter 2 can be adapted here to describe some of the characteristics of adverse-event measures of quality and the screening criteria of which they are constructed:

a) *Specification.*

The adverse event definitions may be fully and explicitly stated in great detail, and in terms which allow any user with sufficient understanding of the terminology involved to decide whether or not an adverse event has occurred. Alternatively, they can be relatively imprecise, and rely on the professional judgements made by a rater who has sufficient clinical knowledge and experience. The former is more likely to be reliable in application than the latter, but it may also be rather inflexible and allow too little room for sensible discretion and clinical judgement, resulting in a less valid measure. In other words, there may be a trade-off between validity and reliability. Some adverse events may be quite heterogeneous and hard to define precisely. For example, criterion 1 in table 3.2 (admission for complications or adverse results of outpatient management) clearly demands clinical expertise to assess the link, if any, between an admission and preceding outpatient care.

b) *Referent.*

The referent of the adverse event definitions is the patient group to which they refer. In the case of some adverse events, this may be all patients. For more specialist adverse events, it may be those patients in a certain specialty, undergoing a certain procedure, or whatever. For example, in table 3.2, criterion 3 (operative consent missing or incomplete) can clearly only refer to patients who have undergone some form of procedure which requires consent, and so this subgroup would form the obvious referent. In practice, adverse-event measures often assume that any adverse event could occur to any patient, because defining the referent can be quite complex, or else they leave the referent undefined.

c) *Monotony*.

In the vocabulary of Donabedian, adverse events are almost always monotonic - in other words, it can generally be said that the less often they occur, the better. It can, however, be argued that some adverse events represent accepted and desirable tradeoffs between complete patient safety and other aspects of quality - such as patient autonomy, or the potential benefits of interventions. For example, criterion 13 in table 3.2 includes patient slips and falls, but it may be that reducing the rate of such accidents below a certain threshold would require unacceptable restrictions on patients' freedom to, for example, sleep

in beds without bedrails and take themselves to the toilet if they wish. Hence, this criterion is not wholly monotonic, and some level of adverse events will represent an optimal tradeoff between patient autonomy and patient safety.

d) *Importance*.

Clearly, not all adverse event definitions are of equal importance. Their importance is primarily a factor of their expected impact or potential impact on patients, though other considerations such as the resulting cost and the organisational effects of the defined adverse event may also be considered. For example, adverse events under criterion 10 in table 3.2 (cardiac arrests and low Apgar scores) are presumably far more significant in terms of likely patient impact than those under criterion 20 (deficiencies in the medical record).

e) *Expected incidence*.

The observed incidence of adverse events may be used as the measure of the quality of care. However, different adverse events may have quite different expected incidences - some may be inherently quite frequent, while others may usually be very rare. A relationship with the *importance* of the adverse event may exist (in that rare adverse events may be perceived as more important than common ones). Clearly, the expected incidence of an adverse event is an important consideration for those using it in quality measurement, since very rare events will require large samples and much effort to identify even small numbers of cases. Conversely, if the expected incidence is very high, collecting information about every instance of the event may be expensive, and the event may be regarded as commonplace and trivial. The concept of the expected incidence of adverse events is related to that of stringency in Donabedian's classification.

f) *Source.*

Donabedian terms criteria normative or empirical; exogenous, endogenous or autogenous; and representative or elitist. Most adverse event definitions are part normative, part empirical - blending professional opinions on events with (usually limited) research findings. They are often exogenous - developed by professionals elsewhere - though they usually have the approval of the professionals involved. They are generally representative,

rather than elitist - defining events which would be deemed adverse by most practitioners, not just those who are specialists or professional leaders. Patients are rarely involved in the process of definition.

Investigating and classifying adverse events

While adverse-event measures of quality are generally based on counts of the numbers of instances of adverse events, those events are also frequently categorised, classified or further subdivided in several ways as part of the process of investigation and analysis.

a) *Importance.*

Individual adverse events of the same type will vary in their importance (which was also identified above as a characteristic of the adverse event definition). Importance is a complex concept, which implicitly combines some of the considerations detailed below (such as effect, causation and avoidability) with such factors as the opportunities for improvement highlighted by the event.

b) *Effect.*

Events are often assessed in terms of their effect on the patient's health. Rating scales which combine the severity and the temporal persistence of effect are often used - ranging, for example, from a minor, temporary effect on health, to a major, permanent effect. Other effects of adverse events are also sometimes considered, particularly the effect on continuing or future healthcare needs (such as extended stays in hospital). The key problem in assessing the effect of adverse events, which is usually addressed (or sidestepped) by relying on implicit review by clinical professionals, is the separation of the effects of the event from the effects of the underlying disease process.

c) *Causation.*

The causative factors in an event are often analysed and categorised. The most basic (and frequently used) distinction is that made between events caused by the healthcare process itself and events caused by factors outside the healthcare process - the rationale being that only those factors within the healthcare process are generally amenable to revision. More

detailed assessments of causation (sometimes called attribution or association) may link the event to specific staff groups, individual clinical professionals, organisational processes, or other factors.

e) *Avoidability*.

Evaluations of the avoidability of individual adverse events are often made, categorising the event on a scale ranging from wholly avoidable, to wholly unavoidable. The concept of avoidability is closely linked to that of causation (in that it is likely that only events caused by healthcare process factors may be deemed avoidable) and to the concepts of acceptability and negligence discussed below.

f) *Acceptability*.

The acceptability of an event is, by definition, a subjective assessment of the extent to which the clinical or organisational practices and actions which led to the event are judged to be in conformity with accepted professional standards. This assessment is generally made by a clinical professional with acknowledged expertise in the area. However, it is fraught with measurement problems, arising from the implicit nature of the review, and the lack of an adequate definition of accepted professional standards. Nevertheless, as a measure of the extent to which practice leading to an event is deemed unacceptable by other clinical professionals, it can be useful.

g) *Existence of negligence*.

The existence of negligence requires a medicolegal rather than a clinical assessment. Legally, to demonstrate negligence requires evidence that on the balance of probabilities the patient suffered some harm which was a result of the negligent actions of those healthcare professionals caring for the patient. Negligent actions are generally defined in the UK as those falling outside practices accepted at the time as proper by a responsible body of medical opinion. Thus, determining the existence negligence is linked to the assessments of effect and acceptability outlined above.

The classification of adverse events in these ways is almost always performed through some form of professional review. The rigour with which those reviews are carried out varies - from those which are simply based on a single professional's personal and implicit assessment of the circumstances, to those which use multiple professional assessments, made with explicit criteria and definitions of the concepts involved. Investigations of the reliability and validity of this review process have indicated that consistent intrarater and interrater reliability are elusive, and that the achievement of reliable judgements may demand more multiple ratings than are practically feasible (Richardson 1972; Goldman 1992; Localio et al 1996). It has also been elegantly demonstrated that reviewers' judgements can be biased by their knowledge of irrelevant case circumstances, which raises serious concerns about the validity of implicit peer reviews (Caplan, Posner and Cheney 1991).

While some authors argue that the value of adverse event detection lies in these detailed analyses and assessments of individual adverse events (Sanazaro and Mills 1991), other suggest that the methodological weaknesses of the mechanisms available to make the judgements required means that the limited advantage gained by performing these analyses is outweighed by their considerable cost (Massanari 1992). Certainly, the limited available literature suggests we should be cautious about using implicit reviews or assessments by clinicians in adverse-event measures of quality because of their low reliability and unproven validity.

Sources of information

By far the most frequently used approach to identifying adverse events in healthcare is to monitor or screen patients' clinical records either during or after the process of care. Information is abstracted from the clinical records by raters or screeners, who use the records to decide whether or not adverse events have occurred and to document and classify those events. There are two important weaknesses in this process. Firstly, the clinical records may be deficient, and as a result adverse events might be missed. Indeed, since the more deficient the medical records are, the harder it will be to identify adverse events, the paradoxical situation could occur in which good, comprehensive records produce a higher adverse event score (and so an indication of a lower quality of care) than sketchy, incomplete records. Secondly, the clinical records are always a summary of events in the patient's care and treatment, rather than a record of every action, conversation and incident. Some adverse events might concern circumstances which are not routinely recorded in the clinical record,

and so reliance on the clinical record as the sole source of information might produce a spuriously low indication of their incidence.

Another source of information on adverse events is the self-reporting of incidents by clinical professionals. Indeed, most hospitals have at least some reporting mechanisms for a range of adverse events such as medication errors, and patient accidents (Williamson and Mackay 1991; Hartwig, Denger and Schneider 1991; and others). However, if the reliability of the clinical records is a concern, the reliability of reporting mechanisms which rely on many different professionals to report adverse events, all of whom may have different personal definitions of what constitutes an adverse event and different degrees of commitment to the self-reporting mechanism, must be even more in doubt. In fact, Craddick and Bader (1983, p11) claim that only 5-10% of adverse events identified through screening the clinical records are identified by self-reporting systems. Thompson and Prior (1992) report that when the results of screening were compared with self-reporting, only 74% of significant adverse events found through screening had been reported. Hartwig, Denger and Schneider (1991) assert that the rate of reporting for medication errors is "probably well below the actual error rate" and point out that this means that any variations in rate may be due as much to changes in the tendency to report them as to any change in the actual incidence of errors, which is certainly a drawback for any measure. However, in another study (O'Neill et al 1993) where adverse events could be reported directly by clinicians via electronic mail, it was found that reporting identified virtually the same number of adverse events as screening, finding 89 versus 85 events in a total of 3,128 admissions, though only 41 events were common to the two sets. The authors argued that well-motivated clinicians were capable of reporting adverse events at least as reliably and validly as screening could detect them, that reporting was cheaper, and that it had the advantage of drawing clinicians into quality measurement rather than making them the passive subject of the process. Nevertheless, the consensus of the literature seems to suggest that incident reporting is likely in most circumstances to miss a proportion of adverse events and to be less consistently and reliably applied than screening for adverse events.

With the increasing availability of information technology in hospitals, some researchers have used available computer systems to identify fairly limited groups of adverse events. For example, Mendenhall (1987a; 1987b) used readily available hospital databases to identify medication

prescribing errors and the overadministration of anaesthetic/analgesic agents. Bates et al (1994) examined a series of 133 adverse events identified through manual reviews of 3,138 patient admission records to assess how many of them were identifiable from computer systems. They found that even when computer systems held only basic demographic data along with test orders, test results and medication prescriptions, 53% of adverse events could be identified. If all physician instructions were available on computer, the proportion identifiable rose to 58%, and if wholly computerised clinical records were in place, 89% of all adverse events could have been detected automatically.

A final, though little used, source of information on adverse events is the direct observation of practice. Though theoretically possible, direct observation is generally too expensive and too intrusive to yield useful information on the incidence of adverse events. However, concurrent review procedures, which involve the periodic screening of patients' records during their hospital stay, inevitably include an element of observation, since the frequent screenings are carried out in the ward environment. The review staff can hardly avoid observing some practice, and their regular contact with the clinical professionals involved in direct patient care often results in information from the clinical record being supplemented with information from observation and staff communication.

Sample definition

Craddick (1984, pV-1) argued that the review of all patient records (in other words, the entire patient population) for adverse events was worthwhile for two reasons. Firstly, she suggested that single, serious adverse events - those which are rare but important - may be missed if only a sample of cases are reviewed. Secondly, she pointed out that any sampling process is open to suggestions of bias, and using adverse event rates from sampled populations allows the validity of comparative analyses to be challenged, on the basis that the sample is not representative of the population of patients as a whole. However, other researchers have used an assortment of sampling strategies to select samples of patients for adverse event review:

a) *Random sampling*.

Random sampling is the simplest sampling strategy, though it may also be the least cost-effective way to identify adverse events. It involves selecting a subgroup of patients at random, whose cases are then reviewed for adverse events. Perhaps its main advantage is that, if sampling is truly random, adverse event rates found in samples can, with appropriate confidence limits, be generalised to the population of all patients.

b) *Stratified sampling.*

Random sampling produces very small samples of groups which form a relatively small part of the whole population, such as minor specialties. These samples may be too small for any useful calculations to be performed. Stratified sampling involves dividing the population of patients into groups according to a set of characteristics (such as specialty, ward, age or whatever) and drawing samples from each group. Thus, small specialties or patient groups can be oversampled (disproportionate stratified sampling) to ensure adequate sample sizes. Rates for the population as a whole can still be calculated, since the structure of the stratified sample and the structure of the population as a whole are known.

c) *Targeted sampling*.

Some researchers have deliberately targeted their sample selection on areas where empirical evidence or shows (or it is assumed) adverse events are more frequent. This form of targeted sample might be made up of a sample of the population as a whole, plus all patients in certain defined high risk groups, such as readmitted patients, long stay patients, patients in particular Diagnosis Related Groups (DRGs), etc. The main rationale for this approach is that it maximises the yield from the screening process - producing more adverse events than a random or stratified sample. Its main disadvantage is that it makes the calculation of rates for comparative analyses or the derivation of population-based estimates of rates very difficult (Jost 1989; Stuart 1989).

d) *Probability sampling*.

Probability sampling is an alternative approach to targeted sampling, which is designed to combine the latter's high yield from screening with the ability to generate population-based estimates of rates and figures for comparative analyses. It involves building a model which uses cumulative records of adverse events to predict where adverse event rates are highest. These predictions are used to guide sampling, and the results of sampling are used to update the model (Ash, Shwartz, Payne et al 1990). Since areas are targeted on the basis of empirical evidence from the model, rather than assumptions or conjecture, probability sampling should be more effective at maximising the yield from case review than targeted sampling. Its main disadvantage is the statistical complexity of the methodology, which reduces the system's transparency to clinicians, and which requires appropriately skilled quality assurance staff to apply the methodology and interpret its results.

In practice, the costs of screening cases for adverse events make the use of whole population screening too expensive for many healthcare organisations to support.

Timing of measurement

Ideally, a patient's care and treatment would be reviewed for adverse events once all elements of that care and treatment were complete, and all effects and results of care were known. However, that delay in measurement conflicts with the need for timely information on adverse events, and with the practicalities of accessing patients' records for review. Two main approaches to undertaking reviews for adverse events have developed:

a) *Concurrent review*.

In concurrent review, the patient's care and treatment is reviewed for adverse events whilst they are in hospital, with reviews taking place shortly after admission and then periodically during the hospital stay. This produces very up-to-date information, and the repeated reviews may help to make the review more reliable. However, it is also more expensive, since each case is reviewed not once but several times.

b) *Retrospective review*.

This involves patients' care and treatment being reviewed for adverse events soon after its completion - usually on or shortly after the date of discharge. This is less costly than concurrent review, and produces relatively up-to-date information. However, since patients' stays in hospital are often very brief, some adverse events may not be observable or recorded until after this retrospective review has taken place.

Construction of measures

Simply detecting and classifying adverse events is not, by itself, enough to provide a useful measure of quality. There are a number of important issues to be considered in converting these raw numbers of adverse events into functional measures of quality:

a) *Denominators*.

To calculate incidence rates of adverse events, some denominator variable is needed. The two commonest denominators used are the number of patients or cases, and the number of inpatient days. The former produces incidence rates will be lower for specialties or areas in which many short-stay patients are treated, but higher for areas in which long lengths of stay are the norm, whereas the latter has the opposite bias. When making comparisons, it is important to recognise the influence of the choice of denominator variable.

b) *Risk adjustment*.

The principles of risk adjustment have already been rehearsed above. Essentially, it involves considering not only whether an adverse event has happened, but whether it could have happened, and using the ratio of actual adverse events to potential adverse events in measures. The main advantage is that it makes little sense to calculate incidence rates for adverse events in which the denominator includes patients who would never have had an adverse event of that type, since the incidence rate will be artificially low as a result. However, deciding whether individual patients were at risk from adverse events is a difficult and unreliable process, and this second layer of rating makes it harder to achieve adequate reliability for these measures (Richards et al 1988).

c) *Multiple classifications of single events*.

Some adverse events may be classifiable as more than one type of event. For example, a missing entry on a drug chart might be classified as either a medical record deficiency or a drug administration error. Clear adverse event definitions can certainly help to reduce the potential for such overlaps, but it can be difficult to remove them altogether. Therefore, adverse-event measures, if they are to be reliably applied, need to define how such potential multiple classifications should be handled.

d) *Multiple adverse events of the same type*.

During a single patient episode, the same type of adverse event can happen more than once (for example, two quite separate medication errors could take place). There are two basic approaches to incorporating such events into a measure: either each individual event is counted (which means that theoretically there is no limit to how many adverse events a single patient might have) or each event type is counted (which means that one or more than one adverse events of a particular type are counted as a single event, and the maximum number of event types for a single patient is the number of different adverse events for which screening takes place).

e) *Multiple adverse events of different types*.

During a single patient episode, a number of different adverse events may be causally or conceptually linked. For example, a hospital acquired infection could lead to a return to theatre for wound debridement. In some adverse-event measures, each causal chain of events would be treated as a single event - usually the initial event in the chain. In other measures, each event would be counted separately. The former approach may understate the incidence of adverse events, but the latter method can lead to apparent double-counting of events which artificially inflates the numbers of events recorded.

f) *Aggregation*.

Many adverse-event measures, having identified the adverse events which occurred during a patient's stay in hospital, then aggregate the results, yielding a single integer (the number of adverse events the patient suffered) or a dichotomous variable (whether or not the patient

suffered any adverse events). This process of aggregation tends to eliminate the distinctions between different types of adverse events, and in some circumstances this may be clinically inappropriate and relatively meaningless.

Conclusions

The framework set out above provides a useful structure against which a wide range of different adverse-event measures can be set in order to compare and contrast them. In practice, many of the issues outlined above have often not been addressed by the developers of adverse-event measures of quality, leaving considerable room for ambiguity and confusion in the way in which those measures are then used. For example, although the definition of denominators or the approach to classifying multiple events are both important issues which have to be faced in defining a quantitative measure based on adverse events, both issues are generally ignored, which does not help to assure the validity and reliability of the resulting measures.

3.3 Using adverse events in quality measurement and improvement

3.3.1 Defining adverse-event measures of the quality of care

Before exploring the findings of some of the studies which have examined the incidence of adverse events in healthcare, using a variety of approaches to measurement, it is important to understand how some of those measures have been developed. It was noted earlier that the definitions of adverse events used by researchers and practitioners varied in two specific ways - the degree of impact on the patient required, and the strength of demonstrated causative relationship with the healthcare process needed - for something to "count" as an adverse event. It will be seen from the examples of adverse-event measures reviewed below that they also differ in two further ways. Firstly, some are intended to be generic, that is they attempt to be applicable to all patients almost regardless of their condition or specialty, while others are more specific to particular patient groups, such as those in a given specialty, with a particular condition, or undergoing a particular intervention. Secondly, some are defined explicitly and in some detail, while others are much less

clearly specified and rely far more on the judgement of the person applying the measure. Some representative examples of adverse-event measures of quality are reviewed below.

Medical Management Analysis

Craddick, who had participated in the Californian Medical Insurance Feasibility Study reported earlier, used the definitions of adverse events developed for that research to design a systematic quality assurance programme based around the detection and analysis of adverse events and aimed at acute hospitals, which she dubbed *Medical Management Analysis* (Craddick 1979). This programme was intended to be more than just a quality measurement system - it included mechanisms for utilisation review, risk management, and medical staff credentialling, as well as offering an organisational model for the structuring of quality assurance committees, the reporting of quality assurance information, and the staffing and resourcing of quality assurance departments.

1.	Admission for complications or adverse results of outpatient management.
2.	Admission for complications or incomplete management of problems on previous hospitalisation.
3.	Operative consent incomplete, missing, or otherwise incorrect.
4.	Unplanned removal, injury or repair of organ or structure during surgery, invasive procedure or vaginal delivery.
5.	Unplanned return to operating room on this admission.
6.	Invasive procedure with tissue removed where pathology report does not match preoperative diagnosis, or non-diagnostic or no tissue removed.
7.	Transfusion required for bleeding/anaemia/other iatrogenic reason, not clinically indicated, or resulting in reaction.
8.	Nosocomial (hospital acquired) infection.
9.	Antibiotic or drug utilisation problems.
10.	Cardiac or respiratory arrest, or low Apgar score.
11.	Transfer from general care unit to special care unit.
12.	Other patient complications.
13.	Hospital-incurred patient incident, such as fall, IV problem, medication error, or skin problem.
14.	Abnormal laboratory, X-ray or other test results not addressed by physicians.
15.	Neurological deficit present at discharge which was not present on admission.
16.	Transfer to another acute care facility.
17.	Death.
18.	Subsequent visit to ER or OPD for complication or adverse results related to this hospitalisation.
19.	Length of stay above certain percentile or allotted days.
20.	Medical record review.
21.	Nursing record review.
22.	Departmental or other problems.
23.	Patient or family dissatisfaction.

Table 3.3. Generic screening criteria used in the Medical Management Analysis programme. Craddick and Bader (1983).

The MMA programme was based around a review of all patient records for adverse patient occurrences (APOs), carried out by quality assurance staff. This review took place both concurrently (while the patient was in hospital) and retrospectively (after discharge). Reviewers used a set of 23 *generic screening criteria*, designed to identify all important APOs in almost any specialty, to screen patients' records for APOs (see table 3.3). For each criterion, more detailed and specific guidance was provided on its interpretation. Information on APOs from other sources - such as incident reports - was also used, but the record screening was the primary source of information. The MMA programme aimed to use this single comprehensive record review to serve a range of purposes, including quality assurance, risk management and utilisation review.

While some of the set of 23 generic criteria were unambiguous indicators that an adverse occurrence had taken place (such the unplanned removal, injury or repair of organ during surgery), many were what Craddick termed clue criteria, merely indicating that a case was worthy of further attention (such as unexpected death). The primary screening was not designed to yield a definitive adverse occurrence rate - its main purpose was to identify a limited subset of records which were worthy of further detailed review by a medical reviewer.

When a record was flagged by one or more of the generic screening criteria on initial screening, it was passed to a medical peer reviewer, who would either confirm or refute the initial judgement that an APO had occurred. If the reviewer confirmed that an APO had occurred, he or she would undertake a structured analysis of the event, assessing the standard of care provided, judging the attribution or involvement of different staff groups in the event, and rating the severity of the event for the patient.

Information from the MMA programme was used in a number of ways. Firstly, individual APOs were followed up, especially when they indicated a serious weakness in individual or organisational performance. Secondly, trends over time in APO rates, often broken down by the severity of the occurrence or the standard of care assessment, were used to compare performance and to identify patterns of suboptimal care. Data from the MMA programme was commonly used by quality assurance committees, risk management staff, and physician credentialling committees, and

overviews of the substantial database of information built up by the programme were presented to the hospital governing body.

Craddick and Bader (1983, p46) asserted that the main advantages of the MMA programme over other approaches to quality assurance were its objectivity; its identification of meaningful patient care issues, including both individual APOs and APO trends; the involvement of physicians and their peers in both assessment and followup action; the emphasis on taking corrective action; and the ability to react and intervene speedily when problems were identified.

Occurrence screening in the Veterans Administration healthcare system

The Veterans Administration (VA) healthcare system runs 172 medical centres in the USA with about 90,000 beds in total. It introduced an occurrence screening programme in 1988 (Goldman 1989; Goldman and Walder 1992) which used a set of 9 generic occurrence criteria (see table 3.4), each of which could be identified through existing computer systems. Cases flagged on one or more criteria were then reviewed by nurse reviewers, who decided whether an adverse event had in fact occurred. A process of peer review was used to determine the causation and appropriate follow-up action for each adverse event.

1. Readmission within 14 days of discharge
2. Admission within 3 days following unscheduled ambulatory care visit.
3. Admission within 3 days following ambulatory surgery procedure.
4. Admission from nursing home within 14 days of discharge from acute care.
5. Transfer from Intermediate Medicine within 14 days of transfer from acute care.
6. Transfer to a special care unit within 72 hours of transfer from special care unit, or within 72 hours of a surgical procedure.
7. Return to operating room in same admission.
8. Cardiac or respiratory arrest.
9. Death.

Table 3.4. Veterans Administration screening criteria. Goldman (1989).

The VA measure can be seen to rely more heavily on the implicit reviews of the quality of care undertaken by nurse and medical reviewers. Indeed, the criteria set out in table 3.4 could be argued to be largely a mechanism for targeted sampling, intended to identify admissions where an adverse event is more likely to have occurred for further review. In use, the automated screening of

discharges flagged 13.3% of all cases for review (57,841 episodes out of a total of 435,000 during six months ending March 1990) of which only 19.5%, or 2.7% of all episodes were confirmed to have an adverse event (Goldman and Walder 1992).

Medicare PRO occurrence screening

While US hospitals make extensive use of adverse-event measures of quality, perhaps the largest single user of these techniques is the Medicare Peer Review Organisation (PRO) programme. This programme is federally mandated to review the quality and cost of care delivered to 31 million Medicare beneficiaries in 7,000 hospitals across the United States. A Peer Review Organisation (PRO) was established in each state to review a carefully defined sample of inpatients admitted to hospitals in the state whose care was being financed by Medicare. The sampling strategy used to select those patients for review was intended to select cases where quality or utilisation problems were more likely to occur (table 3.5).

3%	Random sample of all discharges.
50%	Cases involving an interhospital transfer.
10%	Cases involving a transfer to psychiatric care within same facility.
25%	Cases involving a transfer from acute to nursing home care within same facility.
25%	Cases of readmissions within 31 days of previous admission.
25-100%	Cases in certain "problem" DRGs
25%	Cases with costs above usual DRG prepayment.

Table 3.5. Peer Review Organisation sampling strategy for retrospective review.

The clinical records of all cases in the selected sample are retrieved, and reviewed by trained reviewers (usually with a background in nursing) who use a generic occurrence screening tool to identify potential adverse events. They also undertake reviews of utilisation and appropriateness and data validation. The generic occurrence screening tool used by all PROs is shown in table 3.6.

Adequacy of discharge planning. No documentation of discharge planning or appropriate follow-up care with consideration of physical, emotional and mental status needs at time of discharge.
Medical stability of patient at discharge. BP within 24 hrs of discharge outside limits. Temperature within 34 hrs of discharge over limit. Pulse within 24 hrs of discharge outside limits. Abnormal diagnostic findings not addressed/resolved. IV fluids or drugs given on day of discharge. Purulent/bloody wound drainage within 24 hrs prior to discharge.
Death During or following surgery Following a return to intensive care. In other unexpected circumstances.
Nosocomial (hospital acquired) infection.
Unscheduled return to surgery during same admission for same condition.
Trauma suffered in hospital Unplanned surgery/surgical injury. Fall. Serious complications of anaesthesia. Transfusion error or serious transfusion reaction. Hospital acquired decubitus ulcer/deterioration. Medication error or serious adverse drug reaction. Care or lack of care resulting in serious complications.

Table 3.6. Medicare PRO generic screening criteria.

Any adverse event identified by primary screening is then referred to a physician reviewer, independent of the hospital concerned. If the physician confirms the adverse event, then a complex analysis and rectification process, called the quality intervention plan, commences. The hospital and physician involved are notified of the problem, and can appeal against the PRO assessment of the problem. An assessment of the severity of the adverse event is made, and this is used to build up a profile for both the hospital and the physician concerned. The PRO may withhold payment for the case, may require the hospital or physician to undergo some form of education, may intensify its review of that hospital or physician's cases, may refer details to licensing bodies who allow the hospital to operate and the physician to practice, and ultimately can sanction the hospital and/or the physician by excluding them from the Medicare programme.

Specialty-specific adverse-event measures

The use of a single, generic set of criteria in all acute specialties has been perceived by many developers as a weakness of the approaches. Craddick developed some specialist screening criteria for use in specialties such as anaesthesia, obstetrics, paediatrics and radiology (Craddick 1984). Other researchers have addressed areas such as vascular surgery (Lawrence-Brown et al 1989), accident and emergency (Manning et al 1990), long term geriatric care (Nordal and Ang 1988), obstetrics, orthopaedics and ophthalmology (Walshe, Bennett and Ingram 1995) and others. Shaw (1992) provides suggested adverse event quality indicators for a wide range of specialties.

3.3.2 **The epidemiology of adverse events**

There are many studies of what might be called the epidemiology of adverse events - their incidence, causation and consequences - but they approach the subject from three main perspectives: the study of iatrogenic disease, negligence and malpractice, and quality measurement. Each of these three perspectives is explored below, before some common themes in their findings and conclusions are identified.

Adverse events and iatrogenic disease

Researchers have studied the incidence of specific kinds of iatrogenic disease - such as admissions, deaths, or drug errors - as well as examining the incidence of iatrogenesis more broadly.

Lakshmanan, Hershey and Breslau (1986) studied all admissions to the medical services of a large teaching hospital over a two month period, to identify the incidence and causation of iatrogenic admissions. Records were reviewed by the researchers, who "attempted to be very conservative, and did not include patients where the clinical picture could have been caused by the underlying disease". Among 834 admissions they found 45 (5.4%) which were iatrogenic in origin, resulting from 47 separate iatrogenic events, of which 23 (49%) were deemed to have been avoidable. Most of the iatrogenic events involved drug therapies (35, 78%). Lakshmanan et al's study confirms the findings of earlier research, which suggested that between 3% and 5% of hospital admissions are

iatrogenic in origin, and that pharmaceutical agents are the commonest causative factor (Caranasos, Stewart and Cluff 1974).

Within the acute setting, Trunet et al (1980) examined all admissions to an intensive care unit over a 12 month period, again to assess the incidence of iatrogenic admissions, using a similar process of implicit record review, based on a set of five criteria defining the causation of adverse events. They found that of 325 admissions, 41 (13%) were iatrogenic in origin, and 19 (46%) of these were judged to have been avoidable. Adverse drug reactions and drug prescribing errors caused 23 (56%) of the admissions.

In a ten year study of surgical mortality, McDonald et al (1991) reviewed 543 deaths among 23,557 admissions, and found that 89 (16%) were avoidable. By definition iatrogenic in nature, these deaths resulted most commonly from surgical errors (32 deaths, 36%). The decision about the avoidability of each death was made by a meeting of the clinical team involved, a methodology that McDonald et al acknowledge has obvious weaknesses. In a similar surgical setting, using a process of implicit record review by the researchers, Heywood, Wilson and Sinclair (1989) reviewed 80 deaths from 10,592 admissions, and found that 33 (41%) were avoidable. In the USA, Dubois and Brook (1988) carried out a more rigorous examination of 182 deaths drawn from 12 hospitals which were high or low outliers on the Health Care Financing Administration adjusted mortality statistics. Each death was independently reviewed by three physicians to assess its preventability. Using a majority rule, the three physicians deemed 49 (27%) of the deaths to be preventable; using a unanimity rule, 25 (14%) were judged to be preventable. A range of errors in both diagnosis and management were cited as the causes of the preventable deaths.

There are a number of studies which explore the incidence of medication or drug errors, as a specific but relatively common cause of iatrogenic disease. As long ago as 1964, Cluff, Thornton and Seidl (1964) argued that 5% of admissions resulted from the iatrogenic effects of pharmaceutical therapy, and 20% of patients suffered such effects during their admission to hospital. In a comprehensive review of observational studies of medication errors, Allan and Barker (1990) conclude that there is about one drug error per patient day, though the majority are minor and insignificant deviations from the prescription, and they present a detailed classification of the different types of error. Using

a more stringent definition of an adverse drug event, Bates, Cullen and Laird (1995) studied 4,031 admissions to 11 medical and surgical units in two acute hospitals over a six month period. They found 247 adverse drug events (involving 6.1% of admissions), of which 1% were fatal, 12% life-threatening for the patient, 30% serious and the remainder were less significant in terms of their impact. About 28% of these adverse drug events were preventable. They concluded that preventable adverse drug events were relatively common, and that serious events were more likely to be preventable.

Barr (1955) reports in passing what is perhaps the first study of the prevalence of iatrogenic disease in hospital. Over a period in which about 1,000 patients were admitted to a major US teaching hospital, he found more than 50 major "toxic reactions and accidents consequent to diagnostic or therapeutic measures". He concluded that, since it affected 5% of patients, "iatrogenic disease could be regarded as one of the commonest conditions encountered during the period [of the study]".

Schimmel (1964) undertook a more detailed study of the incidence and causation of iatrogenesis during hospital admissions to general medical services. He reviewed 1,252 admissions over an 8 month period to a teaching hospital medical department, and found that 198 patients (16%) suffered one or more iatrogenic events during their stay. He categorised the 240 events found as reactions to diagnostic procedures (29, 12%), therapeutic drugs (49%), transfusions (31, 13%), or therapeutic procedures (24, 10%); acquired infections (23, 10%); and other hospital hazards (14, 6%). Interestingly, Schimmel deliberately omitted from his study iatrogenesis which resulted from clinical errors or from previous treatment. McLamb and Huntley (1967) studied 240 patients admitted to the general medical department of a general hospital, and found 47 patients (20%) suffered one or more iatrogenic events during their admission - a total of 63 such events were found. Again, drug reactions were the commonest form of iatrogenic event (28, 44%). Both Schimmel (1964) and McLamb and Huntley (1967) used self-reporting mechanisms to gather data on iatrogenic events, which they acknowledge may have underestimated their incidence since some iatrogenic events may not have been reported.

More recently, Steel et al (1981) examined all patients admitted to a teaching hospital medical service over a five month period, using a structured review of all patients' records performed by the

researchers. They found that of 815 admissions, 290 patients (36%) suffered one or more iatrogenic events during their hospitalisation. Of the 497 iatrogenic events found, 208 (42%) resulted from drug therapy, 175 (35%) from diagnostic or therapeutic procedures, and 114 (23%) from other sources, particularly patient slips and falls. In 15 patients (5% of those with iatrogenic events, and 2% of all patients admitted) the iatrogenic event was believed to have contributed to the patients' death.

Couch et al (1981) conducted a one year prospective study focused on serious errors in care in the general surgical service of a large teaching hospital - what the team termed surgical mishaps. Over the period 5,612 admissions were screened for mishaps, with flagged cases being discussed by the clinical team. Only those cases which involved "violation of basic surgical principles" were confirmed as surgical mishaps. The study found 36 (0.6%) patients who suffered a total of 56 serious errors, of which 9 (16%) were diagnostic errors and 47 (84%) were therapeutic. That these cases represented only the most serious instances of avoidable iatrogenic disease can be seen from the poor outcomes for the patients involved - 20 (56%) died, and 5 (14%) were left with serious physical impairment. Couch et al asserted that the errors they found had five main causes, all of which related to physician behaviour and performance: misplaced optimism about patients' state of health or physicians' own ability; unwarranted urgency in undertaking procedures; the urge for perfection resulting in unnecessary surgery; the use of vogue therapies of unproven effect; and insufficient restraint and deliberation before making diagnostic and therapeutic decisions.

These studies of iatrogenic disease are summarised in table 3.7 below, which sets out the setting, methodology and results of each in a common format. It can be seen that while there are major variations in the apparent incidence of iatrogenic events, reflecting the differences in definition discussed earlier, there is some consensus that a substantial proportion of these events (from a quarter to a half) were avoidable.

Study	Setting	Methodology	Episodes examined	All iatrogenic events	Avoidable iatrogenic events	Avoid-ability (%)
Lakshmanan Hershey and Breslau (1980)	Admissions to medical service of teaching hospital	Prospective implicit review of all records by researchers.	834 admissions	45 (5.4%) of admissions	22 (2.6%) of admissions	49%
Trunet, Le Gall, Lhoste et al (1980)	Admissions to intensive care unit	Prospective implicit review of all records by researchers.	325 admissions	41 (13%) of admissions	19 (5.8%) of admissions	46%
McDonald, Royle, Taylor et al (1991)	Mortality in surgical unit	Review of all deaths by clinical team involved to assess avoidability.	543 deaths	n/a	89 (16%) of deaths	16%
Heywood, Wilson and Sinclair (1989)	Mortality in African surgical unit	Review of all deaths by researchers.	80 deaths	n/a	35 (44%) of deaths	44%
Dubois and Brook (1988)	Mortality in 12 selected US hospitals	Independent retrospective review of deaths by three physicians.	182 deaths	n/a	25 (14%) to 49 (27%) deaths	14% to 27%
Bates, Cullen, Laird et al (1995)	Adverse drug events in 2 selected acute hospitals	Prospective stimulated self-report followed by case review	4,031 admission	247 (6.1%) of admissions	69 (1.7%) of admissions	28%
Barr (1955)	Inpatients in general medicine	Methodology not reported.	1000 inpatients	50 (5%) of inpatients	n/a	n/a
Schimmel (1964)	Inpatients in general medicine.	Prospective self-reporting of iatrogenic events by physicians.	1252 inpatients	198 (16%) of inpatients	n/a	n/a
McLamb and Huntley (1967)	Inpatients in general medicine.	Prospective self-reporting of iatrogenic events by physicians.	240 inpatients	47 (20%) of inpatients	n/a	n/a
Steel, Gertman, Crescenzi et al (1981)	Inpatients in general medicine.	Prospective screening of all patients to identify iatrogenic events.	815 inpatients	290 (36%) of inpatients	n/a	n/a
Couch Tilney, Rayner et al (1981)	Inpatients in general surgery.	Prospective screening of all patients to identify iatrogenic events.	5612 inpatients	n/a	36 (0.6%) of inpatients	n/a

Table 3.7. Summary of studies of adverse events and iatrogenic disease.

Adverse events, negligence and malpractice

Three major studies of the incidence of adverse events have been undertaken to explore the likely extent of negligence and the implications for those concerned with the rising levels of medical malpractice litigation.

The first study of the epidemiology of malpractice in healthcare was undertaken in California in 1974, in response to dramatic rises in malpractice litigation and a consequent crisis in the insurance market in the early 1970s. The California Medical Insurance Feasibility Study (CMIFS) set out to examine the incidence and causation of malpractice, with a view to evaluating alternatives to a tort based system for patient compensation (California Medical Association 1977).

A sample of 20,864 inpatient admissions to California hospitals were selected, designed to be representative of the 3 million patients admitted to Californian hospitals during that year. Each patient's medical records were screened for *potentially compensable events* (PCEs), which the study defined as "a temporary or permanent impairment of physical or mental function (including disfigurement) or economic loss in the absence of such impairment, which is caused by healthcare management" (Mills 1978). The screening was performed using a structured list of 20 generic PCE types, which were used to flag records for review by one of the research team. Inconsequential events were not included, and all PCEs were categorised by their nature, severity, causation, and legal liability.

The study found that 970 patients (4.6%) had a PCE, of which the majority (796, 82%) resulted from the adverse effects of treatments or procedures; 144 (15%) involved incomplete or delayed diagnosis or treatment; and 30 (3%) resulted from incomplete protection or prevention. The severity and assessed liability for the PCEs is shown in table 3.8. Overall, 94 patients (9.7% of those with PCEs) died as a result of their PCE, and 165 patients (17% of those with PCEs, and 0.79% of all patients in the study) suffered a PCE for which the researchers judged the healthcare provider was legally liable. The majority of PCEs were caused by therapeutic procedures (641, 66%) or drug therapy (182, 19%), and most (696, 72%) occurred in or originated in the operating theatre (Mills 1978).

Severity of PCE	No of PCEs	No of liable PCEs	Liable PCEs as % of all PCEs
Temporary	776 (80%)	92 (56%)	12%
Minor permanent	63 (6.5%)	15 (9.1%)	24%
Major Permanent	37 (3.8%)	18 (11%)	49%
Death	94 (9.7%)	40 (24%)	43%
All PCEs	970 (100%)	165 (100%)	17%

Table 3.8. Severity and assessed liability of potentially compensable events in the California Medical Insurance Feasibility Study. Mills (1978).

The findings of the California Medical Insurance Feasibility Study (CMIFS) provided the first strong evidence that malpractice was a not uncommon occurrence, affecting almost 1 patient in 100 in the study sample - far more than the proportion of patients who subsequently made legal claims for malpractice.

More recently, the Harvard Medical Practice Study (1990) was established to undertake a more detailed examination of the incidence and causation of malpractice, with a greater focus on the relationship with malpractice litigation. The study had four main aims: to establish a population based measure of the incidence of adverse events and negligent adverse events; to assess the economic effects of these events on patients; to determine what proportion of these events actually lead to litigation; and to examine whether the threat of litigation has any value in deterring malpractice (Hiatt et al 1989; Brennan et al 1991b; Leape et al 1991).

A sample of 30,121 admissions to 51 hospitals in New York State during 1984 was selected, structured to be representative of all patients admitted to hospitals in New York during that year. These records were screened using a generic instrument not dissimilar to that applied by Mills (1978), to identify adverse events - which the study team defined as "unintended injuries caused by medical management rather than by the disease process, and sufficiently serious to lead to prolongation of hospitalisation or temporary or permanent impairment or disability in the patient" (Harvard Medical Practice Study 1990). Each adverse event discovered was reviewed by a team of

physicians using a structured analysis process designed to gather information about its nature, causation, and assessed liability.

Of the 30,121 patient admissions, 1,278 (4.2%) suffered an adverse event during hospitalisation, of which 280 (22%) were judged to have resulted from negligence. Adjusting these figures for the population of admissions to New York hospitals in 1984 indicates that 3.7% of all patients admitted to hospital suffered an adverse event, and that 27.6% of all adverse events were due to negligence. Thus, 1.0% of all patients suffered a negligent adverse event during their admission. While most adverse events resulting in minimal or transient disability, 14% caused or were implicated in the patient's death (and 51% of these were judged to be negligent). Almost half (48%) of adverse events resulted from an operation - including wound infections (14%) and technical complications (13%). The commonest nonoperative causes of adverse events were drug therapy (19%) and diagnostic errors (8.1%).

The Harvard Medical Practice Study undertook a painstaking investigation of the economic costs to patients of the adverse events it discovered, which was complicated by the need to separate out the costs of the adverse event and the underlying illness. They found that most patients (76%) suffering an adverse event returned to normal functioning within 6 months, and their economic losses were minimal. However, for those patients who suffered the effects of the adverse event beyond 6 months, the per capita costs were high, and while 86% of the consequent medical costs were met by health insurance, only 19% of the earnings loss was met by sick leave or disability insurance.

When the results of the study review of 30,121 patient admissions were compared with the New York malpractice claims records, it was found that 47 patients had made malpractice claims relating to their admissions. Of these 47 patients, 18 (38%) had been judged by the study team to have suffered an adverse event, and only 8 (17%) had been rated as a negligent adverse event. Thus, only 1.5% of negligent adverse events actually led to a malpractice claim, and 83% of malpractice claims filed related to care that had been judged to be adequate (Localio et al 1991). In view of these results, it is unsurprising that the study concluded that there was no evidence that the threat of malpractice litigation had any effect on the incidence of adverse events or negligent adverse events.

Researchers in Australia used a very similar methodology to those described above to examine the quality of care and the incidence of adverse events in two states during 1992 (Wilson et al 1995). Although the study was not explicitly focused on levels of negligence and malpractice, it is reviewed here because it was so closely modelled on the Harvard Medical Practice Study (1990). The researchers reviewed 14,179 admissions to 28 hospitals in two Australian states using essentially the same definition and tool for identifying adverse events as the Harvard Medical Practice Study (1990). Instead of using reviewers to judge whether adverse events were negligent, as both earlier studies had done, they instead sought an assessment of how preventable it was, and what disability it caused. They found that 16.6% of patients suffered an adverse event, of whom 46.6% suffered no or minimal disability, 48.8% suffered some moderate temporary or permanent disability, and 4.9% died. Reviewers judged that 51% of adverse events had high preventability, 29.8% had low preventability and 19.0% were not preventable at all.

The three studies are summarised in table 3.9 below. While the Californian and Harvard studies are broadly in consensus, about both the incidence of adverse events and the proportion of negligent adverse events, the Australian study has rather different findings. It suggests that adverse events are almost four times as common in Australia as in the USA, and almost half are judged preventable. There is no obvious explanation for this discrepancy, which could reflect differences in the research design, cultural or other differences in the implicit review process on which all three studies are founded, or differences in the actual quality of care (Brennan 1995).

Study	Setting	Methodology	Episodes examined	All adverse events	Negligent adverse events	Negligence (%)
Mills (1978)	23 hospitals in California	Initial screening for potentially compensable events followed by case review	20,864 admissions	970 (4.6%) of admissions	165 (0.79% of admissions)	17%
Harvard Medical Practice Study (1990)	51 hospitals in New York State	Initial screening using generic adverse-event measure followed by case review	30,121 admissions	1,278 (4.1%) of admissions	280 (0.93% of admissions)	22%
Wilson, Runciman, Gibberd et al 1995	28 hospitals in 2 states in Australia	Initial screening followed by case review using modified Harvard Study measure	14,179 admissions	2,353 (16.6%) of admissions	8.3% of admissions (note - preventability used, not negligence)	51.2%

Table 3.9. Summary of studies of adverse events, negligence and malpractice.

Adverse events and quality assurance

There are many anecdotal reports of experience in using adverse-event measures of the quality of healthcare in quality assurance, some of which are reviewed later in this chapter in an assessment of what can be learned about them from their application. However, despite the widespread use that has been made of these measures, there are rather fewer quantitative reports of their results.

Goldman and Walder (1992) report on the experience of the Veterans Administration healthcare system in the USA in using the adverse-event measure of quality which was described earlier. Over a six month period in 1989-90, computerised screening using 9 criteria identified 57,841 admissions out of a total of 435,000 to be reviewed. On review by quality assurance staff, 10,698 admissions were referred for peer review by medical staff. The quality of care was judged to be acceptable in 80.3% of cases, but 14.5% of cases "might have been handled differently" by experienced and competent practitioners, and 5.2% "should have been handled differently". No separate assessment of the avoidability of these adverse events was made.

Bomhof, Arends and van der Beek (1993) report on a study of adverse events in the ENT department of an acute hospital. Using an adverse-event measure developed with clinicians in the specialty, all admissions during 1989 were screened both by specialists within the department and separately by an external medical reviewer. They found that 16-20% of admissions had one or more adverse events, and that the ENT specialists were much more likely than the external reviewer to identify adverse events which were specific to the specialty. No assessment of the effects or avoidability of the adverse events was made.

Walshe, Bennett and Ingram (1995) used a part-generic and part-specialty specific adverse-event measure to review 1,088 admissions in ophthalmology. They found that while 64.2% of patients had no adverse events, 25.4% had one adverse event, and 10.5% had two or more adverse events. "Clinically relevant" adverse events constituted 37.2% of all events found, and included a variety of surgical complications. Of the 31% of adverse events which were sent to a clinician for peer review, 41.3% were assessed to have had no effect on the patient, 33.9% had a minor effect, and

10.5% had a major effect (some could not be assessed). About 21% of adverse events subjected to peer review were judged to have been avoidable.

Wolff (1996) used a set of 8 generic screening criteria to review all inpatients admitted to a small acute hospital between 1991 and 1994. A total of 15,912 admissions were reviewed, of which 1,465 (9.1%) were found to have one or more adverse events according to the criteria in use. However, all these cases were then subject to peer review by medical staff, and only 155 (0.97% of all admissions and 10.6% of those initially identified by the criteria) were confirmed as adverse events. A range of actions were initiated as a result for the 43.2% of adverse events which were judged to have been preventable in some way.

These studies are summarised in table 3.10 below. It can be seen that comparisons are difficult, particularly between the incidence rates of adverse events, because of the very different definitions of adverse events which were adopted.

Study	Setting	Methodology	Episodes examined	Proportion with at least one adverse event	Avoidable or preventable adverse events	Avoidability (%)
Walshe, Bennett and Ingram (1995)	Admissions in ophthalmology	Prospective screening of patient records using an adverse-event measure	1,088 admission	35.8%	Not all reviewed.	21% (of those reviewed)
Wolff (1996)	Admissions to small acute hospital	Prospective screening using adverse-event measure	15,912 admissions	9.2% on screening; 0.97% on peer review	43.2% of confirmed adverse events	43.2%
Goldman (1992)	Admissions to VA healthcare system	Prospective automated screening using basic criteria, followed by QA staff review and peer review	435,000	10,698 (2.5%) of admissions	485 (0.1%) "should have been managed differently"	4.5%
Bomhof, Arends and van der Beek (1993)	Admissions in ENT in acute hospital	Prospective screening of patient records by both specialists in ENT department and external medical reviewer	921 admissions (not all seen by both reviewers)	16% (specialists review) 20% (external reviewer)	n/a	n/a

Table 3.10. Summary of studies of adverse events and quality measurement.

Conclusions

It must be acknowledged that some of the studies outlined above (summarised in tables 3.7, 3.9 and 3.10) have serious methodological flaws, which limit the generalisability and power of their individual findings. It is also clear that the adoption of different definitions and methods in identifying adverse events influences the results of the studies, and so comparisons, even among those which appear to be methodologically compatible, are invidious. However, three common themes emerge.

a) *Causation.*

While a minority of adverse events can be clearly identified as having resulted from the healthcare process, establishing causation is often complex. While most of the studies cited paid attention to the causation of the adverse events they recorded, some sought stronger and more immediate evidence than others. Most made use of implicit review by clinicians to determine causation, a process which is of unproven reliability and validity. Most viewed causation rather simplistically, focusing largely on determining whether or not the healthcare system was "at fault" rather than exploring the underlying causes and predisposing factors which caused or enabled the event to occur.

b) *Severity.*

All the studies are focused on adverse events which involve (or may involve) some harm to patients, but the severity of these events varies. For example, Steel et al (1981) studied any illness resulting from a diagnostic or therapeutic intervention, however minor, while Couch et al (1981) focused on instances of serious harm. Unsurprisingly, less severe events tend to be commoner, with the result that the severity threshold used in the study has a critical effect on the incidence of adverse events found. The wide variation in the reported incidence of adverse events in the studies reviewed is thought to be largely a result of this definitional variation rather than of any underlying differences in the quality of care.

c) *Avoidability*.

The concepts of avoidability, preventability and (sometimes) culpability are mentioned in many studies, and addressed in different ways. Not all adverse events are avoidable, and some studies approach the problem by trying to identify all events and then separate out those which are avoidable, while others try to exclude unavoidable adverse events and focus solely on those events which are avoidable. In most studies, the avoidability of events is assessed through implicit reviews - again, these are, at best, of unproven reliability.

These studies concur in one key respect. They suggest that adverse events are an important and potentially preventable component of the demand for healthcare. Of all admissions to hospital, the research implies that around 5% result from adverse events, and half of those are preventable. While patients are in hospital, around 20% suffer some form of adverse event, up to half of which are preventable, and a proportion of which involve major or life-threatening illness. It is clear from the research that these patients stay longer in hospital and cost more to treat. Other studies show that about 20% of the deaths which occur in hospital are preventable, and result in part from adverse events. In short, there is ample proof that adverse events are important both in terms of the resources they consume and their impact on the quality of care.

Unsurprisingly, most studies call for more attention to be paid to the incidence, causation and prevention of adverse events. Couch et al (1981) argue that "all hospitals should continuously survey final results ... to determine the prevalence of avoidable misadventure, and evolve surveillance and educational machinery that can minimise them". Steel et al (1981) concur, calling for "new methods of monitoring untoward occurrences in hospitalised patients, especially on medical services". However, most studies pay far more attention to describing and delineating the incidence and nature of adverse events than they do to developing or testing strategies for their prevention. Most simply attribute the preventable or avoidable component of adverse events to the medical staff involved - writing of the "errors and lack of attention" on the part of physicians (Lakshmanan et al 1986), and calling for educational programmes to improve doctors' performance (Trunet et al 1980; Steel et al 1981). In fact, the adverse events identified and documented in the studies reviewed above should be seen in the context of the wider literature on errors and the human

and organisational factors that cause them (Reason 1995) and the strategies for quality improvement which are needed to prevent them (Berwick 1996).

3.3.3 **Using adverse events in quality improvement**

It could be reasonably argued that, despite the relatively limited scientific literature on their development, adverse-event measures of the quality of care have been among the most widely used techniques in healthcare quality assurance. In the terminology of Dick and Hegarty (1971), they have a high "cash validity". Some examples of the use of adverse events in quality assurance are presented below.

Occurrence screening programmes in US healthcare

While there are no definitive statistics available, the vast majority of US healthcare providers make some use of adverse-event measures in their quality assurance, risk management and associated activities. The remarkable size and scope of quality assurance programmes in the USA healthcare system has already been discussed in chapter 2. Methodologies based around the detection and analysis of adverse events play a central role in those programmes. The examples outlined below illustrate the approaches to using adverse event information in two hospital quality assurance programmes and a major hospital chain, and also describe an archetypal quality assurance programme used as a model by many hospitals and published by the American Hospital Association.

The Medical Management Analysis (MMA) programme which was outlined earlier has been applied in at least two hundred US hospitals. The Good Samaritan Medical Center in Phoenix, Arizona was one of the earliest users of the programme, in 1981. It used six full-time screening staff, with nursing backgrounds, to perform concurrent reviews of all patients admitted to the 770 bedded hospital. About 15% of patients were found to have an adverse event, and about one third of these events were passed on to physician reviewers to assess severity, attribution and the standard of care. There were close links between the programme and the risk management department of the hospital. The programme is claimed to have identified a series of important quality problems,

through trends in adverse events, including defective medical supplies, technical problems relating to a particular surgeon's ability, and a series of avoidable obstetric complications resulting from a lack of staff training and supervision. Other hospitals using the MMA programme have reported that it has assisted in identifying and rectifying quality problems related to physician behaviour and knowledge, medical records documentation, pharmacy practice, and excessive departmental workloads (Craddick and Bader 1983, p41).

The Johns Hopkins Health System is a large healthcare provider based around the tertiary teaching hospital of Johns Hopkins in Baltimore, Maryland. It has an extensive and detailed quality assurance programme, which is centred around concurrent quality monitoring for adverse events, carried out by quality assurance staff with a nursing background. These staff visit patients on a daily basis, reviewing both resource utilisation and the quality of care. In each review they use a total of 49 adverse event definitions, subdivided into events concerned with admission, documentation, medical care, surgical care, drug usage, nonclinical events, discharge and, other miscellaneous events. When problems are identified, they are taken up by a physician quality reviewer, who is responsible for taking follow-up action to resolve the issue. Information on the incidence of adverse events is used by hospital quality assurance committees and the governing Board to monitor the quality of care (King and Jones 1989).

The Veterans Administration (VA) healthcare system's extensive use of an adverse-event measure of quality has already been discussed. There have been mixed reports of the effectiveness of the approach in improving the quality of care (Goldman and Walder 1992). For example, the VA Medical Center in Pennsylvania, reporting in positive terms on the process of establishing its successful occurrence screening programme, argued that the planning and implementation of the programme were very important in ensuring its acceptance (Citro et al 1988). They suggested that occurrence screening had brought about timely corrective action through early problem identification, better problem resolution, better use of quality assurance resources, more effective use of clinicians' time, and greater integration of physicians into the quality assurance process. However, in contrast the VA Medical Center in Milwaukee suggested that their occurrence screening programme revealed few significant quality problems and was expensive to undertake (Erdmann 1990).

Integrated Quality Assessment (IQA) programme

The Integrated Quality Assessment (IQA) programme was developed by Longo and colleagues (Longo, Ciccone and Lord 1989). It is, in many ways, an updated equivalent of the MMA programme originated by Craddick and colleagues. The programme is designed to integrate four key US hospital functions - quality assurance, risk management, infection control and utilisation review. At the heart of the programme is a process of concurrent review, in which patients' care and treatment is screened on admission, periodically during their stay, and after discharge. This screening process combines the abstraction of information on adverse events with the assessment of care against predefined quality standards, the measurement of illness severity and resource utilisation, the detection of infection, and focused reviews of other issues.

The IQA programme uses occurrence screening alongside other quality measurement methodologies, rather than relying solely on a single technique. It also allows for greater flexibility in determining the types of adverse events which are screened for, and the way in which the screening process is managed. The authors of the IQA programme argue that its key benefits are the synergies and resource savings of integrating the four functions it contains, and its ability to identify and rectify quality problems while patients are still in hospital.

Occurrence screening in the Medicare PRO programme

The Medicare PRO programme's use of occurrence screening, based on a generic adverse-event measure of quality, was discussed earlier. Over the 1986-88 period, PROs performed over 6.6 million case reviews. Payment to the provider was denied in 4.2% of cases, though this figure ranged from 1.2% in one PRO to 25.5% in another. PROs identified over 87,000 physicians during this period with some level of quality problem, and undertook over 70,000 specific quality interventions. However, very few cases result in serious sanctions - just 83 physicians and one hospital over this period were excluded from the Medicare programme (Institute of Medicine 1990). The PRO programme has been monitored and evaluated continually since its inception, though this evaluation has focused on the timeliness and procedural correctness of the process, rather than its effectiveness in measuring and improving quality (Institute of Medicine 1990, p182). In general, the programme has been criticised for being overfocused on utilisation issues, to the detriment of

quality concerns, for its centralised and inflexible review strategies and systems, and for the ambiguity and conflicts between its two roles - helping providers to improve the quality of care, and sanctioning or excluding poor providers. The generic quality screens used by PROs have been widely criticised as both unreliable (in that different PROs yield quite different results, because they interpret and apply the generic screening tool differently) and insufficiently focused on genuine quality problems. In particular, they have produced high rates of false positives (cases which fail the generic screen but actually have no quality problem) which has resulted in many costly and unnecessary physician reviews. Having co-ordinated an Institute of Medicine evaluation of the programme (Institute of Medicine 1990), Lohr (1990) concluded that "the current system to assess and ensure quality is in general not very effective, and may have serious unintended consequences". Although the PRO programme has since been revised, case review using an adverse event measure of quality remains an important component of its approach (Jencks and Wilensky 1992)

Use of adverse events in risk management

As a result of the high levels of malpractice litigation discussed earlier, most American hospitals have established risk management programmes, designed to minimise the hospital's exposure to risk, by detecting instances of malpractice and acting both to minimise their cost to the hospital and to prevent recurrence (Mills and von Bolschwing 1995). Generally, these use either an occurrence reporting system or some form of occurrence screening to identify adverse events in which there may be an element of malpractice. These events are managed proactively, to try to settle them simply through communication with the patient or by making a small compensatory payment, with the aim of avoiding litigation. Cases which do involve litigation are also managed proactively with the twin aims of minimising legal costs and securing a favourable legal outcome. Perhaps most importantly, the information about adverse events is used to plan and implement prevention strategies, which may involve making organisational or individual changes in practice. There is empirical evidence from a US study that hospitals with risk management programmes are subject to fewer successful legal claims and have lower total awards against them than those which do not use risk management, though it is difficult or even impossible to determine the causes of this relationship (Morlock and Malitz 1991). Recently, most UK healthcare providers have also established risk management programmes, in the face of growing levels of malpractice litigation and changes in their legal liability (Bowden 1990; Clements 1995). While there are few, if any,

examples of UK healthcare providers screening for adverse events as part of their approach to risk management, almost all have established new reporting systems under which clinicians are expected to report adverse events.

Canada

While adverse event measures of quality have not been as widely used in Canada as in the USA, a survey of practice in one province suggested that 25% of healthcare providers had occurrence screening programmes in place (Barrable 1992), and their are a number of anecdotal accounts of the use of such approaches (Carlow 1988) For example, Nordal and Ang (1988) describe the development of an occurrence screening programme at the Queen Elizabeth Hospital in Toronto where, they claim, it has had "tremendous value as a tool for reducing our exposure to risk situations, ensuring appropriate utilisation of hospital resources, and, most importantly, improving the quality of patient care".

Australia

The first Australian institution to establish a quality assurance programme based on adverse event analysis was the Royal North Shore Hospital in Sydney. Initially, the hospital adopted the MMA programme outlined earlier almost in its entirety. However, dissatisfaction with the costs involved and antagonism from the hospital's medical staff towards the programme soon led to its redesign. A much smaller team of screening staff was established, to undertake screening of a targeted sample of patients in certain groups with a higher risk of adverse events. The mechanisms for peer review in specialties adopted from the MMA programme had not worked well, and a smaller central team of medical assessors was established. With these revisions, the programme has became more effective and efficient (Stuart 1989). Other Australian hospitals have also experimented with occurrence screening, albeit on a smaller scale than the Royal North Shore Hospital project. The Royal Perth Hospital has undertaken a series of occurrence screening studies in individual specialties (Lawrence-Brown and Manning 1989; Manning and Lawrence-Brown 1990) which have concluded that the approach is "worthwhile, productive and rewarding" (Lawrence-Brown and Manning 1989). At the Royal Adelaide Hospital, a comprehensive screening instrument covering both adverse events and other issues was developed and piloted, but the project lacked the cooperation and involvement of many medical staff, and unsurprisingly its findings were inconclusive (Burr 1990).

In a rather different setting, the 200-bed Wimmera Base Hospital established an occurrence screening programme in 1988, and reported that it had identified a series of important quality problems, resulting in changes to protocols for transfusion, better preoperative assessment of fitness for anaesthesia, and reduced misprescribing of diuretics (Wolff 1992; Wolff 1995). Wolff concluded:

> "Occurrence screening provides an efficient, integrated and coordinated approach to quality assurance in hospitals. Occurrence screening using a limited number of criteria and retrospective review has the potential to provide a simple, inexpensive and effective medical quality control system suitable for medium size hospitals." (Wolff 1992)

United Kingdom

Systems based on the detection and analysis of adverse events in healthcare have existed in the National Health Service (NHS) for some time. The longstanding Confidential Enquiry into Maternal Deaths (Department of Health 1991a) and the more recent National Confidential Enquiry into Perioperative Deaths (NCEPOD 1989) are both studies of a single type of adverse event - unexpected or potentially avoidable deaths. Most hospitals have systems for recording certain sorts of adverse events, such as patient accidents and medication errors. Many pharmacy departments routinely review patient records to identify drug prescribing or administration problems and intervene to change practice. However, these systems have been fragmented, isolated, and often little used in quality measurement or improvement.

The first trial of an adverse event based quality measure took place at Hove Hospital, a small acute hospital in Sussex (Stevens and Bennett 1989). A sample of 250 cases were screened retrospectively for adverse events, using a set of generic screening criteria based on those developed by Craddick. The trial, which was primarily directed at testing the methodology, found that over 90% of cases in the sample, which was drawn from three specialties, could be screened for adverse events in less than 10 minutes, and that 20% of cases in the sample had one or more adverse events. It concluded that the approach "could be practically applied in a British hospital" (Stevens and Bennett 1989).

Other British hospitals have also made some use of occurrence screening, in their growing clinical audit and quality assurance programmes. For example, Bromley Health Authority established an extensive system of occurrence screening, based closely on the MMA model (Lyall 1990). Using 11 nursing staff to perform concurrent reviews of all patients throughout the district's acute services. Writing shortly after the system's introduction, Lyall (1990) reported that improved recordkeeping and fewer patient complaints had resulted from the occurrence screening programme. In a smaller study in Bath, Lewis and Charny (1992) developed and used specialty-specific adverse-event screening criteria, and outside the hospital setting, Smith (1992) reported the small scale application of an adverse-event measure in general practice.

A research project was established at the Royal Sussex County Hospital in Brighton in 1989 to test the use of adverse events in healthcare quality assurance in the UK (Walshe, Bennett and Ingram 1995). It had four main aims:

a) To investigate the reliability and validity of adverse-event measures in the measurement of the quality of healthcare in acute services.

b) To establish whether the information produced by adverse-event measures of quality could be analysed and fed back to clinicians in a timely and relevant format which they found useful in medical audit and quality assurance activities.

c) To assess whether the provision of information about adverse events, through adverse-event measures of quality, identifies areas where quality improvements are needed, and whether those quality improvements are made.

d) To identify the costs involved in applying adverse-event measures of quality in the setting of a British district general hospital.

The development of adverse-event measures commenced in January 1990, and screening for adverse events began in February 1990. Screening for adverse events continued until the end of the

project in April 1992. During that 26 month period, adverse-event measures were developed and applied in a total of 12 different specialties, for periods ranging from 4 months to 26 months. In aggregate, over 13,000 patient episodes were screened for adverse events during the project's life, involving 12 specialties (Walshe, Bennett and Ingram 1995). The research reported in later chapters of this thesis was carried out as part of this project.

3.4. **Evaluation of adverse-event measures of quality**

3.4.1 **Merits and demerits of adverse-event measures**

Philosophical and conceptual issues

Our attitudes to approaches to quality measurement are shaped by a combination of theoretical and methodological innovation, practical experience and empirical findings, and developments within the healthcare system itself. At times, certain methodologies and theories become in vogue, then lose popularity as the collective consensus about their value shifts. For example, in the USA criterion-based quality assessment methodologies were very widely used in the early and mid-1970s, then fell into disrepute as concerns about their reliability and effectiveness grew, and then returned to favour again in the late 1980s and 1990s. In the late 1970s and early 1980s, problem-focused approaches to quality became widely preferred, and adverse-event measures of quality came to predominate in quality assurance activities.

However, in the late 1980s, they too began to be criticised for a number of reasons by both researchers and practitioners in the field - especially by those espousing the ideas and methods of continuous quality improvement (O'Leary 1991; Laffel and Blumenthal 1989). In particular, it was suggested that adverse-event measures focus attention on inspection for quality problems (after the event), rather than on the prevention of quality problems. It was argued that they promote a view of healthcare workers which sees them as potentially deficient and incompetent rather than as generally committed to high quality, and that these measures tend to present quality problems as the result of individual practitioners' behaviour rather than as the result of system or process mechanisms. It was also asserted that adverse-event measures focus attention on bringing outliers - cases or practitioners different from the norm - into the mainstream, rather than making the quality of care in the mainstream improve (in the language of CQI, shifting the distribution), and that the resulting quality assurance activities address individual cases rather than aggregate trends

There is certainly some justification for these arguments, and they are explored further below. However, it also seems that they confuse the properties of adverse-event measures of quality with

the characteristics of their application in a particular setting - the US healthcare system of the 1980s. It is undoubtedly true that adverse-event measures of quality have been used as described above, in an adversarial and inspection-based quality paradigm of the sort that Berwick (1989) and others regard as unhelpful. However, the measures themselves are not inherently so directed - it has been the application of the information they yield in monitoring physicians and hospitals that perhaps deserves criticism, rather than the measures themselves.

It can also be argued that the concept of using adverse event information is actually central to CQI. The oft-quoted maxim "every defect is a treasure" (Berwick 1989) embodies the cardinal concept of CQI, that quality problems represent opportunities for quality improvement, and so their discovery should be valued and prioritised. The pioneers of CQI outside healthcare, use adverse event information extensively in measuring systems' performance and identifying avenues for process improvement. Deming (1986, p203), offers a series of potential quality indicators in healthcare, of which two thirds are adverse event based. Kritchevsky and Simmons (1991) make the importance of adverse event information in CQI very clear. They maintain that to improve quality, an understanding of the nature of quality problems within a system must be matched by monitoring of the incidence of those problems, study of their causes, and evaluation of the efficacy of changes to the system in reducing their incidence.

In conclusion, adverse-event measures are an important (and, arguably, a crucial) part of quality assessment. Like any measure, they can be misused and misapplied, but their value to healthcare quality assurance should be assessed on their genuine merits and demerits (discussed below) rather than on instances of their misapplication.

Merits of adverse-event measures

Developers and users of adverse-event measures of quality have cited a wide range of perceived merits (Craddick 1979; Craddick and Bader 1983; Bennett and Walshe 1990; Wolff 1992; and others):

a) *Systematism.*

The occurrence screening process outlined in previous sections of this chapter is highly systematic. The centralised process of case review by screening staff provides a controllable and manageable mechanism for data collection in which sample selection, data collection methods, and other aspects can all be systematically determined and varied as required. It also facilitates rigorous training and regular monitoring of screening staff (since relatively small numbers of staff are involved) which should help to establish and maintain high standards of reliability.

b) *Integration*.

Since occurrence screening can combine data collection for infection control, utilisation review, quality assurance and risk management, it is often perceived or presented as an efficient and integrated methodology, able to serve multiple separate but related data needs.

c) *Flexibility.*

Despite the centralisation of the data collection mechanism, occurrence screening can make use of a wide range of different adverse event definitions, tailored for specific specialties or patient groups. Detailed studies on specific sorts of adverse events can be performed, and new adverse-event measures can easily be incorporated. This flexibility allows occurrence screening systems to respond rapidly to new data needs or to clinician requirements.

d) *Objectivity*.

Since adverse-event measures are designed to detect incidents which are generally agreed to represent or indicate a poor quality of care, the resulting information is often claimed to be more objective than that gathered by less clearly defined and explicit methodologies.

e) *Use of clinician time*.

When adverse-event measures are used, clinicians are generally not involved in the process of primary review. Usually, clinicians are involved in deciding what adverse-event measures will be used, and then in the subsequent review of those cases where adverse events are found - in assessing the individual incidents, and in planning and implementing

quality improvements. This, it can be argued, minimises the amount of clinician time which must be devoted to quality assurance, and makes the most effective use of that time in activities which genuinely require the clinician's expertise.

f) *Timeliness*.

Both concurrent and retrospective review for adverse events produce information close to the time when events actually occurred which, it is argued, makes review processes more effective and the alteration of practice more attainable. Indeed, proponents of concurrent review assert that through the real-time detection of adverse events, their effects on patients can be minimised and potential subsequent adverse events can be prevented.

g) *Patient focus*.

Adverse-event measures of quality are patient focused, gathering data about all aspects of the care provided to patients. Some authors suggest that many adverse events result from interdepartment or interprofession rather than intradepartment or intraprofession problems. A profession or department centred quality measure might miss those quality problems, because they lie on the periphery of each individual department's or profession's concerns. By contrast, a patient focused measure can identify these interdepartment or interprofession issues. It is also suggested that adverse-event measures, since they are centred on undesirable events which impact (or potentially impact) on patients themselves may provide a more valid measure of quality than other measures which are more driven by clinicians' interests or provider concerns and activities.

h) *Causation focus*.

Most adverse-event measures of quality make some use of an investigation of the causes of adverse events. It is contended that this attention to the processes and circumstances which cause events makes it easier to identify suitable improvements to systems or processes to prevent similar events in the future.

i) *Corrective action focus*.

Quality assurance programmes using adverse-event measures of quality generally have formal mechanisms for recording the causes of adverse events and the action taken to rectify or prevent them. It has been claimed that this focus on taking and recording corrective action makes these programmes more likely to succeed in producing and maintaining quality improvements.

j) *Liability limitation*.

There is frequently a link between adverse-event measures of quality and parallel risk management programmes, which are aimed at reducing medical negligence liability. Thus the liability implications of adverse events are actively considered, and potential liability is a factor in planning and implementing quality improvements. Even where such a link does not exist, it is argued that the focus on adverse events (and their causation and prevention) is likely to have some limiting effect on levels of liability.

k) *Trend identification*.

The ability to link together series of adverse events and thereby to discover significant trends or patterns (sometimes called *trending*) is often cited by users and developers of adverse-event measures of quality as an important advantage, in that it can reveal quality problems which might otherwise remain undetected. In particular, it is argued that systematic screening for adverse events across many departments or specialties can highlight causal links between isolated events in different areas, or can reveal a series of less serious adverse events in time to allow action to be taken to prevent a potentially much more serious event.

l) *Cost savings*.

If adverse-event focused quality programmes are successful in preventing future adverse events, they are likely to reduce overall healthcare costs. Within the healthcare organisation, many adverse events will be followed by reworking or remedial action, which is costly (for example, the return to theatre of a patient where the first operation was not successful). Moreover, the liability limitation effects described above should also reduce the healthcare organisation's insurance or litigation costs.

Demerits of adverse-event measures

While the developers and users of adverse-event measures of quality have asserted they offer many advantages, their critics have suggested that there are important deficiencies in such measures which limit their usefulness (Sanazaro and Mills 1991; O'Leary 1991; Laffel and Berwick 1992; and others):

a) *Negativity*.

Perhaps the most common criticism of adverse-event measures is that they focus attention and quality improvement efforts on the negative (what has been labelled *disquality*), to the exclusion of other important facets of quality. This may bias our assessment of quality in a number of ways - for example, towards technical quality and away from interpersonal or art-of-care quality issues. It is also argued that the negativity of adverse-event measures arouses defensiveness and self-justification, and leads, almost inevitably, to a quality assurance process which is negative, punitive, and adversarial in its orientation.

b) *Outlier bias.*

It has been argued that adverse-event measures tend to focus attention on outliers - patients or clinicians whose care lies outside the patterns of the majority - rather than on quality issues concerning the practices and treatment of all practitioners. This may mean that poor quality is tolerated simply because it is endemic or part of accepted practice. Critics contend that this overattention to the periphery (rather than the centre) of practice reduces the ability of quality assurance to bring about meaningful quality improvements.

c) *Infrequent events*.

Serious adverse events are usually relatively infrequent occurrences. This means that detecting them may require the screening of many patient records where no event is found - so the cost of each actual event detected is quite high. The value for quality assessment of the negative information yielded by screening (that most patients do not have a serious adverse event) may be relatively limited. It also means that, even when a number of events are found, the small numbers involved make statistical comparative analysis practically

impossible. This infrequency of significant events is sometime referred to as the *small-n* problem.

d) *Difficulty defining events.*

Adverse-event measures are relatively easy to design for areas of healthcare in which patients with single pathologies are treated with clearly defined interventions to achieve explicit and measurable goals - examples might include obstetrics, or elective surgery. They are much harder to define for areas in which patients have multiple pathologies some of which may be chronic in nature, and receive many overlapping interventions - examples might include general medicine or psychiatry. In these areas, it may be very difficult to define meaningful adverse events in terms sufficiently explicit to permit screening. It may also be much more difficult, once adverse events are identified, to establish their causes.

e) *Patient safety bias*.

Adverse-event measures are sometime characterised as defining quality solely or primarily in terms of patient safety, to the exclusion of other important facets of quality, particularly appropriateness. It is argued that this overemphasis on safety may discourage worthwhile, desirable but risky interventions, and may stifle innovations in practice which would be to the long-term good of patients.

f) *Individual case bias.*

Although the advocates of adverse-event measures suggest that rates and trends in events can and should be used to detect and address quality issues, critics assert that they are not so used. They argue that too much attention is given to assessing the causation of and assigning responsibility for individual adverse events, drawing attention away from underlying systemic quality problems.

g) *Low reliability*.

Although the reliability of adverse-event measures has not been well researched, a number of authors have suggested that they are not particularly reliable, especially when definitions of adverse events are vague or non-specific, and much is left to the rater's discretion or

professional judgement. The evidence for the reliability of adverse-event measures is reviewed in detail below.

h) *Low validity*.

Some critics simply contend that there is little or no inherent relationship between the incidence of adverse events and the quality of care (though this depends on what is meant by quality of care). They argue that adverse-event measures may be able to evaluate levels of iatrogenic risk, or potential liabilities, but they are not good indicators of quality. The issues surrounding validity, and the available evidence, is reviewed in detail below.

h) *Reliance on implicit reviews*.

Many adverse-event measures make extensive use of professional review, frequently with no explicit criteria to guide those professional judgements. These implicit reviews are used to assess causation, to assign responsibility, to plan further action, and so on. However, these multitiered review systems can be very expensive to implement (especially if the primary review produces large numbers of cases for professional review, or if multiple reviews are required). There is also a well-established body of evidence that such reviews are, even in the best of circumstances, not necessarily reliable or valid. Hence the information they yield, at some considerable cost, may be of little value.

i) *Cost*.

The costs of using adverse-event measures of quality are dependent on the method and circumstances of their application. However, since some proponents of these measures have advocated concurrent review of the whole population of patients it is not surprising that some critics have labelled the process as unduly expensive. The limited empirical evidence available as to the costs of adverse-event measures is reviewed in more detail below, and the costs must be balanced against the potential costs savings outlined above.

Many of the advantages and disadvantages of adverse-event measures of quality outlined above are highly subjective, and open to the different interpretations of the proponents and critics of such measures. However, the validity and reliability of these measures do permit a more scientific,

objective and quantitative evaluation, and the available evidence on these important aspects of the measures' worth is outlined in the following sections.

3.4.2 **Validity of adverse-event measures of quality**

In chapter 2, the concept of validity was defined and explored, and a number of approaches to instrument validation were discussed. The importance of validating a measure in the context or environment in which it is to be used, and the desirability of using a number of validation techniques rather than a single approach were both highlighted. In this section, the relatively limited empirical evidence available for the face and content validity, criterion-related validity and construct validity of adverse-event measures is critically reviewed.

Face and content validity

Face validity is a measure of whether an instrument seems reasonable, and produces reasonable data, from the viewpoint of its users. Content validity is a measure of whether the items within an instrument adequately reflect the conceptual definition of its scope. Neither face nor content validity can be assessed empirically. Instead, the opinions of expert panels or user groups are sought, and this information is used to explore the face and content validity. In this relatively subjective form of validity analysis, it is important that a broad and representative range of opinion is sought; that the instruments used to gather opinions are themselves carefully designed and applied; and that a number of approaches (rather than a single approach) to assessing face and content validity are employed, so that the results of different approaches can be compared.

Disappointingly, the literature yields no substantial published studies of the face or content validity of any adverse-event measures of healthcare quality. Panniers and Newlander (1986) report briefly that they tested the content validity of the APO inventory (an adverse-event measure which was discussed in earlier), using an expert panel consisting of one doctor and three nurses. Regrettably, they do not offer any specific results from this exercise. Richards et al (1988) used a panel of three doctors to rate the "adversity" of each element of the APO inventory, and reported that "all of the weights obtained were negative, and the physicians generally agreed with one another in their

evaluations". The small size of the expert panels used and the brevity with which their findings are reported severely limit the usefulness of these two studies.

Criterion-related validity

Criterion-related validity is a measure of the relationship between measurements made using an instrument and an external variable (the criterion, sometimes called the gold standard) with which it is expected to correlate. In assessing the criterion-related validity of an adverse-event measure of healthcare quality, the most important and difficult issue is the selection of an appropriate criterion. The few researchers who have studied the criterion-related validity of adverse-event measures have mostly used some form of implicit professional assessment of the quality of care as their criterion. While this is obviously simpler to perform than identifying a separate explicit measure of the quality of care as the criterion, the acknowledged low validity and reliability of implicit professional judgements (Brook and Appel 1973; Hulka 1979; Goldman 1992; Caplan, Posner and Cheney 1991) present serious methodological difficulties, and must be considered in interpreting the results of these studies.

Barnes and Moynihan (1988) used data from the SuperPRO project, which rereviews a sample of records from each Peer Review Organisation (PRO), in an attempt to "assess both the accuracy and the efficiency of the generic screens in identifying deficiencies". They compared the results of the primary screening for adverse events with the professional opinion of a physician reviewer on whether a quality problem existed (see table 3.11), and found the primary screening had a sensitivity of 48.5% and a specificity of 72.8%. They also found that the adverse events missed by the primary screening tended to be the less important ones, and that some adverse event definitions in the PRO generic screens were much more effective at identifying genuine quality problems than others. Barnes and Moynihan (1988) concluded that while "generic screens perform well in identifying the most serious quality problems", modifying some of the generic screens and providing more specific guidelines on their interpretation would improve their validity.

		Professional assessment by physician reviewer		Totals
		Quality problem(s) identified	No quality problems identified	
Primary screening by nurse reviewer	Adverse event(s) found	162 (5.0%)	792 (24.4%)	954 (29.3%)
	No adverse events found	172 (5.3%)	2,126 (65.4%)	2,298 (70.7%)
Totals		334 (10.3%)	2,918 (89.7%)	3,252 (100.0%)

Table 3.11. Comparison of the results of primary screening and physician assessment. Barnes and Moynihan (1988)

The study is interesting, but methodologically weak. Firstly, the range of adverse events identified by the PRO generic screens is collapsed to a single dichotomous variable - the presence or absence of an adverse event or events, both for the primary screening and the secondary review. This means that the identification of quite different adverse events by primary screening and secondary review would falsely suggest agreement, and so the quantitative measures of validity may be inaccurate. Secondly, there is a complete absence of rigour in defining and measuring the criterion variable to which primary screening was compared - that is, the professional assessment of "whether a quality problem existed" (Barnes and Moynihan 1988). Moreover, since physician reviewers were aware of the findings of the primary screening, the two can hardly be said to be independent. These flaws mean Barnes and Moynihan's conclusions, though a useful contribution to the specific debate about the value of the PRO programme, must be treated with some caution.

The Institute of Medicine (1990, p185) presented further evidence of the relationship between primary screening and the presence of confirmed quality problems in the PRO programme, for a sample of 6.3 million cases reviewed by PROs up to June 1989. For those cases where an adverse event was identified in primary screening, between 7.5% and 71.3% were subsequently shown through the PRO professional review mechanisms to have a confirmed quality problem (see table 3.12). These findings provide some support for Barnes and Moynihan's conclusions - particularly their finding that different adverse event definitions have different degrees of validity.

Generic screen	Percent with adverse event identified at primary screening	Percent of primary screening adverse events confirmed as quality problem
Adequacy of discharge planning	3.1%	71.3%
Medical stability of patient on discharge	12.5%	10.6%
Deaths	1.5%	7.5%
Nosocomial infections	7.8%	35.7%
Unscheduled return to surgery	1.0%	7.6%
Trauma suffered in hospital	4.9%	20.8%

Table 3.12. Rates of adverse events and confirmed quality problems in the PRO programme
Base for table: 6,309,839 cases reviewed up to June 1989 by PROs.
Institute of Medicine (1990, p185).

Brennan, Localio and Laird (1989) examined the validity of an adverse-event measure of quality which they had themselves developed for use in the Harvard Medical Practice Study. They performed multiple reviews of 360 cases, each of which was reviewed separately by two medical records administrators and by a senior physician. A total of 672 pairs of reviews were performed, and the results are shown in table 3.13. For the 360 cases (which were deliberately selected to include many cases with adverse events) the sensitivity was 84.5% and the specificity was 71.8%. However, adjusting for the estimated incidence of adverse events in the population of patients from which the sample had been drawn, the estimated sensitivity was 21% and the estimated specificity was 99.5%. Brennan, Localio and Laird (1989) argued that "the careful completion of the screening process by senior physicians was the best approximation to a gold standard that we could develop", and concluded that despite the very high false positive rate signalled by a sensitivity of 21%, their adverse-event measure was valid for wider use in the Harvard Medical Practice Study.

Like Barnes and Moynihan (1988), Brennan, Localio and Laird (1989) collapsed the range of adverse events in their measure to a single dichotomous variable for validity assessment, and used a single physician review as the criterion variable, and so there must be some concern about the generalisability of their results.

		Senior physician review		Totals
		Criteria present	Criteria not present	
Medical records administrator review	Criteria present	403 (60.0%)	55 (8.2%)	458 (68.1%)
	Criteria not present	74 (11.0%)	140 (20.8%)	214 (31.8%)
Totals		334 (49.7%)	195 (29.0%)	672 (100.0%)

Table 3.13. Comparison of the results of medical record administrator and senior physician reviews. Brennan, Localio and Laird (1989).

Bates et al (1995) reviewed 3,137 admissions to an acute hospital, using a generic adverse-event measure applied by medical records staff. Records which they identified as containing an adverse event were they reviewed by a medical reviewer, who determined whether an adverse event had occurred or not, whether it was preventable and how serious it was using his or her professional judgement in an implicit review. Unlike the other studies detailed above, Bates et al (1995) did not collapse the results of screening to a single variable, but presented sensitivity and specificity data separately for each criterion in their adverse-event measure. In table 3.14, these results are presented alongside odds ratios derived from a logistic regression with the presence or absence of an adverse event as the dependent variable and the various screening criteria as the independent variables.

Criterion	Sensitivity	Specificity	Odds ratio (95% CI)
Prior hospitalisation	68	56	2.5 (1.9 - 3.2)
Readmission to any hospital	28	80	1.5 (1.1 - 2.0)
Hospital incurred trauma	16	97	3.6 (2.3 - 5.5)
Transfer to special care unit	12	98	2.8 (1.7 - 4.6)
Adverse drug reaction	10	99	6.7 (3.9 - 12.0)
Death	9	97	1.8 (1.1 - 3.0)
Cardiorespiratory arrest	7	98	not reported
Return to operating room	7	99	2.7 (1.2 - 6.1)
New neurologic deficit at discharge	6	99	2.5 (1.2 - 5.4)
MI, CVA or PE during or after an invasive procedure	5	99	2.8 (1.3 - 5.8)
Treatment or operation because of damage subsequent to an invasive procedure	4	99	not reported
Transfer to another hospital	4	98	not reported
Other	12	96	2.1 (1.3 - 3.3)

Table 3.14. The sensitivity and specificity of screening criteria in detecting adverse events identified through physician reviews. Bates, O'Neil, Petersen et al (1995).

These results suggest that most criteria had poor sensitivity, and would produce large numbers of false positive results, though it should be noted that most were significantly correlated with the presence of adverse events in the logistic regression. Again, the reliance of the study on implicit professional review by a single clinician as the criterion against which validity is assessed is a serious weakness. In addition, since the clinician reviewers only reviewed those cases where an adverse event was found in screening, they were not truly blinded to the results of the first screening. In order to allow sensitivity and specificity calculations, a clinician reviewer did review a random sample of 25% of screen negative cases, and this was used to extrapolate to the 75% of screen negative cases that had not been reviewed, but again the absence of blinding to the results of screening is unfortunate.

The only published research into the validity of adverse-event measures which does not use implicit professional reviews as the criterion variable was carried out as part of a much larger examination of the differences between teaching and non-teaching hospitals (Richards et al 1988). Richards and colleagues applied an adverse-event measure (the APO inventory), a severity of illness measure (the Severity of Illness Index) and a number of condition-specific criterion-based quality measures for myocardial infarctions, gallbladder disease, benign prostatic hypertrophy, and a series of other

conditions. These criterion-based measures worked by applying a set of standards for the management of the given condition to each case, and scoring the case on the degree of compliance with the standard found. The instruments were applied to a sample of about 25,000 patients from 45 hospitals. Correlation coefficients between the APO inventory and the criterion-based quality measures ranged from 0.03 to 0.65 (see table 3.15). Richards et al (1988) suggested that stronger correlations were observed for the more serious conditions because the APO inventory was made up of relatively serious and clinically significant adverse events, which were very unlikely to occur in some diseases, and which only occurred to a small minority of patients.

Condition-specific criterion-based quality of care measure	No of cases in sample	Correlation with APO inventory score (weighted)
Myocardial infarction	680	0.63
Gallbladder disease	678	0.22
Benign prostatic hypertrophy	618	0.16
Asthma	616	0.06
Gastroenteritis	553	0.03
Upper gastrointestinal haemorrhage	619	0.65

Table 3.15. Correlations between condition-specific criterion-based quality of care measures and APO inventory scores. Richards, Lurie, Rogers et al (1988).

There is surprisingly little published evidence for the criterion-related validity of adverse-event measures of healthcare quality. Of the few existing studies, all but one are marred by their dependence for the criterion variable on implicit reviews performed by individual physicians with few or no guidelines and little attention to review reliability. Nevertheless, three important points can be drawn from this research. Firstly, four of the studies cited above suggest that adverse-event measures may suffer from high false positive rates - identifying cases as having adverse events when in fact they contain no real quality problem. Although the studies are not conceptually or methodologically clear about what a quality problem consists of, this apparent *hair trigger effect* is worthy of further investigation. Secondly, some studies indicate that different individual adverse event definitions may have quite different validity characteristics. This implies that the validity of adverse-event measures may be crucially dependent on the mixture of adverse events they contain,

and that attempts to develop or test adverse-event measures should pay attention to the individual items included within the instrument rather than relying on aggregate scores. Thirdly, it seems from one study that adverse-event measures of quality may exhibit a *floor effect*, providing a valid measure of quality only for those patients whose illness is relatively severe. With these provisos, the existing research supports the cautious conclusion that adverse-event measures can be valid indicators of the quality of care.

Construct validity

Construct validity is a measure of how well an instrument supports or conforms with theories or constructs. Exploring the construct validity of adverse-event measures of quality is difficult, because there are few established theories and constructs about the distribution and effects of adverse events for researchers to test. In fact, three theoretical constructs have been examined by researchers:

a) *Length of stay and resource usage*.

It can reasonably be expected that a correlation should exist between adverse-event measures of quality and measures of resource usage such as patients' length of stay in hospital or costs of treatment. Patients who suffer adverse events are likely to require further treatment for the effects of those events - increasing their stay in hospital, and increasing the costs of treatment. Conversely, patients who stay longer in hospital have greater exposure to the risk of suffering adverse events, and so should have higher adverse event rates. While the effects of these two relationships would be difficult to separate, both should result in a positive correlation between the incidence of adverse events and length of stay or costs of treatment.

Panniers and Newlander (1986) applied an adverse-event measure, the APO inventory, to 426 patients with myocardial infarctions, and demonstrated that patients who had suffered adverse events both stayed significantly longer in hospital and had significantly higher total treatment costs (for both $p < 0.001$). In a second, related study of a further 354 patients with the same diagnosis, Panniers (1987) again found significant correlations between APO inventory scores and both length of stay and treatment costs. Richards et al (1988), whose

extensive study is outlined above, also found a "mild correlation" between APO inventory scores and length of stay, and concluded that "longer lengths of stay are associated with more adverse occurrences" (Richards et al 1988). Thus, the findings of three separate research studies all support the construct validity of the adverse-event measures they used.

b) *Severity of illness on admission*.

Patients who are more severely ill are generally more difficult to treat and complex to manage. It seems likely that in this more complex treatment, the opportunities and risks of adverse events would also be higher, so a positive correlation between admission severity and adverse-event measures of quality might be expected.

Panniers (1987) showed that for her sample of 354 patients with myocardial infarctions, there was a significant relationship between APO inventory scores and severity of illness on admission, as measured by the AS-SCORE severity of illness measure ($p < 0.001$). Schumacher, Parker, Kofie et al (1987) reviewed 752 cases drawn from 7 hospitals, using both the APO inventory and the Severity of Illness Index (SII). They demonstrated that patients' adverse-event measure scores and their severity of illness scores were related ($p < 0.0001$), and that more severely ill patients tended to have more adverse events. Richards et al (1988) concurred, finding a significant correlation between APO inventory scores and Severity of Illness Index scores (Richards et al 1988). Again, all three research studies support the construct validity of the adverse-event measures they tested. However, both the severity of illness measures used (AS-SCORE and SII) include some hospital complications, which may result in these measures being influenced by adverse events which occur during an admission. Thus, the positive correlations observed in these studies may be at least in part spurious.

c) *Hospital characteristics*.

There is widespread evidence that different types of hospital provide different levels of healthcare quality, and that factors such as the hospital's size, facilities, teaching status and location can be important predictors of quality. It might reasonably be expected that

adverse-event measures of quality would show patterns of variation across institutions which echo those of other studies of healthcare quality.

Richards et al (1988) found few significant differences in APO inventory scores across groups of teaching and non-teaching hospitals. However, for some particular conditions, such as myocardial infarction and gastrointestinal haemorrhage, mild trends were observed, with more adverse events occurring in teaching hospitals. Brennan et al (1991a) used the database of the Harvard Medical Practice Study (containing 31,000 patient records) to examine patterns in the incidence of adverse events across 51 hospitals in New York during 1984. A large variation in adverse event rates across hospitals was found (ranging from 0.2% to 7.9%, mean 3.2%). All adverse events had been reviewed to determine negligence, and the proportion of adverse events deemed negligent varied from 1% to 60% (mean 24.9%). They found that primary teaching hospitals had significantly higher adverse event rates, and rural hospitals had significantly lower rates. They also observed that the proportion of adverse events due to negligence was lower in teaching hospitals and in proprietary hospitals, and higher in nonprofit and public hospitals. Brennan et al (1991b) concluded that their measures of adverse event rates and negligent adverse event rates "may represent an important improvement on existing measures of quality".

In all three of the constructs outlined above, the available research evidence supports the construct validity of adverse-event measures as indicators of the quality of care.

Conclusions of existing validity research

The most striking feature of the literature on the validity of adverse-event measures is the absence of any substantial studies of face or content validity - a serious and important omission, which future research should endeavour to rectify. The investigations of criterion-related and construct validity, though not especially numerous or rigorous, are generally encouraging, in that they suggest that adverse-event measures do provide a valid indication of the quality of healthcare. However, they also identify a number of deficiencies in the measures that they test, which are deserving of further investigation.

This review of the validity literature also indicates that it is probably as difficult to make valid general statements about the validity of all adverse-event measures as it would be for all general health status measures, or for any other family of different instruments based on a particular methodology. The findings of studies which demonstrate the adequacy (or inadequacy) of a specific adverse-event measure cannot simply be generalised to all such measures, because as the reports cited above demonstrate, different measures (and different items within measures) have different validity properties. We can, however, draw conclusions from these studies about the potential validity of adverse-event measures. In that respect, this review of the literature suggests that such measures can potentially provide a valid insight into the quality of healthcare.

Table 3.16 summarises the methods, results and conclusions of the validity research referred to above.

Study	Dimension of validity	Methods	Results and conclusions
Panniers and Newlander (1986)	Face/ content	Consulted "expert panel" made up of 1 doctor and 3 nurses.	Results not reported.
Panniers and Newlander (1986)	Construct	Applied APO inventory to sample of 426 patients in DRGs 121 and 122 (myocardial infarction) from one hospital, and collected details of length of stay and treatment costs.	Significant correlation between APO inventory score and both length of stay ($p < 0.001$) and treatment costs ($p < 0.001$).
Panniers (1987)	Construct	Applied APO inventory and AS-SCORE to sample of 354 patients in DRGs 121 and 122 (myocardial infarction) from one hospital, and also collected details of length of stay and treatment costs.	Significant correlation between APO inventory score and AS-SCORE ($p < 0.001$).
Schumacher, Parker, Kofie et al (1987)	Construct	Applied APO inventory and Severity of Illness Index to 752 patients from 7 hospitals.	Significant correlation between APO inventory score and SII ($p < 0.0001$).
Barnes and Moynihan (1988)	Criterion-related	Used 3,252 records from SuperPRO project, drawn from PROs all over USA. Applied PRO generic screens for primary review; used single physician assessment for secondary review.	Primary screening had sensitivity of 48.5%, specificity of 72.8%. Missed adverse events tended to be less clinically significant ones. Some adverse event definitions were more valid than others.
Richards, Lurie, Rogers et al (1988)	Face/ content	Used panel of 3 doctors to rate "adversity" of each element of APO inventory.	Results not specifically reported. All ratings obtained were negative and "generally agreed".
	Criterion-related	Applied APO inventory, Severity of Illness Index, and condition-specific quality measures to 25,000 patients from 45 hospitals, and also collected administrative/demographic and clinical details of cases.	APO inventory score correlated with criterion-based quality measures for more serious conditions.
	Construct		APO inventory score correlated with length of stay and with SII. Some variations in APO inventory score observed across different types of hospitals.

Table 3.16. Summary of research into the validity of adverse-event measures of healthcare quality (continued overleaf).

Study	Dimension of validity	Methods	Results and conclusions
Brennan, Localio and Laird (1989)	Criterion-related	Applied own adverse-event measure twice each to 360 cases selected to contain a high proportion of cases with adverse events; undertook a separate single physician assessment of cases.	For general population of cases, estimated measure had sensitivity of 21% and specificity of 99.5%.
Institute of Medicine (1990)	Criterion-related	Used 6,309,839 cases reviewed by PROs up to June 1989 to compare percentages of cases with an adverse event identified at primary screening and percentage of those adverse events subsequently confirmed as quality problems.	Between 7.5% and 71.3% of adverse events were confirmed as quality problems - rate varied by type of adverse event.
Brennan, Hebert, Laird et al(1991)	Construct	Used 31,000 patient records drawn from 51 hospitals to compare adverse event rates and negligent adverse event rates across hospitals.	adverse event rate and negligent adverse event rate varied substantially across hospitals. Primary teaching hospitals had higher adverse event rates; rural hospitals had lower adverse event rates. Proportion of adverse events due to negligence lower in teaching and proprietary hospitals; higher in nonprofit and governmental hospitals.
Bates, O'Neil, Petersen et al (1995)	Criterion-related	Applied adverse-event measure to 3,137 admissions; undertook a separate clinician review of those screened positive to assess presence, seriousness and preventability of adverse event.	Sensitivity of screening criteria ranged from 68% to 4%; specificity ranged from 56% to 99%. No criteria had both high sensitivity and high specificity.

Table 3.16. Summary of research into the validity of adverse-event measures of healthcare quality (continued from previous page).

3.4.3 **Reliability of adverse-event measures of quality**

In chapter 2, the different approaches to defining and measuring the reliability of instruments were discussed, and the importance of careful reliability assessment was highlighted. In an apposite comment, Richards et al (1988) observed that "although reliability analyses do not make for glamorous research, positive answers to the questions posed by these analyses are essential for confidence in results obtained from any subsequent work". The available empirical evidence on the reliability of adverse-event measures is outlined below.

Interrater reliability

Interrater (or interobserver) reliability measures whether, when the same test is applied to the same respondent or subject by different raters, the same results are produced. A number of researchers have examined the interrater reliability of adverse-event measures, generally by arranging for multiple reviews of patients' case records by different screening staff, and then comparing the results of screening.

Panniers and Newlander (1986) used two reviewers, both qualified and experienced nurses, who separately applied a slightly modified form of the APO inventory (which has been described in section III.1.3) to a sample of 200 cases from 426 patients with myocardial infarctions. The kappa statistic (Fleiss 1981, p146) was used to assess the extent to which observed agreement exceeded that which would be expected simply through chance. Of the 15 adverse events in the APO inventory, 10 showed raw agreement levels of 99-100% on whether or not the event had occurred. For the remaining 5 items, raw agreement levels ranged from 72% to 96% (kappa from 0.29 to 0.83). Reliability data for judgements of whether a patient was at risk from an adverse event (the denominator section of the APO inventory) were not reported. Panniers and Newlander concluded that the APO inventory was reliable.

Schumacher et al (1987) used seven different raters to review 752 cases (each being reviewed two or three times) drawn from seven teaching hospitals, also using the APO inventory. The APO inventory is a risk adjusted score, composed of a numerator (the count of actual adverse events) and a denominator (the count of adverse events for which the patient was at risk). Schumacher compared agreement for each pair of raters for the numerator, the denominator and the APO inventory score as a whole, using Pearson's correlation coefficient. The mean correlation coefficient for the APO inventory score was 0.33, with individual rater pairs' correlations ranging from -0.05 to 0.58. Correlations for denominator scores were higher than those for numerators or the APO inventory score itself. Schumacher and colleagues concluded that the APO inventory was not reliable.

Unfortunately, Schumacher et al (1987) used the Pearson correlation coefficient alone to assess reliability - a measure which is really an indicator of linear association, not agreement, and which takes no account of deviations of scale and bias. Brennan and Silman (1992) advise that the Pearson correlation coefficient "is not applicable as a measure of between observer variability"; Main and Pace (1991) concur, recommending the use of intraclass correlation methods to measure agreement for continuous variables instead. In addition, by only citing the correlation coefficient on aggregate scores rather than on the individual items of the APO inventory, Schumacher and colleagues concealed the potentially different reliabilities of different items within the APO inventory. In view of these flaws, it is questionable whether their conclusions can be justified on the basis of their published results.

As part of a larger study described above, Richards et al (1988) applied the APO inventory to 516 cases drawn from 5 hospitals, each of which was reviewed by a pair of raters. Kappa statistics were calculated for each item within the APO inventory - both the adverse event items and the at risk items (indicators of whether a particular patient was at risk of a certain type of adverse event). Agreement for the at risk items ranged from 0.28 to 0.83 (mean 0.50), while agreement for the adverse event items ranged from -0.18 to 0.73 (mean 0.33). Richards and colleagues also examined the reliability of the overall APO inventory score (in the form of a weighted sum of their own design). They found that within-patient variability of this APO inventory score was substantially less than the overall variability of the sample (51.5% of cases had within-patient variability < 0.33 SD; 77.8% had within-patient variability < 1 SD). They also constructed an analysis of variance model, using diagnosis, hospital, patient and rater as the independent variables and the APO score as the dependent variable. The model accounted for 91% of the variability of APO scores, and little of the variability was attributable to differences between raters (2%) while most was attributable to patient (63%), diagnosis (9%) and interactions of patient, diagnosis and hospital variables. Richards et al (1988) concluded that "reliability analyses for the APO as a whole suggest that it is at best only a moderately reliable measure".

The Harvard Medical Practice Study (1990, p5-26) investigated the reliability of the adverse-event measure developed and validated (Brennan, Localio and Laird 1989) for use in the project. A randomly selected sample of 282 cases (about 1% of all cases reviewed for the study) was separately

screened using the adverse-event measure by a medical records administrator (MRA) and the MRA supervisor. When the decisions made by the medical records administrator and the MRA supervisor about whether or not a case had an adverse event were compared, a raw agreement rate of 93.6% was calculated, with a kappa statistic of 0.85. This suggests a relatively high and acceptably reliable degree of agreement, though combination of raters decisions about a number of different adverse events into a single dichotomous variable (whether or not an adverse event occurred in a case) may overstate the level of agreement.

While several studies of interrater reliability exist, all but one made use of the same adverse-event measure (the APO inventory). Bearing in mind the unusual construction of this measure, it may be difficult to draw general conclusions about the interrater reliability of adverse-event measures from these studies. It is also notable that a clear consensus about the interrater reliability of adverse-event measures does not emerge from the four studies: one suggests reliability is very high, two suggest it is moderately good, and one concludes that it is poor. More extensive testing of the interrater reliability of adverse-event measures is certainly needed if a definitive conclusion is to be reached.

Intrarater reliability

Intrarater (or intraobserver) reliability measures whether, when the same test is applied twice by the same rater to the same subject, the same results are produced. It is an important test of reliability, complementing information on interrater reliability. Unfortunately, no studies of the intrarater reliability of adverse-event measures have been identified in the literature. This is, therefore, an area which future studies of these measures should address.

Internal consistency

Measures of internal consistency show how well the individual items within a measure which are meant to be measuring a particular trait or characteristic are correlated. High inter-item correlation suggests that the measure is indeed measuring a particular trait. Low inter-item correlation suggests that the measure may actually be measuring different traits, and further analysis might be indicated to identify and separate these characteristics.

Adverse-event measures generally count or flag the occurrence of a number of different sorts or types of adverse event, with each item within the measure acting as a counter or flag for a particular type of adverse event. In these circumstances, correlations between items in the measure would often not be expected to occur, unless there was some causal link between the different types of adverse event. For most pairings of items within an adverse-event measure, such causal links will not exist, and so it is not appropriate to seek high interitem correlations or to use such correlations as a benchmark of reliability. Measures of internal consistency are not very useful in assessing the reliability of adverse-event measures, and indeed no examples of studies addressing internal consistency were found.

Conclusions of existing reliability research

This review of existing investigations of the reliability of adverse-event measures of healthcare quality is brief, because of the scarcity of published studies. The studies reviewed provide some indications of the reliability of the measures they tested, but no clear consensus emerges from their conclusions. Considering the extensive use made of adverse-event measures in the US healthcare system, and their growing deployment in other countries such as the UK, it appears that there is a pressing need for further research in this area.

Table 3.17 summarises the methods, results and conclusions of the reliability research referred to above.

Study	Dimension of reliability	Methods	Results and conclusions
Panniers and Newlander (1986	Interrater reliability	Used 2 raters to apply modified form of APO inventory to sample of 200 cases from 426 patients with myocardial infarctions.	Raw agreement of 99-100% for 10 items of APO inventory; other 5 items ranged 72-96% (kappa 0.29 to 0.83). Concluded APO inventory was reliable.
Schumacher, Parker, Kofie et al (1987)	Interrater reliability	Used 7 raters to apply APO inventory to 752 cases (each being reviewed 2 or 3 times) drawn from 7 hospitals.	Pearson correlation coefficients cited, measuring association between raters. Mean correlation for APO score was 0.33 (ranged from -0.05 to 0.58). Concluded APO inventory insufficiently reliable.
Richards, Lurie, Rogers et al (1988)	Interrater reliability	Used multiple raters to apply APO inventory to 516 cases drawn from 5 hospitals, each reviewed by 2 raters.	Kappa statistics for APO numerator items had mean of 0.33 (ranged -0.18 to 0.73); for APO denominator items mean was 0.50 (ranged 0.28 to 0.83). For APO score, found within-patient variability much less than overall variability. ANOVA showed differences between raters responsible for 2% of APO score variability. Concluded APO inventory "at best moderately reliable".
Harvard Medical Practice Study (1990)	Interrater reliability	Used multiple raters to apply own adverse-event measure to 282 cases (random 1% sample of total study), each reviewed by 2 raters.	Raw agreement on presence/absence of adverse event in each case of 93.6%, kappa of 0.85. Concluded measure was sufficiently reliable for use in study.

Table 3.17. Summary of research into the reliability of adverse-event measures of healthcare quality.

3.5 **Conclusions**

While there is much anecdotal evidence of the value of adverse-event measures of quality, and of quality assurance programmes which use such measures, there is limited empirical research evidence to support many of the bold contentions of either their advocates or their critics which were described earlier.

The body of research on the validity of these measures is generally cautiously positive about their value as measures of the quality of healthcare, though important concerns about their behaviour do emerge. The validity of these measures has not been adequately explored in any area, but particular aspects of validity where further research is especially necessary include the face and content validity of adverse-event measures, and their construct validity. It is evident that examining the criterion-related validity of these measures presents methodological problems.

The reliability research into adverse-event measures is equally disappointing in both its scope and quality. Little consensus emerges from the few existing studies about the reliability of adverse-event measures - indeed, different researchers seem to have achieved markedly different reliability results using identical instruments, which may point to the importance of rater training. Both the interrater and intrarater reliability of adverse-event measures require further investigation.

If the state of current knowledge about adverse-event measures of quality is mapped out against the framework for the structural evaluation of quality measurement techniques developed by the Institute of Medicine (1990, p312) and discussed earlier, the results are concerning. The reliability, validity, sensitivity and specificity of these measures have been inadequately explored. However, other attributes identified as important in any evaluation by the Institute of Medicine have barely been addressed by researchers concerned with adverse-event measures. Characteristics such as appropriateness, clarity, acceptability, concordance, inclusiveness, clinical adaptability, flexibility, patient responsiveness and documentation have not been explicitly examined, though some anecdotal accounts of the use of adverse-event measures provide an imperfect understanding of certain attributes.

Even less evidence is available for the key process evaluation attributes enumerated by the Institute of Medicine (1990) as important characteristics of the application of any quality measurement techniques - pretesting, dynamism, evaluation, comprehendability, manageability, nonintrusiveness, appealability, feasibility, computerisation, and executability. Again, our limited knowledge provides some comprehension of the behaviour of adverse-event measures in these dimensions - but the areas of ignorance or supposition again outweigh those of empirically derived and tested, objective veracity.

Few could argue with Sanazaro and Mills (1991) who wrote:

> "The appeal [of occurrence screening] is clearcut: it purports to be a simple method of simultaneously serving the purposes of both quality assessment and risk management. However, its use for this dual purpose is yet another instance of widespread adoption of inadequately tested technology in quality assessment. Users of generic screening criteria tend to ignore the fact that the screens have never been shown to have any intrinsic value other than to identify charts that may or may not contain a clinical event attributable to care. Consequently, use of such screens for quality assessment may have serious limitations." (Sanazaro and Mills 1991)

The task for researchers and others interested in using and applying adverse-event measures in healthcare is clear. Before these measures can be used with confidence, their validity, reliability and general utility must be properly evaluated, with particular attention being given to those areas in which there is little or no existing research. Then, once those methodological foundations have been established, the behaviour and application of adverse-event measures in healthcare organisations should be examined, and their effectiveness within quality assurance programmes should be evaluated.

Chapter 4

Content and face validity of adverse-event measures of quality

---

4.1 Introduction

In chapter 3, the available evidence for the validity of adverse-event measures of quality was presented. When that relatively limited evidence is placed in context, against the very widespread use of these measures in healthcare quality assurance programmes in the USA, and their growing use in the UK and elsewhere, it is fair to conclude that no aspect of validity could be said to have been overresearched. Indeed, some important aspects of validity have scarcely been examined at all, at least in the published literature.

It was noted that no substantial studies of the face or content validity of any adverse-event measures of healthcare quality were found in the published literature. This suggests that the face and content validity of adverse-event measures of quality is perhaps the most important single research need identified by the literature review. Two research studies intended to address this research need are reported in this chapter.

a) *Survey of clinician opinion*

The primary aim of this study was to explore the face and content validity of a generic adverse-event measure of quality by gathering the opinions of clinical professionals who had the skills and experience necessary to provide an informed view of its validity, through a postal questionnaire survey. The study also gathered other information from the clinicians surveyed on other secondary aspects not directly concerned with validity, such as the expected behaviour of the measure in practice, the practicability of collecting the data it required, and the potential for improvements in its design and specification.

b) *Interviews with clinicians*

The primary aim of this study was to support and supplement the largely quantitative information gathered by the questionnaire study of clinician opinion with a more qualitative understanding of the issues of face and content validity, gained through individual interviews with a small number of clinicians. It was particularly intended to identify themes or matters which might have been omitted or overlooked by the questionnaire study due to its prestructured format, and to explore the place of adverse-event measures of quality in the quality measurement and quality improvement process in healthcare.

Both these studies drew on an adverse-event measure of quality which had been developed for use within the RSCH project. The generic adverse-event measure used in both the studies of clinician opinion can be found in appendix 4.1. This measure was developed by clinicians within the RSCH project, based on information collected on measures used elsewhere and developed by others and on the particular interests and views of the clinicians involved in the project. Control over the development of the measure thus lay at least in part outside this study, and the clinicians involved did not see the maximisation of the measure's validity and reliability as their primary goals, though these issues were discussed. As a result, it can be argued that the studies using this measure provide a pragmatic and realistic test of the validity and reliability of such measures, which is likely to be generalisable to other similar settings, rather than an ideal or maximal estimate of validity and reliability which might be hard to replicate in practice.

It should also be noted that the process of developing this and other measures continued over the life of the RSCH project, with small changes being made to definitions and interpretations quite frequently. This could complicate the interpretation of the data drawn from the RSCH project itself and used as the basis of the study of construct validity, and particular caution should be exercised in assessing temporal trends.

4.2 Survey of clinician opinion

4.2.1 **Aims of survey**

The primary aim of the survey of clinician opinion was to assess the content and face validity of the adverse-event measure which was developed for use in the Royal Sussex County Hospital occurrence screening project (and which is contained in appendix 4.1).

Content validity and face validity were both defined in chapter 2, and set in context among other dimensions of validity. Content validity refers to how adequately the items within a measure reflect the conceptual definition of its scope. Assessing the content validity of an adverse-event measure of quality thus involves ascertaining whether the criteria making up the measure are real indicators of quality, and particularly whether any important criteria are omitted (in other words, whether the measure adequately reflects the scope of adverse events). It can be assessed through the use of some form of expert panel made up of people with an understanding of both the measure and the concept it is designed to measure. Face validity refers to whether "on the face of it" the items making up a measure and the measure as a whole seem reasonable in the view of users or potential users of the measure. Again, for an adverse-event measure of quality it can be assessed through the use of some form of expert panel as described above. Face validity is, in essence, a measure of the acceptability of the measure under test by those who use it or might use it, and so it needs to be evaluated by seeking the opinions of users or potential users.

Content and face validity are clearly related and overlapping concepts (for example, a measure which does not reflect the scope of the concept being measured is unlikely to be acceptable to users and so will have a low face validity). They are both measured by gathering and analysing opinion. This study set out to assess both content and face validity by gathering the opinions of clinical professionals who had the skills and experience needed to assess content validity and who also, as potential users of the measure, were able to provide information on its face validity.

The secondary aim of the survey of clinician opinion was to gather the opinions of clinical professionals on other important aspects of the adverse-event measure of interest - such as its

expected behaviour in use, and its practicability - and to improve the measure if the survey results indicated that improvements were necessary and feasible. To these ends, the study set out to establish professionals' opinions on the expected incidence of adverse events defined in the adverse-event measure, the availability of information on these adverse events in patient records, the severity of effect of these adverse events on patients, and the potential for improvement in the design of the adverse-event measure. Textual comments on all aspects of the measure were also sought, in an effort to ensure that any important issues or themes not specifically addressed in the questionnaire would be uncovered.

4.2.2 **Method**

Survey design

It was not have been feasible to gather opinions from a relatively large number of clinical professionals on the validity and other properties of the adverse-event measure by any means apart from a postal questionnaire. In any case, the structured nature of the data required for the assessment of face and content validity was well suited to a questionnaire-based methodology. Nevertheless, the potential disadvantages of self-completion questionnaire-based surveys were appreciated. In particular, the risk that important issues, not specifically mentioned in the questionnaire, might be implicitly excluded from the study despite their importance was recognised. This consideration was addressed by the interview study which is described in section 4.3.

Five issues were identified which the questionnaire needed to address for each screening criterion (and the associated adverse event definition) within the adverse-event measure to be tested.

a) *Relationship to the quality of care*.

The respondent's assessment of the relationship between the criterion and the quality of care delivered to the patient needed to be assessed, either directly or indirectly.

b) *Availability of information in patient records*.

The anticipated availability of information on adverse events in patient records is an important indicator of the practicability of each criterion in the adverse-event measure. If the necessary information is not well documented in patient records, this may have important implications for the reliability of any measure based on that information (and hence may impose a practical limit on the measure's validity). Therefore, the likely availability of the necessary information needed to be assessed.

c) *Expected incidence of adverse events*.

The costs of detecting very infrequent adverse events will be high (since most screenings will produce a negative result), and the reliability with which they can be identified may be lower (since screening for them may seem tedious and unrewarding to those involved). For these reasons, it might in some circumstances be reasonable to exclude screening criteria which are focused on such infrequent events from an adverse-event measure, and so the expected incidence of the adverse events included in the measure to be tested needed to be investigated. Of course, in deciding to include or exclude a certain type of adverse event, the frequency of that event would have to be considered alongside other factors - like the effect of that event on the patient, discussed below.

d) *Effect on patient*.

Different adverse events will inevitably have varying degrees of clinical significance for patients' health status. It is reasonable that adverse events with a greater impact on patients' health status might be regarded as more important. If an aggregate score were to be derived from the adverse-event measure, some form of weighting based on the importance of different criteria would be required, and this might be based on an assessment of the effect on patients of these events. In any case, the severity of effect of the events in the adverse-event measure should be examined, since it might be argued that events with little or no effect on patients are not good indicators of the quality of care delivered to patients.

e) *Potential for improvement of the measure*.

Another important indicator of the general adequacy or suitability of each criterion within the adverse-event measure would be the extent to which professionals perceived some potential for its improvement. Criteria which are well specified and meet with professionals' approval might be expected to be rated as having little improvement potential, while those with important perceived defects might be rated as having much more room for improvement (unless those defects were so severe as to render the criterion unimprovable in respondents' eyes). Therefore, assessing the potential for improvement of each criterion would provide a further indication of its validity, and would also yield specific suggestions for the improvement of individual criteria within the measure and of the measure as a whole.

The adverse-event measure to be tested in the questionnaire study contained 20 screening criteria. For each criterion, some details of its definition needed to be given, and a number of questions needed to be asked. It was therefore clear from an early stage in the study that the questionnaire itself might necessarily be quite long and bulky, and that it might take quite some time for respondents to complete.

Since it was anticipated that most British clinicians would be unfamiliar with the concepts of adverse events and quality measures based on adverse events, some explanation of the background and terminology would be necessary, as would an explanation of the purpose of the study itself.

Pilot study

The pilot questionnaire, accompanying letter and supporting information developed for the study is shown in Appendix 4.2. A single page was used for each criterion in the adverse-event measure, with details of the criterion at the top of the page and seven questions beneath. Four questions used linear visual analogue scales, two used three point multiple choice scales, and one sought a dichotomous (yes or no) response. Respondents were also invited to write textual comments on any aspect of each criterion.

A pilot study was undertaken in July 1990 using these materials. Questionnaires were distributed to six subjects with a range of clinical and research skills who were not directly involved in the RSCH occurrence screening project. Each questionnaire was accompanied by a letter and supporting

materials (including a paper on occurrence screening). In addition, each pilot subject also received a single sheet questionnaire seeking information on the time taken to complete the questionnaire, details of particular difficulties encountered, views on the questionnaire design, and suggestions for improvements.

The pilot study highlighted three main issues which required attention before the main study could commence:

a) *Questionnaire length*.

All respondents commented on the length of the pilot questionnaire, and suggested that it needed to be shorter if clinicians were to be expected to persist with its completion. Some suggested that too much information about each criterion was given. It was observed that the questionnaire design did not make it immediately clear that the same questions were being asked about each criterion, and that awareness of this might make completion quicker. The physical bulk of the questionnaire was also criticised. Respondents had taken between 30 and 70 minutes to complete the questionnaire.

b) *Question complexity/design*.

Three questions in the pilot questionnaire addressed the validity of each criterion. Question 1 asked directly for a scaled rating of the relationship between the criterion and the quality of care, while questions 4 and 5 duplicated question 1, asking about whether detailed case review would reveal lapses in the quality of care and would show a better or worse overall quality of care. Several respondents found questions 4 and 5 particularly difficult to understand, and observed that they duplicated each other, and question 1. The use of analogue scales was criticised by some respondents as making completion slower and adding nothing to the questionnaire's ease of use.

c) *Supporting materials/information*.

Respondents suggested that the amount of information in the questionnaire itself on each criterion was too great for subjects to assimilate unless they were particularly well motivated

to complete the questionnaire. They also suggested that the letter and supporting information contained too much information. The risk that the paper about occurrence screening which was included in the supporting materials might bias or predispose subjects towards adverse-event measures was highlighted.

In the light of these important findings from the pilot study, the questionnaire was substantially redesigned. Firstly, by condensing the descriptions of screening criteria and placing questions in a tabular format, the length of the questionnaire was reduced to four pages. This allowed it to be produced in the form of a four-sided (two sheet) A4 sized leaflet. The tabular question format also made it clearer that the same questions were being asked about each criterion, and made it easier for the respondent to compare his or her own answers for different criteria. Secondly, questions four and five from the first questionnaire were eliminated, because of the problems outlined above. Finally, the supporting materials were reduced in length, the paper on occurrence screening was removed, and a single page background information leaflet was substituted.

Having made these changes, the questionnaire was informally repiloted with the same subjects. The revisions were wholly welcomed, and the questionnaire was judged suitable for use in the main study. The revised questionnaire, accompanying letter and information, are shown in Appendix 4.3.

Despite the substantial improvements made between the pilot and main study questionnaires, the length of the questionnaire (and the length of time it would take to complete) was still a matter of some concern to the researcher. Seeking five items of data about each of 20 screening criteria inevitably involved 100 responses from each respondent, however the questionnaire was physically presented. If, making rather optimistic assumptions, each question took a mere 15 seconds to complete, filling in the questionnaire would take 25 minutes. Reading and assimilating the information needed to complete the questionnaire might take the same length of time again. Thus, it would still require very good motivation on the subject's part to complete the questionnaire, and the response rate to the questionnaire might be poor.

Because of these concerns, alternative study designs which would reduce the burden on the subject were considered. Two main alternatives existed. Either the number of items of data per criterion

could be reduced (by eliminating some questions) or the number of criteria on which data was sought could be reduced (by partitioning the measure and using questionnaires which only included some of the screening criteria). The former solution was rejected for two reasons. Firstly, the loss of information involved in eliminating some of the questions in the questionnaire was deemed unacceptable. Secondly, it was believed that the actual time saving for the subject that would result would be relatively limited, since he or she would still need to read the information about each criterion before responding to the reduced set of questions about that criterion. Answering fewer questions about each criterion would not produce a proportionate reduction in the time taken to complete the questionnaire.

The latter solution was also rejected, because of its potential effect on the power of the study to demonstrate the validity of the measure being tested. Any assessments of content validity which were derived from questionnaires which only included half of the measure being tested would inevitably be misleading. To provide an informed view of the measure (and especially of its content validity), clinicians needed to be presented with the whole measure, rather than a part of it. The number of criteria in the adverse-event measure could not be reduced simply to suit the needs of validity testing, since that would itself reduce its validity.

Therefore, it was decided that the main study should proceed using the revised questionnaire shown in appendix 4.3. In order to encourage subjects to complete the questionnaire, it would be accompanied by a letter from a clinician with whom they could identify (if possible, a member of the same specialty), seeking their cooperation. In addition, subjects would be offered a copy of the study report if they completed the questionnaire, as an incentive to completion. Non-respondents would be identifiable (since the questionnaires would be numbered) and they would be contacted a second time to encourage them to respond. Finally, response rates would be monitored prospectively throughout the study, in order to identify any difficulties with response rates as early as possible.

Main study

It was not feasible to attempt to draw a random sample which was statistically representative of all senior medical staff in the UK, or of all medical staff, or of all clinical professionals with an interest

or involvement in the healthcare processes addressed by the adverse-event measure. Nor, indeed, was it necessary to seek such a representative sample in order to undertake an adequate assessment of the face and content validity of the adverse-event measure. As has already been noted (see chapter 2), face and content validity are commonly assessed through expert panels. Having established the skills which an expert panel might be expected to hold, the primary concern should be that the sample group (and the subgroup of those who respond) should contain clinical professionals with those skills, and not that they should necessarily be representative of their profession as a whole.

The sample of clinicians to be used in the main study was identified on the basis of five main considerations or requirements: clinical knowledge, training and experience; an understanding of the science of measurement and of quality assessment in healthcare; independence from the RSCH occurrence screening project; achieving a representative geographic and specialty spread; and the demand that the relatively complex questionnaire would place on subjects' time. It was recognised that no single group of clinicians would possess all the desired attributes detailed above, and that the unavoidable size and length of the questionnaire might discourage its completion by many clinicians. Therefore, three different groups were selected:

a) *Public health physicians*.

A group of 201 public health physicians, made up of all those holding posts as Directors of Public Health in Districts or Regions within England, was identified. This group was chosen because public health physicians have skills and training in both clinical medicine and in epidemiology, statistics and the science of measurement. They were therefore regarded as being particularly well suited to assessing the validity of the adverse-event measure. They also provided a wide geographic spread, representing every area of England.

b) *Practising clinicians - Worcester Health Authority*.

A group of 59 consultants, made up of all those currently practising in clinical medicine within Worcester Health Authority in the West Midlands, was identified. This group was chosen because it provided a diversity of clinical specialists from every major specialty, and

was sufficiently physically distant from the site of the RSCH occurrence screening project to ensure that the clinicians involved would have little or no previous knowledge of the project.

c) *Practising clinicians - leaders in clinical audit*.

A group of 22 consultants currently practising in clinical medicine and acting as medical leaders of clinical audit programmes within the South East Thames region was identified. The group was made up of all consultants who were either Chairs of the District Medical Audit Advisory Committee (DMAAC) in their own District or were members of the Regional Medical Audit Advisory Committee (RMAAC) for the South East Thames region. This group was chosen because it provided an additional diversity of clinical specialists from various major specialties, and because it also provided the opinions of clinicians with particular experience and skills in quality assessment in healthcare.

Questionnaires, accompanied by the background information and supporting letters described above, were distributed to clinicians in group A by post in March 1991. Non-respondents were contacted again by letter, with a further copy of the questionnaire, in May 1991. For group B, questionnaires were distributed by post in January 1991, and non-respondents were contacted again shortly afterwards. For group C, questionnaires were distributed in May 1991, and non-respondents were contacted again shortly afterwards.

Returned questionnaires were immediately identified (through their questionnaire number) and the response was noted on the list of study subjects used to identify non-respondents. The Statistical Package for the Social Sciences (SPSS-PC) was used to record the numerical and categorical responses from each questionnaire, and to produce the analyses and statistics detailed below.

A wordprocessor was used to record, classify and tabulate the textual comments made by respondents to the questionnaire, and these are also discussed below. The questionnaire was designed to encourage respondents to comment on the adverse-event measure, on individual screening criteria within the measure, and on the design and implementation of the questionnaire study itself. Respondents' comments were sought both to enrich the quantitative data gathered by the questionnaire and to help identify themes or issues which the questionnaire had not addressed.

All comments from respondents were transcribed and categorised, and a full listing of these comments can be found in appendix 4.4.

4.2.3 **Results and discussion**

Response rates

The final response rates to the questionnaire study, from each group and from the sample as a whole, are shown in table 4.1. It can be seen that the response rate from public health physicians was substantially higher than the response rates from the two groups of practising clinicians. Of the practising clinicians, the response rate of those with a known interest in clinical audit was higher than the response rate from ordinary practising clinicians in Worcester.

Group	Group size	Number returned	Response rate (%)
Public health physicians	201	132	65.7%
Practising clinicians - Worcester Health Authority	59	10	16.9%
Practising clinicians - leaders in audit	22	8	36.4%
All groups	282	150	53.2%

Table 4.1. Response rates to questionnaire study.

The sample of clinicians used in the questionnaire was selected, as has already been noted earlier, because they were felt to have the skills and knowledge required of an expert panel commenting on the validity of an adverse-event measure of quality. Since some of the sample did not respond to the questionnaire, it is necessary to examine the response rates and their causes, and to consider their potential effects on the representativeness and meaning of the responses received to the questionnaire. The critical issue, in considering the effects of lowered response rates, is whether

there are differences between respondents and non-respondents which are germane to the purpose or aims of the study.

Contact with many non-respondents from all three groups was established, through the process of mailing additional questionnaires to and receiving uncompleted returns from subjects who did not respond. Subjects who did not wish to respond to the questionnaire were asked to return blank questionnaires so that they would not be troubled by reminders, and many of these blank returns were accompanied by notes or letters explaining their reasons for non-completion. In addition, a few non-respondents were contacted by telephone during the study. From these sources, it appeared that there were three main reasons for non-response:

a) *Time availability.*

Many non-respondents indicated that they could not spare the time required to complete the questionnaire. As was discussed above, the questionnaire was a substantial one, and required a considerable investment of subjects' time if it was to be completed. Senior medical staff with a clinical commitment often declined to complete the questionnaire for this reason.

b) *Interest in quality measurement*.

Some subjects indicated that the subject of quality measurement or clinical audit did not interest them, felt it did not involve them, or believed they knew too little about the subject to answer any questionnaire relating to it.

c) *Attitude to questionnaires*.

All subjects were clinicians working within the NHS. A number of subjects observed that they received many questionnaires, from other clinicians, professional bodies, and outside organisations such as pharmaceutical companies. In the face of the many questionnaires they received, some subjects indicated a reluctance to complete any questionnaires at all.

The different response rates obtained from the three study groups provide some additional evidence that the reasons cited above for non-response are correct. Public health physicians might be

expected to have more office-based time than their clinical colleagues because of the nature of their responsibilities; to be more interested in the measurement of quality than their clinical colleagues because of their training in epidemiology and the science of measurement and their involvement in areas such as needs assessment and quality specification where such measures would be useful; and to be better disposed towards questionnaires since they are often themselves involved in originating and managing questionnaire studies. It is therefore unsurprising that the response rate for public health physicians (65.7%) was far higher than that for the other two groups.

The third group - of practising clinicians who were leaders in clinical audit - might be expected to be similar to the second group - practising clinicians from Worcester Health Authority - in most respects. However, they might be expected to have a greater than average interest in healthcare quality, since they were all involved in leading the clinical audit programme. This supposition is supported by the fact that the response rate from these clinicians involved in audit (36.4%) was twice that of the clinicians from Worcester Health Authority (16.9%), though it was still much lower than that of the group of public health physicians.

Response rates in the two smaller groups - both formed of practising clinicians - were particularly low, and therefore give some cause for concern about the potential for biases resulting from differences between the groups of respondents and non-respondents. The reasons for non-response discussed above provide some reassurance, but it is still important that the nature of potential biases is explored and, where possible, the data from respondents is used to assess the extent of bias. There are two main forms of systematic bias which might be introduced by the lowered response rates in the two smaller groups:

a) *Bias towards those favouring the measure*.

If clinicians who were favourably disposed towards the adverse-event measure were more likely to respond than those who were not, a misleadingly positive view of clinical opinion about the measure might be established. Responses might not represent the full range of clinical opinion about the measure. It is very difficult to assess the extent of this form of bias. However, the reasons for non-response cited above provide some reassurance (since opposition to the measure was not featured among them). In addition, the analysis of textual

comments from respondents can provide some evidence that respondents did at least represent a full spread of opinion about the measure.

b) *Bias towards clinicians not practising clinical medicine*.

The higher response rate from public health physician subjects than from practising clinician subjects means that while 71% of the sample were public health physicians, they made up 88% of respondents. Overall, the actual numbers of practising clinicians responding to the study were small (n=18). It might be argued that public health physicians have less direct clinical experience and understanding, and might therefore rate the adverse-event measure differently. Since the respondent group is undoubtedly biased towards public health physicians, the effect of this bias can best be examined by comparing the responses of public health physicians and practising clinicians to identify any significant differences.

With the proviso that the effects of these two biases are examined in analysing and discussing the results of the questionnaire study, the overall response rate of 53.2% from the three groups outlined earlier provides a base of data with which the analysis and interpretation of the face and content validity of the adverse-event measure can be performed with a reasonable degree of confidence.

Validity of adverse-event measure as a measure of quality

Each respondent was asked to rate the validity of each screening criterion in the adverse-event measure on a scale of zero to ten, on which zero indicated that the events identified by the screening criterion were "not at all related to the quality of care that patients receive" while ten indicated that the events were "very closely related to the quality of care that patients receive".

The results are summarised in table 4.2. For each screening criterion, the mean, mode and median responses are given. The table also shows the standard deviation of the distribution of responses to provide an indication of the spread of responses.

The mean, modal and median responses for each criterion provide indicators of the consensus view of the validity of that criterion from the clinicians who participated in the study. The standard

deviation provides an indication of the spread of the distribution of responses, and hence of the level of agreement among the clinicians involved in the study.

Crit no	Criterion title	No of responses	Validity ratings			
			Mean	Mode	Median	Standard deviation
1	Adm for adv results o/p mgt	149	6.75	8	7	2.45
2	Readmission for comp prev adm	150	7.34	8	8	1.97
3	Error in operative consent	150	5.75	10	6	3.13
4	Unpl rem/inj/repair in surg	149	7.78	10	8	2.51
5	Unpl return to theatre	150	6.97	8	7	2.00
6	Path/hist varies from diag	149	5.83	7	6	2.46
7	Prob of transfusion	148	6.39	8	7	2.61
8	Hosp acquired infection	150	6.33	8	7	2.33
9	Medication error/reaction	149	8.03	10	8	1.88
10	Cardiac/resp arrest in hosp	148	3.99	1	4	2.71
11	CVA/MI/PE in hosp after surg	149	5.32	5	5	2.52
12	Unexp transfer to spec care	150	5.27	7	6	2.70
13	Pt related clinical complcn	147	7.15	8	8	2.39
14	Non-clin problem/incident	146	7.24	8	8	2.07
15	Neuro deficit devel in hosp	144	5.38	7	5	2.66
16	Unexp patient death	148	5.50	5	5	2.72
17	Medical record deficiency	149	6.77	7	7	2.26
18	Nursing record deficiency	149	6.78	8	7	2.33
19	Pt/family dissatisfaction	150	6.68	8	7	2.20
20	Discharge related problems	150	6.60	8	7	2.16
All criteria		2974	6.40	8	7	2.59

Table 4.2. Analysis of the ratings by all respondents of validity of all screening criteria in adverse-event measure.

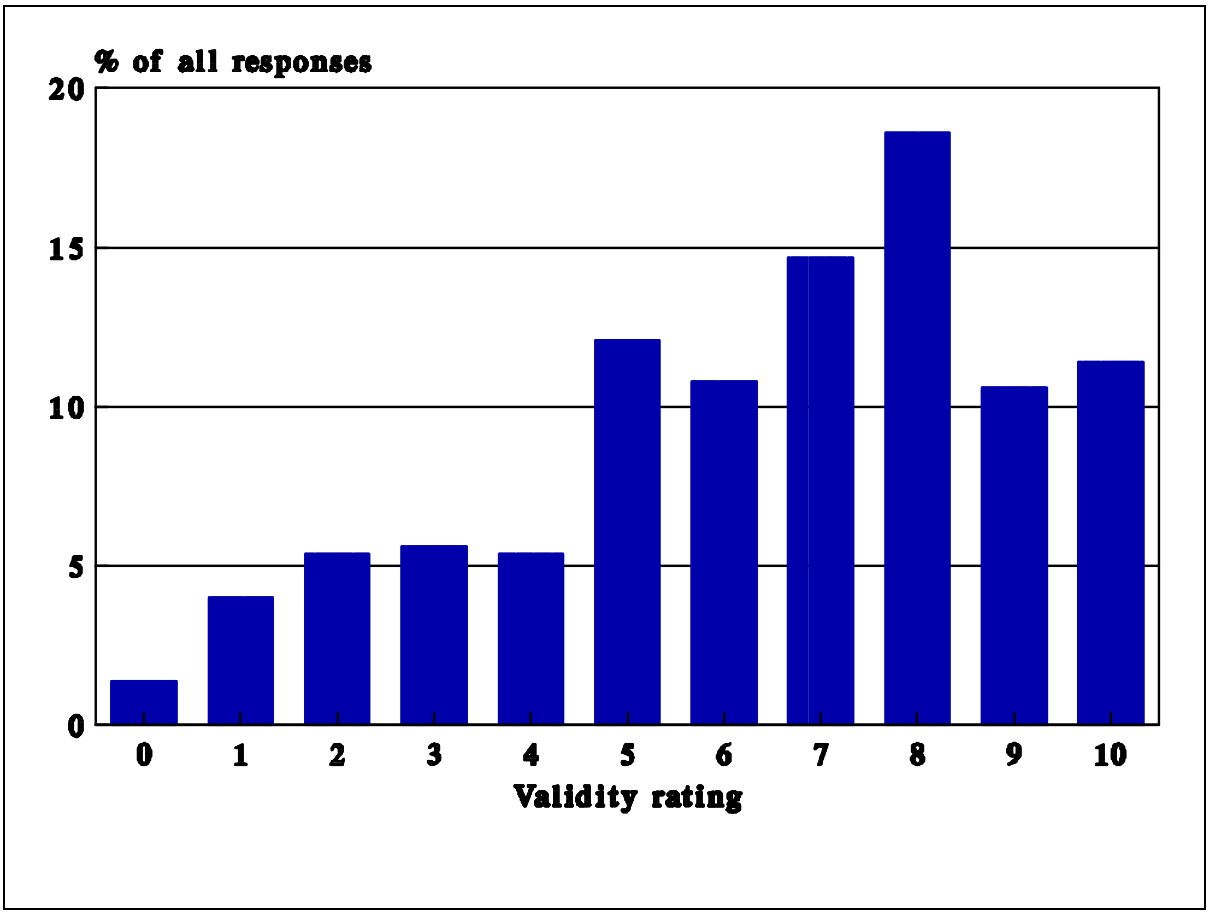


Figure 4.1. Distribution of validity ratings for all criteria by all respondents.

The graph in figure 4.1 shows the distribution of validity ratings by all respondents for all criteria. It is visually evident that the distribution is weighted towards the upper end of the validity rating scale, with a modal response of eight, and with 66.1% of all ratings being six or more. This suggests that the adverse-event measure was, in general, considered valid by respondents, since responses to the upper end of the validity rating scale predominated.

However, it is clear from table 4.2 that different criteria within the adverse-event measure received quite different ratings from respondents (means range from 3.99 to 8.03, modes range from one to ten). Those differences, as well as informing us about the validity of the criteria relative to one another, can also be used to make a judgement about whether each criterion (and, indeed, the measure as a whole) can be deemed a valid measure of the quality of care.

For example, the criterion whose validity was rated highest by respondents was criterion nine - which identified incidents of medication errors and reactions. It can be argued that this criterion has very strong face validity, since almost no-one would suggest that such adverse events are not important to the quality of care. Its mean validity rating was 8.03 (with a modal response of ten - the maximum validity).

In contrast, the criterion whose validity was rated lowest by respondents was criterion ten - which identified incidents of cardiac or respiratory arrest occurring during the patient's stay in hospital. Here, the face validity can be argued to be low, since many patients are admitted to hospitals for observation because of the risk of cardiac or respiratory arrest, and since most such arrests in hospital result from the progress of the patient's disease rather than from the effects of poor quality care. Its mean validity rating was 3.99 (with a modal response of one - almost the lowest rating respondents could give).

If we accept that criterion nine's rating of 8.03 (mode of ten) constitutes acceptable validity, and that criterion ten's rating of 3.99 (mode of one) constitutes unacceptable validity, the next step is to consider the distribution of other criteria across the range between these two values. As can be seen from figure 3.2, criterion nine is a low outlier. Most other criteria have mean validity ratings of above 5.0, and 13 out of 20 criteria have means above 6.0. The mean rating for all criteria is 6.4. Similarly, table 4.2 shows that the modal validity rating for most criteria was seven or more, and that the modal rating across all criteria was seven. These ratings provide reasonable evidence that the respondents to the questionnaire study considered most criteria within the adverse-event measure to be valid indicators of the quality of care.

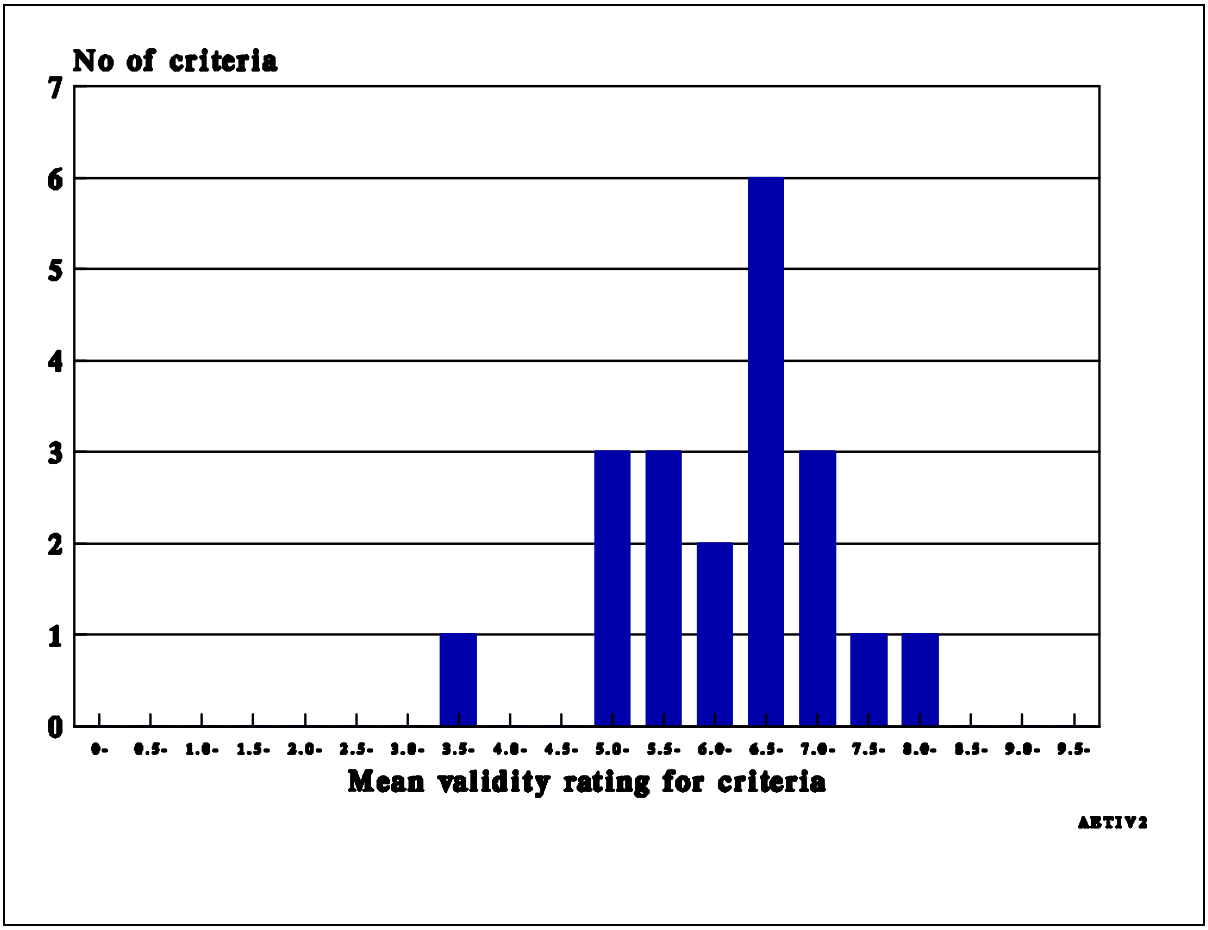


Figure 4.2. Distribution of mean validity ratings for all criteria.

The level of agreement among respondents about each criterion can be measured by the standard deviation (shown in table 4.2). Differences between the mean, median and modal responses for each criterion may also indicate disagreement about its validity. Two criteria show particular disagreement: criterion ten (cardiac/respiratory arrest while in hospital) and criterion three (error in operative consent). For the former, the low level of agreement seems to result from a minority of respondents rating its validity much more highly than the majority who gave it a low validity rating. For the latter, the low level of agreement seems to result from a genuinely wide spread of opinion among respondents about whether this essentially documentation-oriented criterion is a valid indicator of the quality of care.

Expected incidence of adverse events

Each respondent was asked to rate the expected incidence of each screening criterion in the adverse-event measure on a scale of zero to ten, on which zero indicated that the events identified by the screening criterion would be expected to happen to no patients, while ten indicated they would be expected to happen to all patients. This information was sought in order to assess the yield of adverse events which the screening criteria within the adverse-event measure might be expected to produce. It was also sought because, unlike other information gathered by the questionnaire, it could be compared with empirical evidence of actual incidence rates from the RSCH occurrence screening project. If the clinician ratings corresponded to the empirical findings, this would support the use of the clinician ratings of other aspects of the measure which could not be tested empirically.

The results are summarised in table 4.3. For each screening criterion, the mean response is given, along with the associated ranking of the criterion (ranging from one for the highest expected incidence criterion to 20 for the lowest expected incidence). Empirical incidence rates and associated rankings are also given, from a sample of 8,504 inpatients in several specialties who were screened using this adverse-event measure during the RSCH occurrence screening project.

The mean ratings for each criterion given by respondents cannot be directly interpreted as simple incidence rates, because of the nature of the scale on which the ratings were made. Although it was intended that zero represented a 0% incidence rate, and ten represented a 100% incidence rate, the rankings in between were used by respondents to record relative rather than literal incidence rates, because of the poor granularity offered by the scale. For example, if the scale had been used literally, a response of one would have represented a 10% incidence, and respondents who expected incidences somewhere between 0% and 10% would have been forced to opt for either one or the other. Since many adverse events are relatively rare occurrences, the scale used provided insufficient resolution for respondents to estimate literal rather than relative incidence rates, and in practice respondents used the scale to indicate relative expected incidence rates. In retrospect, a different response scale for this question would have yielded additional useful information, although this issue was not raised by the participants in the pilot study. For example, a simple percentage scale could have been used, on which respondents rated expected incidence on a scale of 0 to 100.

Crit no	Criterion title	Expected incidence ratings			Actual incidence data		Rank differ -ence
		No of responses	Mean response	Relative ranking	Incidence (%)	Relative ranking	
1	Adm for adv results o/p mgt	136	1.87	11	1.28	12	-1
2	Readmission for comp prev adm	136	2.13	10	3.62	4	6
3	Error in operative consent	135	1.69	12	1.35	11	1
4	Unpl rem/inj/repair in surg	135	1.39	18	2.18	9	9
5	Unpl return to theatre	137	1.53	15	0.83	13	2
6	Path/hist varies from diag	137	2.75	7	0.32	15	-8
7	Prob of transfusion	135	1.56	13	0.47	14	-1
8	Hosp acquired infection	139	2.84	5	2.25	8	-3
9	Medication error/reaction	137	2.64	8	3.08	6	2
10	Cardiac/resp arrest in hosp	136	1.41	17	0.27	17	0
11	CVA/MI/PE in hosp after surg	136	1.55	14	0.24	18	-4
12	Unexp transfer to spec care	137	1.47	16	0.31	16	0
13	Pt related clinical complcn	135	2.87	4	2.85	7	-3
14	Non-clin problem/incident	134	3.34	3	9.92	3	0
15	Neuro deficit devel in hosp	133	1.29	19	0.14	20	-1
16	Unexp patient death	137	1.18	20	0.19	19	1
17	Medical record deficiency	137	5.37	1	14.81	2	-1
18	Nursing record deficiency	135	4.32	2	26.70	1	1
19	Pt/family dissatisfaction	136	2.61	9	1.95	10	-1
20	Discharge related problems	137	2.80	6	3.35	5	1
All criteria		2720	2.33	-	-	-	-

Table 4.3. Analysis of the ratings by all respondents of expected incidence of all screening criteria in adverse-event measure, compared to empirical findings on actual incidence of all screening criteria in adverse-event measure from a sample of 8,504 patients screened during the RSCH occurrence screening project.

It is immediately evident from table 4.3 that the rankings provided by the respondents to the questionnaire study correspond quite closely with the empirically derived rankings from the RSCH occurrence screening project. Indeed, if the expected and empirically derived incidence rankings are compared, using the Spearman rank correlation coefficient (Hays 1974, p788), $r_s = 0.826$, which indicates a strong degree of correlation ($p < 0.001$). An alternative statistical test which can also be employed to measure the correlation between the two sets of rankings is Kendall's Tau coefficient (Siegel and Castellan 1988, p245). For the data in table 4.3, $\tau = 0.662$, which indicates that a strong and statistically significant correlation exists ($p < 0.0001$). This empirical confirmation of the clinician opinions sought in the questionnaire study is encouraging, and strengthens the validity of the clinician ratings of other aspects of the adverse-event measure which cannot be empirically verified.

Unsurprisingly, the adverse events which clinicians correctly expected to be most frequent were those which were also deemed least serious (see the discussion of severity of effect ratings below). Events such as medical and nursing record deficiencies and non-clinical problems and incidents (criteria 17, 18 and 14) were rightly thought to be the commonest adverse events included in the adverse-event measure.

It is interesting to note the criteria for which expected and empirical incidence rankings diverge. Respondents overestimated the frequency of criterion six (pathology/histology results vary from diagnosis) and 11 (CVA, MI or PE after surgery). However, they underestimated the relative frequency of criteria two (readmission resulting from previous care) and four (unplanned removal injury or repair during surgery).

Availability of required information in records

Each respondent was asked to rate the likely availability in patients' casenotes (including both medical and nursing records) of information about the events specified in each screening criterion in the adverse-event measure. The scale used for rating availability ranged from zero (information never found in records) to ten (information always found). This question aimed to identify potential practical difficulties which might emerge in applying the adverse-event measure, and to gauge the risk that an absence of information would produce results which showed an artificially low incidence rate for some criteria within the measure. Clearly, if some criteria require information which is not routinely recorded in patients' records, the measure might be improved by removing or adjusting those criteria.

Crit no	Criterion title	No of responses	Availability of information ratings			
			Mean response	Modal response	Median response	Standard deviation
1	Adm for adv results o/p mgt	149	5.55	5	5	2.57
2	Readmission for comp prev adm	149	6.88	8	7	2.13
3	Error in operative consent	150	7.10	10	8	3.08
4	Unpl rem/inj/repair in surg	149	7.76	8	8	2.09
5	Unpl return to theatre	150	8.02	9	9	1.99
6	Path/hist varies from diag	149	7.66	9	8	2.05
7	Prob of transfusion	148	7.12	8	8	2.20
8	Hosp acquired infection	150	6.33	8	7	2.33
9	Medication error/reaction	148	6.30	8	7	2.36
10	Cardiac/resp arrest in hosp	149	8.39	10	9	2.33
11	CVA/MI/PE in hosp after surg	149	8.00	10	9	2.18
12	Unexp transfer to spec care	150	8.08	10	9	2.25
13	Pt related clinical complcn	147	6.33	8	7	2.16
14	Non-clin problem/incident	146	4.23	5	4	2.38
15	Neuro deficit devel in hosp	146	6.86	8	7.5	2.41
16	Unexp patient death	147	8.21	10	9	2.65
17	Medical record deficiency	146	8.18	10	10	2.72
18	Nursing record deficiency	147	8.06	10	10	2.81
19	Pt/family dissatisfaction	150	3.11	2	2	2.42
20	Discharge related problems	150	3.97	2	4	2.51
All criteria		2969	6.81	10	8	2.82

Table 4.4. Analysis of the ratings by all respondents of the availability of information in medical and nursing records for all screening criteria in adverse-event measure.

Table 4.4 shows the respondents' ratings of information availability for each criterion within the measure. For each screening criterion, the mean, mode and median responses are given. The table also shows the standard deviation for each criterion.

It is evident from the mean and modal responses shown in table 4.4, that for most screening criteria in the adverse-event measure, respondents believed that the necessary information would be available in the medical and nursing records. Seven criteria have a modal response of ten (the maximum, denoting information always available), and 16 criteria have a modal value of eight or more. Indeed, as the graph in figure 4.3 shows, the modal response for all criteria was the maximum value of ten, and 52% of all ratings of the availability of information were eight or more.

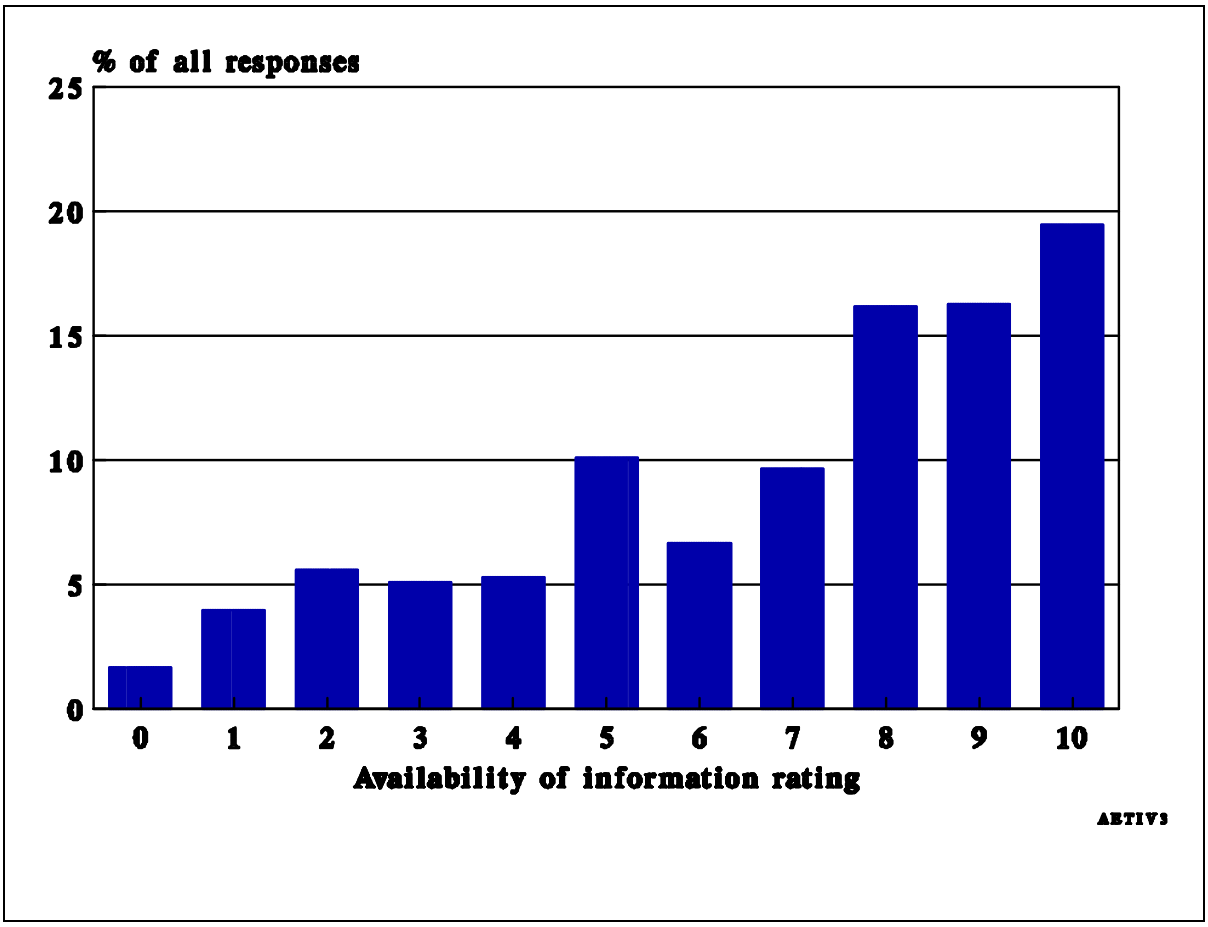


Figure 4.3. Distribution of availability of information ratings for all criteria by all respondents.

The lowest ratings of the likely availability of information were given to criteria one (admission for adverse results of outpatient management), 14 (non-clinical problems and incidents), 19 (patient or family dissatisfaction) and 20 (discharge related problems). For these criteria, the mean ratings ranged from 3.11 to 5.55 (modal response two to five). In each case, the probable reasons for respondents' lower ratings are fairly clear. Either the criteria related to events which occur outside the hospital environment which may therefore not be documented in hospital records, or they related to events which though important may not be routinely recorded in patient records.

Severity of effect of adverse events

Respondents were asked to rate the likely severity of effect of the adverse event defined in each screening criterion in the adverse-event measure, in terms of its impact on patients' health. The rating scale ranged from zero (no effect on patient's health) to ten (very serious effect on patient's health). The severity of effect of adverse events on patients' health is clearly important to the validity of the adverse-event measure (since events with little or no effect may not be good indicators of the quality of healthcare). However, the information was also sought because aggregating the results of any adverse-event measure involves combining scores for different screening criteria (which identify quite different adverse events) into a single score. It is simplistic to weight all adverse events equally, since some are clearly much more important than others. To develop a weighting system requires some quantitative assessment of the relative importance of different screening criteria. Clinicians' ratings of the severity of effect of adverse events on patients' health might form the basis for such a weighting system.

Table 4.5 shows respondents' ratings of the severity of effect of each screening criterion within the measure. For each screening criterion, the mean, mode and median responses are given. The table also shows the standard deviation for each criterion.

The mean clinician ratings shown in table 4.5 reflect clinicians' priorities and concerns. The highest ratings were given to those adverse events which have direct and serious effects on patients' physical well-being, such as unexpected death, cardiac and respiratory arrest, and CVA/MI or PE after surgery (criteria 16, 10 and 11). Non-clinical events, though they may be of great significance to patients, received generally lower ratings. It is indicative of respondents' priorities that some documentation-related adverse events such as medical and nursing record deficiencies (criteria 17, and 18) are rated as having a more serious effect on patients' health than either patient or family dissatisfaction or discharge related problems (criteria 19 and 20).

Crit no	Criterion title	No of responses	Rating of severity of effect on patients' health.			
			Mean response	Modal response	Median response	Standard deviation
1	Adm for adv results o/p mgt	138	5.61	5	6	2.14
2	Readmission for comp prev adm	142	6.22	8	7	2.14
3	Error in operative consent	144	3.15	1	2	2.70
4	Unpl rem/inj/repair in surg	142	6.93	5	7	2.20
5	Unpl return to theatre	141	6.65	7	7	2.03
6	Path/hist varies from diag	137	5.76	5	6	2.41
7	Prob of transfusion	139	6.52	8	7	2.27
8	Hosp acquired infection	138	5.80	5	6	1.76
9	Medication error/reaction	141	6.23	5	6	2.22
10	Cardiac/resp arrest in hosp	142	9.03	10	10	1.60
11	CVA/MI/PE in hosp after surg	145	8.73	10	9	1.34
12	Unexp transfer to spec care	139	7.64	8	8	1.88
13	Pt related clinical complcn	135	6.33	5	6	1.90
14	Non-clin problem/incident	132	4.54	5	5	2.05
15	Neuro deficit devel in hosp	136	7.31	8	8	2.00
16	Unexp patient death	136	9.55	10	10	1.68
17	Medical record deficiency	140	4.51	5	5	2.24
18	Nursing record deficiency	140	4.16	5	4	2.28
19	Pt/family dissatisfaction	141	3.85	5	4	2.28
20	Discharge related problems	141	4.06	2	4	2.27
All criteria		2789	6.13	5	6	2.71

Table 4.5. Analysis of the ratings by all respondents of the severity of effect on the patients health of adverse events defined by screening criteria in adverse-event measure.

Potential for improvement in the adverse-event measure

Respondents were asked to indicate, for each screening criterion within the adverse event measure, whether or not they felt that the criterion could be "altered to make it related more closely to the quality of care that patients receive". The purpose of this question was to gauge respondents' attitude towards the adverse-event measure as it stood, and to pinpoint particular screening criteria which consensus opinion indicated should be revised. The question also sought textual comments from respondents on specific improvements, which are analysed later in this section.

Table 4.6 shows respondents' opinions on the potential for improvement of each screening criterion. The table gives the number and proportion of respondents replying that the criterion could or could

not be improved for each criterion. Of course, respondents might indicate that they saw no potential for improvement in a criterion for two reasons: either because they felt it was already very well defined, was a valid indicator of quality, and they were satisfied with it as it stood; or because they felt it was highly inappropriate, was not a valid indicator of quality, and should be removed altogether rather than modified. It would therefore be unwise to interpret a high "could not be improved" rating as an indication of respondents' satisfaction with or approval of a criterion without also considering their ratings of its validity.

Crit no	Criterion title	No of responses	Could not be improved		Could be improved	
			Number	Percent	Number	Percent
1	Adm for adv results o/p mgt	136	85	62.5	51	37.5
2	Readmission for comp prev adm	134	93	69.4	41	30.6
3	Error in operative consent	130	106	81.5	24	18.5
4	Unpl rem/inj/repair in surg	135	96	71.1	39	28.9
5	Unpl return to theatre	134	105	78.4	29	21.6
6	Path/hist varies from diag	128	95	74.2	33	25.8
7	Prob of transfusion	131	94	71.8	37	28.2
8	Hosp acquired infection	127	96	75.6	31	24.4
9	Medication error/reaction	130	89	68.5	41	31.5
10	Cardiac/resp arrest in hosp	127	102	80.3	25	19.7
11	CVA/MI/PE in hosp after surg	130	100	76.9	30	23.1
12	Unexp transfer to spec care	129	107	82.9	22	17.1
13	Pt related clinical complcn	123	90	73.2	33	26.8
14	Non-clin problem/incident	122	92	75.4	30	24.6
15	Neuro deficit devel in hosp	125	95	76.0	30	24.0
16	Unexp patient death	122	102	83.6	20	16.4
17	Medical record deficiency	125	98	78.4	27	21.6
18	Nursing record deficiency	124	99	79.8	25	20.2
19	Pt/family dissatisfaction	126	99	78.6	27	21.4
20	Discharge related problems	123	97	78.9	26	21.1
All criteria		2561	1940	75.8	621	24.2

Table 4.6. Analysis of the opinions of all respondents on the potential for improvement of screening criteria in adverse-event measure.

Overall, respondents indicated that that there were no improvements they would wish to make to most criteria, replying in 75.8% of cases that they would leave the criterion as it was. For most criteria, between 70% and 85% of respondents would leave the definition as it stood. The criteria with the greatest perceived potential for improvement were those dealing with admissions and readmissions (criteria 1 and 2) and medication errors and reactions (criteria 9). The improvements suggested in these cases are detailed below.

Comparing the opinions of different respondent groups

In section 4.2.2, the potential biases arising from the different response rates from the three groups, and the preponderance of public health physicians among respondents, were discussed. In particular, it was suggested that public health physicians might rate the adverse-event measure differently from practising clinicians, and so their overrepresentation among respondents might give a misleading impression of overall clinical opinion.

In order to test this hypothesis, respondents to the questionnaire were divided into two groups: public health physicians (study group A described earlier), and practising clinicians (study groups B and C). For each respondent, a mean response to each question on the questionnaire was calculated (by taking the mean of his or her responses to that question for all 20 criteria). This mean response effectively represents the respondent's rating of the adverse-event measure as a whole on that question. These respondent means were calculated for each of the four questions asked about each criterion - the relationship to quality, the expected incidence, the availability of information in records, and the severity of effect. The distributions of respondent means for the two groups were then compared for each of the questions for which detailed analyses have already been described - the rating of validity, the estimate of expected incidence, the rating of likely availability of information, and the expected severity of effect.

The results of this analysis are shown in table 4.7. For each question, information about the distributions of respondent means for the two groups is given and compared. Firstly, for each group the mean and standard deviation of the distribution of respondent means are given. Then, two statistics which compare the two groups are cited. The Wilcoxon-Mann-Whitney test is a non-parametric test which shows whether two groups are drawn from the same population (Siegel and Castellan 1988, p128). The Kolmogorov-Smirnov two sample test also tests whether two groups are drawn from the same population, but is sensitive to differences in central location, dispersion and skewness (Siegel and Castellan 1988, p144). For each test, the probability that the two groups are drawn from different populations is cited.

Question	Public health physicians (n = 132)		Practising clinicians (n = 18)		Wilcoxon Mann-Whitney test p value	Kolmogorov -Smirnov 2 sample test p value
	Mean	Standard deviation	Mean	Standard deviation		
Validity as measure of quality	6.48	1.29	5.85	1.52	0.059	0.011
Expected incidence	2.34	1.02	2.46	1.25	0.869	0.651
Availability of information	6.81	1.15	6.89	1.22	0.991	0.974
Severity of effect	6.12	1.20	5.97	1.40	0.529	0.330

Table 4.7. Comparison of respondent mean distributions between groups of public health physicians and practising clinicians.

Table 4.7 shows that for three of the four questions asked in the questionnaire - those relating to expected incidence, availability of information, and severity of effect - there was virtually no difference between the distributions of respondent means for the two groups. This means that public health physicians and practising clinicians gave similar mean ratings to the adverse-event measure on these questions.

However, the distribution of respondent means for ratings of the validity of the adverse-event measure as a measure of quality do show some significant differences. On average, public health physicians rated the validity as a measure of quality of the measure slightly higher than the practising clinicians (6.48 versus 5.85) and both the Wilcoxon-Mann-Whitney and Kolmogorov-Smirnov two sample test suggest the difference between the two distributions may be significant ($p = 0.059$ and $p = 0.011$ respectively).

Of course, the respondent means used in the above analysis could hide important differences in the way public health physicians and practising clinicians rated individual screening criteria within the adverse-event measure. Such differences can be examined by repeating the analysis of table 4.7 for each criterion individually.

Crit no	Criterion title	Public health physicians (n = 132)			Practising clinicians (n = 18)			Wilconxon Mann-Whitney test p-value	Kolmogorov-Smirnov 2 sample test p-value
		Mean	Rel rank	Std dev'n	Mean	Rel rank	Std dev'n		
1	Adm for adv results o/p mgt	6.57	13	2.46	8.06	2	1.98	0.008	0.113
2	Readmission for comp prev adm	7.34	4	1.92	7.39	3*	2.40	0.528	0.787
3	Error in operative consent	6.05	14	2.95	3.61	20	3.63	0.005	0.014
4	Unpl rem/inj/repair in surg	7.84	2	2.46	7.33	5	2.87	0.472	0.977
5	Unpl return to theatre	6.91	8	2.02	7.39	3*	1.82	0.4683	0.998
6	Path/hist varies from diag	5.77	15	2.44	6.28	8	2.63	0.406	0.970
7	Prob of transfusion	6.58	12	2.43	4.94	14*	3.49	0.059	0.070
8	Hosp acquired infection	6.60	11	2.07	4.78	16*	2.69	0.004	0.009
9	Medication error/reaction	8.00	1	1.92	8.28	1	1.60	0.680	0.991
10	Cardiac/resp arrest in hosp	3.86	20	2.57	4.94	14*	3.54	0.330	0.287
11	CVA/MI/PE in hosp after surg	5.34	18	2.45	5.17	12*	3.02	0.769	0.833
12	Unexp transfer to spec care	5.26	19	2.60	5.39	11	3.47	0.773	0.554
13	Pt related clinical complcn	7.19	5	2.34	6.83	6	2.79	0.722	0.756
14	Non-clin problem/incident	7.35	3	1.97	6.44	7	2.60	0.179	0.192
15	Neuro deficit devel in hosp	5.50	17	2.59	4.41	18	2.98	0.129	0.226
16	Unexp patient death	5.66	16	2.62	4.33	19	3.18	0.067	0.134
17	Medical record deficiency	6.99	7	2.04	5.17	12*	3.03	0.014	0.019
18	Nursing record deficiency	7.05	6	2.09	4.78	16*	2.98	0.002	0.010
19	Pt/family dissatisfaction	6.83	9	1.93	5.61	10	3.50	0.352	0.032
20	Discharge related problems	6.71	10	2.07	5.78	9	2.65	0.216	0.886

* denotes tied rankings.

Table 4.8. Comparison of respondent ratings of the validity of screening criteria within the adverse-event measure.

Since the ratings of validity obtained from respondents were the primary aim of the questionnaire, and since it was only in the ratings of validity that some differences seemed to exist between public health physicians and practising clinicians, it is appropriate to examine these ratings in more detail. To this end, table 4.8 shows an analysis of validity ratings for each screening criterion by the two groups. For each criterion, the mean rating, relative ranking and standard deviation of the distribution of ratings for the two groups is given. The same two statistical tests (Wilcoxon-Mann-Whitney test and Kolmogorov-Smirnov two sample test) are used to compare the two groups.

Table 4.8 shows that, while there are some differences in the validity ratings given to individual criteria by the two groups of respondents, in most cases the ratings are fairly similar. Indeed, although the value of the mean validity ratings of the two groups for particular criteria may differ, it is clear from table 4.8 that in most cases the relative rankings of the criteria (where 1 is the most valid and 20 is the least valid indicator of the quality of care) are similar. If the relative rankings are compared using the Spearman rank correlation coefficient adjusted for ties (Siegel and Castellan 1988, p239), $r_s = 0.602$ ($p < 0.01$). If Kendall's Tau coefficient adjusted for ties is calculated (Siegel and Castellan 1988, p249) using the ranking data in table 4.8, $\tau = 0.399$ ($p = 0.007$). These results suggest that there is a strongly significant correlation between the validity rankings provided by the two groups.

It is interesting to note the screening criteria on which the public health physicians' and practising clinicians' opinions of validity diverge. The main area of contention is the validity of adverse events which are essentially documentation-related (criteria three, 17 and 18). Public health physicians saw these adverse events as more valid indicators of the quality of care than their clinical colleagues. Perhaps because of their epidemiological training and responsibilities, public health physicians also regarded hospital acquired infections (criterion eight) as more valid indicators of quality than did the practising clinicians. More puzzlingly, practising clinicians gave a higher validity rating to admissions resulting from the effects of outpatient management (criterion one) than the public health physicians.

Respondents' comments on the validity of the adverse-event measure

Many respondents' commented on the validity of the adverse-event measure as an indicator of quality (and, often, about the validity of adverse-event measures in general). The range and variety of comments received indicated that a broad spectrum of opinion existed, and comments were neither overwhelmingly in favour of such measures nor overwhelmingly against them. For example, some respondents praised the measure:

"As screening criteria I think most of the twenty are very useful".

"Most of the clinical indicators are valid in terms of quality of care ... I thought the measures were well derived."

However, others clearly had doubts about the practicality of the measure, its relevance, and the attitudes to quality improvement which it might encourage:

"The whole approach horrifies me - bad apples versus continuous quality improvement. Too much emphasis on outliers - it hasn't worked in US."

"Nearly all the criteria are too loosely defined at the present. They need to be much more specific."

"This questionnaire convinces me even more that this type of audit is costly, has a low payback rate, and is threatening and about blame. 90% of the clinical issues will have a perfectly rational explanation. Let us learn from the American experience."

The main areas of concern about the use of adverse-event measures in general which were raised by respondents generally reiterated some of the potential problems with such measures which were enumerated and described in chapter 2. The primary themes which emerged from respondents' comments were:

a) *Negativity.*

A number of respondents expressed concern that the focus on negative events inherent in any adverse-event measure would skew the medical or clinical audit process towards the detection of errors and mishaps, and produce an atmosphere of blame and recrimination among clinical staff. Several respondents cited the American experience with such measures as evidence that this might happen.

b) *Cost.*

Some respondents questioned the cost of gathering adverse-event data across large numbers of patients, especially when some events would be very rare and so individual instances of those events would be expensive to detect.

c) *Accuracy*.

While the ability of the measure to identify instances of poor quality care was generally accepted, several respondents felt that many adverse events identified by the measure would, on examination, turn out to be normal variations in clinical practice or patient outcomes. Some saw the adverse-event measure as being primarily of use in identifying cases for subsequent review by clinicians rather than as a measure in itself.

d) *Generic nature*.

The adverse-event measure which respondents were given was generic in nature, designed to be applicable to acute inpatient care in a wide range of specialties or settings. A number of respondents identified this as a potential problem, suggesting that patient factors (such as age, sex, general health, etc) and diagnosis or procedure-specific factors (such as the nature and severity of disease and the complexity of procedures or treatments) were so important in determining whether or not adverse events had occurred that they needed to be explicitly taken into account in the definition of such measures.

e) *Availability of information*.

Some respondents suggested that the adverse-event measure was heavily reliant on good quality medical records, and that some of the screening criteria it contained defined adverse events which might not be routinely documented in patients' medical records.

Respondents also made some specific criticisms of the adverse-event measure itself (in other words, comments which were specific to this measure rather than being directed at adverse-event measures in general), relating to how it had been constructed and defined. Two main points were raised:

a) *Inappropriate groupings*.

The most frequently occurring criticism was that the individual screening criteria grouped together adverse events which were dissimilar in causation, in their nature, and in the effect they might have on patients. This made rating their validity, effect on patients' health and other aspects more difficult, and might reduce the validity of the measure as a whole. Some specific examples are outlined below. It was suggested that subdividing some screening criteria into two or more separate criteria would be the best way to deal with this problem.

b) *Inadequate definition*.

A number of respondents suggested that criteria needed to be more clearly defined, and that the terms used in existing definitions were vague or were open to more than one interpretation. It was suggested that greater precision and less subjectivity would be achieved if more examples were given, and if definitions were stated in more detail. Of course, each criterion within the measure did have a much more detailed definition, which was not given to the clinicians participating in the survey because it would have presented them with much more information to read and comprehend before filling in what was already a very time-consuming questionnaire.

It was, to some degree, reassuring that no wholly new criticisms of adverse-event measures emerged from respondents' comments. The issues raised were all ones which had been considered and raised during the development of the adverse-event measure, or which had been identified by other developers and users of such measures.

Respondents' comments on the validity of individual criteria

The questionnaire sought comments from respondents on the criteria which made up the adverse-event measure. In particular, respondents were asked to suggest ways in which the criteria could be improved. Respondents provided a number of suggestions for improvements to individual screening criteria within the adverse-event measure, and a range of comments on their individual validity and suitability. All these comments are listed in full in Appendix 4.4. Below, for each screening criterion, a brief summary of the main comments, criticisms and proposed improvements is given:

a) *Admission for adverse results of outpatient management*.

This criterion was generally seen as very broad, covering a wide range of different sorts of adverse events, and liable to pick up false positives. More definition of the concepts involved was recommended, with particular attention to the issue of causation.

b) *Readmission for complications relating to previous admission*.

Comments were similar to those for criterion one above, with respondents identifying problems in defining the events concerned and their causation as the main limitation to the use and validity of this criterion. Suggestions included better definition of the factors involved in readmission, particularly its timing and causation. It was pointed out that in some specialties, such as elderly medicine, discharge home and subsequent unplanned readmission if necessary may actually represent better quality of care than the retention of the patient in hospital.

c) *Error in obtaining consent to operative procedure.*

This criterion was seen as a useful proxy for quality in other areas (in that poor documentation of a legal requirement such as operative consent may suggest a lack of attention to detail in other areas), though many respondents raised what they saw as the more important issue of patients' understanding of procedures and their informed consent. Of course, current documentation does not provide the information needed to assess whether patients' understand the treatments to which they are consenting.

d) *Unplanned removal/injury/repair of organ/structure during surgery*.

Many respondents pointed out that unplanned actions might be taken because of the findings of exploratory surgery (though in fact the definition of this criterion explicitly excluded that kind of unplanned action). However, replacing the word unplanned with unintended or accidental might make the meaning clearer. Some respondents suggested this was a rare event, and that some technical errors of this sort were very difficult to avoid.

e) *Unplanned return to theatre.*

This criterion attracted relatively little criticism, though the suggestions for improvements tended to echo those set out for criterion four above. Some modification to make the causal relationship between the first and second procedure clearer was suggested.

f) *Pathology/histology varies from diagnosis*.

Many respondents commented on this criterion, and suggested that the two main groups of adverse events - one concerned with pathology or histology results, and the other with postmortem results - should be separated into two criteria. It was pointed out that postmortems often produce a different diagnosis of the cause of death from the antemortem diagnosis (which may suggest that this is a valuable criterion to include). Some respondents identified a need to define more clearly the effect of this variation on treatment or clinical management, and to focus attention on those cases where a significant difference existed.

g) *Transfusion problems: reactions, complications, usage.*

This criterion was one of those criticised by many respondents as incorporating three distinctly different types of adverse event, which are only linked by the fact that they involve the use of blood and blood products. Some respondents suggested trying to omit adverse events which are probably unavoidable (such as reactions to transfusions) and focusing on avoidable complications and misuse of the service. it was also suggested that misuse needed to be more clearly defined.

h) *Hospital acquired infection*.

The main concern raised by respondents in relation to this criterion was the issue of definition. Hospital acquired infections were recognised as clinically important, relatively common and often preventable. Respondents suggested trying to define more clearly the separation between hospital acquired and pre-existing infections, and classifying infections according to their site or type. Some respondents pointed out that many less serious hospital acquired infections may not be actively treated or even recorded.

i) *Antibiotic/drug utilisation problems*.

This criterion grouped together medication reactions, and prescribing and administration errors. Like criterion seven above, it was mainly criticised because it grouped adverse events which were really quite different from each other together. Also like criterion seven, it was suggested that most drug reactions are probably unavoidable, while prescribing or administration errors were preventable. Many respondents suggested separating the criterion into two or more criteria.

j) *Cardiac or respiratory arrest in hospital.*

The primary criticism voiced by some respondents of this criterion was that they believed most hospital cardiac and respiratory arrests to be unavoidable, resulting primarily from the progress of the patients' disease. However, others suggested that avoidable factors did exist, and might be incorporated into the criterion. Some respondents indicated that the management of such arrests was a more appropriate focus for the criterion.

k) *CVA or MI or PE following surgery*.

Some respondents suggested that the three events linked together in this criterion would be better separated into different criteria. In particular, pulmonary embolus was seen as largely avoidable, while cerebrovascular accidents and myocardial infarctions were seen as resulting from the patients' pathology or disease process.

l) *Unexpected transfer to special care/higher dependency unit.*

This criterion is designed to identify sudden transfers to high dependency or intensive care units which might result from an unexpected deterioration in a patient's condition. Some respondents wanted more attention to be given to the reasons for the transfer in the definition, while others pointed out that a transfer at an appropriate point in a patient's stay might actually represent timely detection of problems and a high quality of care.

m) *Patient related clinical complications occurred.*

This criterion, which links together a number of quite diverse clinical complications, was criticised by some respondents for that diversity. Some suggested splitting it into several criteria, or limiting it to a much smaller set of complications. The complications included in the criterion were generally recognised as being valid attributes of the quality of care.

n) *Patient-related non-clinical problems/incidents occurred.*

Respondents made similar comments about the diversity of this criterion as were described above for criterion 13. Interestingly, some respondents saw the events grouped together in this criterion as more avoidable than the clinical issues detailed in criterion 13.

o) *Neurological deficit on discharge not present on admission.*

In this case, many respondents felt that the criterion needed to be more specific about the nature of the neurological deficits involved, and their causes.

p*)* *Unexpected death.*

The primary problems with this criterion identified by respondents were the absence of a clear definition of unexpected, and the potential for overlap with many other criteria. Respondents wanted to adjust the definition to focus it more closely on avoidable deaths.

q) *Medical record deficiency.*

Most respondents combined their comments on this criterion and criterion 18 (below). Many suggested that a clearer definition of what the medical record should contain was needed if inadequacies were to be reliably identified, and proposed checklists of items which should be present. Respondents seemed to view medical and nursing record quality as important issues in themselves as well as being proxies for more clinical quality attributes. Some suggested that there was a need to distinguish between minor and more severe record deficiencies.

r) *Nursing record deficiency.*

See comments above for criterion 17 (point (q)).

s) *Evidence of patient and/or family dissatisfaction*.

It was interesting to note that respondents - who were all consultant medical staff - did not generally regard patient dissatisfaction as an important attribute of quality. Some suggested that adverse events identified by this criterion could be trivial and that non-clinical dissatisfaction was not especially relevant. Respondents also pointed out that most dissatisfaction is not documented in patient records, and so could not be identified by the adverse-event measure.

t) *Discharge related problems*.

Respondents generally felt that this criterion was a useful indicator of quality, though they highlighted the fact that it combined problems caused by hospital care (such as poor discharge planning) with those resulting from poor community care (such as shortage of nursing home provision) and suggested that these items might be separated.

Respondents' comments on the questionnaire design and completion

Respondents were largely complimentary about the actual questionnaire design and layout, and seemed to have found it and the accompanying letter comprehensible and relatively user-friendly. The only design issue which a few respondents raised was the placing of screening criteria definitions on the back page of the four page questionnaire booklet. A few respondents observed that they had not realised that these definitions were there until they were part or all of the way through completing the questionnaire.

Quite a number of respondents felt that they lacked some or all of the knowledge needed to fill in the questionnaire - particularly the ratings of availability of information, expected incidence of adverse events, and effect on patients' health which were sought for each screening criterion. Some felt uncomfortable that their answers were subjective and questioned how repeatable the study would be. It is interesting that although the question about the expected incidence of adverse events probably aroused the most concern and self-doubt in respondents, as has already been shown, their ratings on this question were actually well matched to the actual incidence of adverse events in a

relatively large sample of actual patients. This empirical confirmation of the accuracy of clinicians' ratings on one question may support their accuracy on other questions.

The use of zero to ten rating scales throughout the questionnaire attracted little comment, and seemed to be well accepted, except on the question about the expected incidence of adverse events. Here a number of respondents drew attention to the problems of rating on this scale, and indicated that they had interpreted the scale as an ordinal rather than a ratio scale. The need for a different scale on this question has already been discussed.

The length of time it would take to complete the questionnaire was expected to deter some potential respondents. This was confirmed by the comments of some respondents who did complete the questionnaire. They observed that it needed some determination to complete it, that it took longer than they had expected (and, indeed, than our estimate had suggested it would take), and it required good motivation to complete.

4.2.4 **Conclusions**

The aim of the questionnaire study of clinician opinion, set out in section 4.2.1, was to assess the content and face validity of the adverse-event measure, and to gather the opinions of clinical professionals on other aspects of the measure which were relevant to its validity.

Despite the relatively low response rate to the questionnaire study, the causes of non-completion do not in general detract from the ability of respondents to act as an expert panel in assessing the validity of the measure, or from the value of their comments on other aspects of the measure relevant to its validity.

The study found that the adverse-event measure was generally regarded as a valid measure of the quality of care given to acute hospital inpatients. Some individual criteria within the measure were regarded as more or less valid than others, and the measure could be made more valid in the opinion of some respondents by omitting some criteria, and restructuring or adding some others. However,

overall respondents indicated a generally high level of confidence in the validity of the measure. In their quantitative ratings and in their written comments they generally confirmed the face and content validity of the measure.

The study also found that respondents were able to predict, with reasonable accuracy, the relative incidence of adverse events found empirically with a sample of 8,504 patients screened using the adverse-event measure. This empirical confirmation of respondents' ratings provided some reassurance that respondents' ratings in other areas were also meaningful.

In addition, the study showed that respondents believed that, for most criteria within the adverse-event measure, the information needed to determine whether events had happened or not would be available in the patients' medical and nursing records. Although respondents to the study came from two distinct clinical professional groups - practising clinicians and public health physicians - the ratings of these two groups on all aspects of the measure were generally consistent with each other.

Respondents suggested a wide range of alterations to the adverse-event measure which might improve it, but overall they endorsed the measure as a valid measure of the quality of care.

4.3 Interviews with clinicians

4.3.1 **Aims of interview study**

Qualitative enquiry, using data collected through interviews, observation, or document analysis, allows the researcher to study issues in depth and detail, without being constrained or limited by predetermined categories of analysis (Patton 1990, p14). While quantitative methods facilitate the comparison and statistical aggregation of data, qualitative methods can produce detailed and insightful information about the subject of the research. Though its results may be less generalisable (and indeed generalisability is not one of its aims), qualitative research can identify themes or issues which quantitative enquiry, by its prestructured nature, excludes.

The primary aim of the interview study of clinician opinion on the use of adverse-event measures of quality was to support and supplement the largely quantitative information on the validity of adverse-event measures gathered and analysed by the parallel questionnaire study (reported in section 4.2). The interview study was designed to provide a more qualitative understanding of clinicians' perceptions of adverse-event measures of quality and their application in quality measurement and quality improvement, and to enable clinicians participating in the study to raise issues or themes which had not been addressed by the questionnaire study. It has been suggested that the simultaneous or sequential use of qualitative and quantitative research methods and the integration of their results, known as triangulation, provides stronger and more valid research results and conclusions than using one or other method by itself (Field and Morse 1985, p16).

While the broad aim of the interview study was to provide a qualitative understanding of clinicians' perceptions of adverse-event measures of quality and their application in quality measurement and quality improvement, it was undertaken with two main objectives in mind:

a) *To identify themes or issues omitted by the questionnaire study.*

The questionnaire study was focused on a series of issues related to the validity of the adverse-event measure being tested. In particular, it gathered clinicians' ratings of the validity of criteria within the measure, the expected incidence of adverse events and the severity of their effect on patients, the availability of information about adverse events in patients' records, and clinicians' suggestions for improvements in individual criteria or in the measure as a whole. While respondents' views on other aspects of adverse-event measures in general were solicited, it is in the nature of a postal questionnaire study, especially one as long and complex to complete as this one, that it provides little encouragement for respondents to address themes or issues outside its existing boundaries. Therefore, one objective of the interview study was to identify any themes or issues relating to the general validity and utility of adverse-event measures which did not emerge from the questionnaire study.

b) *To explore the place of adverse-event measures of quality in the quality measurement and quality improvement process in healthcare.*

It was noted in chapter 2 that one does not simply validate a measuring instrument. Rather, one validates a measuring instrument in relation to the purpose for which it is being used, and in the context or environment in which it is being deployed. While the questionnaire study was designed to validate the adverse-event measure, it was not intended to yield much information about the potential application of such measures, and their place in quality measurement and quality improvement. Therefore, a second objective of the interview study was to gather clinicians' opinions on the place of adverse-event measures in the current British healthcare environment, their relationship to other quality measures, and their application.

4.3.2 **Method**

In quantitative research, sampling strategies are generally driven by the objective of producing generalisable results, which can be used and applied by other people in a range of contexts or situations. As a result, they often use relatively large, random samples. By contrast, qualitative research is driven by the need to develop a deep understanding of the phenomenon under study within its own context. As a result, it typically focuses on relatively small samples, selected purposefully. The intention of such purposeful sampling is to select information-rich cases or subjects, which are judged likely to provide sufficient information about the phenomenon or issue under study (Patton 1990, p168).

For this interview study a purposeful sampling strategy was chosen. It was felt that information-rich subjects were likely to have a clinical background and experience in the practice of clinical medicine, an interest in quality assessment in healthcare, and some understanding of the use of adverse-event measures. In addition, they would need to be willing to commit time to being interviewed for the study. On this basis, it was decided to ask members of the Regional Medical Audit Advisory Committee (RMAAC) for the South East Thames region to participate in the study. This group contained a total of 10 senior medical staff, all of whom held senior posts in clinical

medicine, and who were (by virtue of their voluntary membership of this committee) interested and involved in quality assessment and quality improvement in healthcare. Small sample sizes of this kind are common in qualitative research, where the main criteria used to determine the sample size are its appropriateness (whether it fits the needs of the research study) and its adequacy (whether the data obtained is sufficient, and whether saturation is achieved in data collection) (Morse 1991).

A number of general themes were chosen to be explored with each interviewee, using the preliminary analysis of the textual responses in the questionnaire study and information from the literature review. These themes were then used to form the basis of a semi-structured interview schedule (a copy of which can be found in appendix 4.5). It was not intended that this schedule would be rigidly adhered to, but it was designed to guide the interviewer and to be used as the main contemporaneous record of the interview.

The open structure of the questions put to interviewees and the broad nature of the themes identified for discussion during interviews were both intended to encourage interviewees to raise issues which were of concern or interest to them, and to enable issues relating to the design or application of adverse-event measures of quality which might not have previously been encountered to be raised. After all, the primary aim of the interview study was not to confirm the findings of the questionnaire study or the existing literature - it was to identify any issues which might have been overlooked in those findings.

The six themes chosen for discussion in each interview were:

a) *The general advantages and disadvantages of adverse-event measures.*

This theme was used at the start of each interview to initiate a general discussion of the merits and demerits of adverse-event measures of quality, in which interviewees could raise any perceived advantages or disadvantages of these measures, without their opinions being directed or structured by the interviewer in any way. It was the most general and least specific of the six themes, since the remaining five themes addressed during the interview all focused on particular aspects of these measures.

b) *The similarities and differences between adverse-event measures and more traditional audit or quality assurance mechanisms*.

This theme was felt to be worthy of exploration because both the literature review and informal opinion from clinicians suggested that some important methodological similarities existed between adverse-event measures of quality and traditional audit mechanisms such as mortality and morbidity meetings or death and complication reviews. These similarities might be important aids in explaining adverse-event measures to clinicians and in influencing them to use such measures in quality assurance. It was therefore important to seek interviewees' views of any similarities or differences, without directing them to particular examples or instances.

c) *The utility of adverse-event information in measuring quality for individual patients or for groups of patients*.

In order to encourage interviewees to consider how useful they would find adverse-event information, this theme was raised during each interview. The term utility was used because it emphasised the practical issues in measurement and because the broad meaning of utility allowed interviewees to raise issues relating to validity, reliability, costs and almost any other aspect of the measure.

d) *The utility of adverse-event information in creating or promoting changes in practice and the quality of care.*

This theme was discussed with each interviewee, in combination with (c) above, in order to examine what interviewees felt would be the results of using adverse-event information. It was included in order to encourage interviewees to look beyond the process of measurement and to consider how adverse-event information might actually be put to use in producing changes in the quality of care. Again the term utility was used because its broad and relatively unspecific meaning allowed interviewees to raise a wide range of issues around the effects of adverse-event information on practice.

e) *The suitability of adverse-event measures of quality for different areas or specialties in inpatient care.*

It has been suggested, both in the literature review and in responses to the questionnaire study, that adverse-event measures of quality are less well suited to areas of healthcare where patients may have multiple problems and undergo multiple concurrent interventions (such as geriatrics or general medicine) and are more suited to areas where patients are generally healthy apart from a single problem which is treated by a single intervention (such as some surgical specialties). This theme aimed to explore any differences in the suitability of such measures without directing interviewees to the reasoning outlined above.

f) *Factors which might either facilitate or obstruct the application of adverse-event measures of quality*.

This last, broad theme sought interviewees' opinions on the organisational, structural and managerial factors which might either help the introduction of a quality assurance system based on adverse-event measures of quality or might hinder it. The theme aimed to explore interviewees' perceptions of the role of such measures in healthcare organisations in the NHS, and to seek their views of the way such measures would fit into the culture and ethos of such organisations.

A letter explaining the purpose of the interview was sent to all members of the Regional Medical Audit Advisory Committee (RMAAC) for the South East Thames region. It sought their agreement to be interviewed, and identified the general themes that the interview would seek to address. This letter was designed to give potential interviewees the opportunity to think about the issues involved before being interviewed. All interviews were conducted by telephone, because of the logistic difficulties of arranging meetings with interviewees who were geographically located across south east England. Interviews varied in length from 20 minutes to over an hour. All members of the group agreed to be interviewed, but in practice establishing times when they were able to make themselves available for the interview proved difficult in some cases. As a result, interviews were only conducted with 6 of the 10 members of the group. These interviews were conducted between November 1991 and March 1992.

Each interview commenced with a brief explanation of the aims of the interview. Then the themes identified on the interview schedule were explored in turn, allowing the interviewee to determine the

direction and subject of the interview. During each interview the interviewer made notes on a copy of the interview schedule, and these notes were then written up immediately the interview ended. Contemporaneous notes were used as the interview record, rather than tape recording and transcription, because the use of telephone interviewing made the keeping of such detailed notes relatively easy, and it was judged that verbatim transcripts would not add significantly to the value of the interview record as a data source in the context of the study aims.

The concepts of validity and reliability in quantitative research are paralleled by the concepts of credibility and replicability in qualitative research. Credibility is concerned with the intellectual rigour with which the study is undertaken, the qualifications and experience of the researcher, and the adherence to key concepts in qualitative research like naturalistic inquiry. Replicability is concerned with the extent to which a study is adequately documented and recorded, so that decisions and judgements made by the researcher are made explicit and the study could be reproduced by another researcher using the report of the study (Patton 1990, p460). For this interview study, the information given here is intended to provide the evidence of its credibility and replicability that a reader might require.

The notes from all the interviews were analysed manually. First, each interview was numbered and the notes of each interview were read through and coded to identify sections of the notes which related to particular themes, concepts or issues. These themes, concepts or issues were in part derived from the interview schedule, but were supplemented by some themes which emerged from the interview notes themselves. Then notes relating to specific themes, concepts or issues were grouped together, to facilitate the description of interviewees' attitudes and beliefs on each theme and the development of a coherent or common view where that was possible. In the analysis which follows, these themes and issues are reported and discussed. All quotations from interview notes are followed by the number of that interview.

4.3.3 **Results and discussion**

General advantages and disadvantages of adverse-event measures

In all, the interviewees mentioned seven different advantages and seven different disadvantages of adverse-event measures, which are summarised in table 4.9 below. The most frequently mentioned advantages of adverse-event measures of quality were their systematic nature and their focus on taking action and preventing future adverse events (each of which was cited twice), while the most frequently mentioned disadvantage was the potential for bias in adverse-event measures (cited by four interviewees). These advantages and disadvantages are discussed in more detail below.

Advantages	Disadvantages
Action/problem focus	Potential for bias
Systematic nature	Difficulty of attribution/causation
Objective nature	Timing of measurement
Validity	Indirect measure of quality
Efficiency	Clinician motivation problems
Provides aggregate data	Poor quality documentation
Done by additional staff	Needs additional staff

Table 4.9 Advantages and disadvantages of adverse-event measures of quality cited by interviewees.

The advantages of adverse-event measures of quality which were cited by interviewees are described in their own words in table 4.10.

Systematism	"..advantages of occurrence screening are mainly its systematisation of events audit - which is a long-standing part of audit in most specialties. ...[it] would satisfy a need for a systematic approach to audit." [4]
	"Huge advantage of occurrence screening is that it picks up events systematically.." [6]
Objectivity	"..it is clear cut, with readily recognisable sources of information.." [1]
Validity	"..if criteria are good, the information is directly related to quality of care." [2]
Efficiency	"It can obtain a much better pickup rate of adverse events at reduced cost." [5]
Action focus	"..it is easy to identify the action that needs to be taken from the information gathered." [1]
	"..it picks up events systematically and can be used to stop them happening again - prevention." [6]
Aggregate data	"..you build up a picture of near misses - minor incidents that no-one knows how frequently they occur. Particularly in [specialty] which is a sharp-end specialty where lots happens in a short space of time need a way to monitor minor incidents." [3]
Staffing	"One advantage of occurrence screening is that data is collected by an audit assistant - data needs to be collected by audit assistants as clinicians haven't got enough time." [2]

Table 4.10. Advantages of adverse-event measures of quality cited by interviewees

Several of the advantages cited by interviewees were clustered around what might be termed the technical worth of adverse-event measures of quality - aspects such as their systematism, objectivity, validity, and efficiency. Other perceived advantages included the action or problem focus of adverse-event measures (in other words, the fact that having detected an adverse event, it provides a natural focus for action to prevent future such events and so to improve the quality of care).

The aggregate information about rates of adverse events which adverse-event measures provide was seen as a particular advantage by one interviewee, since that kind of information was not available elsewhere. When adverse-event measures of quality are employed, the data is generally gathered by specifically designated staff rather than by clinicians themselves. This means that little additional workload is placed on clinical staff - which one interviewee saw as another advantage.

The disadvantages of adverse-event measures of quality which interviewees described are detailed in table 4.11.

Potential bias	"..the focus on adverse aspects of care, and only on those criteria selected by clinicians (so if not selected, then not picked up - perhaps bias in the criteria?)." [2] "Bias towards the things on the list. Also a danger that list is subjective and reflects value judgements on what is or is not adverse or appropriate." [4] "..Occurrence screening is not going to audit everything - in particular, just because an event is not listed in the criteria does not mean it doesn't matter. There would be some danger in relying on occurrence screening as the only method of audit." [5] "..audit is about more than just adverse events (just because no adverse events does not mean perfect care)." [6]
Indirect measure	"..these things are an indirect measure - a proxy - for real quality of care (need a test - serum quality!)." [3]
Timing	"Problems are picked up after they have happened (horse has bolted) though it depends on at what position/time screening takes place." [6]
Causation	"..doing it without attributing blame could be difficult." [1]
Poor documentation	"...these incidents are not always documented (often not), so a self-reporting system would be needed. Problems with reliability and honest reporting would be a problem especially if not confidential." [3]
Motivation	"Occurrence screening is not enough for audit by itself - it would get boring. Have had some contact with [clinicians] in [district where occurrence screening being used] and they don't like it! With a centralised system there is the danger that you lose the sense of clinical ownership." [2]
Staffing	"..it needs an external agent to collect data (eg the nurse screener or whoever) which would not be available in [specialty]. Could use [existing member of staff] perhaps to perform this function for a limited number of cases." [1]

Table 4.11. Disadvantages of adverse-event measures of quality cited by interviewees

The most frequently mentioned single disadvantage cited by interviewees was the potential for bias in adverse-event measures of quality. Several interviewees suggested that because a measure might include certain adverse events but not include others, it might solely represent clinicians' judgements of what is or is not adverse. The omission of important adverse events from the measure, and the resultant lack of attention to instances of those events, was seen as a problem.

It was interesting to note, however, that the negative focus of adverse-event measures (their attention to what are, essentially, instances of poor quality) was not mentioned as a disadvantage by interviewees (except, in passing, by interviewee 2 above). Although this perceived disadvantage is given prominence in the literature and was raised by a number of respondents to the questionnaire study, it did not seem to be as important to this group of interviewees. One interviewee suggested that adverse-event measures were essentially indirect rather than direct measures of the quality of healthcare.

While the action or problem focus of adverse-event measures had been cited by some interviewees as an advantage, one interviewee felt that the timing of data collection meant that adverse events would be noted after the event rather than prevented, and another believed that deciding on the causation of adverse events without seeming to be allocating blame would be difficult. Another interviewee suggested that the poor quality of medical record documentation made the application of adverse-event measures practically difficult. Interestingly, only one interviewee suggested that clinicians might not want to use adverse event information to review the quality of care they deliver (or might at least want other sorts of information as well).

Although one interviewee (cited above) had seen the fact that adverse-event measures are usually applied by specially designated staff rather than by clinicians themselves as an advantage, another perceived it as a disadvantage, because such staff would not be available.

Similarities to and differences from traditional quality assurance methods

It has been commonplace for medical staff to review individual cases or groups of cases in mortality and morbidity meetings or death and complication reviews. The cases chosen for such review have often been selected because they involve some sort of potentially avoidable adverse event. Case presentations have generally been followed by some form of discussion, in which senior clinicians provide informal guidance on the future management of similar circumstances.

All interviewees perceived strong parallels between these traditional medical approaches to quality assurance and adverse-event measures of quality, and regarded adverse-event measures of quality as a kind of continuation of more informal current practices. For example:

> *"In [specialty] we already have anecdotal information on adverse events - which gets discussed at [specialty] meetings, coffee, etc." [1]*
> *"Occurrence screening is like traditional audit in that it is notes-based and some of the criteria pick up common traditional audit issues..." [2]*
> *"[Specialty] have been using adverse events for years in audit.." [3]*
> *"Strong similarities with traditional case audits in [specialty]." [4]*
> *"Occurrence screening is very similar to some traditional audit methods - M&M meetings etc." [5]*
> *"There are obvious parallels with traditional M&M meetings..." [6]*

Interviewees described five main differences between adverse-event measures and the traditional methods of quality assurance - the degree of systematism and formality in the method, the objectivity of the method, the range of events included, the inclusion of an outside view, and the potential to delegate data collection. In each case the difference seemed to be regarded as a positive one by the interviewees (in other words, the adverse-event measures were seen as better than traditional methods because of it). Interestingly, none of the interviewees raised differences between adverse-event measures and traditional audit methods in which the latter were seen as superior. These differences are described in more detail in table 4.12

Systematism	"The main difference with occurrence screening would be that the data collected would be more formal and more systematic." [1] "Main difference .. is the systematic nature of occurrence screening." [4] "The difference is that occurrence screening is more systematic and sophisticated.." [6]
Objectivity	"It is different from traditional approaches to audit in that there are much clearer definitions of things - they are less subjective." [5]
Range	"The range of definitions of occurrence screening is much wider, encompassing far more than the strictly clinical matters." [5]
Sensitivity	"It is important to know about near misses, which occurrence screening does and other approaches do not." [5]
Outside view	"M&M meetings take place within the profession - occurrence screening brings in a healthy outside view." [6]
Clerical support	"You can define occurrences to a degree that delegating initial screening to clerical support becomes feasible." [5]

Table 4.12. Differences between adverse-event measures and traditional methods of quality assurance identified by interviewees

The links between adverse-event measures of quality and some traditional quality assurance methods seemed to be used by interviewees to conceptualise the former's use and application, and the similarities appeared to help in gaining their acceptance of adverse-event measures. While the interviewees did see some important differences between adverse-event measures and traditional quality assurance methods, those differences were advantages rather than disadvantages.

Utility in measuring quality for individual patients or groups of patients

All interviewees were asked how the information from adverse-event measures of quality might be used to assess the quality of care for individual patients and for groups of patients. Adverse event information on both individual patients and groups of patients was seen as valuable and helpful,

though most interviewees believed that aggregate information about the rates of adverse events in groups of patients would be more useful because it would be more reliable:

> *"Feel occurrence information is useful in aggregate to pinpoint problem areas. Useful in individual cases for case discussion and for prevention. Both sorts of data are needed." [2]*
> *"Rates of occurrences would be main useful indicator rather than individual events - eg how often an incident occurred (5 reports and start to do investigation)." [3]*
> *"Data from individual events might be misleading and might be prematurely interpreted. Group data - rates of adverse events - would be much more useful in comparing practice and identifying problems." [4]*
> *"Aggregate data would be useful for saying how often x happened, unless there is a specific single problem would tend to look at how often things happened - whether acceptable. Need aggregate data to argue for more resources." [5]*

However, one interviewee raised some of the difficulties involved in interpreting rates of adverse events, and pointed out that the review of individual cases and groups of cases had some advantages:

> *"Most useful part of occurrence screening would be the review of individual events or groups of events. Rates and so on would be less useful. There is a risk that patients have unrealistic expectations of care and some adverse events are unavoidable - so rates less meaningful than study of individual cases. Problems is that some events would happen very rarely - low incidence - and the nature of the events affects their importance clinically." [6]*

Utility in creating or promoting changes in practice and the quality of care

Interviewees had a variety of ideas of how they might use information about individual adverse events or about groups or rates of adverse events to create and promote changes in practice or the quality of care. Some believed that the simple availability of information would have a beneficial influence on individual practitioners' patterns of practice and behaviour, and this would bring about improvements in the quality of care:

> *"I believe that this information would create a heightened self-awareness which would reduce event rates (self-knowledge makes people change)." [3]*
> *"Suggest that constant repetition of information - drip drip effect - would be effective." [4]*

Other interviewees saw the process of changing individuals' behaviour as more complex and challenging than this, and suggested that to change practice required more than simply the provision of information:

> *"For occurrence screening to bring about change, very important to bring people who need to change into the setting of criteria initially - get people to define adverse occurrences in advance, and set target percentages for each occurrence. Important that the system is not just used to criticise but also to praise." [5]*

Some saw adverse-event information as a starting point for investigations which would focus on changing the way processes or systems were organised and managed. From their viewpoint, the information would be used to change systems or processes rather than to alter any individual's behaviour:

> *"I recently found [an adverse event] which when investigated turned out to result from a string of organisational problems ... we have now changed the way things are organised. The point is that it was not a problem of individual behaviour, it was an organisational problem." [1]*

Suitability for different areas or specialties in inpatient care

When asked about the appropriateness of adverse-event measures for particular specialties or areas, interviewees were divided on whether they were equally appropriate in all areas of inpatient care. Some interviewees felt they might be developed and applied in any specialty:

> *"Occurrence screening is well suited to most specialties..." [3]*
> *"I think occurrence screening could work across the board in most specialties ..." [5]*
> *"I think that occurrence screening could be applied equally to every specialty - there are adverse events in all specialties worthy of screening." [6]*

In contrast, other interviewees argued that adverse-event measures were better suited to the more acute inpatient areas:

"Occurrence screening is best used in monotonal disease areas, surgery (especially for the under 75's), for younger patients in general, and is good for clinical audit as well. It is not well suited to ... specialties where multiple problems/pathologies are common, because it is harder to separate disease process and effects of care." [2]
"Occurrence screening is better suited to the acute areas ... less well suited to the long term chronic care work." [4]
"... it would be less suitable for long stay patients because of the lower numbers and the difficulties of setting standards." [5]

Interestingly, all interviewees said that they would (or did) use adverse-event measures in their own specialties, although they represented a full spread of specialties including both surgical and medical and acute and non-acute areas:

"Can see ways of using occurrence screening in [specialty]. Think it is a good idea, worth trying out. Suggest some possible criteria for use in [specialty] [list of suggestions followed]." [1]
"I would use occurrence screening in audit, but only in combination with other audit tools. Suggest one third occurrence screening and two thirds outcomes work (in [specialty] at least). Occurrence screening would be useful for some big issues we are not currently doing enough about which prolong hospitalisation - such as pressure sores, constipation." [2]
"I would certainly use occurrence screening in own specialty/hospital ... Currently am working out a pilot system based on a standard checklist ... Occurrence screening is very like the things we have already been doing - in favour of it." [3]
"I would like to use occurrence screening as one of the main planks of audit in my own department, though I would want to retain the element of 'unstructured' free discussion which currently exists." [4]
"We are currently doing it! Clerk pulling notes of patients in outpatients clinic to check for failures to record height and weight, non-availability of results. We would use (do use) occurrence screening as part of our audit, included with other techniques." [5]
"At present we only use adverse events in audit as a 'trigger' for further investigations or studies - not routinely. I believe we should use occurrence screening more - gathering data and feeding it back through our medical audit assistant. Proper feedback when events happen is essential, to get people to start to work out strategies for avoiding/preventing adverse occurrences. Occurrence screening certainly has a defined place in medical audit, though it is not all that is needed - it might make up only 20% of audit activity/studies." [6]

While the interviewees were all clinicians with some interest and involvement in clinical audit, it was interesting to note that they showed considerable enthusiasm for using adverse-event measures

in their own audit or quality assurance activities, in combination with other approaches to measurement and monitoring.

Factors which might facilitate or obstruct application of measures

Interviewees identified a number of factors which might facilitate or assist the introduction of adverse-event monitoring in a department or specialty, including the atmosphere and relationships, the leadership provided by senior clinicians, the extent of participation in the process, and the degree to which it could be made interesting to those involved. These factors, and others which might obstruct such an introduction, are set out in table 4.13 below.

Atmosphere:	"To set it up needs a good rapport in the department - and a level of interest in knowing about every adverse occurrence. " [6]
Leadership	"Need to convince the consultants to participate then juniors would follow." [3] "In getting occurrence screening to work the key thing would be to motivate and secure conviction of consultants to do it." [4]
Participation	"To get occurrence screening to work well it would be important to involve people in setting standards/criteria themselves, to review the standards/criteria regularly, and to ensure it is not just criticism." [5]
Interest	"Occurrence screening would be best accepted if it was [part of] a rotating audit - covering different topics each month rather than looking at the same things again and again." [2]
Lack of interest	"Problems in occurrence screening include getting people excited (essential part of audit) and establishing ownership." [2] "Problems might result from high occurrence rates (get used to it) or low occurrence rates (not enough to sustain interest)." [6]
Punitive use	"Very important that no aspect of blame/punitive nature/disciplinary factors/etc attached to occurrence screening or events wouldn't be reported." [3]
Restrictiveness	"This approach might be too 'disciplined' for some - though it might usefully encourage focus on smaller areas, better defined." [4]

Table 4.13. Factors which might facilitate or obstruct the application of adverse-event measures of quality identified by interviewees.

Some of the factors identified by interviewees which might obstruct the introduction of adverse-event monitoring were essentially the reverse of those factors identified above. For example, the difficulties of sustaining interest were seen as an obstacle by some. However, two new factors which might impede adverse-event monitoring were identified - the risk that the process might be used punitively to assign blame and even to take disciplinary action against clinicians, and the possibility that the structured nature of adverse-event monitoring might seem restrictive to clinicians who are used to a more unstructured and less methodical approach:

4.3.4 **Conclusions**

The interview study set out to identify any themes or issues relating to the validity of adverse-event measures in general which might have been overlooked by the questionnaire study. The structure of the interviews gave interviewees the opportunity to raise almost any aspect of adverse-event measures and their application in healthcare, including (but not restricted to) aspects of their validity.

Interviewees generally regarded adverse-event measures as valid indicators of the quality of care, and cited a number of merits of such measures which directly impinged on their validity and their wider utility. However, they also expressed concerns about the potential for bias in adverse-event measurement, especially if adverse-event measures were the only indicators of quality to be used. The validity issues raised by interviewees (both positive and negative) closely matched those raised by respondents to the questionnaire study, suggesting that the two studies support and confirm each others' findings. Perhaps the clearest signal that interviewees regarded adverse-event measures as valid and useful indicators of the quality of care was that they all, without exception, indicated that they would wish to use, planned to use or were using these measures themselves in their own clinical practice.

The interview study afforded the opportunity to explore other aspects of adverse-event measures, beyond their validity. It was clear that interviewees regarded these measures as having much in common with traditional medical approaches to quality assurance, but saw them as superior to the latter for a number of reasons. Interviewees felt that the aggregate data which results from adverse-event measurement, giving rates and trends in adverse events, would be particularly useful in identifying quality problems and comparing practice. Interviewees had differing ideas of how they might use adverse-event data to influence and change clinical practice, and while some thought that the simple availability of data would bring about change others believed that a more proactive and interventionist application of the data would be needed. Interviewees saw adverse-event measurement as relevant to the whole range of inpatient care, though most interviewees perceived it

as being easier to apply in areas of medical care where the effects of the disease process and of the healthcare interventions can be separated. A number of factors important to the introduction of adverse-event measurement were identified, such as the support of clinical leaders, the existence of good working relationships, the participation of clinical staff in designing or applying the measure, and the stimulation of clinicians' interest in the process.

All the clinicians interviewed for this study were already involved in clinical audit or quality assurance, and so might be thought to be predisposed towards the general idea of healthcare quality measurement, but there was no predisposition towards this particular form of quality measurement, using adverse events. It was, therefore, significant that all interviewees wanted to use adverse-event measures in the quality assessment of their own clinical practice, often in combination with other forms of quality measures.

Chapter 5

Construct validity of adverse-event measures of quality

5.1 Introduction

The concepts of construct validity were discussed in chapter 2, where it was noted that the examination of construct validity can be of particular benefit in situations where the criterion-related validity of a measure is difficult to test because a suitable criterion variable is not available. It has already been demonstrated in chapter 3 that studies of the criterion-related validity of adverse-event measures have struggled to identify suitable criterion variables, with which those measures should be expected to correlate if they are valid. Most studies have relied upon some form of implicit, professional assessment of quality as their gold standard, although there is evidence that such assessments are themselves low in validity and reliability. A few studies have used other quality measures (such as standards-based quality assessment tools for given patient groups) as their criterion variable, but it can be argued in these cases that the criterion measure and the measure under test actually measure different dimensions of the quality of care, and that a correlation between them would not necessarily be expected.

In these circumstances, an examination of the construct validity of an adverse-event measure is both methodologically more correct and appropriate and more likely to provide an insight into the behaviour of the measure being studied than an attempt to measure criterion-related validity. The aim of the analysis of adverse-event measure data from the RSCH project was to examine whether or not the data supported or was consistent with a number of constructs or theories about adverse events and the quality of care in hospitals. If the data supported these constructs, it would provide some further evidence that the measures themselves were valid.

Three main constructs have been examined by previous research studies - namely that patients with adverse events would stay longer in hospital and use more resources; that patients who are more severely ill would be more likely to have adverse events; and that the incidence of adverse events would vary from hospital to hospital. Each of these constructs has been explored by more than one research study. However, there are many other potential constructs which could be used to assess the construct validity of adverse-event measures, so there is a need for research to confirm construct validity for those constructs which have already been tested and to extend the evidence for construct validity by identifying and testing new constructs.

This chapter presents a study of the construct validity of the adverse-event measures of quality used in the RSCH occurrence screening project. Firstly, a number of constructs or theories to be tested are defined. Then, a number of univariate, bivariate and multivariate statistical techniques are used to test whether the data collected during the RSCH project supports these constructs.

5.2 Method

5.2.1 Constructs to be tested

The first step in testing the construct validity of a measure is to identify the constructs or hypotheses about the measure's behaviour which are to be tested. In the case of adverse-event measures, there is little existing direct empirical evidence relating to their behaviour, and few established theories about them, though a small number of constructs have been tested in past research studies. For this study, the following constructs were chosen, and the rationale for the selection of each construct is stated below:

a) *Rates of different types of adverse event, and aggregate adverse-event rates for a generic adverse-event measure, will vary across specialties.*

Patients in different specialties have different diseases and illnesses, and undergo different diagnostic and therapeutic processes. While they may have certain types of adverse event in common, it is likely that the incidence of adverse events will vary from specialty to

specialty, as will the number of adverse events per patient. For example, hospital acquired infections might be expected to be more common in general surgery than in ophthalmology, because of the nature of the surgical procedures each specialty involves, and so the rates of this adverse event in these specialties would be expected to differ.

b) *Patients who have adverse events will stay longer in hospital than those with no adverse events, and patients who have multiple adverse events will stay longer than those with a single adverse event.*

When an adverse event results in some additional morbidity for patients and consequently requires some additional diagnostic and therapeutic actions over and above the needs of the patient on admission, it is likely to result in an increased length of stay. Alternatively, patients with longer lengths of stay may be exposed to the risk of certain adverse events for longer, and may consequently be more likely to have adverse events. For both these reasons, patients with adverse events are likely to stay longer in hospital and those with multiple adverse events are likely to stay longer still. For example, a patient who suffers an adverse event like a slip or fall while in hospital may require further diagnostic tests and examinations to check for injuries, and those injuries may require treatment in their own right, hence prolonging the patient's admission.

c) *Different types of adverse event will result in different degrees of prolongation of the hospital length of stay.*

It was argued above that patients with adverse events are likely to stay longer in hospital because of the diagnostic and therapeutic consequences of the adverse event. Different types of adverse event will have quite different consequences, and are so likely to have different effects on patients' length of stay. For example, adverse events like medical record deficiencies would be unlikely to result in a longer length of stay, while an adverse event such as a postoperative pulmonary embolism would almost certainly extend the length of stay. If this construct were confirmed, it would support the contention that adverse events cause prolonged hospital stays rather than the converse (that prolonged hospital stays result in a higher level of adverse events through greater exposure to risk).

d) *Patients who die in hospital will have had more adverse events during their stay than those who are discharged alive.*

Patients who die in hospital will commonly have been more severely ill, and their care will often have been more complex to manage. Certain adverse events will be more common in patients whose illnesses are severe and complex, and so patients who die in hospital are likely to have had more adverse events than those who do not. Alternatively, some adverse events may be so serious that they contribute to or cause the patient's mortality, with the same result that patients who die in hospital are likely to have had more adverse events than those who do not.

e) *Patients who are admitted as emergencies will have more adverse events than those admitted electively.*

Emergency admissions are commonly more complex to manage than elective admissions, because they present less predictable health problems which require faster diagnosis and treatment and make immediate and unscheduled demands on healthcare services. As was argued above, certain adverse events will be more common in patients whose management is more complex, and so emergency admissions are likely to have more adverse events than elective admissions.

f) *Elderly patients will have more adverse events than younger patients.*

Like patients who die in hospital and patients who are admitted as emergencies, patients who are elderly are commonly more complex to manage than younger patients because they have more comorbidities and are less able to tolerate the effects of treatment. As argued above, certain adverse events will therefore be more common among elderly patients.

It should be noted that some of these constructs have already been explored in some of the research studies reviewed in chapter 3. Specifically, a relationship has already been identified between adverse-event measures and length of stay, and between adverse-event measures and severity of illness (a concept related to the constructs listed under d-f above). However the other constructs listed have not been examined previously in the published literature.

5.2.2 **Sources and nature of data from the RSCH project**

The data analysed and presented in this chapter was collected for the RSCH occurrence screening project, which was described in chapter 3. In that project, 14,815 inpatient admissions from 12 different specialties were screened using a number of different adverse-event measures over the period from February 1990 to April 1992. Each adverse-event measure consisted of a number of screening criteria, as shown by the example in Appendix 4.1.

For each inpatient admission, a set of demographic and administrative data was collected from the hospital's Patient Administration System (PAS). Soon after the patient was discharged, their medical and nursing casenotes were screened by one of the project staff. Using an adverse-event measure, the screener checked the notes to see whether the adverse events contained within the measure had occurred during the admission. For each screening criterion within the adverse-event measure, the screener recorded whether or not that type of adverse event had occurred. If it had occurred, the screener went on to record a further data set describing the adverse event (the contents of that data set are not used or detailed here). Table 5.1 lists the data set collected for each patient, including some data items which were derived from other data within the set. The data set outlined in table 5.1 forms the basis for the analyses reported in this chapter.

Data item	Description	Source
Unit number	Unique patient identifier code	PAS
Surname	Patient's surname	PAS
Forename	Patient's forename	PAS
Sex	Patient's sex, coded as 1 (male) or 2 (female)	PAS
Date of birth	Patient's date of birth	PAS
Date of admission	Date on which patient was admitted to hospital	PAS
Date of discharge	Date on which patient was discharged from hospital	PAS
Admission method	How patient was admitted (range of codes for admissions from waiting list, A&E, via GP etc)	PAS
Admission category	Whether patient was an NHS or private patient	PAS
Discharge method	How patient was discharged (range of codes for died, discharged with/against medical advice, etc)	PAS
Discharge destination	Where patient was discharged to (home, another hospital, nursing home, etc)	PAS
Consultant	Code for consultant responsible for admission	PAS
Specialty	Code for specialty of consultant admitting patient	PAS
Ward(s)	Code(s) for ward(s) on which patient stayed	PAS
Screener	Code for member of project staff who screened the admission	Screening
Screening data	A series of dichotomous variables, one for each screening criterion in the adverse-event measure being used, for each of which 0 represented no adverse event and 1 represented an adverse event. Different adverse-event measures were used with different specialties, and so the number and meaning of these dichotomous variables varied from admission to admission.	Screening
Age on admission	Age in years on date of admission	Derived
Admission type	Type of admission, categorised simply as emergency or elective	Derived
Discharge type	Type of discharge, categorised simply as alive or dead	Derived
Length of stay	Length of stay in days	Derived
Number of screenings	Number of screening criteria against which the admission was screened	Derived
Number of positive screenings	Number of positive screenings (eg number of criteria for which an adverse event was found)	Derived
Number of negative screenings	Number of negative screenings (eg number of criteria for which no adverse event found)	Derived
Screening percentage score	Number of positive screenings expressed as a percentage of number of screenings	Derived
Age range	Age on admission, categorised into three groups (0-50 years, >50 to 70 years, >70 years)	Derived
Length of stay range	Length of stay, categorised into four groups (0 days, 1-5 days, 6-10 days, 11+ days)	Derived

Table 5.1. Data set collected for each inpatient admission in the RSCH project.

In the analyses of construct validity presented in this chapter, data from the RSCH project on all admissions in eight specialties was used. The data for these eight specialties represented a total of 12,676 admissions (or 86% of the 14,815 admissions screened during the project and held on the project database). The remaining 2,139 admissions were excluded from the analysis because they involved a range of other specialties or groups for which there were insufficient admissions to support analysis at the specialty level.

The resulting data set is described in table 5.2. It can be seen that the total of 12,676 screened admissions was not spread evenly across the eight specialties studied. The data set contained more admissions for some specialties, notably obstetrics, ophthalmology and trauma and orthopaedics, than for others, largely because screening in those specialties took place over a longer time span during the project or because these specialties admitted higher volumes of patients. In six of the eight specialties being studied, the generic adverse-event measure described in chapter 4 and appended in Appendix 4.1 was used. These specialties therefore formed an important subgroup for analysis, consisting of a total of 7,633 admissions from six specialties all screened with the same adverse-event measure. In the other two specialties (obstetrics and accident and emergency) adverse event measures developed specifically for those specialties had been used.

Specialty	No of admissions screened	Adverse-event measure used in screening
Accident and emergency	1,031	Special measure designed for A&E short stay admissions
ENT	800	Generic adverse-event measure plus specialty addition.
General surgery	549	Generic adverse-event measure plus specialty addition
Gynaecology	566	Generic adverse-event measure plus specialty addition
Obstetrics	4,012	Special measure designed for obstetrics
Ophthalmology	2,252	Generic adverse-event measure plus specialty addition
Trauma and orthopaedics	2,945	Generic adverse-event measure plus specialty addition
Urology	521	Generic adverse-event measure plus specialty addition

Table 5.2. Data set from RSCH project used in analyses of construct validity.

5.2.3 **Statistical techniques used**

It is evident from tables 5.1 and 5.2 that the RSCH project provided a rich but complex data set for analysis. The data contains a large number of variables, including categorical, ordinal, interval and ratio data items. None of the data items could be expected to be parametrically distributed and some categorical data items would be expected to be very unevenly distributed, with low relative frequencies for some categories.

In order to test the constructs described in section 5.1 using the RSCH project data a combination of univariate, bivariate and multivariate statistical techniques was used. Firstly, each construct was tested using appropriate univariate and bivariate techniques, including both parametric and non-parametric methods. Secondly, a multiway frequency analysis using a loglinear model was performed, to examine the relationships between a number of key variables in the RSCH data set. A summary of the univariate and bivariate analytical techniques and data sets used to test each construct is given in table 5.3 below.

	Construct	Data set used	Analytical techniques
a)	Adverse event rates will vary between specialties.	Sample of 7,633 admissions in 6 specialties all screened using the same generic adverse-event measure.	χ^2 tests used for each criterion within the measure, to establish whether significant differences in rates existed between specialties. Overall adverse event rates compared by Kruskal-Wallis analysis of variance by ranks and one way analysis of variance (ANOVA).
b)	Adverse event rates will correlate with length of stay.	Whole sample of 12,676 admissions in 8 specialties	For each specialty, t-test and Mann-Whitney test used to compare length of stay for patients with and without adverse events; also Spearman rank correlation coefficient used to measure association between adverse event rate and length of stay.
c)	Different types of adverse events will have different effect on length of stay.	Sample of 4,012 admissions in obstetrics all screened using an obstetric adverse-event measure.	For each criterion within the obstetric adverse-event measure, t-test and Mann-Whitney test used to compare length of stay for patients with and without that adverse event.
d)	Patients who die will have had more adverse events than those who do not.	Whole sample of 12,676 admissions in 8 specialties	For each specialty, t-test and Mann-Whitney test used to compare the numbers of adverse events among patients who died and those who were discharged alive.
e)	Emergency patients will have more adverse events than elective patients.	Whole sample of 12,676 admissions in 8 specialties	For each specialty, t-test and Mann-Whitney test used to compare the numbers of adverse events among patients who were admitted electively and patients who were admitted as emergencies.
f)	Elderly patients will have more adverse events than younger patients.	Whole sample of 12,676 admissions in 8 specialties	For each specialty, age on admission was compared for patients with 0, 1 and 2 or more adverse events using Kruskal-Wallis analysis of variance by ranks and one way analysis of variance (ANOVA).

Table 5.3. Summary of analytical techniques and data sets used to test construct validity.

The statistical techniques used varied from construct to construct, so in the presentation of results in section 5.3.1 below the testing of each construct is reported separately, along with those details of the statistical methods used which are specific to that construct. The results of the multiway frequency analysis are reported separately and subsequently in section 5.3.2, to allow the discussion to compare and contrast those results with the findings of earlier univariate and bivariate testing for individual constructs.

5.3 Results and discussion

5.3.1 Univariate and bivariate analyses of construct validity

Analysis of rates of adverse events across specialties

In section 5.1 it was proposed that the rates of different types of adverse event and aggregate adverse-event rates for a generic adverse-event measure would be found to vary across specialties. The rationale for this construct was that patients in different specialties have different diseases and illnesses, and undergo different diagnostic and therapeutic processes. While they may have certain types of adverse event in common, it is likely that the incidence of adverse events will vary from specialty to specialty, as will the number of adverse events per patient.

To test this construct, the sample of 7,633 admissions in 6 specialties, which were screened using the same generic adverse-event measure, was examined. For each of the 20 screening criteria within that measure (which can be found in Appendix 4.1), χ^2 tests were used to establish whether significant differences in incidence rates existed between different specialties. Because some adverse events were quite rare, in many of the χ^2 tests expected frequencies for some cells fell below 5, which is sometimes cited in statistics texts as a threshold below which the χ^2 should not be used because assumptions made in approximating the statistic to the χ^2 distribution are violated. However Everitt (1992, p39) and others argue that this restriction is arbitrary, and not based in mathematical or empirical evidence. They suggest that 2 x c tables (such as these) can always be tested if no expected frequency falls below 1, and can usually be used if no expected value is below 0.5. In none of the χ^2 tables did any expected values fall below 0.5, and only in one table (for criterion 16) were any expected values below 1.0.

Where significant differences in incidence rates of adverse events across specialties were found, the standardised adjusted residuals were used to identify the specialty groups which were significantly different from the others. These statistics are approximately normally distributed with a mean of 0 and standard deviation of 1.0, so standardised adjusted residuals of ± 2.58 or greater are significant at the 0.01 level (Everitt 1992, p47).

The results of testing are presented in table 5.4 below. Each row in the table contains a summary of the crosstabulation (and associated statistics) of specialty by screening result for a single criterion within the generic adverse-event measure. For each specialty, the incidence rate of that type of adverse event is given (shown as a percentage of patients who had that adverse event during their hospital stay), along with the standardised adjusted residual. The incidence rate for all specialties is also given, for the purposes of comparison. The last two columns in the table contain the χ^2 statistic for that crosstabulation and its significance level.

Crit no	Criterion title	Specialty adverse event incidence rates (%) and standardised adjusted residuals (in square brackets)							χ^2 statistics	
		ENT	Gynae-cology	Ophth-almlgy	T & O	Gen surgery	Urology	All specs	χ^2	P value
1	Adm for adv results o/p mgt	1.75 [1.1]	3.56 [4.8]	2.70 [-5.2]	1.57 [1.5]	1.28 [-0.1]	1.54 [0.4]	1.33	43.29	< .0001
2	Readmission for comp prev adm	4.25 [0.7]	13.88 [12.9]	2.49 [-3.9]	2.63 [-4.3]	4.93 [1.4]	3.65 [-0.2]	3.82	179.12	< .0001
3	Error in operative consent	2.63 [3.5]	0.71 [-1.3]	1.11 [-0.9]	1.06 [-1.5]	1.82 [1.1]	1.54 [0.5]	1.30	15.81	.0074
4	Unpl rem/inj/repair in surg	1.50 [-1.4]	1.25 [-1.6]	4.53 [9.1]	0.92 [-6.0]	1.28 [-1.5]	2.11 [-.01]	2.18	86.23	< .0001
5	Unpl return to theatre	0.13 [-2.4]	0.36 [-1.3]	1.02 [1.0]	0.61 [-1.8]	2.73 [5.0]	0.15 [0.8]	0.85	32.83	< .0001
6	Path/hist varies from diag	0.13 [-1.1]	0.71 [1.6]	0.00 [-3.3]	0.34 [0.0]	1.82 [6.2]	0.19 [-0.6]	0.34	46.84	< .0001
7	Prob of transfusion	0.13 [-1.6]	0.18 [-1.2]	0.00 [-4.1]	1.02 [4.9]	0.55 [0.1]	0.77 [0.8]	0.51	30.90	< .0001
8	Hosp acquired infection	0.25 [-4.1]	0.53 [-2.9]	0.49 [-6.9]	4.47 [9.8]	4.20 [3.0]	1.34 [-1.5]	2.32	126.55	< .0001
9	Medication error/reaction	2.13 [-1.8]	1.25 [-2.7]	1.60 [-5.1]	5.05 [7.4]	4.93 [2.4]	1.34 [-2.5]	3.18	72.39	< .0001
10	Cardiac/resp arrest in hosp	0.00 [-1.3]	0.00 [-1.1]	0.09 [-1.3]	0.34 [2.5]	0.00 [-1.0]	0.38 [1.1]	0.18	9.72	.0837
11	CVA/MI/PE in hosp after surg	0.00 [-1.5]	0.00 [-1.2]	0.04 [-2.3]	0.51 [3.6]	0.36 [0.6]	0.19 [-0.3]	0.25	15.68	.0078
12	Unexp transfer to spec care	0.13 [-0.9]	0.18 [-0.5]	0.09 [-2.0]	0.14 [-1.8]	2.01 [8.0]	0.38 [-0.5]	0.28	65.72	< .0001
13	Pt related clinical complcn	1.00 [-3.6]	0.00 [-4.4]	0.22 [-9.4]	6.25 [12.6]	6.20 [4.4]	0.96 [-2.9]	3.09	214.69	< .0001
14	Non-clin problem/incident	6.38 [-4.0]	6.41 [-3.3]	4.18 [-11.6]	17.47 [15.9]	8.94 [-1.2]	10.17 [-.02]	10.44	274.53	< .0001
15	Neuro deficit devel in hosp	0.00 [-1.2]	0.00 [-1.0]	0.00 [-2.2]	0.38 [3.8]	0.00 [-1.0]	0.19 [0.2]	0.16	15.44	.0086
16	Unexp patient death	0.00 [-1.0]	0.00 [-0.8]	0.00 [-1.9]	0.27 [3.1]	0.18 [-.05]	0.00 [-.08]	0.12	11.03	.0508
17	Medical record deficiency	18.50 [3.3]	4.98 [-6.7]	11.90 [-4.3]	14.44 [-.04]	34.06 [13.4]	11.32 [-2.2]	14.62	235.70	< .0001
18	Nursing record deficiency	43.13 [14.3]	26.87 [2.2]	5.02 [-24.2]	33.82 [17.7]	17.67 [-3.1]	10.94 [-6.8]	23.04	842.96	< .0001
19	Pt/family dissatisfaction	1.38 [-1.4]	2.31 [0.5]	1.51 [-2.0]	2.59 [2.9]	2.01 [0.0]	1.54 [-0.8]	2.01	10.41	0.064
20	Discharge related problems	1.00 [-4.3]	1.07 [-3.5]	1.16 [-7.7]	6.65 [10.7]	6.20 [3.2]	2.69 [-1.3]	3.72	150.41	< .0001

Table 5.4. Variation in rates of adverse events across specialties

It is clear that the results in table 5.4 support the construct that adverse event rates vary across specialties. For 17 of the 20 criteria in the generic adverse event measure, significant differences existed in the incidence rates found in different specialties ($p < 0.01$). Though the purpose of this analysis is to test the construct, rather than to seek to explain the variations in rates across specialties which were observed, for many of the criteria within the generic adverse-event measure the information presented in the table conforms with the known characteristics of the specialties involved. For example, higher rates of discharge related problems (criterion 20) were observed in trauma and orthopaedics and general surgery than in ENT or ophthalmology, which probably reflects the nature of the patients and disease processes involved. The former specialties deal with elderly patients admitted more frequently as emergencies with major health problems which change their ability to cope at home either progressively or more immediately. It would therefore be expected that discharging these patients would be more complex and difficult, and that a higher frequency of discharge-related problems would be observed, as was indeed the case.

The construct under examination suggested not only that individual adverse event rates for single criteria within the adverse-event measure would vary across specialties, but also that the aggregate adverse-event rates for the generic adverse-event measure as a whole would vary across specialties. In order to test this, two analyses were undertaken on the same sample of 7,633 admissions in 6 specialties. The generic adverse-event rate was calculated for each admission, by counting the number of screening criteria within the generic adverse-event measure for which an adverse event had been recorded for that admission. The rate therefore had a minimum of 0 and a theoretical maximum of 20 (since there were 20 screening criteria in the generic adverse-event measure). For a small number of admissions, data was missing for one or more of the criteria in the generic adverse-event measure. These cases were excluded from the analysis and so the actual sample used consisted of 7,210 admissions

Firstly, a non-parametric Kruskal-Wallis analysis of variance by ranks was undertaken. This tests whether a number of samples (in this case, the admissions in each of 6 specialties) have the same median adverse-event rate (or, more exactly, are drawn from populations with identical median adverse-event rates) (Siegel and Castellan 1988, p206). Secondly, a one way analysis of variance

(ANOVA) was used to test whether the mean adverse event rates in the 6 specialties differed significantly from each other. Both tests were used because, while the data certainly violates the parametric assumptions underlying the one-way ANOVA test, the power of the Kruskal-Wallis analysis of variance by ranks is more limited.

The Kruskal-Wallis analysis of variance by ranks indicated that significant differences in adverse-event rates between specialties existed ($\chi^2 = 532.52$, $\chi^2 = 641.89$ when corrected for ties; in both cases $p < 0.0001$). Similarly, the one-way ANOVA test also indicated that the mean adverse-event rates in specialties were significantly different ($F = 139.38$, $p < 0.0001$). Some descriptive statistics illustrating the differences in adverse event rates between specialties are presented in table 5.5 below.

Specialty	Adverse event rate				
	No of admissions	Mean	Standard deviation	Minimum	Maximum
ENT	800	0.844	0.866	0	5
Gynaecology	562	0.642	0.745	0	5
Ophthalmology	2249	0.357	0.611	0	4
T & O	2930	1.004	1.160	0	8
General surgery	548	1.005	1.134	0	7
Urology	521	0.524	0.713	0	3

Table 5.5. Descriptive statistics illustrating differences in adverse-event rates between specialties.

Both the Kruskal-Wallis analysis of variance by ranks and the ANOVA test provide procedures for identifying the source of any statistically significant difference found, by examining the significance of differences between pairs of groups within the analysis. In the Kruskal-Wallis test, the absolute difference between the mean rankings of groups is normally distributed for large samples, and can be tested using the standard normal distribution as long as adjustments are made to take account of the multiple comparisons being made (Siegel and Castellan 1988, p213). For the ANOVA test, the Scheffé method allows a similar multiple pairwise comparison of means. These procedures were each used to test for significant differences (at the $p < 0.05$ level) between all combinations of pairs of groups in the sample, and the results are shown in table 5.6 below

	ENT	Gynae-cology	Ophth-almology	T & O	Gen surgery	Urology
ENT		§ ‡	§ ‡	§		§ ‡
Gynaecology			§ ‡	§ ‡	§ ‡	
Ophthalmology				§ ‡	§ ‡	§ ‡
T & O						§ ‡
Gen surgery						§ ‡
Urology						
‡ indicates Kruskal-Wallis test shows significant difference between two groups ($p < 0.05$) § indicates Scheffe method shows significant difference between two groups ($p < 0.05$)						

Table 5.6. Pairwise comparison of group means/medians for significant differences in adverse-event rates between specialties.

Interestingly, there is a high level of agreement between the two statistical methods. Significant differences in adverse event rates exist between most pairs of specialties. Only in three pairs - general surgery and ENT; general surgery and trauma and orthopaedics; and urology and gynaecology - was no significant difference found.

In conclusion, the evidence from these analyses strongly supports the construct that rates of different types of adverse event and aggregate adverse-event rates vary across specialties.

Analysis of the relationship between adverse event rates and length of stay

The second construct to be examined proposed that patients who have adverse events would stay longer in hospital than those who do not, and that patients with multiple adverse events would stay longer in hospital again than those with a single adverse event. Two rationales for this construct were put forward. Firstly, when an adverse event results in some morbidity for the patient, the additional diagnostic and therapeutic actions required may require an extension to their stay in hospital. Secondly, patients who stay longer in hospital will be more exposed to the risks of some types of adverse event and so more likely to have them.

To test this construct, the whole sample of 12,676 admissions in 8 specialties was used, and two analyses were undertaken. Firstly, to examine whether patients with adverse events stayed in hospital longer than those with no adverse events, the length of stay of patients in these two groups was examined and compared using both t-tests and Mann-Whitney tests. The results are presented in table 5.7 below.

Specialty	No of admissions	Patients with 0 adverse events		Patients with 1 or more adverse events		t-test	Mann-Whitney test
		No	Mean length of stay	No	Mean length of stay		
Accident and emergency	1031	613	1.07	418	1.01	0.755	0.448
ENT	800	253	2.40	547	2.98	0.002	0.008
General surgery	549	168	6.42	381	9.36	0.137	< 0.001
Gynaecology	566	204	1.49	362	1.72	0.051	0.130
Obstetrics	3960	1721	2.18	2239	4.04	< 0.001	<0.001
Ophthalmology	2252	1386	2.61	866	3.10	< 0.001	< 0.001
Trauma and orthopaedics	2945	1202	7.19	1743	14.46	< 0.001	< 0.001
Urology	521	251	6.23	270	7.61	0.073	0.680

Table 5.7. Comparison of length of stay for patients with and without adverse events.

Secondly, to examine whether patients with multiple adverse events staying in hospital longer than those with a single adverse event, the length of stay of patients in these two groups was examined and compared using both t-tests and Mann-Whitney tests. The results are presented in table 5.8 below.

Specialty	No of admissions	Patients with 1 adverse events		Patients with 2 or more adverse events		t-test	Mann-Whitney test
		No	Mean length of stay	No	Mean length of stay		
Accident and emergency	1031	352	0.92	66	1.56	0.200	0.332
ENT	800	282	2.74	265	3.23	0.091	0.056
General surgery	549	214	6.14	167	13.50	< 0.001	< 0.001
Gynaecology	566	258	1.70	104	1.75	0.833	0.286
Obstetrics	3960	1370	3.35	869	5.11	< 0.001	< 0.001
Ophthalmology	2252	642	2.86	224	3.80	< 0.001	< 0.001
Trauma and orthopaedics	2945	940	10.93	803	18.58	< 0.001	< 0.001
Urology	521	194	6.21	76	11.18	0.004	0.004

Table 5.8. Comparison of length of stay for patients with single and multiple adverse events.

It is evident from the two tables that patients with adverse events have significantly longer lengths of stay than those who have no adverse events, and that multiple adverse events are associated with a further prolongation of the hospital stay. Only in two specialties - gynaecology and accident and

emergency - was no association found. In others, such as obstetrics, trauma and orthopaedics, ophthalmology and general surgery there was a marked increase in length of stay in patients with adverse events.

The differences between specialties which are shown in tables 5.7 and 5.8 warrant some discussion. Firstly, the much larger sample size available in some specialties (such as obstetrics, orthopaedics and ophthalmology) makes it easier to demonstrate statistical significance, even for numerically small differences in mean lengths of stay. Secondly, some specialties such as accident and emergency and gynaecology have short lengths of stay for most or all cases and so the impact of adverse events may be less detectable. Length of stay is measured in whole days, and so the prolongation of length of stay by an adverse event has to be proportionately greater in a specialty with short lengths of stay than it would in a specialty with longer lengths of stay for that prolongation to be recorded. However, it may also be that the differences in presenting diseases and diagnostic and therapeutic processes between specialties, already referred to above, account for the different degrees of association between adverse events and length of stay that were observed.

In conclusion, these two analyses both support the construct that patients who have adverse events stay longer in hospital than those who do not, and that patients with multiple adverse events stay longer in hospital again than those with a single adverse event. However, they also indicate that conformance with the construct varies across specialties.

Analysis of relationship between different types of adverse event and length of stay

The third construct to be tested also related to the relationship between length of stay and adverse events. It proposed that patients who had undergone different types of adverse events would have different degrees of prolongation of their hospital stay, because some adverse events had more serious consequences for patients' health than others. If this construct were confirmed it would support the contention that adverse events cause increased length of stay (or that both are caused by some third factor, such as severity), rather than that increased length of stay exposes patients to greater risk of adverse events and so may result in higher scores on adverse event measures.

In order to examine the effect of an individual adverse event on lengths of stay, it is necessary to separate out that effect from the effects of any other adverse events which may have also occurred during a patient's admission. Since patients quite commonly have multiple adverse events, this may be difficult to do without undertaking a multivariate analysis capable of identifying separately the contributions that each type of adverse event makes to the patient's length of stay. However, such an analysis would be problematic because of the large number of variables to be included (one for each criterion in the adverse-event measure being used).

However, if a relatively large sample of admissions is available for analysis, an alternative approach can be adopted which permits the continued use of bivariate statistical approaches. The analysis can be restricted to patients who had either no adverse events or a single adverse event under a single criterion. In other words, their score on the adverse-event measure would be either 0 or 1. Using this approach, bivariate statistical techniques can be used to compare the lengths of stay of patients with no adverse events at all with the lengths of stay of patients who had a single adverse event under a given criterion. Effectively, this involves comparing a single base group of all patients with no adverse events with a series of subgroups of patients with one and only one adverse event, of a particular type. Of course, this approach has the disadvantage that if some adverse events rarely or never occur alone, the subgroup of patients with one and only one adverse event of that type may be very small or might not exist. There is also some risk of bias, since patients with single adverse events may not be typical of all patients with adverse events (and may particularly be those patients with less severe adverse events). Both these limitations need to be borne in mind in interpreting the results of such an analysis.

To test this construct, the sample of 4,012 admissions in obstetrics was examined, and as described above the analysis was restricted to those admissions with either 0 or 1 adverse events. For each of the 14 criteria within the adverse-event measure that was used in obstetrics (which can be found in Appendix 5.1), the lengths of stay of patients with and without an adverse event under that criterion were compared using the parametric t-test and the non-parametric Mann-Whitney test. Because the patients with an adverse event under that criterion were known to have had only that adverse event (and no others), any difference in lengths of stay could be said to be associated with that type of

adverse event alone, rather than being confounded by other adverse events which might also have occurred during that admission.

The results are presented in table 5.9 below. For each criterion within the obstetric adverse-event measure, the table shows the number of admissions with no adverse events and their mean length of stay, and the number of admissions with a single adverse event under that criterion, and their mean length of stay. The number of admissions with no adverse events and their mean length of stay remains the same, since this base group is constant in each analysis, while the number of admissions with a single adverse event under the criterion and their mean length of stay obviously varies from criterion to criterion. The table shows the significance of differences in the lengths of stay between these two groups, using the t-test and the Mann-Whitney test. It also shows the 95% confidence intervals for the estimate of the mean length of stay of patients in the group with a single adverse event. This allows the significance of differences in the lengths of stay of patients with single adverse events of different types to be explored.

Crit no	Criterion title	Patients with no adverse event		Patients with adverse event				t-test	Mann-Whitney test
		No	Mean LoS	No	Mean LoS	Lower 95% CI	Upper 95% CI	p value	p value
1	SROM	1721	2.18	62	2.40	1.98	2.82	0.337	0.035
2	Induction	1721	2.18	140	3.91	2.78	5.04	0.003	<0.001
3	Probs labour/delivery	1721	2.18	29	2.79	2.39	3.19	0.009	<0.001
4	Caesarean section	1721	2.18	249	6.59	5.73	7.45	<0.001	<0.001
5	Probs Caesarean section	1721	2.18	3	6.67	3.40	9.94	0.114	0.004
6	Perinatal probs	1721	2.18	52	3.48	2.45	4.51	0.018	<0.001
7	Maternal postnatal probs	1721	2.18	98	2.88	2.55	3.21	<0.001	<0.001
8	Drug-related probs	1721	2.18	8	2.50	1.07	3.93	0.678	0.508
9	Mother/family dissatisfaction	1721	2.18	7	2.42	1.39	3.46	0.659	0.327
10	Non-clin probs/incident	1721	2.18	192	2.42	2.20	2.65	0.076	<0.001
11	Record deficiency	1721	2.18	142	2.54	1.91	3.17	0.265	0.213
12	Antenatal anaemia	1721	2.18	117	1.82	1.58	2.06	0.016	0.612
13	Probs of anaesthesia	1721	2.18	4	2.25	1.76	2.74	0.807	0.333
14	Probs of pain relief	1721	2.18	205	2.45	2.23	2.68	0.051	<0.001

Table 5.9. Comparison of length of stay for patients with and without adverse events in obstetrics.

The data in table 5.9 suggests that most types of adverse event in obstetrics are associated with some increase in the length of stay in hospital. The greatest differences in lengths of stay were observed for adverse events which had clear clinical sequelae which would be expected to increase the length of stay. For example, the occurrence of induction of labour, caesarean sections, perinatal complications and postnatal maternal complications were all associated with significant increase in the length of stay. Other adverse events which might be expected to have little impact on the length of stay were indeed associated with smaller and sometimes non-significant increases in the length of stay. For example, record deficiencies, non-clinical problems and incidents, and the antenatal presence of maternal anaemia all fell into this category.

The significance of differences in length of stay between groups with different types of adverse event (rather than differences between each group and the base group of patients with no adverse events) can be assessed using the confidence interval estimates for the mean lengths of stay cited in the table. Where the confidence intervals between a given two groups do not overlap, their means can be said to be significantly different. The confidence interval and mean data given in table 5.9 is also displayed in figure 5.1 below. It can be seen that, while some groups (particularly those containing low numbers of patients, such as those for criteria 5, 6, 8 and 9) have wide confidence intervals, most do not, and many instances of non-overlapping confidence intervals exist.

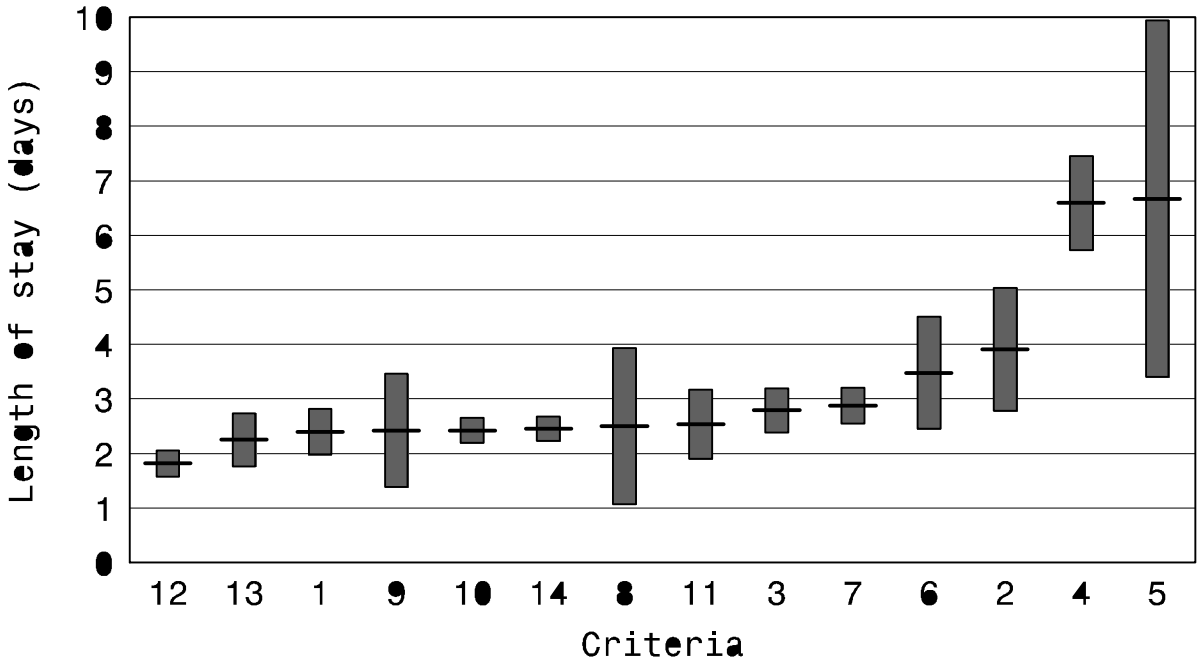


Figure 5.1. Mean lengths of stay and associated 95% confidence intervals for groups of patients with a single adverse event (criteria numbers refer to table 5.9).

Some caution should be exercised in drawing conclusions from this data, because of the limitations outlined earlier. The subgroups of patients used in this analysis often contained quite small numbers, and many patients with adverse events were excluded from the analysis because they had multiple adverse events. For example, 162 patients had an adverse event recorded under criterion 6 (perinatal problems) but only 52 of these had only this adverse event and so were eligible for inclusion in the analysis. However, it can be concluded that the data from this analysis of admissions in obstetrics supports the construct that different types of adverse events will be associated with different degrees of prolongation of the hospital stay. Clinically significant adverse events were found to be associated with substantially greater increases in length of stay than more trivial and non-clinical adverse events, and some of the differences between groups with different types of adverse event were significant. This also tends to support the contention that the association between adverse events and increased length of stay described earlier represents an effect of the adverse event itself, and that it is the consequences of the adverse event which cause the increased length of stay. The alternative explanation of this association would be that patients who happen to be in hospital longer have a greater chance of encountering an adverse event during their

stay, but if this were so we would expect to see broadly similar degrees of association and increases in length of stay for all types of adverse events, and this was not the case.

Analysis of relationship between adverse event rates and death among patients

The fourth construct outlined in section 5.2.1 was that patients who died in hospital would have had more adverse events than those who were discharged alive. This might be expected to be true for two reasons. Firstly, patients who die have usually been more severely ill and more complex to manage than the norm, and so the likelihood of some adverse events occurring might be expected to be higher. Secondly, some serious adverse events might contribute to or even cause a patient's death.

To test this construct, the whole sample of 12,676 admissions in 8 specialties was used. In each specialty, the adverse event rate for patients who were discharged alive was compared with the adverse event rate for patients who died in hospital, using both t-tests and the non-parametric Mann-Whitney test. It must be noted that deaths in hospital are rare events, and can be virtually unknown in some specialties. This means that this approach to testing the construct is only likely to be of use in specialties with relatively high death rates, or when extremely large samples of patients are used.

The results are presented in table 5.10. For each specialty it shows the numbers of admissions ending in the patient being discharged alive or dying, the mean adverse event rate for admissions in these two groups, and where possible the significance of differences in adverse event rate reported by the two statistical tests.

Specialty	No of admissions	Patients discharged alive		Patients died in hospital		t-test p-value	Mann-Whitney test p-value
		No	Mean no of adverse events	No	Mean no of adverse events		
Accident and emergency	1031	1024	0.48	1	0.0	-	-
ENT	800	798	1.13	0	-	-	-
General surgery	549	548	1.23	1	2.0	-	-
Gynaecology	566	566	0.88	0	-	-	-
Obstetrics	3960	3949	0.89	0	-	-	-
Ophthalmology	2252	2245	0.52	0	-	-	-
Trauma and orthopaedics	2945	2901	1.02	40	2.80	< 0.001	< 0.001
Urology	521	520	0.70	1	2.0	-	-

Table 5.10. Comparison of number of adverse events per admission for patients discharged alive or dead in specialties

In several of the specialties for which data is presented above no deaths occurred at all during the period when screening was taking place. Effectively, there were only sufficient deaths in one specialty, trauma and orthopaedics, to test the construct. In this specialty patients who died in hospital had almost three times as many adverse events as those who were discharged alive, a difference which both statistical tests indicate was significant. We can conclude that the construct is supported by the data from this specialty, but should be cautious about generalising from the limited data available.

Analysis of relationship between adverse event rate and method of admission

The next construct, which was described in section 5.2.1, was that patients admitted as emergencies would have more adverse events than those admitted electively. The rationale for this was that emergency admissions are likely to be more complex and more heterogeneous than elective cases. The former present unpredictable and often unstable health problems which require immediate diagnosis and treatment and may make unscheduled but urgent demands on healthcare services. The latter present more predictable and generally stable health problems, for which a planned sequence of diagnostic and therapeutic interventions can be identified. For these reasons, we might expect that some sorts of adverse event would be commoner among patients admitted as emergencies.

To test this construct, once again the whole sample of 12,676 admissions in 8 specialties was used. In each specialty, the adverse event rate for patients admitted electively was compared with the adverse event rate for patients admitted as emergencies, using both t-tests and the non-parametric Mann-Whitney test. The results are presented in table 5.11 below. For each specialty it shows the numbers of elective and emergency admissions, the mean number of adverse events for admissions in these two groups, and the significance of differences in numbers of adverse events reported by the two statistical tests.

Specialty	No of admissions	Elective admissions		Emergency admissions		t-test p-value	Mann-Whitney test p-value
		No	Mean no of adverse events	No	Mean no of adverse events		
Accident and emergency	1031	4	0.25	1027	0.48	0.429	0.492
ENT	800	641	1.06	157	1.41	0.001	0.001
General surgery	549	235	1.19	314	1.27	0.501	0.077
Gynaecology	566	15	1.13	551	0.88	0.268	0.180
Obstetrics	3960	4	1.5	3955	0.89	0.712	0.517
Ophthalmology	2252	1986	0.49	254	0.71	0.001	0.001
Trauma and orthopaedics	2945	1121	0.87	1812	1.14	< 0.001	< 0.001
Urology	521	225	0.73	295	0.67	0.418	0.511

Table 5.11. Comparison of number of adverse events per admission for patients admitted electively and as emergencies

Although the table presents data for all specialties, for the sake of completeness, it is clear that the lack of elective admissions in three specialties - accident and emergency, obstetrics and gynaecology - makes comparisons of adverse event rates between elective and emergency admissions in those specialties relatively meaningless. The low numbers of elective admissions recorded in accident emergency and obstetrics are unsurprising, given the nature of the specialties concerned. The low numbers of elective admissions in gynaecology reflects the fact that the Royal Sussex County Hospital only admitted gynaecology patients as emergencies, while the elective gynaecology service was provided by another hospital. Obstetric admissions, even those for planned induction of labour or caesarean section, were generally recorded by the hospital as not elective.

In the remaining five specialties, three conform to the construct that emergency admissions would have more adverse events than elective admissions. In ophthalmology, trauma and orthopaedics and ENT emergency admissions had, on average, 45%, 31% and 33% more adverse events than elective admissions, a difference which is statistically significant in each case. However, in the remaining two specialties, of general surgery and urology, no significant difference emerges. It should be noted that the latter two specialties are those with the smallest sample sizes, and therefore the least statistical power to detect any difference. With this point in mind, it is reasonable to conclude that the data supports the construct that emergency admissions will have more adverse events than elective admissions.

Analysis of relationship between adverse events and patients' age

The final construct to be tested, outlined in section 5.2.1, was that elderly patients would have more adverse events than younger patients. Like the analyses reported above, of the relationship between numbers of adverse events and patients' discharge status and admission method, this construct is essentially concerned with the impact on adverse event rates of the complexity of patients' health problems and the severity of illness they represent. Elderly patients are more likely to have comorbidities, may be less able to tolerate the effects of treatment, and may take longer to recover from the effects of ill-health. These factors are likely to make their care more complex to manage, and might therefore be expected to make some sorts of adverse event more common, and so a relationship between adverse event rates and patients' ages would be expected.

To test this construct, the whole sample of 12,676 admissions in 8 specialties was used. In each specialty, age on admission (in years) of patients with 0, 1 and 2 or more adverse events was examined and compared, using the one way analysis of variance (ANOVA) and the non-parametric Kruskal Wallis analysis of variance by ranks. The results are presented in table 5.12 below. For each specialty, it shows the mean age on admission (in years) of patients with 0, 1 and 2 or more adverse events during their admission. The Scheffé method for multiple pairwise comparisons of means was used to identify whether the differences between mean ages on admission for groups with 0 and 1 adverse events and 0 and 2 or more adverse events were significant (at the $p < 0.05$ level). Where significant differences existed, the relevant mean values in the table are marked with an asterisk.

Specialty	No of adm	Patients with 0 adverse events		Patients with 1 adverse event		Patients with 2 or more adverse events		One way ANOVA	Kruskal-Wallis analysis of variance
		No	Mean age	No	Mean age	No	Mean age		
Accident and emergency	1031	613	45.5	352	46.7	66	46.7	0.722	0.626
ENT	800	253	42.4	282	44.9	265	43.7	0.379	0.368
General surgery	549	168	54.3	214	55.0	167	57.3	0.424	0.291
Gynaecology	566	204	29.0	258	28.4	104	28.6	0.740	0.911
Obstetrics	3960	1721	28.3	1370	28.1	869	28.4	0.161	0.243
Ophthalmology	2252	1386	65.8	642	67.2	224	65.0	0.301	0.656
Trauma and orthopaedics	2945	1202	49.4	940	54.2 *	803	59.8 *	< 0.001	< 0.001
Urology	521	251	60.0	194	60.0	76	65.4	0.094	0.040

Note: Asterisk denotes mean age significantly different from that for patients with 0 adverse events ($p < 0.05$, Scheffé method).

Table 5.12. Comparison of age on admission for patients with 0, 1 and 2 or more adverse events in specialties, and results of tests of the statistical significance of differences in age on admission

The results of this analysis suggest that for most specialties there is no significant association between patients' ages and their adverse event scores. In accident and emergency, ENT, obstetrics, gynaecology and ophthalmology no relationship was observed. In general surgery and urology, patients with more adverse events tended to be older, but the difference was not statistically significant. As observed above, the sample sizes for these two specialties were small, and any difference would need to be quite large for it to be statistically significant. However, in one specialty - trauma and orthopaedics - a clear and statistically significant association between adverse event scores and age on admission was detected. Patients with 1 and 2 or more adverse events were on average about 5 and 10 years older on admission respectively than those with no adverse events.

These results need to be interpreted with some caution. They suggest that the construct, that adverse events will be more common among elderly patients is not generally supported. While an association was found in one specialty, it is worthy of note that no association was found in other specialties, despite quite large sample sizes. For example, it might be expected in obstetrics that elderly mothers might have more adverse events than younger mothers, but this was apparently not the case. Similarly, the association between age and adverse events found in trauma and orthopaedics might have been expected to be replicated in ophthalmology, since both are surgical specialties which deal with many elderly and very elderly patients, but it was not.

5.3.2 **Multivariate analysis of construct validity**

The bivariate statistical techniques used in section 5.3.1 to explore the construct validity of various adverse-event measures of healthcare quality have the advantage of being relatively easy both to undertake and to interpret. However, analyses of the relationships between two variables can be seriously misleading if both of the variables being studied are associated with another variable or variables which are not included in the analysis. Bivariate methods are well suited to experimental research designs, in which randomisation can eliminate the risks of bias or confounding and can ensure that only the relationship between the dependent and independent variable being studied is material, but they are less well able to cope in non-experimental or observational research studies.

Multivariate statistical methods are particularly well suited to analysing data sets containing a number of variables, where few if any assumptions can be made about the nature of any relationships between them. They can be used in situations in which there are multiple dependent and independent variables and where the classification of variables as either dependent or independent is not clear cut. They can also be used to identify and quantify shared or overlapping variance among variables (which occurs when independent variables are associated with each other). Moreover, while some multivariate methods require the data to meet certain conditions (generally multivariate normality, linearity and homoscedasticity), others, particularly those used to analyse categorical data, are relatively free of such restrictions.

However, multivariate analyses are often difficult to interpret. Firstly, their meaning is often less intuitively evident to the user, the results are less easily related directly to the real world represented by the variables in the data set being analysed, and data errors or other faults in the analysis are generally harder to spot. Secondly, the results of some multivariate methods can be quite sensitive to the methods chosen and the strategy for their application to the problem. It has been suggested that a "judicious mix of multivariate and univariate statistics" is necessary to make a comprehensive analysis of a multivariate data set (Tabachnick and Fidell 1989, p7).

In order to explore the construct validity of an adverse event measure using multivariate methods, it is necessary to examine the relationships among a number of variables, some of which are categorical or nominal (such as whether a patient was an emergency or elective admission), while others are ordinal or even ratio scale variables (such as the length of stay, or the number of adverse events). Because some of the data is categorical, and few assumptions can be made about the distribution of the ordinal and ratio scale variables in the data set, it is appropriate to use a multivariate method designed for categorical data sets, known as multiway frequency analysis or loglinear analysis.

Multiway frequency analysis can be used to examine the relationships between a number of categorical variables without presupposing that particular variables are dependent or independent. It has two main purposes. Firstly, it is used to explore the relationships or associations which exist between the variables. Secondly, it can be used to build a model which allows the values of one variable to be estimated from the values of others, and which contains only those variables which have a significant effect on the variable whose value is being estimated.

Multiway frequency analysis is essentially an extension of the principles underlying the familiar χ^2 test of association to the multivariate situation. A linear model is created of the logarithm of expected cell frequencies in the multiway table produced by cross-tabulating every variable by every other variable. The logarithms of expected cell frequencies are used (rather than the frequencies themselves) in order to transform the multiplicative formula for calculating expected cell frequencies familiar from the χ^2 procedure to an additive one suitable for the general linear model used in many multivariate analytical approaches. The formula used to predict the observed

cell frequencies contains a term (usually referred to as an effect) for each possible one-way, two-way, three-way and higher order association for all variables, pairs of variables, triplets of variables, and so on. For each possible combination of variable categories within an association, a parameter estimate is calculated. When the formula contains an effect for every possible association in this way, it can predict all the observed cell frequencies exactly using the parameter estimates, and the model is said to be saturated.

For each effect in the model (or each association), the significance of the partial association (that is, the degree of association adjusted for all other associations) can be calculated. The partial association represents the variance which can be ascribed solely to that effect rather than being shared with other effects. Effects that have non-significant partial associations contribute little to the model and can be eliminated without impairing its ability to estimate the observed cell frequencies sufficiently accurately. When all the non-significant effects have been eliminated, the model can be used to predict one variable from other variables.

Multiway frequency analysis makes no assumptions about population distributions, and can be applied without limitations to categorical data and continuous variables which have been transformed to discrete categories. The only limitation to using the technique is the size of expected frequencies in each cell. When expected frequencies are low, they reduce the power of the technique to detect associations or effects which are significant. In other words, they increase the chance that associations which are actually significant will be classified as non-significant. If the number of cases in the sample is low compared with the number of variables and categories within variables in the data set, or if the distribution of events across categories is highly skewed indicating that some categories are very rare, there is a risk that expected frequencies will be low in some cells. The rule of thumb proposed by some authors is that expected cell frequencies for all two way associations should be examined to make sure that all are greater than one and no more than 20% are less than five. Of course, by eliminating variables or collapsing categories, the cells with low expected frequencies can be removed or combined, but this also affects the power of the analysis by removing its ability to detect associations involving the eliminated variables or relationships concerning the collapsed categories.

A multiway frequency analysis was first undertaken on all the screened admissions in the eight specialties studied. This data set, consisting of a total of 12,676 admissions, was described in table 5.2. The variables from the data set chosen for inclusion in the multiway frequency analysis were those which had been used in the bivariate analyses of section 5.3.1, and they are listed in table 5.13 below. The three ratio scale variables - length of stay, age and adverse event score - were each converted to categorical variables by grouping data values to a relatively small number of categories. One variable used in the bivariate analyses, the discharge type, was excluded from the multiway frequency analysis. This variable had only two categories (discharged alive or died), but since deaths in hospital were very rare (the sample of 12,676 admissions only contained 43 deaths) the inclusion of this variable would have created many cells with unacceptably low expected frequencies. It was anticipated that, with this set of variables for analysis, the multiway frequency analysis would provide evidence on four of the six constructs set out in section 5.1.2. It could not inform the consideration of the constructs concerning the effects of different types of adverse event and the relationship between adverse event rate and discharge type because the necessary variables were not part of the data set analysed.

Variable name	Variable title	Data type	Levels
SPEC	Specialty	Categorical	8 levels in total
SEX	Patient sex	Categorical	2 levels (male, female)
ADMTYPE	Admission type	Categorical	2 levels (elective, emergency)
LOSGRP	Length of stay	Ratio scale, converted to categorical	4 levels (0 days, 1-5 days, 6-10 days, 11+ days)
AGEGRP	Age on admission	Ratio scale, converted to categorical	3 levels (0-50 years, 51-70 years, 71+ years)
AEGRP	Adverse event score	Ratio scale, converted to categorical	4 levels (0, 1, 2, 3+ adverse events)

Table 5.13. Variables used in the multiway frequency analysis of 12,676 admissions from eight specialties screened for adverse events.

The results of the first multiway frequency analysis undertaken for the whole data set are presented in table 5.14 below. The first section of the table shows that only the first, second and third order effects were significant. This means that no significant higher order effects (associations involving

four or more variables) were found. The second part of the table lists the significant effects, with their partial associations (all effects with a significance of $p < 0.05$ are listed).

Tests that K-way effects are zero:

K	DF	L.R. Chisq	Prob	Pearson Chisq	Prob	Iteration
1	17	35529.657	.0000	496094.081	.0000	0
2	108	26599.612	.0000	31346.738	.0000	0
3	334	1068.383	.0000	14994.674	.0000	0
4	533	310.356	1.0000	307.801	1.0000	0
5	417	162.933	1.0000	149.926	1.0000	0
6	126	20.606	1.0000	15.346	1.0000	0

Tests of PARTIAL associations.

Effect Name	DF	Partial Chisq	Prob	Iter
AEGRP*LOSGRP*SPEC	63	128.771	.0000	8
LOSGRP*SPEC *AGEGRP	42	70.923	.0035	7
SPEC *AGEGRP*SEX	14	146.841	.0000	8
LOSGRP*SPEC *ADMTYPE	21	160.631	.0000	8
LOSGRP*AGEGRP*ADMTYPE	6	39.368	.0000	8
SPEC *AGEGRP*ADMTYPE	14	99.515	.0000	8
SPEC *SEX*ADMTYPE	7	35.002	.0000	8
AGEGRP*SEX*ADMTYPE	2	22.675	.0000	8
AEGRP*LOSGRP	9	786.140	.0000	13
AEGRP*SPEC	21	433.923	.0000	11
LOSGRP*SPEC	21	2761.916	.0000	12
AEGRP*AGEGRP	6	20.102	.0027	13
LOSGRP*AGEGRP	6	572.496	.0000	15
SPEC *AGEGRP	14	4584.666	.0000	15
LOSGRP*SEX	3	23.285	.0000	12
SPEC *SEX	7	4396.296	.0000	12
AGEGRP*SEX	2	418.327	.0000	13
AEGRP*ADMTYPE	3	11.652	.0087	12
LOSGRP*ADMTYPE	3	167.791	.0000	11
SPEC *ADMTYPE	7	5118.443	.0000	15
AGEGRP*ADMTYPE	2	76.351	.0000	13
SEX*ADMTYPE	1	6.847	.0089	13
AEGRP	3	5255.669	.0000	2
LOSGRP	3	15421.426	.0000	2
SPEC	7	7081.866	.0000	2
AGEGRP	2	4489.366	.0000	2
SEX	1	1902.232	.0000	2
ADMTYPE	1	1379.070	.0000	2

Table 5.14. Results of multiway frequency analysis of 12,676 admissions from eight specialties screened for adverse events.

It can be seen that there are eight significant third order effects, representing complex associations between three variables, and that six of these eight third order effects involve the specialty variable. Furthermore, five of the fourteen second order effects also involve the specialty variable. The presence of both a third order effect involving adverse event rate and specialty (with length of stay)

and a second order effect involving adverse event rate and specialty support the first construct outlined in section 5.1.2, that rates of adverse events would vary across specialties.

Because third order effects are difficult to interpret, and because one variable (specialty) is involved in most of the third order effects and many lower order effects, these results suggest that it is necessary to undertake separate multiway frequency analyses for each specialty rather than a single analysis across all specialties.

Further multiway frequency analyses were then undertaken for each specialty separately. These analyses can be found in full in appendix 5.2, but their results are summarised in table 5.15. It shows, for each separate multiway frequency analysis of data for a single specialty, whether the results supported the constructs set out in section 5.1.2, and whether other significant associations between variables were also found.

Specialty	(b) Patients with adverse event/multiple adverse events will stay in hospital longer AEGRP* LOSGRP	(e) Patients admitted as emergencies will have more adverse events than those admitted electively AEGRP* ADMTYPE	(f) Elderly patients will have more adverse events than younger patients AEGRP* AGEGRP	Other significant associations between variables
Accident and emergency				Length of stay and age. Age and sex.
ENT	*	*		Length of stay and age. Length of stay and admission type. Age and admission type.
Gynaecology	*			
Ophthalmology	*		*	Length of stay and age. Age and sex. Length of stay and admission type. Age and admission type. Sex and admission type.
Orthopaedics	*		*	Length of stay and age. Length of stay and sex. Age and sex. Length of stay and admission type. Age and admission type. Sex and admission type.
General surgery	*	*		Length of stay and age. Length of stay and admission type. Age and admission type.
Urology				Length of stay and age. Age and sex. Age and admission type.

Table 5.15. Summary of the results of separate multiway frequency analyses for each specialty.

The results in table 5.15 provide rather qualified support for the constructs being tested. They certainly suggest that an association between adverse event rates and length of stay exists - the first of the three constructs examined. However, for the second and third constructs they indicate that associations between adverse event rates and admission type and adverse event rates and age only exist in some specialties. It should be remembered that the power of these analyses to detect significant associations is low because of the relatively small numbers of cases in some specialties,

and the asymmetric distribution of cases resulting in many cells with low expected frequencies.

It is particularly useful to note the commonly occurring other associations listed in the final column of the table, since they provide some indications of the associations which might have confounded the earlier bivariate analyses presented in section 5.3.1. It is not surprising that associations between both length of stay and age and between age and sex were frequently found. The association between age and admission type commonly reported is less intuitively easy to explain. Of course, these multiway frequency analyses only indicate that an association exists; they do not provide information on the direction or strength of the association without either further analysis of the parameter estimates or the undertaking of separate bivariate analyses such as those already concluded for the constructs being tested.

5.4 **Conclusions**

It was noted at the outset of this chapter that this examination of construct validity was particularly important because, in the absence of meaningful criterion variables it was difficult or impossible to explore the criterion-related validity of adverse-event measures of quality. In this situation, an examination of construct validity was both more methodologically appropriate and likely to provide greater insight into the behaviour of the measure or measures being tested.

The series of analyses of adverse event data from the RSCH project presented above leave little doubt of the construct validity of the adverse-event measures tested. As table 5.16 below demonstrates, each construct was supported by data in at least some specialties, and most were widely supported by data from several different specialties. Overall, five of the six constructs tested were confirmed by the data from the RSCH project. Only one construct was not supported by the data.

	Construct	Summary of results
a)	Adverse event rates will vary between specialties.	Confirmed. Both rates of individual types of adverse events and generic adverse event rates found to vary significantly across specialties.
b)	Adverse event rates will correlate with length of stay.	Confirmed. Patients with an adverse event stayed significantly longer in hospital than those without adverse events in 5 out of 8 specialties. Patients with multiple adverse events stayed significantly longer than those with just one adverse event in 5 out of 8 specialties. Specialties where no difference found tended to be those with very short mean lengths of stay. Also supported by multiway frequency analysis.
c)	Different types of adverse events will have different effect on length of stay.	Confirmed in obstetrics. Patients with clinically significant adverse events found to stay longer in hospital than those with more minor or non-clinical adverse events.
d)	Patients who die will have had more adverse events than those who do not.	Confirmed in trauma and orthopaedics, the only specialty in which there were sufficient deaths to test the construct.
e)	Emergency patients will have more adverse events than elective patients.	Confirmed in 3 of the 5 specialties with a mixture of elective and emergency admissions. Specialties where no difference found were those with small sample sizes. Supported in some specialties by multiway frequency analysis.
f)	Elderly patients will have more adverse events than younger patients.	Only confirmed in one specialty - trauma and orthopaedics -and not supported in 7 other specialties by bivariate analysis. Significant association found in two specialties - trauma and orthopaedics and ophthalmology - in multiway frequency analysis.

Table 5.16. Overview of the results of construct validity analysis using RSCH project data.

However, the multiway frequency analysis presented in section 5.3.2 also provided some reasons to be cautious in the interpretation of adverse-event measure data. It highlighted the number of other associations which existed in the RSCH project data (such as those between age and sex, age and length of stay, length of stay and admission type, and so on). While these do not affect our conclusions on the validity of the adverse-event measures being tested, they make the interpretation of such data more difficult and they particularly impede the making of valid comparisons between specialties, clinical teams, wards or other groupings.

Chapter 6

Reliability of adverse-event measures of quality

6.1 Introduction

Chapter 3 reviewed the literature on adverse-event measures of quality and presented an analysis of the small number of studies of the reliability of these measures which have been undertaken. It was demonstrated that, given the widespread use of adverse-event measures of quality, there was surprisingly little evidence that the reliability of these measures had been adequately assessed. Indeed, since the few published studies reached markedly different conclusions about the reliability of adverse-event measures of quality, even when the same measure was being tested, it was clear that the reliability of such measures needed to be investigated further.

The approach taken to testing the reliability of any measurement instrument is shaped in part by the design and characteristics of the instrument itself, as was noted in chapter 2. The construction and definition of adverse-event measures of quality were discussed in some detail in chapter 3. An adverse-event measure is usually made up of a number of criteria, each of which relates to a particular type of adverse event. When the measure is used to review the care provided to a patient, a series of dichotomous data values is produced, indicating for each criterion whether or not that type of adverse event was found. A summary score is also sometimes constructed, totalling the number of adverse events found. Many of the reliability studies reported in chapter 3 only assessed the reliability of the latter summary score rather than the reliability of each individual criterion within the adverse-event measure. Though this provides a useful overall estimate of the reliability of the measure, it risks overestimating its reliability since quite different sets of adverse events recorded by a rater can produce the same summary score value. Moreover, if the results of reliability testing are to be used to improve the measure to make it more reliable, reliability results for each criterion individually are needed. For these reasons, it is advisable to report the reliability

of adverse-event measures in both ways - with statistics for each criterion individually and with statistics for the summary score based on the total number of adverse events found.

The reliability of a measurement instrument can be tested experimentally, by undertaking specific studies in which the instrument is repeatedly applied and the results are compared. Such studies have the advantage of being capable of isolating the variability in results which is attributable to interrater variation from other sources of variability, and so providing a credible and valid estimate of their reliability. However, it is difficult to ensure that such reliability studies are undertaken in the same conditions in which the measure would normally be used, and there is a risk that differences in rater training, the time available to raters to collect data, rater motivation and skill level, and other areas could lead to the resulting estimates of reliability being inflated.

An alternative approach to testing the reliability of a measurement tool is to analyse the data gathered by different raters using the tool and to test whether differences exist which might be attributable to rater variation. This is methodologically and statistically more complex, since there may be many known and unknown confounding factors which cause the variations observed in the data, and while the known confounders can be controlled, the unknown sources of bias cannot. However, since this approach is based on data collected from the use of the measurement tool in a more realistic setting, the resulting estimates of reliability may have greater external validity.

Since both these approaches to assessing reliability have their merits, two sets of investigations of the reliability of a number of adverse-event measures were undertaken for this study. Firstly, a number of specific experimental studies were carried out, in which a set of patient admissions was reviewed repeatedly by different screeners and the results of these screenings were compared. Secondly, an analysis was undertaken of the adverse-event data from the RSCH project to examine whether there were variations in the rates of adverse events associated with differences in raters or screeners. This chapter reports the findings from both sets of investigations.

6.2 Experimental studies of interrater and intrarater reliability

6.2.1 Aims of interrater and intrarater reliability studies

The aim of the interrater and intrarater reliability studies reported in this section was to assess the reliability of a number of adverse-event measures of quality experimentally through the repeated application of the measures to the same patient admissions. It was noted in chapter 3 that estimates of reliability are known to be significantly affected by the conditions under which the study takes place and by the extent and quality of training provided to the raters or screeners applying the measures under test. The studies reported here were designed to provide a realistic estimation of the reliability of these measures, obtained in conditions which approximated their normal use in an acute hospital setting. They took place as part of the routine of data collection for the RSCH project, using the same staff and methods. No special training or other measures were undertaken to prepare for the reliability studies, and the process of screening or reviewing patients' records, abstracting data about adverse events and recording it was performed in the usual manner, using the same forms and computer systems.

6.2.2 Methods

Four separate but linked studies of the reliability of adverse-event measures of quality were undertaken using the staff and resources of the RSCH occurrence screening project, in three specialties - ENT, ophthalmology and obstetrics. In each of these specialties, an adverse-event measure of quality had been developed and was already in routine use. In ENT and ophthalmology, this measure was based on the generic adverse-event measure described in chapters 4 and 5. In obstetrics, a measure developed specifically to reflect the healthcare process in that specialty had been developed (and was also described in chapter 5). The reliability studies were undertaken on these measures as they stood, although, as the description of the development process in chapter 3 makes clear, they had not necessarily been designed to maximise reliability. The intention was to provide a realistic estimate of the reliability of adverse-event measures such as these, rather than to establish a theoretical reliability which might be difficult to match in actual practice.

In each study, the reliability of the adverse-event measure under test was investigated by applying the measure twice to a series of patient admissions, and then comparing the results from these two applications of the measure to assess the degree of agreement between them. In three studies of interrater reliability, the first and second applications of the measure were undertaken by different individuals with no knowledge of each other's ratings. In one study of intrarater reliability, the first and second applications of the measure were undertaken by the same person, but with an interval of time between the two applications sufficient to ensure that the person would not be able to remember the details of patient admissions or the results of the first application of the measure when applying it for the second time. In all four studies, the reliability of the adverse-event measures being tested was then assessed by calculating various agreement statistics. The studies were undertaken in three specialties (ENT, ophthalmology and obstetrics), with interrater reliability being tested in all three and intrarater reliability being tested in obstetrics only.

In two specialties - ENT and ophthalmology - the interrater reliability of the adverse-event measures in use in those specialties was investigated by arranging for a sample of patient admissions to be screened for adverse events twice, by different screening staff on each occasion. These studies both took place during the period February to April 1992. Not all patient admissions during that period in the two specialties were included in the study, but to avoid any case-selection bias either all admissions on any given day were subjected to rescreening, or none were. The first screening took place as part of the routine data collection process of the project. The second screening was then undertaken by a different member of staff, with no knowledge of the results of the first screening and no communication with the member of staff who undertook the first screening. The project staff involved in these two studies were, of course, aware that the reliability of their application of the adverse-event measures was being tested, and this could have influenced their behaviour and use of the measure. The results of the first and second screening of each admission were then compared.

In the third specialty - obstetrics - both the interrater reliability and the intrarater reliability of the adverse-event measure in use was investigated. To measure the intrarater reliability of the measure, a consecutive series of patient admissions was identified which had taken place during the period September to November 1991 and had been screened at that time using the adverse-event measure. This sample of patient admissions was restricted to those which had been screened by one particular

individual member of the project staff, and the study took place in January to March 1992. The patients' casenotes were retrieved in chronological order of admission date, and each admission was then screened for a second time by the same member of the project staff, with no access to the results of the first screening. Each patient admission was screened for the second time approximately four months after it had first been screened. Given that the member of project staff had, in the intervening period, screened hundreds of other patients' admissions, it was judged that she would not be able to remember the results she had recorded when she first screened these cases four months previously. It should be noted that the member of project staff was not aware when undertaking the first screening of these admissions that they would subsequently be the subject of this intrarater reliability study, so her behaviour and use of the measure on first screening was not affected by the study. The results of her first and second screening were then compared

The interrater reliability of the adverse-event measure used in obstetrics was then investigated by arranging for the sample of patient admissions used in the intrarater study to be screened once again. This third screening was not undertaken by one of the project screening staff, but by a doctor qualified in obstetrics who undertook a brief training programme on how to interpret and apply the adverse-event measure. The doctor undertaking this third screening had no knowledge of the results of the first and second screening of the admission undertaken by one of the project staff. The results of the doctor's screening were then compared with each of the previous screening results (from September-November 1991 and January-March 1992) separately, resulting in two sets of interrater comparisons.

The four studies of interrater and intrarater reliability are summarised in table 6.1 below (note that the obstetric interrater reliability study involved the comparison of two different screenings of a set of obstetric cases by a member of project staff with that undertaken by a doctor qualified in obstetrics and effectively produced two sets of data, and so is listed as C and D).

Study	Specialty	Investigation	No of cases	First screening	Second screening
A	ENT	Interrater reliability	146	Undertaken by a member of project screening staff, during study period of Feb-April 1992.	Undertaken by a different member of project screening staff, on same day as first screening.
B	Ophthalmology	Interrater reliability	125	Undertaken by a member of project screening staff, during study period of Feb-April 1992.	Undertaken by a different member of project screening staff, on same day as first screening.
C	Obstetrics	Interrater reliability	108	Undertaken by a member of project screening staff during period Sept-Nov 1991.	Undertaken by a doctor qualified in obstetrics during the period Jan-March 1992.
D	Obstetrics	Interrater reliability	110	Undertaken by a member of project screening staff during the period Jan-March 1992.	Undertaken by doctor qualified in obstetrics during the period Jan-March 1992.
E	Obstetrics	Intrarater reliability	110	Undertaken by a member of project screening staff during period Sept-Nov 1991.	Undertaken by the same member of project staff during the period Jan-March 1992, approx 4 months after first screening.

Table 6.1. Summary of studies of interrater and intrarater reliability.

Each study of interrater and intrarater reliability produced a paired data set containing the results of two applications of the adverse-event measure being tested to a set of patient admissions. The adverse-event measures whose reliability was being tested consisted of a number of screening criteria, each defining a particular type of adverse event. The results of screening a patient admission using a measure was therefore a series of dichotomous data values (adverse event occurred or no adverse event occurred), one for each screening criterion in the measure. In addition, a summary measure was constructed, totalling the number of adverse events found for a given patient admission.

In order to assess the reliability of the adverse-event measures being tested, the two sets of data from separate screenings of the same patient admission needed to be compared and the degree of agreement measured. It was noted earlier that comparisons should be made both for each criterion individually within the measure, and for the total number of adverse events found for a given patient admission. These two forms of comparison require different statistical approaches.

Comparing the results obtained for each criterion in an adverse-event measure produces a crosstabulation of the results from each screening, as shown in table 6.2 below.

.

		Second screening: No adverse event	Second screening: Adverse event
First screening	No adverse event	a	b
	Adverse event	c	d

Table 6.2. Comparison of the results of repeated screening for a single adverse event criterion.

The percentage agreement can thus be calculated as:

$$\frac{a+d}{n} \times 100$$

where n is the total number of cases. However, some degree of agreement would be expected through chance alone, and so the raw percentage agreement does not provide a useful or comparable measure of agreement. The commonly accepted measure of agreement is the P (kappa) statistic which measures how much the observed level of agreement exceeds that which would be expected by chance (Siegel and Castellan 1988, p284):

$$\kappa = \frac{p_o - p_c}{1 - p_c} = \frac{\frac{a+d}{n} - \frac{(a+b)(a+c)+(b+d)(c+d)}{n \times n}}{1 - \frac{(a+b)(a+c)+(b+d)(c+d)}{n \times n}}$$

The maximum value of P is 1, representing perfect agreement, while a value of 0 represents no more agreement than would have been expected through chance alone. Negative values of P can occur, indicating that there is even less agreement than would be expected by chance, but the meaning of such a finding is not straightforward, and the scale of negative P values is difficult to interpret. A commonly used benchmark for the use of the P statistic (Brennan and Silman, 1992) is set out in table 6.3 below, but it should be noted that this represents a pragmatic and rather arbitrary appraisal of the statistic's meaning with no particular foundation in statistical reasoning. The probability that the P statistic is greater than zero can be calculated, to identify whether the degree of agreement found is significantly greater than that expected by chance alone.

Value of P statistic	Strength of agreement
< 0.20	Poor
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Good
0.81 - 1.00	Very good

Table 6.3. Interpretation of the P statistic (Brennan and Silman, 1992).

In this study, the P statistic was calculated for each criterion within the adverse-event measures tested, and the probability that the value of P was significantly greater than 0 was calculated. In addition, a value of P was also calculated for the measure as a whole, by aggregating the numbers of agreements and disagreements, as shown in table 6.2, across all criteria in the measure. It should be noted that P has some drawbacks, particularly when the levels of agreement to be expected through chance are high because of the underlying prevalence of the characteristic being measured. In these circumstances, the value of P is attenuated and so the benchmarks set out in table 6.3 may be harder to meet (Thompson and Walter 1988). In some circumstances when the prevalence of the characteristic being measured is low, the P statistic cannot be calculated because when table 6.2 contains only a single non-zero row or column, $p_c = 1$ and so the denominator in the formula for P becomes zero.

The total number of adverse events found for a patient admission is the commonly used summary score. It has been common practice to use the Pearson product-moment correlation coefficient to assess agreement for ratio scale or continuous data items, as the review of previous studies of the reliability of adverse-event measures in chapter 3 demonstrated. However, the Pearson correlation coefficient is a measure of association rather than agreement, and it does not take account of differences of scale or bias. Moreover, it would be surprising if two measures of the same characteristic were not correlated, and so the correlation coefficient and associated significance test results are arguably irrelevant to the issue of agreement. It should not be used to assess agreement (Bland and Altman 1986).

An alternative approach to assessing agreement for ratio scale or continuous data has been proposed by Bland and Altman (1986). They recommend the use of a combination of graphical techniques and simple calculations involving the examination of mean differences between the two ratings or values. By calculating the estimated limits of agreement based on the standard deviation of the differences, and confidence intervals for those limits of agreement, a numerical estimate of the extent of agreement is obtained. This enables us to state that, for example, we can expect that in 95% of cases the numerical difference between the two measurements will not exceed a given level. A judgement then needs to be made about the operational significance of the numerical difference observed, and the implications for the use of the measure being tested. When comparing repeated applications of the same adverse-event measure, we would hope to see a small numerical difference between the two results and would wish the mean difference to be close to zero (indicating that neither result was consistently higher or lower than the other).

In this study, scatterplots were used to examine graphically the relationship between the number of adverse events found for patient admissions on the first and second screenings, and the associated Pearson correlation coefficients were calculated, taking into account the above provisos about their interpretation. Then, the differences between ratings and limits of agreement were calculated, as described by Bland and Altman (1986).

6.2.3 **Results and discussion**

The results from each of the interrater reliability studies - in ENT, ophthalmology and obstetrics - are presented in turn below. Then, the findings from the intrarater study in obstetrics are explored. Finally, a comparison of the results in each of the studies is made.

Interrater reliability in ENT

A total of 146 patient admissions were screened twice by different members of the project staff, and the results are presented in table 6.4 below. For each screening criterion, the table shows how many patient admissions were recorded as no adverse event (N-N) by both screeners; as adverse event by

one and no adverse event by the other (N-V and V-N); and as an adverse event by both screeners (V-V). It then shows the crude percentage agreement, and the value and significance of the P statistic. As was noted earlier, in some cases P cannot be calculated because the cross-tabulation of screeners' results contains only one non-zero row or column, resulting in an infinite denominator in the equation for the P statistic.

Crit no	Criterion title	N-N	N-V	V-N	V-V	Agreement (%)	P	Significance of P
1	Adm for adv results o/p mgt	141	1	3	1	97	0.32	< 0.001
2	Readmission for comp prev adm	139	1	1	5	99	0.83	< 0.001
3	Error in operative consent	122	8	13	3	86	0.15	< 0.001
4	Unpl rem/inj/repair in surg	145	0	1	0	99	-	-
5	Unpl return to theatre	146	0	0	0	100	-	-
6	Path/hist varies from diag	146	0	0	0	100	-	-
7	Prob of transfusion	146	0	0	0	100	-	-
8	Hosp acquired infection	145	0	1	0	99	-	-
9	Medication error/reaction	132	8	4	2	92	0.21	< 0.01
10	Cardiac/resp arrest in hosp	146	0	0	0	100	-	-
11	CVA/MI/PE in hosp after surg	145	1	0	0	99	-	-
12	Unexp transfer to spec care	145	1	0	0	99	-	-
13	Pt related clinical complcn	140	1	2	3	98	0.66	< 0.001
14	Non-clin problem/incident	128	5	0	13	97	0.82	< 0.001
15	Neuro deficit devel in hosp	146	0	0	0	100	-	-
16	Unexp patient death	146	0	0	0	100	-	-
17	Medical record deficiency	79	20	17	30	75	0.43	< 0.001
18	Nursing record deficiency	42	18	16	70	77	0.52	< 0.001
19	Pt/family dissatisfaction	138	1	1	6	99	0.85	< 0.001
20	Discharge related problems	140	1	4	1	97	0.27	< 0.001
E1	Early adm for elect proc	146	0	0	0	100	-	-
E2	Prep-op problems	129	4	1	12	97	0.81	< 0.001
E3	Probs supply of theatre eqpt	146	0	0	0	100	-	-
E4	Missing/incompl audit sheet	128	5	9	4	90	0.31	< 0.001
All	All criteria	3206	75	73	150	96	0.65	< 0.001

Table 6.4. Agreement statistics by criterion for interrater study in ENT.

The first point to note is that for twelve of the 24 adverse-event criteria used in the measure being tested in ENT, P could not be calculated as noted above, largely because no adverse events were recorded in the sample of 146 patient admissions. Many of the criteria in the adverse-event measure being tested related to relatively rare adverse events, and so it was not a cause for concern that no events were found in a fairly small sample. However, this means that the study is effectively

not able to test the reliability of those criteria within the measure beyond calculating raw percentage agreements as shown in the table.

However, for the remaining 12 criteria, P varied from 0.15 to 0.85. In each case, P was significantly greater than zero, indicating that the level of agreement was significantly better than that expected through chance alone. Comparing the P values in table 6.4 with the benchmarks set out in table 6.3, it can be seen that 5 of the 12 criteria have P values rated as "good" or "very good" (eg above 0.6) while only one falls into the "poor" category (less than 0.2). Taking the measure as a whole, the overall P statistic (based on an aggregation of the results from all 24 criteria) was 0.65. These results suggest that the overall reliability of the measure is good, but that some criteria within the measure are much less reliable than others. There is clearly potential to improve the reliability of the measure by either adjusting the definitions of those criteria or by removing them from the measure.

Figure 6.1 below contains a scatterplot of the number of adverse events found on the first screening against the number of adverse events found on the second screening for the 146 admissions in ENT. Each data point on the scatterplot represents a single patient admission. Where one or more data points overlap on the graph, a "sunflower plot" is used, with each petal or line representing one overlapping case, so that the clustering of data points can be seen.

The graph shows agreement between the number of adverse events on first and second screening for many data points, though there are a number of data points representing substantial disagreements (for example, one case with 0 adverse events on first screening and 3 on second screening). Overall, the number of adverse events found on first and second screening was the same for 51.4% of cases, and was within a range of ±1 adverse event for 89.9% of cases. The Pearson correlation coefficient was 0.66, indicating moderately strong association (though the limited usefulness of this statistic in assessing agreement was noted earlier).

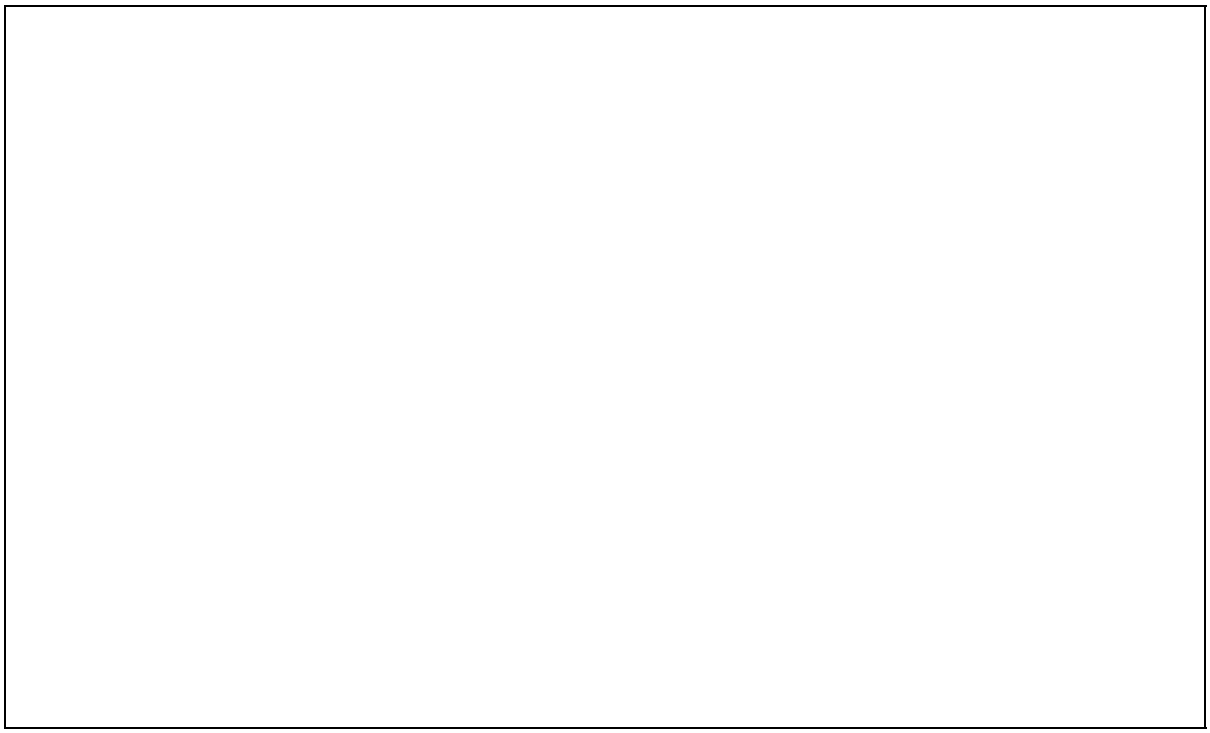


Figure 6.1. Scatter plot of number of adverse events found on first and second screening for interrater study in ENT.

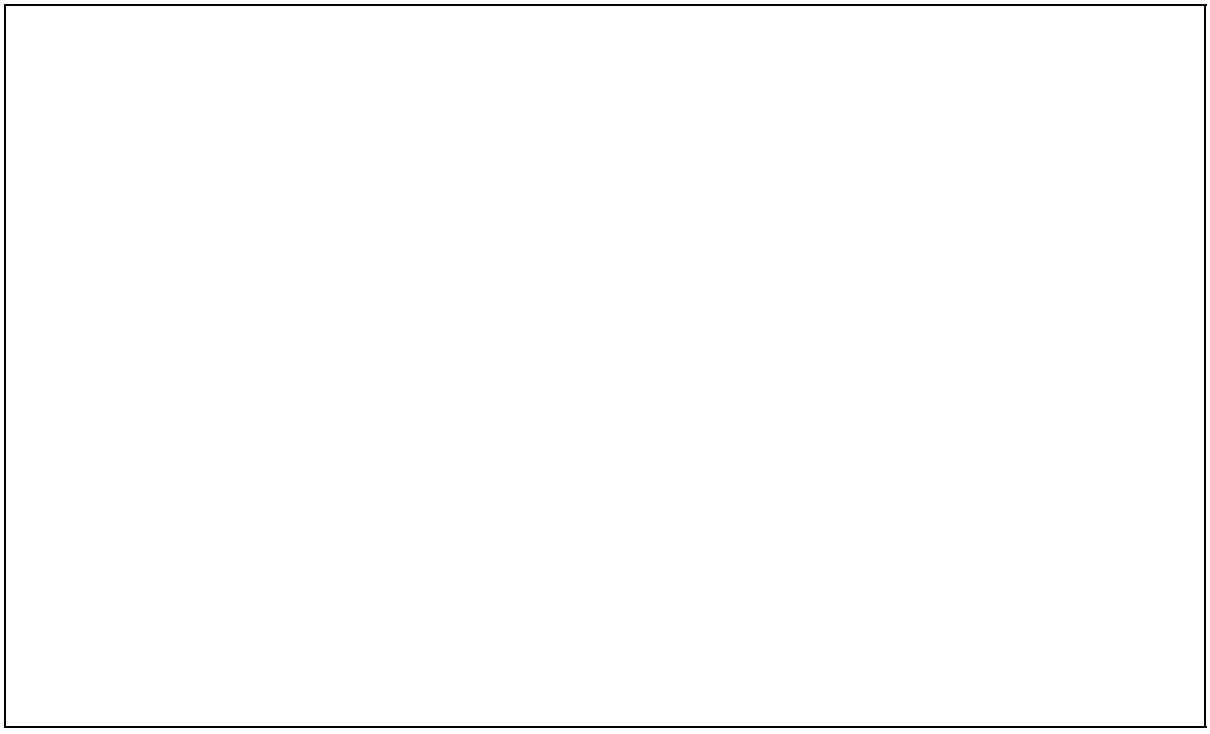


Figure 6.2. Difference in number of adverse events on first and second screening plotted against mean number of adverse events on first and second screening for interrater study in ENT.

Figure 6.2 above shows a second scatterplot, of the difference between the number of adverse events found on first and second screening against the mean of the two values. It can again be seen that the degree of agreement is relatively good, with most data points on or close to the 0 line on the vertical axis. No correlation between the mean and difference is visually evident, indicating that the level of agreement appears to be constant across cases with differing numbers of adverse events. The mean difference is 0.01, suggesting that there was no tendency for the adverse-event score from the second screening to be either consistently higher and lower than that from the first screening, and so no systematic bias exists. The estimated limits of agreement, within which 95% of all differences should fall, are -1.83 to +1.81 (with 95% confidence intervals for these estimates of -2.09 to -1.57, and 1.55 to 2.07 respectively). This means that we could expect that the number of adverse events found on a second application of this adverse-event measure would be within about ±1.83 of the number found on the first application in 95% of cases. In the worst case, taking the outer 95% confidence limits, we might conclude that the difference between adverse-event scores on two separate screenings would not exceed 2 in about 95% of cases. Whether this level of reliability is acceptable depends in large part on the purpose to which the data is to be put, and this is discussed later in this section.

Interrater reliability in ophthalmology

A total of 120 patient admissions were screened twice by different members of the project staff, and the results are presented in table 6.5 below. Once again, the table shows for each criterion how many patient admissions were recorded as no adverse event by both screeners (N-N); as an adverse event by one screener but not the other (N-V and V-N); and as an adverse event by both screeners (V-V). The crude percentage agreement, and the value and significance of P are also shown.

Crit no	Criterion title	N-N	N-V	V-N	V-V	Agreement (%)	P	Significance of P
1	Adm for adv results o/p mgt	120	0	0	0	100	-	-
2	Readmission for comp prev adm	118	0	0	2	100	1.00	< 0.001
3	Error in operative consent	117	1	1	1	98	0.49	< 0.001
4	Unpl rem/inj/repair in surg	106	3	4	7	94	0.63	< 0.001
5	Unpl return to theatre	116	0	2	2	98	0.66	< 0.001
6	Path/hist varies from diag	120	0	0	0	100	-	-
7	Prob of transfusion	120	0	0	0	100	-	-
8	Hosp acquired infection	118	2	0	0	98	-	-
9	Medication error/reaction	115	3	2	0	96	-0.02	-
10	Cardiac/resp arrest in hosp	120	0	0	0	100	-	-
11	CVA/MI/PE in hosp after surg	120	0	0	0	100	-	-
12	Unexp transfer to spec care	120	0	0	0	100	-	-
13	Pt related clinical complcn	120	0	0	0	100	-	-
14	Non-clin problem/incident	117	1	2	0	98	-0.01	-
15	Neuro deficit devel in hosp	120	0	0	0	100	-	-
16	Unexp patient death	120	0	0	0	100	-	-
17	Medical record deficiency	94	10	9	7	84	0.33	< 0.001
18	Nursing record deficiency	42	18	16	70	77	0.52	< 0.001
19	Pt/family dissatisfaction	113	0	5	2	96	0.43	< 0.001
20	Discharge related problems	119	0	1	0	99	-	-
OP1	Probs of cataract extraction	118	0	1	1	99	0.66	< 0.001
OP2	Specific post-op complications	96	4	11	9	88	0.48	< 0.001
OP3	Surg proc with tissue biopsy	119	1	0	0	99	-	-
OP4	Prob rel to drug usage	115	2	2	1	97	0.32	< 0.001
OP5	Oph nursing record review	93	10	10	7	83	0.31	< 0.001
OP6	Sore throat after GA	111	0	2	7	98	0.87	< 0.001
All	All criteria	2907	55	68	116	96	0.63	< 0.001

Table 6.5. Agreement statistics by criterion for interrater study in ophthalmology.

As in the results presented for ENT, the small sample size and the low incidence of some types of adverse events prevented the calculation of P for 12 of the 26 criteria used in the measure. For 12 of the remaining 14 criteria, P varied from 0.31 to 1.00, but two criteria had P values of -0.01 and -0.02 suggesting no agreement beyond that expected by chance. These results place 5 of the criteria in the "good" or "very good" agreement categories from table 6.3, 7 in the "fair or moderate agreement" categories and 2 in the "poor agreement" category. For the measure as a whole, the P statistic based on an aggregation of results from all 26 criteria was 0.63. Clearly, the overall reliability of the measure is good, but there are some criteria within it which perform substantially less well than others. Once again, there is potential to improve the reliability of the measure by

either adjusting the definitions of those criteria with poor reliability or by removing those criteria from the measure.

Figure 6.3 below presents a scatterplot of the total number of adverse events found on first and second screening for the sample of 120 patient admissions, each admission being represented as a data point on the graph with "sunflowers" used to show where multiple data points overlap. Overall, in 55.2% of cases the same number of adverse events were found on first and second screenings, and in 92.8% of cases the number of adverse events found was within a range of ±1 adverse event. The Pearson correlation coefficient was 0.57, though once again the limited value of this statistic should be noted.

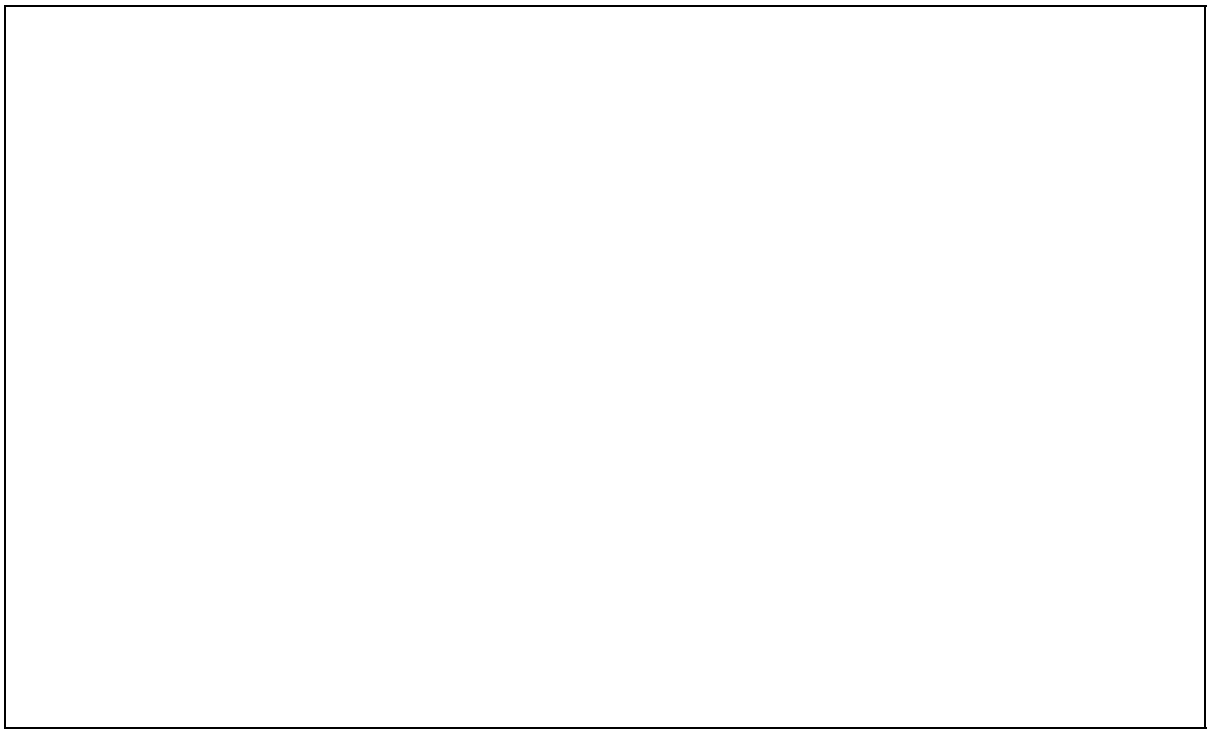


Figure 6.3. Scatter plot of number of adverse events found on first and second screening for interrater study in ophthalmology.

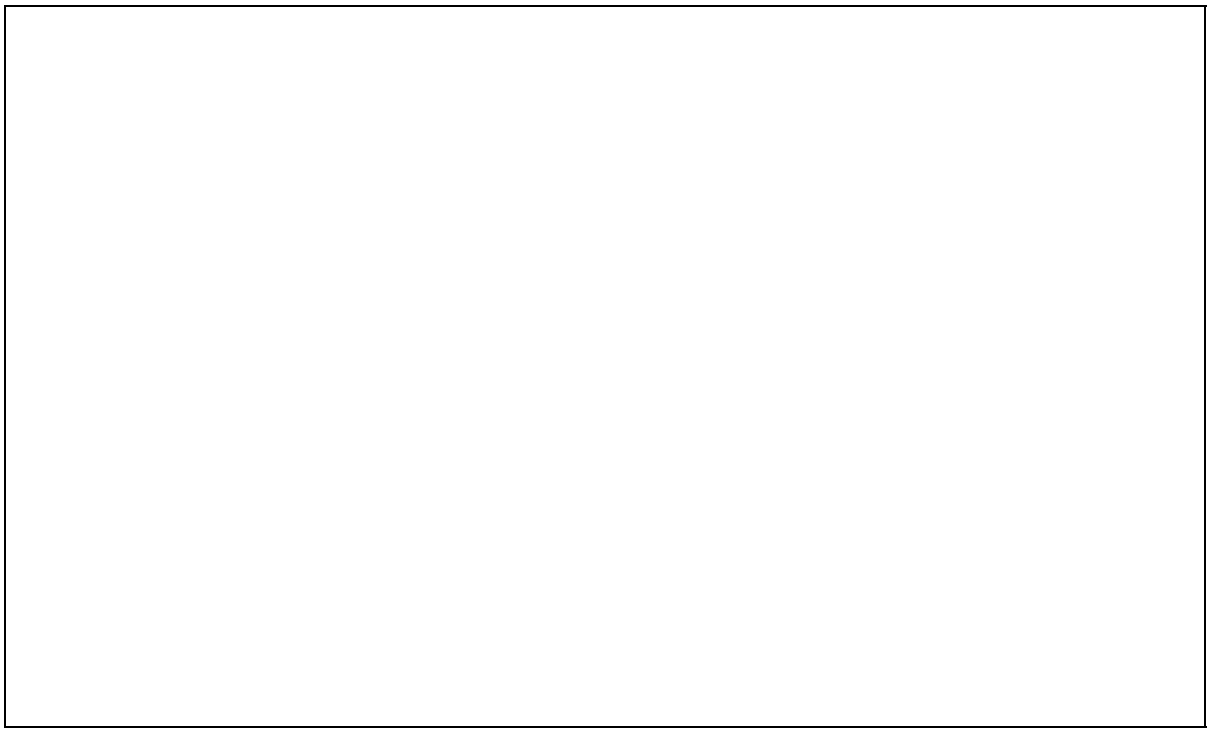


Figure 6.4. Difference in number of adverse events on first and second screening plotted against mean number of adverse events on first and second screening for interrater study in ophthalmology.

Figure 6.4 shows a second scatterplot, of the difference in the number of adverse events found on first and second screening against the mean of the two values. It shows relatively good agreement, with most data points on or close to the 0 line on the vertical axis (denoting no difference, or complete agreement). The mean difference is 0.16, suggesting that little or no systematic bias exists.

The estimated limits of agreement, within which 95% of all differences should fall, are -1.79 to 1.47 (with 95% confidence intervals for these estimates of -2.04 to -1.54 and 1.22 to 1.72 respectively). In other words, we could expect that in 95% of cases, a second application of the adverse-event measure used in ophthalmology will result in an adverse-event score which is between 1.79 less than the first to 1.47 more than the first. The scatterplot in figure 6.4 is slightly asymmetric along the x axis, which might suggest a weak relationship between the magnitude of the difference and the mean. As a result these estimated limits of agreement might be too wide for lower numbers of adverse events, and too narrow for higher numbers.

Interrater reliability in obstetrics

A total of 117 patient admissions in obstetrics were used in the interrater and intrarater reliability studies in obstetrics. Each admission was screened three times - twice by a member of the project staff and a third time by a doctor qualified in obstetrics who first received a brief training programme on how to interpret and apply the adverse event measure. It was therefore possible to make two assessments of the interrater reliability of the adverse event measure used in obstetrics, by comparing the results of each of the two project staff screenings in turn with the result of the screening by a doctor with obstetric experience. The results of these two comparisons are reported below. In practice, not all 117 of the patient admissions were screened three times, for a variety of logistic reasons to do with the availability of records, the project staff and the doctor involved in the study. As a result, there were 108 admissions for which there was data from both the first project staff screening and the clinician screening; and 110 admissions for which there was data from both the second project staff screening and the clinician screening.

Table 6.6 shows the results obtained by comparing the first screening by the member of project staff with the doctor's screening. The table shows, for each criterion, how many patient admissions were recorded as having no adverse event by both screener and doctor (N-N); as having an adverse event by the screener but not having an adverse event by the doctor (V-N); as having no adverse event by the screener but as having an adverse event by the doctor (N-V); and as having an adverse event according to both the screener and the doctor (V-V). The crude percentage agreement and the value and significance of P are also shown.

Crit no	Criterion title	N-N	N-V	V-N	V-V	Agreement (%)	P	Significance of P
1	Management of SROM	97	4	2	5	94	0.59	< 0.001
2	Elective induction of labour	88	7	1	12	93	0.71	< 0.001
3	Problems of labour/delivery	90	15	1	3	86	0.25	< 0.001
4	Caesarean section	87	2	1	18	97	0.91	< 0.001
5	Problems of Caesarean section	102	3	1	2	96	0.48	< 0.001
6	Perinatal problems/complications	98	7	2	1	92	0.15	-
7	Post-natal problems/complications	98	6	0	4	94	0.55	< 0.001
8	Drug-related problems	75	31	0	1	71	0.04	-
9	Mother/family dissatisfaction	107	1	0	0	99	-	-
10	Non-clinical prob/incidents	87	3	13	5	85	0.31	< 0.001
11	Obstetric record review	75	20	4	9	77	0.31	< 0.001
12	Prob of anaesthesia	107	1	0	0	99	-	-
13	Prob of pain relief	96	2	4	6	94	0.64	< 0.001
14	All criteria	1207	102	28	66	91	0.46	< 0.001

Table 6.6. Agreement statistics by criterion for interrater study in obstetrics, comparing results from first screening by project staff and screening by doctor.

Table 6.7 shows the same analysis, this time for the comparison of the second project staff screening with the doctor's screening. Again the table shows, for each criterion, how many patient admissions were recorded as having no adverse event by both screener and doctor (N-N); as having an adverse event by the screener but not having an adverse event by the doctor (V-N); as having no adverse event by the screener but as having an adverse event by the doctor (N-V); and as having an adverse event according to both the screener and the doctor (V-V). The crude percentage agreement and the value and significance of P are also shown.

Crit no	Criterion title	N-N	N-V	V-N	V-V	Agreement (%)	P	Significance of P
1	Management of SROM	95	4	6	5	91	0.45	< 0.001
2	Elective induction of labour	91	9	0	10	92	0.65	< 0.001
3	Problems of labour/delivery	89	11	3	7	87	0.43	< 0.001
4	Caesarean section	89	1	1	19	98	0.94	< 0.001
5	Problems of Caesarean section	104	4	1	1	95	0.27	< 0.01
6	Perinatal problems/complications	101	6	0	3	95	0.48	< 0.001
7	Post-natal problems/complications	96	3	4	7	94	0.63	< 0.001
8	Drug-related problems	75	31	2	1	69	0.01	-
9	Mother/family dissatisfaction	109	1	0	0	99	-	-
10	Non-clinical prob/incidents	79	1	23	7	78	0.29	< 0.001
11	Obstetric record review	59	16	22	13	65	0.17	-
12	Prob of anaesthesia	109	1	0	0	99	-	-
13	Prob of pain relief	97	1	5	7	95	0.67	< 0.001
14	All criteria	1193	89	67	80	89	0.45	< 0.001

Table 6.7. Agreement statistics by criterion for interrater study in obstetrics, comparing results from second screening by project staff and screening by doctor.

For table 6.6 on the first project staff/doctor screening comparison, P varies from 0.04 to 0.91. Of the 12 criteria for which P could be calculated, only 3 fall into the "good" or "very good" agreement categories from table 6.3, while 7 have "fair" or "moderate" agreement and 2 "poor" agreement. Overall, the P statistic for the adverse-event measure as a whole was 0.46. In table 6.7, for the second project staff/doctor comparison, the results are similar. Here, P varies from 0.01 to 0.94. Again, 12 criteria had P values, of which only 4 fall into the "good" or "very good" agreement categories, while 6 have "fair" or "moderate" agreement and 2 "poor" agreement. In this case, the P statistic for the adverse-event measure as a whole was 0.45.

It is evident from tables 6.6 and 6.7 that the interrater reliability results for the adverse-event measure used in obstetrics was considerably lower than that found in either ENT or ophthalmology. This could indicate that either the measure itself is less reliable, or that the two screeners using the measure were applying it or interpreting it differently. There is some evidence to support the latter conclusion, since across most criteria the doctor found more adverse events than the project staff. This tendency was particularly pronounced for the criterion concerning drug-related problems (such as missed doses, prescribing and administration errors, etc), where the doctor found far more

adverse events. It may be that the doctor was able to use his clinical knowledge to interpret findings from the records, or it may be that this systematic bias reflects significant differences in the interpretation of the definition of the adverse-event measure or of information in patient records by the doctor and the member of project staff.

Figures 6.5 and 6.6 below show scatterplots of the total number of adverse events found by different screenings. Figure 6.5 compares the first project staff screening with the doctor's screening, for the sample of 108 patients analysed in table 6.6 above. Figure 6.6 compares the second project staff screening with the doctor's screening, for the 110 patients reported on in table 6.7. In both figures, each data point represents one admission and multiple overlapping data points are represented by "sunflower" plots.

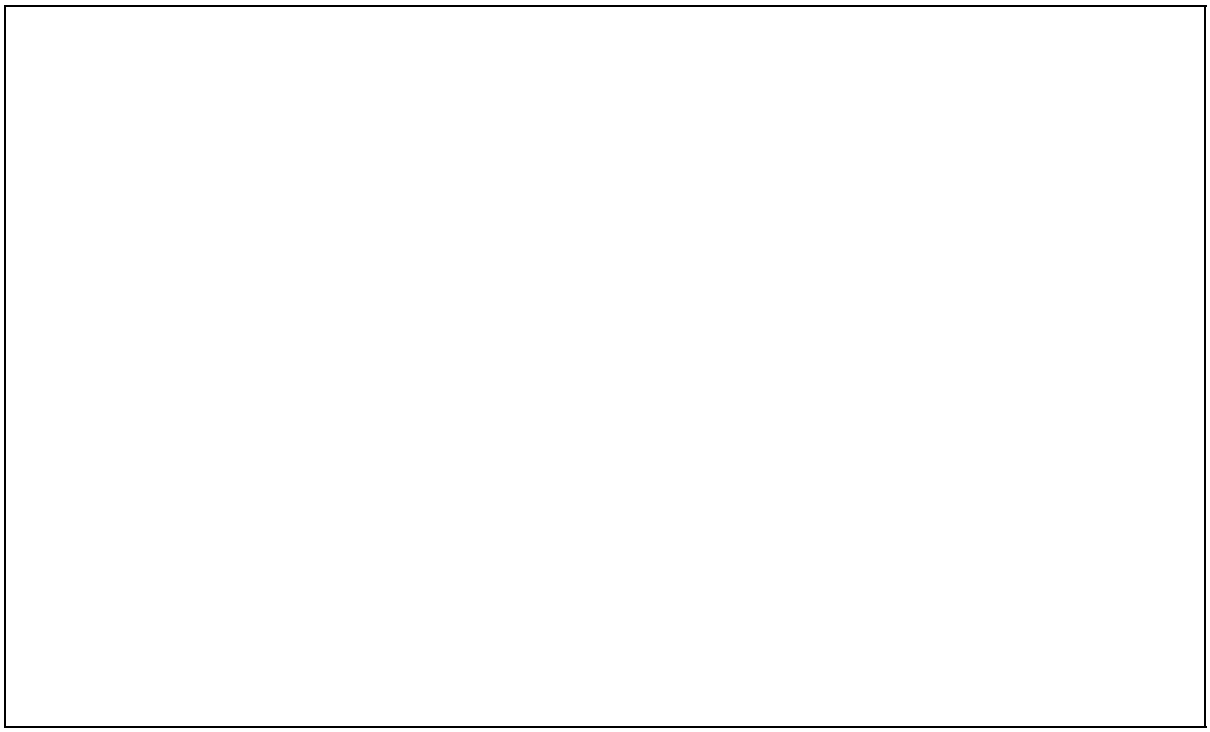


Figure 6.5 Scatter plot of number of adverse events found on first project staff screening and doctor's screening for interrater study in obstetrics.

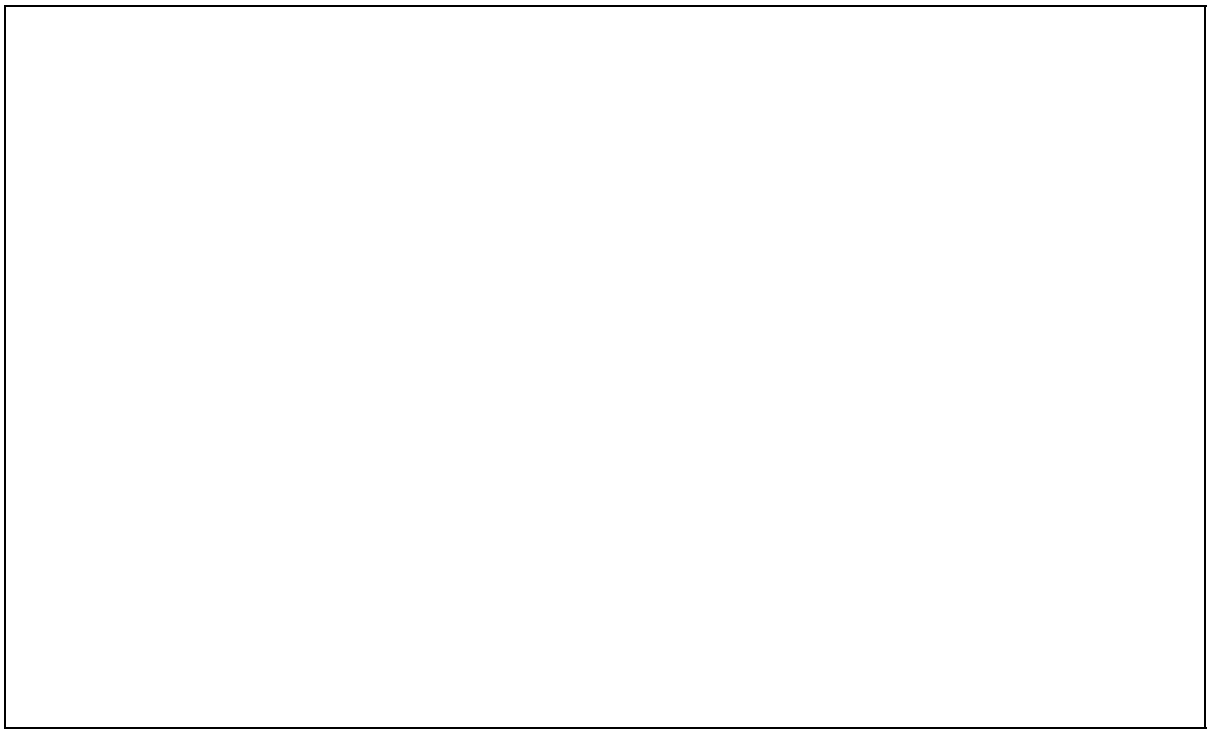


Figure 6.6 Scatter plot of number of adverse events found on second project staff screening and doctor's screening for interrater study in obstetrics.

These two scatterplots show much more disagreement than previous figures (6.1 and 6.3) for ENT and ophthalmology. They confirm the existence of a consistent bias, with the doctor finding more adverse events than the member of project staff. When the first project staff screening and the doctor's screening are compared, only in 35.9% of cases did they agree about the number of adverse events found; in 70.1% of cases, the two ratings were within a range of ±1 adverse event. Comparing the second project staff screening and the doctor's screening gives similar results - they agreed on the number of adverse events in 29.9% of cases, and were within a range of ±1 adverse event in 71.0% of cases. The Pearson correlation coefficients for these two comparisons were 0.61 and 0.57 respectively.

Figures 6.7 and 6.8 below show two further scatterplots for the comparisons of project staff screening results with the doctor's screening results. The difference between the number of adverse events found is plotted against the mean of the two values.

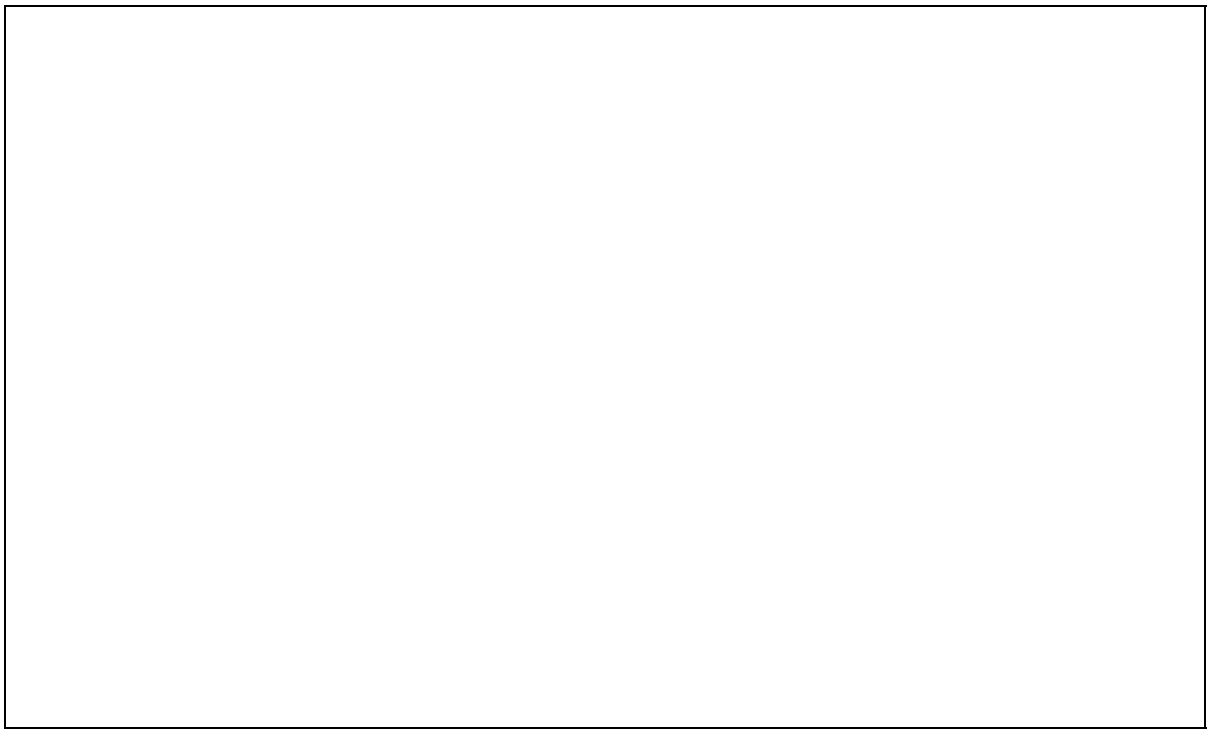


Figure 6.7 Scatter plot of difference in number of adverse events on first project staff screening and doctor's screening plotted against mean number of adverse events for interrater study in obstetrics.

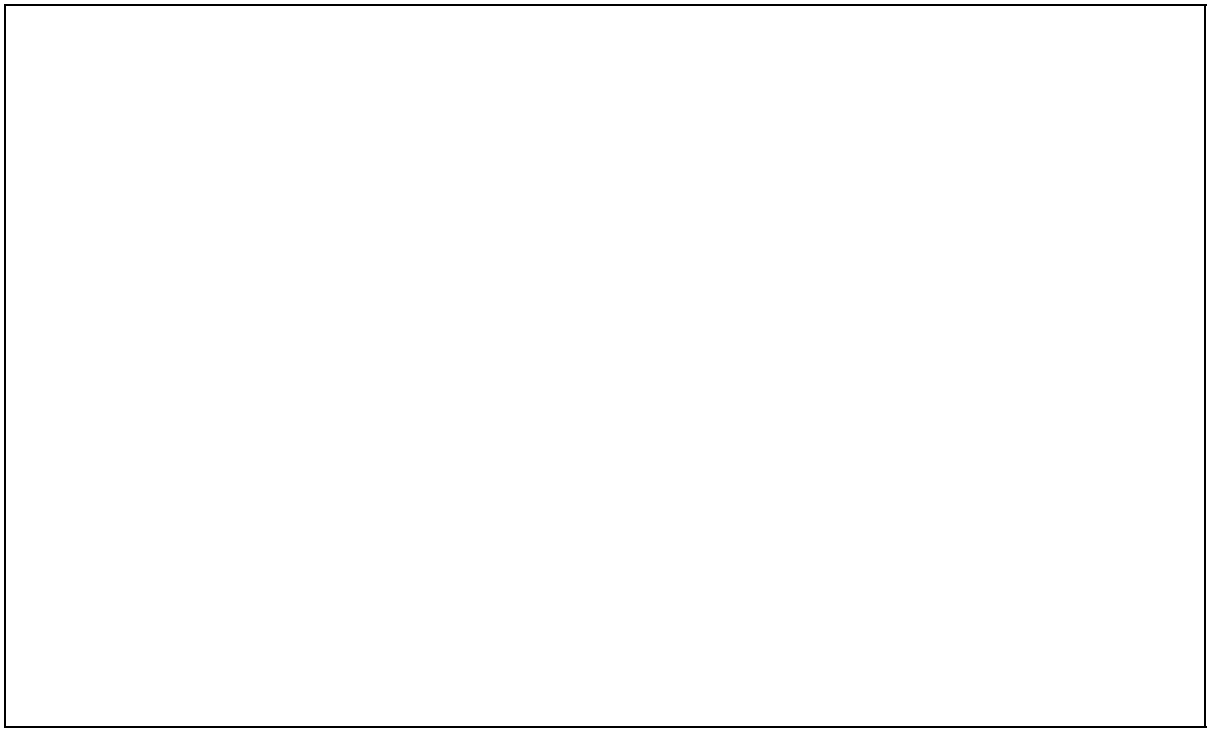


Figure 6.8. Scatter plot of difference in number of adverse events on second project staff screening

and doctor's screening plotted against mean number of adverse events for interrater study in obstetrics.

These plots both show a far less consistent and reliable agreement between the two screenings than was seen in figures 6.2 and 6.4 for the studies in ENT and ophthalmology. The mean differences are 0.69 and 0.20 respectively, suggesting that some systematic bias may exist, with the doctor identifying more adverse events than the member of project staff. The estimated limits of agreement, within which 95% of all differences should fall, are also wider than those found in ENT and ophthalmology: -1.56 to 2.94 for the comparison of first project staff screening and doctor's screening (95% confidence intervals -1.94 to -1.18 and 2.56 to 3.32), and -2.27 to 2.67 for the second project staff/doctor' screening comparison (95% confidence intervals -2.68 to -1.88 and 2.26 to 3.08).

Intrarater reliability in obstetrics

A total of 110 patient admissions were screened twice by a single member of the project staff. The first screening and second screening were separated by about 4 months, a period judged sufficient for the member of staff concerned, who was concurrently involved in screening in several other specialties, to recollect any details of the cases or the results of the first screening when screening them for the second time. The results are presented in table 6.8 below. For each screening criterion, the table shows how many patient admissions were recorded as no adverse event (N-N) on both the first and second screening; as adverse event on first screening but not on second (V-N); as adverse event on second screening but not on first (N-V); or as adverse event on both screenings (V-V). It also gives the crude percentage agreement and the value and significance of P.

Crit no	Criterion title	N-N	N-V	V-N	V-V	Agreement (%)	P	Significance of P
1	Management of SROM	97	6	2	5	92	0.52	< 0.001
2	Elective induction of labour	97	0	3	10	97	0.85	< 0.001
3	Problems of labour/delivery	98	9	1	2	91	0.25	< 0.001
4	Caesarean section	90	1	0	19	99	0.96	< 0.001
5	Problems of Caesarean section	106	1	2	1	97	0.39	< 0.001
6	Perinatal problems/complications	105	2	3	0	95	-0.02	-
7	Post-natal problems/complications	99	7	0	4	94	0.51	< 0.001
8	Drug-related problems	107	2	0	1	98	0.49	< 0.001
9	Mother/family dissatisfaction	110	0	0	0	100	-	-
10	Non-clinical prob/incidents	77	14	3	16	84	0.56	< 0.001
11	Obstetric record review	69	28	6	7	69	0.14	-
12	Prob of anaesthesia	110	0	0	0	100	-	-
13	Prob of pain relief	94	5	3	8	93	0.63	< 0.001
14	All criteria	1259	75	23	73	93	0.56	< 0.001

Table 6.8. Agreement statistics by criterion for intrarater study in obstetrics.

For the 12 criteria for which P could be calculated, values varied from -0.02 to 0.96. Using the benchmarks set out in table 6.3, 3 criteria are placed in the "good" or "very good" agreement categories, 7 in the "fair" or "moderate" categories and 2 in the "poor" agreement category. For the measure as a whole, the P value was 0.56, suggesting moderately good agreement.

Figure 6.9 below presents a scatterplot of the total number of adverse events found on first and second screening for the sample of 110 patient admissions, each admission being represented as a data point on the graph with "sunflowers" used to show where multiple data points overlap. Overall, in 46.8% of cases the same number of adverse events were found on first and second screenings, and in 87.3% of cases the number of adverse events found was within a range of ±1 adverse event. The Pearson correlation coefficient was 0.70, though once again the limited value of this statistic should be noted.

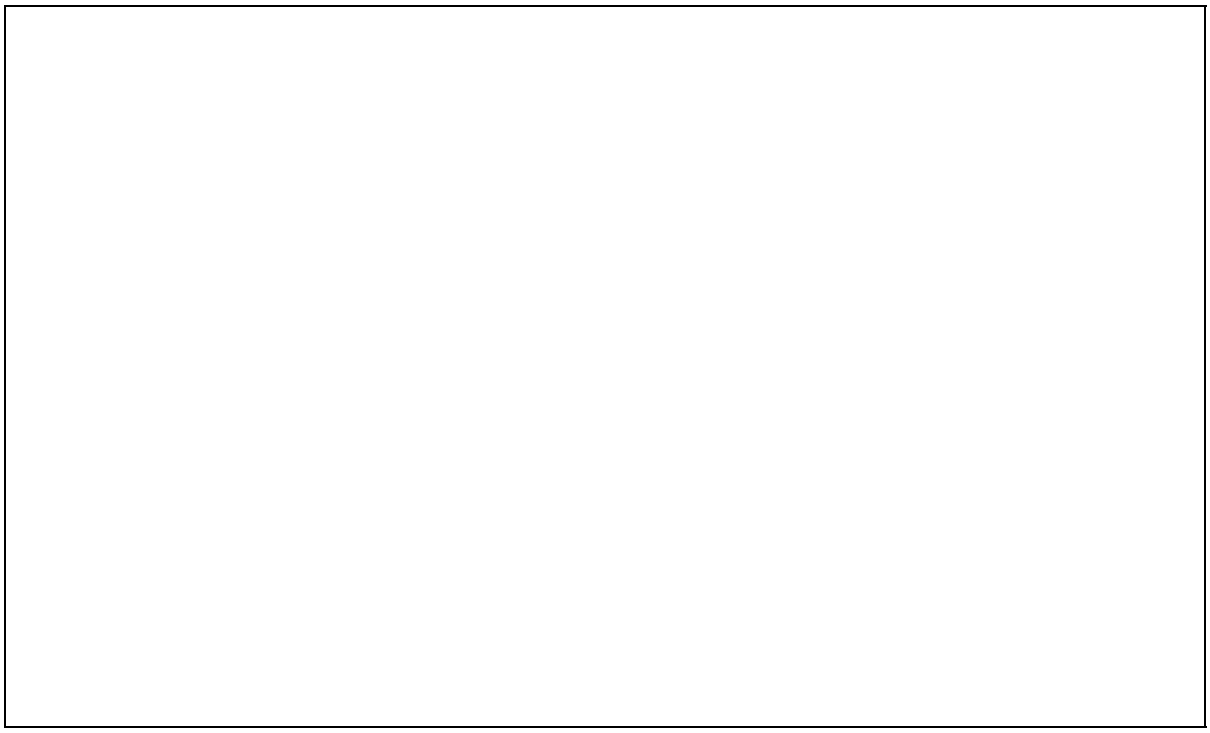


Figure 6.9. Scatter plot of number of adverse events found on first and second screening for intrarater study in obstetrics.

Figure 6.10. Difference in number of adverse events on first and second screening plotted against mean number of adverse events on first and second screening for intrarater study in obstetrics.

Figure 6.10 presents a second scatterplot, of the difference in the number of adverse events found on first and second screening against the mean of the two values. Agreement is moderately good, though some systematic bias is visually evident, with a tendency for the second screening to find more adverse events than the first. Indeed, the mean numerical difference between first and second screening was 0.48.

The estimated limits of agreement, which contain 95% of all differences, are -0.88 to 2.32 (with 95% confidence intervals of -1.19 to -0.57 and 2.01 to 2.63 respectively). In other words, we could expect that in 95% of cases the difference between the first and second screening would be in the range -0.88 to 2.32.

Summary of results of interrater and intrarater studies

The series of tables and figures presented above provide a detailed but complex account of the levels of agreement found in a number of different but related studies. In order to draw some general conclusions about the reliability of the adverse-event measures tested , table 6.9 presents a summary of the results from all the studies, in a form which supports the making of comparisons across them. For each study, the table lists some basic information about the study (cross-referenced to table 6.1 which described the studies in more detail); the levels of agreement found for criteria within the adverse-event measures being tested, using the statistic; and the level of agreement on overall adverse-event scores using the Pearson correlation coefficient and estimated limits of agreement.

Study (see table 6.1 for further details)	A	B	C	D	E
Specialty	ENT	Ophthal-mology	Obstetrics	Obstetrics	Obstetrics
Reliability testing undertaken	Interrater	Interrater	Interrater	Interrater	Intrarater
No of patient admissions in sample	146	120	108	110	110
Overall P statistic	0.65	0.63	0.46	0.45	0.56
Proportion of criteria with "good" or "very good" agreement (P > 0.60)	5/12	5/12	3/12	4/12	3/12
Proportion of criteria with "poor" agreement (P ≤ 0.20)	1/12	2/12	2/12	2/12	2/12
Pearsons correlation coefficient for adverse event scores	0.66	0.57	0.61	0.57	0.70
Mean numerical difference between adverse event scores	0.01	-0.16	0.69	0.20	0.48
Proportion of cases where adverse event scores agree	51.4	55.2	35.9	29.9	46.8
Proportion of cases where adverse event scores within range of ±1 adverse event	89.9	92.8	70.1	71.0	87.3
Estimated limits of agreement	-1.83 to 1.81	-1.79 to 1.47	-1.56 to 2.94	-2.27 to 2.67	-0.88 to 2.32
Worst case limits of agreement based on outer 95% confidence intervals	-2.09 to 2.07	-2.04 to 1.72	-1.94 to 3.32	-2.68 to 3.08	-1.19 to 2.63

Table 6.9. Summary of results from studies of interrater and intrarater reliability of adverse-event measures of quality

The findings summarised in table 6.9 in part confirm the results of previous studies of the reliability of adverse-event measures of quality, reviewed in chapter 3, though a more complex and detailed picture of the reliability of these measures also emerges. Taken together, they suggest that the measures tested are of moderate to good reliability, though there is undoubted scope for improvement. The overall P agreement statistic ranged from 0.45 to 0.65, and in each of the studies there were only one or two criteria within the adverse-event measures tested for which agreement was poor (P ≤ 0.20). The low incidence rates for some adverse events prevented the calculation of P for some criteria, and for all the adverse-event measures tested will have constrained the value that P might reach, perhaps understating the reliability of the measures.

The best reliability results were obtained in interrater tests in ENT and ophthalmology, followed by the intrarater test in obstetrics. The least good reliability results were found in the interrater studies

in obstetrics. It seems likely that both the definition of the adverse-event measures used and the way they were applied will have affected these results. It might be argued that the generally lower reliability found for the obstetric adverse-event measure suggests it was less well designed and specified than the largely generic adverse-event measure used in ENT and ophthalmology. It could also be said that the low reliability found in the two interrater studies in which a doctor undertook one of the screenings suggests that different raters, with different clinical backgrounds and training programmes may produce very different results with obvious adverse consequences for the reliability of measurement. We might attribute the differences found to the brevity and implied inadequacy of the doctor's training to undertaking the task (compared with the extensive experience of the member of project staff who undertook the other screening). However, it could also be true that the doctor's more extensive clinical knowledge and experience enabled him to identify adverse events which the member of project staff would overlook.

Overall, the reliability statistics for the five studies were broadly in line with each other, suggesting that the reliability of these adverse-event measures may not vary much from specialty to specialty, and so these results could be extrapolated with caution to other specialties.

It was interesting to note that in the single intrarater reliability study in obstetrics, the second screening by the member of project staff yielded more adverse events on average than the first. Because the study was organised retrospectively, the first set of screening data was collected as part of the routine working of the RSCH project, with no foreknowledge that those cases would be subject to this reliability study. However, the second set of data was collected by the member of project staff in the knowledge that the reliability of the measure, and her application of it, was being examined. When the two were compared, the second screening data contained more adverse events than the first screening data. This provides some evidence to support the contention that the awareness of reliability testing in studies such as these may change rater behaviour (perhaps making them more careful and thorough, and so causing them to find more adverse events), and so affect the estimates of reliability that are obtained. It illustrates the value of observational studies of reliability, which make use of routine data, such as those presented later in this chapter.

6.2.4 **Conclusions**

These studies indicate that the adverse-event measures tested were of moderate to good reliability, in the settings in which they were examined. Whether they are sufficiently reliable, depends on the purpose to which they are to be put, and the degree of error or variation which can be tolerated. While this judgement is inevitably subjective, two points should be borne in mind. Firstly, these measures are generally intended to be used across groups of patients, to identify deficiencies in the quality of care, rather than to make decisions about individual patients. The presence of some random variation, which might make the latter difficult, need not be as great a concern in the former, since the statistical importance of that variation will diminish as the group size grows. Secondly, the reliability of these measures must be set alongside the reliability of other approaches to assessing the quality of care. It was noted in chapter 2 that the definitional difficulties of quality measurement, and the practitioner-led tradition of development in quality measurement had left many quality measures poorly defined, researched and validated. In this context, the moderately good reliability of these adverse-event measures may be viewed quite positively.

It is clear from these studies that the reliability of adverse event measures of quality can be improved, by adjusting their definitions and by paying greater attention to rater training. The wide range of reliability statistics for individual criteria within the measures suggest that some criteria definitions needed to be revised, or removed from the measures. While the impact of changes to the criteria which make up the adverse event measures on the validity of those measures should be taken into account, there is certainly potential to maximise reliability by paying particular attention to those areas where low levels of agreement were found. It may be possible to make the definitions clearer and less ambiguous, and so to improve reliability. These studies also highlight the important contribution that raters' backgrounds, existing knowledge, and training are likely to make to the reliability of data collected using an adverse-event measure. The use of a doctor, with significantly greater clinical knowledge and skills but significantly less training and experience in abstracting data on adverse events from patients' records, resulted in rather poor reliability results.

Given that the reliability of adverse-event measures of quality seems to vary from setting to setting, and to be influenced by many factors apart from the construction and definition of the measures

themselves, there is a strong argument for incorporating ongoing reliability testing into any routine use of such measures. In other words, while research studies such as these may be helpful in testing the reliability of adverse-event measures of quality and providing some insight into the factors which contribute towards a high level of reliability, it may be unwise to rely upon them for assurance that in practice, the use of these measures will be reliable. It may be advisable to undertake periodic sampled rescreening in order to monitor reliability, and to use mechanisms like regular workshops, case presentations and refresher training among raters to maintain reliability.

6.3 Analysis of interrater variation in the RSCH project data

6.3.1 **Aims of study of interrater variation**

The aim of the study reported in this section was to assess the reliability of a number of adverse-event measures of quality by examining the data collected through the RSCH occurrence screening project, which was outlined in chapter 3. The data set produced by that project, which was available for analysis in this study, was described in section 5.2.2

In that project, 14,815 patient admissions in 12 specialties were screened using a number of different adverse-event measures. Each patient admission was screened once, by one of four members of the project staff. This study set out to examine the data from screening, and to identify any differences in the numbers and types of adverse events recorded during screening which might be attributable to screener variation.

6.3.2 **Method**

The data set available for study from the RSCH occurrence screening project has already been described in chapter 5. During the life of the project between February 1990 and April 1992, 14,815 patient admissions were screened by four members of the project staff. These staff were all qualified nurses, and had all undergone a training programme in using the adverse-event measures employed in the project. The data was collected using a purpose designed computer database,

which automatically recorded the identity of the member of staff who screened each patient admission. However, this automatic recording was only introduced part way through the project, around January 1991. As a result, of the 14,815 patient admissions screened during the project, only 6,708 admissions had the identity of the screener recorded and were therefore suitable for analysis in this study. These 6,708 admissions were spread across 12 specialties, but were not distributed evenly. In some specialties there were small numbers of cases, which could not support a meaningful statistical comparison of adverse-event rates between screeners, and so analysis was restricted to those specialties in which a sample of at least 400 patient admissions was available for study. An overview of the resulting data set of 6,095 admissions in 8 specialties is presented in table 6.10, which shows the numbers of patient admissions screened by each member of project staff, analysed by specialty.

Specialty	Member of project staff								All staff	
	A		B		C		D			
	No	%	No	%	No	%	No	%	No	%
Accident and emergency	37	5.9	84	13.3	237	37.6	273	43.3	631	10.4
ENT	-	-	116	14.5	239	29.9	445	55.6	800	13.1
Gynaecology	8	1.4	68	12.0	191	33.8	298	52.7	565	9.3
Obstetrics	151	10.6	10.0	41.8	597	41.8	537	37.6	1428	23.4
Ophthalmology	36	3.5	33	3.2	346	33.5	619	59.9	1034	17.0
General surgery	-	-	73	13.3	235	42.8	241	43.9	549	9.0
Trauma and orthopaedics	95	14.7	12	1.9	324	50.2	215	33.3	646	10.6
Urology	31	7.0	45	10.2	181	41.0	185	41.9	442	7.3
All specialties	358	5.9	574	9.4	2350	38.6	2813	46.2	6095	100

Table 6.10. Numbers of patient admissions screened by project staff, analysed by specialty.

Table 6.10 shows that the screening workload was not distributed evenly across the four project staff involved in screening during the life of the project (referred to as A, B, C and D in the table), with

each screening about a quarter of the patient admissions as might be expected. One member of staff (B) was part-time, and another member of staff (A) left the project part way through the period being studied, and this in part accounts for the variations evident in table 6.10 which shows, for example, that screener D screened almost four times as many patient admissions as screener B. However, regardless of these differences in overall workload, some substantial and significant variations in the distribution of cases between screeners are also evident from specialty to specialty (for table 6.10, $\chi^2 = 544.5$, $p < 0.0001$). For example, screener B, who screened 9.4% of all patient admissions, screened only 1.9% of patient admissions in trauma and orthopaedics. Screener A, who screened 5.9% of cases overall was responsible for screening 10.6% of cases in obstetrics. In part, some of these variations result from an interaction between the timing of staff changes and the dates at which screening commenced or ceased in various specialties. However, they probably also reflect the project staff's working practices. Though the project staff were in theory all allocated to work in all specialties, in fact some degree of specialisation developed, with some screeners taking the lead in particular specialties. The non-random distribution of patient admissions across the four members of project staff which resulted should be borne in mind in the analysis of interrater variations which follows, since it could have introduced biases which would affect those findings.

In order to examine the variation in the numbers and types of adverse events recorded by screeners, a number of bivariate and multivariate analyses were performed. Firstly, the variation in the overall adverse event rate (the number of adverse events found on screening each patient admission) between different members of project staff was assessed for each specialty, using both parametric and non-parametric techniques (the one way analysis of variance and the Kruskal Wallis analysis of variance by ranks, and Kendall's coefficient of concordance). Secondly, the variation in rates of different types of adverse events (recorded for different criteria within the adverse-event measure) was examined in two specialties - ophthalmology and ENT - by comparing the event rates recorded by different project staff for each screening criterion in the adverse event measure used in those specialties. For each criterion in the measure, χ^2 tests were used to establish whether adverse event rates varied significantly among screeners, and standardised adjusted residuals were used to identify where these differences occurred. Thirdly, multiple regression was used to identify whether linear relationships existed between the overall numbers of adverse events recorded for patient admissions and a number of other variables including patient demographics, length of stay and the identity of

the project staff involved in screening. The proportion of variation in the numbers of adverse events which could be attributed to differences between project staff was calculated. Each of these statistical approaches is described in more detail in the following sections. In some cases, these techniques have already been described and used in chapters 4 or 5.

6.3.3 **Results and discussion**

Variation in the overall number of adverse events found by members of project staff

Firstly, the sample of 6,095 patient admissions in 8 specialties described above was used to calculate the mean numbers of adverse events found per patient admission screened by the four members of the project staff (designated, as before, as A, B, C and D) in each specialty. The results are presented in table 6.11. Two statistical techniques were used to test whether the differences in numbers of adverse events found by members of the project staff were statistically significant - the parametric one way analysis of variance (ANOVA) and the non-parametric Kruskal-Wallis analysis of variance by ranks. Both approaches were used because the data did not conform to the parametric assumptions underlying the ANOVA test, but this technique has greater power to detect statistically significant differences. The ANOVA test was able to identify which differences between members of project staff were significant (using the Scheffé method for multiple pairwise comparisons). The results of these statistical tests are also presented in table 6.11.

Specialty	Screener	No of cases	Mean number of adverse events found	95% confidence intervals for mean	Significant differences (Scheffé method)	One way ANOVA p-value	Kruskal-Wallis analysis of variance p-value
Accident and emergency	A	37	0.45	0.27 - 0.64			
	B	84	0.43	0.28 - 0.58			
	C	237	0.31	0.25 - 0.38			
	D	273	0.55	0.46 - 0.63			
	All	631	0.44	0.39 - 0.49	C&D	< 0.001	< 0.001
ENT	A	-	-	-			
	B	116	0.78	0.63 -0.94			
	C	239	0.81	0.69 - 0.93			
	D	445	1.40	1.30 - 1.50			
	All	800	1.13	1.06 - 1.20	B&D, C&D	< 0.001	< 0.001
Gynaecology	A	8	1.88	0.30 - 3.45			
	B	68	0.74	0.50 - 0.97			
	C	191	0.76	0.66 - 0.86			
	D	298	0.98	0.88 - 1.08	A&B, A&C,		
	All	565	0.89	0.82 - 0.96	A&D	< 0.001	0.004
Obstetrics	A	151	1.05	0.84 - 1.26			
	B	143	0.90	0.73 - 1.08			
	C	597	0.71	0.64 - 0.78			
	D	537	0.97	0.88 - 1.06			
	All	1428	0.86	0.81 - 0.92	A&C, D&C	< 0.001	0.001
Ophthalmology	A	36	0.28	0.12 - 0.43			
	B	33	0.70	0.41 - 0.98			
	C	346	0.42	0.36 - 0.49			
	D	619	0.52	0.46 - 0.58			
	All	1034	0.49	0.44 - 0.53	None	0.023	0.044
General surgery	A	-	-	-			
	B	73	1.33	1.01 - 1.65			
	C	235	1.14	0.97 - 1.32			
	D	241	1.30	1.14 - 1.46			
	All	549	1.24	1.13- 1.35	None	0.369	0.088
Trauma and orthopaedics	A	95	1.34	1.07 - 1.60			
	B	12	1.00	0.62 - 1.38			
	C	324	0.53	0.45 - 0.61			
	D	215	1.05	0.92 - 1.18			
	All	646	0.83	0.74 - 0.90	A&C, D&C	< 0.001	< 0.001
Urology	A	31	1.26	0.89 - 1.62			
	B	45	0.71	0.47 - 0.95			
	C	181	0.57	0.47 - 0.67			
	D	185	0.67	0.56 - 0.78	A&B, A&C,		
	All	442	0.67	0.60 - 0.75	A&D	< 0.001	0.001

Table 6.11. Comparison of numbers of adverse events found analysed by project staff and by specialty

In six of the eight specialties investigated, there were significant differences in the numbers of adverse events recorded by different members of the project staff, according to both the ANOVA and Kruskal Wallis tests. In one specialty - ophthalmology - there were differences of borderline significance, and only in general surgery were no significant differences found. Of course, the statistical significance of the differences in numbers of adverse events found is a product of both the actual size of the differences and the sample size in each specialty, and it is important to recognise that some statistically significant differences may be of limited operational significance. For example, in accident and emergency the mean number of adverse events found by the project staff ranged from 0.31 to 0.55, and the differences though statistically significant may not be particularly important.

It can also be seen from table 6.11 that some of the more extreme values, where one member of project staff was significantly out of step with his or her colleagues, occurred in areas where the person concerned had only screened a small number of patient admissions. For example, in gynaecology, staff member A's results are very different from his/her colleagues, but this member of staff only screened 8 patient admissions in this specialty. It might be argued that in these cases, the member of project staff was still learning to use the adverse event measure in the specialty concerned, and this might have adversely affected the reliability of screening. Incidentally, it can be seen from table 6.11 that although the 95% confidence intervals for A, B, C and D overlap, the Scheffé method finds significant differences (at the $p < 0.05$ level) between A&B, A&C, and A&D. The fact that the F statistic from the one way ANOVA test is highly significant ($p < 0.001$) means that one or more of the pairwise comparisons must be significant (Hays 1994, p458). This apparent inconsistency can be attributed to the formula for calculating confidence intervals performing poorly with small samples such as this, where screener A had only screened 8 cases (Hays 1994, p223).

There are specialties where statistically and operationally significant differences exist in the number of adverse events found which, although they might result from a number of other causes apart from screener variation, should give some cause for concern. For example, in ENT the bulk of patient admissions were screened by project staff C and D, yet D recorded on average 73% more adverse events per patient admission than C.

It is apparent from table 6.11 that in a number of specialties, significant variations exist in the number of adverse events found by different members of the project staff. These variations could represent differences in interpretation of the adverse-event measure; different levels of skill, expertise or clinical knowledge; or differences in practices (such as the use made of various parts of the clinical record). However, they could also be an artefact, resulting from the non-random distribution of cases for screening across members of project staff described earlier or from other known or unknown confounding variables.

One way to explore these variations further is to compare the variations in different specialties, in order to identify whether any systematic biases existed, with, for example, one member of project staff consistently finding more adverse events or fewer adverse events than his or her colleagues across a number of specialties. To this end, table 6.12 presents the ranked performance of each member of project staff across all specialties. In each specialty the four members of project staff have been ranked in order of the mean number of adverse events found per patient admission, with the highest number ranked 1 and the lowest number ranked 4. In two specialties, only three of the staff were involved in screening and so they are ranked from 1 to 3. The table also shows the mean rank for each member of the project staff.

Specialty	Project staff - ranked by mean no of adverse events (1 = highest mean; 4 = lowest mean)			
	A	B	C	D
Accident and emergency	2	3	4	1
ENT	-	3	2	1
Gynaecology	1	4	3	2
Obstetrics	1	3	4	2
Ophthalmology	4	1	3	2
General surgery	-	1	3	2
Trauma and orthopaedics	1	3	4	2
Urology	1	2	4	3
Mean rank across specialties	1.67	2.50	3.38	1.88

Table 6.12. Project staff ranked by mean number of adverse events found per patient admission, analysed by specialty.

It is almost self-evident from the table that some systematic bias exists. For example, D had the highest or second highest mean number of adverse events in all but one specialty, and an average ranking of 1.88, while C had the lowest or second lowest mean number of adverse events in all but

one specialty, and an average ranking of 3.38. The rankings of the four members of project staff in the eight specialties were used to calculate Kendall's coefficient of concordance, W, which is a commonly used as a measure of the degree of agreement among a number of raters but in this case serves as a measure of the tendency for project staff to have consistently higher or lower rankings across specialties (Siegel and Castellan 1988, p270). The W statistic ranges from 0 (indicating no agreement or consistent pattern of ranking) to 1 (indicating complete agreement or absolutely consistent rankings). For the data in table 6.12, $W = 0.33$ (showing significant levels of agreement or consistency in rankings, $p < 0.05$) which suggests that there is some statistically significant pattern in the rankings, with some project staff consistently being ranked higher than others. In other words, the visual impression from table 6.12 that staff members A and D tended, across all specialties, to find more adverse events than B and C is statistically confirmed.

Variation in rates of adverse events found by project staff in ophthalmology and ENT

It has already been noted earlier in this chapter that the summary score of the number of adverse events per patient admission can conceal important variations in the application of the adverse-event measure, since two sets of quite different circumstances and events can produce the same summary score. In order to explore whether differences existed in the way that the project staff used these measures, a further analysis was undertaken of the data from two specialties - ophthalmology and ENT. The incidence of adverse events was calculated separately for each criterion within the adverse-event measures used in these specialties, for each member of project staff, and compared using χ^2 tests to establish whether significant differences existed. The use of these tests, and of standardised adjusted residuals to identify the sources of significant variations, was discussed in chapter 5.

		Adverse event incidence rates (%) and standardised adjusted residuals (in square brackets) by member of project staff					χ^2 statistic
		A	B	C	D	All	P value
1	Adm for adv results o/p mgt	0.0 [-0.4]	0.0 [-0.4]	0.6 [0.7]	0.3 [0.4]	0.4	0.883
2	Readmission for comp prev adm	8.3 [2.5]	3.0 [0.3]	2.6 [0.6]	1.6 [-1.6]	2.2	0.056
3	Error in operative consent	2.8 [1.8]	0.0 [-0.4]	0.6 [0.0]	0.5 [-0.5]	0.6	0.347
4	Unpl rem/inj/repair in surg	2.8 [-0.8]	3.0 [-0.7]	7.8 [2.1]	4.9 [-1.4]	5.7	0.198
5	Unpl return to theatre	0.0 [-0.7]	0.0 [-0.6]	0.9 [-0.6]	1.5 [1.1]	1.2	0.673
6	Path/hist varies from diag	0.0	0.0	0.0	0.0	0.0	-
7	Prob of transfusion	0.0	0.0	0.0	0.0	0.0	-
8	Hosp acquired infection	0.0 [-0.3]	0.0 [-0.3]	0.3 [0.0]	0.3 [0.2]	0.3	0.974
9	Medication error/reaction	0.0 [-0.7]	0.0 [-0.7]	1.7 [0.7]	1.3 [-0.2]	1.4	0.720
10	Cardiac/resp arrest in hosp	0.0	0.0	0.0	0.0	0.0	-
11	CVA/MI/PE in hosp after surg	0.0 0.0	0.0	0.0	0.0	0.0	-
12	Unexp transfer to spec care	0.0 [-0.3]	0.0 [-0.3]	0.0 [-1.0]	0.3 [1.2]	0.2	0.718
13	Pt related clinical complcn	0.0 [-0.2]	0.0 [-0.2]	0.0 [-0.7]	0.2 [0.8]	0.1	0.880
14	Non-clin problem/incident	0.0 [-1.0]	9.1 [2.5]	1.2 [-1.9]	2.9 [1.3]	2.4	0.018
15	Neuro deficit devel in hosp	0.0	0.0	0.0	0.0	0.0	-
16	Unexp patient death	0.0	0.0	0.0	0.0	0.0	-
17	Medical record deficiency	2.8 [-1.5]	21.2 [2.1]	6.1 [-3.2]	12.6 [2.9]	10.3	0.001
18	Nursing record deficiency	0.0	0.0	0.0	0.0	0.0	-
19	Pt/family dissatisfaction	0.0 [-0.9]	0.0 [-0.9]	2.6 [0.7]	2.1 [-0.1]	2.1	0.599
20	Discharge related problems	0.0 [-0.4]	0.0 [-0.4]	0.9 [1.3]	0.3 [-0.9]	0.5	0.633
OP 1	Probs related to cataract extraction	2.8 [0.3]	9.1 [1.7]	4.0 [0.6]	3.1 [-1.1]	3.6	0.303
OP 2	Specific ophthalmic comps	0.0 [-1.6]	3.0 [-0.9]	7.2 [0.4]	7.1 [0.5]	6.8	0.311
OP 3	Ophthalmic proc with tissue biopsy	0.0 [-0.3]	0.0 [-0.3]	0.3 [0.5]	0.2 [-0.3]	0.2	0.954
OP 4	Ophthalmic problems indic by specific drug usage	2.8 [0.5]	3.0 [0.6]	0.6 [-1.9]	2.1 [1.4]	1.6	0.271
OP 5	Oph nursing record review	5.6 [-0.4]	12.1 [1.1]	3.8 [-3.1]	9.1 [2.7]	7.3	0.014

Table 6.13. Variation in adverse event incidence rates in ophthalmology analysed by member of project staff.

Table 6.13 above shows the results of this analysis for ophthalmology. There were a total of 1,034 patient admissions in this specialty in the sample of 6,095 admissions described earlier. For each

screening criterion, the table shows the adverse event incidence rates - effectively the proportion of cases screened in which an adverse event was found under that screening criterion. Given that the number of adverse events found under a given criterion is always 0 or 1, this rate can also be interpreted as the mean number of adverse events per case (in other words, an incidence rate of 10% is the same as a mean number of events of 0.1). Incidence rates are presented rather than mean numbers of events for reasons of convenience, because the incidence of events for individual criteria is low, and a rate of 0.2% is easier to read than a mean of 0.002. The adverse event incidence rates are presented both for each member of project staff and for all staff. The table also shows the results of the χ^2 test performed on the numbers of adverse events found to show whether the differences in the incidence of adverse events for each criterion among members of project staff are significant.

Where a significant difference was found, the standardised adjusted residuals (given in square brackets) can be used to help identify which members of project staff were the source of that significant difference. The standardised adjusted residuals are approximately normally distributed with a mean of 0 and standard deviation of 1.0, so residuals of ±2.58 are significant at the 0.01 level (Everitt 1992, p47). Testing significance at the relatively conservative 0.01 level is advisable, given the repeated use of these tests in the analysis.

In interpreting table 6.13, it should be borne in mind that many of the cells in the crosstabulations used in χ^2 tests had very low expected frequencies, which may cause the significance of any differences to be underestimated (Everitt 1992, p39). However, it can be seen that significant variations in the adverse event incidence rates detected by different members of the project staff were only found for one of the 25 criteria listed in the table (at $p < 0.01$). For only four criteria did one or more of the standardised adjusted residuals reach or exceed ±2.5 (remembering that the 0.01 significance level is ±2.58), and it is interesting to note that in each case a different member of project staff was involved. Overall, the results presented earlier in table 6.11, which suggested that the numbers of adverse events found by different project staff in ophthalmology did not differ significantly, are confirmed and supported by this analysis of data for individual screening criteria within the adverse-event measure used in ophthalmology.

		Adverse event incidence rates (%) and standardised adjusted residuals (in square brackets) by member of project staff				χ^2 statistic
		B	C	D	All	P value
1	Adm for adv results o/p mgt	0.9 [-0.8]	1.7 [-0.1]	2.0 [0.7]	1.8	0.693
2	Readmission for comp prev adm	1.7 [-1.5]	5.4 [1.1]	4.3 [0.0]	4.3	0.266
3	Error in operative consent	3.4 [0.6]	4.2 [1.8]	1.6 [2.1]	2.6	0.105
4	Unpl rem/inj/repair in surg	0.0 [-1.4]	2.1 [0.9]	1.6 [0.2]	1.5	0.309
5	Unpl return to theatre	0.0 [-0.4]	0.0 [-0.7]	0.2 [0.9]	0.1	0.671
6	Path/hist varies from diag	0.9 [2.4]	0.0 [-0.7]	0.0 [-1.1]	0.1	0.052
7	Prob of transfusion	0.9 [2.4]	0.0 [-0.7]	0.0 [-1.1]	0.1	0.052
8	Hosp acquired infection	0.0 [-0.6]	0.4 [-0.6]	0.2 [-0.2]	0.3	0.751
9	Medication error/reaction	1.7 [-0.3]	3.8 [2.1]	1.3 [-1.7]	2.1	0.107
10	Cardiac/resp arrest in hosp	0.0	0.0	0.0	0.0	-
11	CVA/MI/PE in hosp after surg	0.0	0.0	0.0	0.0	-
12	Unexp transfer to spec care	0.0 [-0.4]	0.4 [1.5]	0.0 [-1.1]	0.1	0.309
13	Pt related clinical complcn	1.7 [0.8]	2.1 [2.0]	0.2 [-2.5]	1.0	0.045
14	Non-clin problem/incident	6.9 [0.2]	7.5 [0.9]	5.6 [-1.0]	6.4	0.602
15	Neuro deficit devel in hosp	0.0	0.0	0.0	0.0	-
16	Unexp patient death	0.0	0.0	0.0	0.0	-
17	Medical record deficiency	14.7 [-1.2]	16.3 [-1.0]	20.7 [1.8]	18.5	0.193
18	Nursing record deficiency	24.1 [-4.5]	16.7 [-9.8]	62.2 [12.2]	43.1	< 0.001
19	Pt/family dissatisfaction	1.7 [0.3]	1.3 [-0.2]	1.3 [-0.1]	1.4	0.936
20	Discharge related problems	0.0 [-1.2]	0.4 [-1.1]	1.6 [1.8]	1.0	0.177
ENT 1	Early adm for elective proc	0.0	0.0	0.0	0.0	-
ENT 2	Problems in pre-op period	14.7 [-0.5]	13.4 [-1.5]	18.4 [1.8]	16.4	0.204
ENT 3	Problems with theatre equipment/supplies	0.0 [-0.6]	0.4 [0.6]	0.2 [-0.2]	0.3	0.751
ENT 4	Missing/incomplete ENT records	5.2 [-2.5]	4.6 [-4.4]	18.4 [5.8]	12.4	< 0.001

Table 6.14. Variation in adverse event incidence rates in ENT analysed by member of project staff.

Table 6.14 contains a similar analysis for the 800 patient admissions in ENT that were contained in the sample of 6,095 admissions described earlier. Again, it shows the adverse event incidence rates for each member of staff and for all staff, criterion by criterion. It also presents the results of χ^2 tests

and the standardised adjusted residuals. Results are only presented for three members of project staff (B, C and D) since no cases in this specialty were screened by staff member A.

The same cautions apply, about the conservative effect of low expected frequencies on our estimates of the significance of differences, as they did for table 6.13. In this analysis, there are only two criteria for which significant differences exist ($p < 0.01$) but in both cases the differences are highly significant, with threefold variations in the actual incidence rates, and standardised adjusted residuals of as much as 12.2. Of the remaining 22 criteria, none show significant differences (at $p < 0.01$). It will be remembered that table 6.11 showed that the summary score, the number of adverse events per patient admission, varied significantly across members of project staff in ENT. This analysis suggests that the source of that variation may have been largely attributable to two screening criteria within the adverse-event measure used. For these two criteria, reliability, for whatever reason, was particularly poor, but for all the other 22 criteria in the measure it seems that reliability was much better. In a sense, this table highlights the risks of relying on analyses of the summary score, and the benefits of exploring the use of adverse-event measures criterion by criterion. It seems likely that the significant variation in performance across members of the project staff could be reduced by making some changes to the adverse event measure, or by undertaking further training focused on the criteria where variations exist.

Multivariate analysis of the relationship between adverse event rates, patient characteristics and project staff

It was noted earlier that the bivariate analyses presented above could be affected by a number of known or unknown confounding variables which might lead us to ascribe variations in rates of adverse events to differences among project staff when in fact they resulted from other characteristics in the data, or might also lead us to conclude that no significant variations existed between project staff when in fact they did, but were suppressed by other variables. In order to address these concerns, a multivariate analysis of the data set of 6,095 patient admissions described earlier was undertaken, with the aim of identifying the extent to which the number of adverse events recorded was correlated with a range of variables in the data set, and identifying the contribution to that correlation made by variables representing the identity of the screener.

Given the stated aim, it was decided, though with some reservations outlined below, that the most appropriate multivariate technique to use was multiple regression (Tabachnick and Fidell 1989, p123), with the number of adverse events as the dependent variable and a range of other variables in the data set as independent variables. Multiple regression is an extension of bivariate linear regression, in which a number of independent variables are used to build a linear equation to predict the value of a dependent variable. In this application of the technique, with the number of adverse events as the dependent variable and a range of other variables including the identity of the screener as well as other case or patient characteristics as independent variables, the predictive power of the overall equation was of subsidiary interest. Since the intention was to explore the role of screener variation, the main focus of the analysis was the component of any correlation identified which was attributable to the screener variables rather than to other characteristics.

Multiple regression assumes multivariate normality, linearity and homoscedasticity. It also requires that the variables contained in the data set are not multicollinear. In practice, few data sets conform entirely to these underlying theoretical assumptions but the technique is robust enough to cope with some divergence (Tabachnick and Fidell 1989, p131). However, in the RSCH data set, there were a number of variables which needed to be entered into the regression analysis as independent variables but which clearly did not conform to these assumptions. Moreover, there was no real justification for expecting a linear relationship between the dependent and independent variables. Although data transformations were used to address some of these problems, these limitations should be borne in mind in interpreting the results presented below.

The variables listed in table 6.15 were used as the basis of a number of multiple regression analyses, with the number of adverse events as the dependent variable. The two categorical variables - specialty and project staff identity - were transformed into a series of dichotomous variables in order to allow them to be included in the multiple regression, though this process of course produced a number of collinear variables which presented some difficulties in the analysis. Initial analyses of the data demonstrated significant skews and so logarithmic transformations were used on two variables - length of stay, and number of adverse events to improve the multivariate normality of the data set, though plots of predicted and residual values continued to show skewness. The correlation matrix was examined to ensure that the data set was not multicollinear, and it was

confirmed that no bivariate correlation coefficient was above the suggested limit of 0.9 (Tabachnick and Fidell 1989, p87). Indeed, most correlations were below 0.2, though notable exceptions included the correlation between OBS and SEX (-0.40) and between OPH and ADMTYPE (0.50). The highest correlations were observed between the dichotomous variables constructed from the categorical variables for specialty and project staff identity. For example, the correlation between variables C and D was 0.74. These correlations resulted in some collinearity which at times necessitated the exclusion of one or more variables from some analyses described below.

Variable	Meaning	Notes
NPOS	Number of adverse events	Transformed to improve normality: LNPOS = $\log_{10}$(NPOS). Used as dependent variable in regression analyses.
SPEC	Specialty.	Categorical variable, transformed to a series of dichotomous variables: AE, ENT, GYN, OBS, SURG, TAO, UROL.
SEX	Sex of patient (0 = female, 1 = male)	-
ADMTYPE	Admission type (0 = elective, 1 = emergency).	-
LOS	Length of stay (days)	Transformed to improve normality: LLOS = $\log_{10}$(LOS).
AGEONADM	Age of patient on admission (years).	-
SCREENER	Identity of project staff who screened patient admission.	Categorical variable transformed to a series of dichotomous variables: A, B, C, D.

Table 6.15. Description of RSCH data set used in multiple regression analyses.

Initially, a standard multiple regression was performed including all the variables listed in table 6.15 and using 6,081 of the 6,095 cases in the data set (14 cases were excluded because of missing data on one or more of the variables). The results are shown in table 6.16, which lists the four dichotomous variables representing the project staff identity (A, B, C and D) and each of the other independent variables included in the regression analysis and for each one shows its regression coefficient and standardised regression coefficient in the resulting equation; the squared semipartial correlation coefficient, sr^2,(which is the part of R^2 which is uniquely attributable to that variable)

and the F statistic and its significance showing whether the coefficients and sr^2 are significantly different from zero.

Variable	Coefficient B	Standardised coefficient ϑ	Squared semipartial correlation sr^2 (unique)	F statistic	Significance of F statistic
A	0.016	0.018	0.0003	2.09	0.145
B	-0.040	-0.057	0.0029	20.53	< 0.001
C	-0.061	-0.143	0.0177	126.70	< 0.001
D	-	-	0	-	-
AE	-0.039	-0.058	0.0020	14.30	< 0.001
ENT	0.075	0.122	0.0061	43.39	< 0.001
GYN	0.040	0.055	0.0024	16.84	< 0.001
OBS	-	-	0	-	-
SURG	0.043	0.059	0.0019	13.47	< 0.001
TAO	0.052	0.076	0.0031	21.81	< 0.001
UROL	0.061	0.076	0.0030	21.62	< 0.001
SEX	0.003	0.008	0.0001	0.39	0.533
ADMTYPE	0.020	0.046	0.0009	6.57	0.010
LLOS	0.174	0.284	0.0539	385.08	< 0.001
AGEONADM	< 0.001	0.025	0.0003	2.44	0.119
			0.0946		
n = 6081	R^2 = 0.151 Adjusted R^2 = 0.149 R = 0.389		F = 77.023 Significance of F = < 0.001		

Table 6.16. Standard multiple regression of all variables on log of number of adverse events, for patient admissions in all specialties.

It can be seen that between them, all the variables in the data set were able to account for only 14.9% of the variation in the log of numbers of adverse events, though a significant multiple correlation was found. Three of the four project staff variables were included in the regression equation (D was excluded, because of collinearity) and two had coefficients which were statistically significantly different from zero.

The values of sr^2 represent the unique contribution of each variable to the value of the overall correlation coefficient R^2 (or more exactly, the amount by which R^2 is reduced if that variable was deleted from the regression equation). If there is no correlation whatsoever between the independent variables in the regression equation, the values of sr^2 should sum to R^2 but in practice the sum of sr^2 is usually smaller than R^2. The difference represents shared variance (contributed by

combinations of two or more collinear independent variables). In some circumstances, the sum of sr^2 can exceed R^2, when R^2 is small, and then the interpretation of the sum of sr^2 is problematic (Tabachnick and Fidell 1989, p151).

With the above provisos, the values of sr^2 can be summed to give the proportion of variation in the number of adverse events which is attributable to a particular group of variables. The Sum of sr^2 for the project staff variables in the data set is 0.021; in comparison the sum of sr^2 for specialty variables is 0.019, and the value of sr^2 for log length of stay was 0.054. This means that the screener variables contributed 2.1% of the variation in log number of adverse events, the specialty variables 1.9%, and the log length of stay variable 5.4%. In other words, while this regression analysis had a relatively low power to predict the dependent variable, the log number of adverse events, only a small proportion of the overall correlation found was attributable to screener variables. Most resulted from the other independent variables included in the regression analysis, particularly the log length of stay variable, and from the shared variance attributable to unspecified combinations of independent variables.

This analysis was then repeated for each of the specialties in turn. This permitted the eight dichotomous variables representing specialty to be eliminated from the multiple regression, improving the ratio of cases to variables and so increasing the power of the analyses to identify statistically significant correlations. More practically, this step recognised that previous analyses of the RSCH data set (see, for example, the multiway frequency analysis in chapter 5) had found very different results in different specialties, and had concluded that analyses across all specialties were of limited value because the differences between specialties (in both the clinical content of care and in the adverse-event measures used) were so important. The results of these further analyses of each specialty in turn are listed below in table 6.17. Once again, because of the collinearity of the project staff variables A, B, C and D, one of these variables was commonly excluded from the regression analysis. Since it is primarily the contribution of variables A, B, C and D to the multiple correlation that is of interest, table 6.17 presents the regression data for these variables and the overall correlation data for each analysis.

Specialty	Variable	Coefficient B	Standardised coefficient β	Squared semipartial correlation sr^2 (unique)	F statistic	Significance of F statistic
Accident and emergency	A	-0.017	-0.025	0.0006	0.371	0.543
	B	-0.037	-0.030	0.0051	3.262	0.071
	C	-0.060	-0.177	0.0268	17.294	< 0.001
	D	-	-	-	-	-
n = 630	R^2 = 0.033 Adjusted R^2 = 0.022 R = 0.182		F = 3.048 Significance of F = 0.004			
ENT	A					
	B	-0.001	-0.002	< 0.0001	0.003	0.958
	C	-	-	-	-	-
	D	0.133	0.309	0.0749	67.697	< 0.001
n = 798	R^2 = 0.125 Adjusted R^2 = 0.118 R = 0.353		F = 18.771 Significance of F = < 0.001			
Gynaecology	A	0.118	0.072	0.0051	2.938	0.087
	B	-0.069	-0.116	0.0124	7.178	0.008
	C	-0.046	-0.112	0.0114	6.591	0.011
	D	-	-	-	-	-
n = 563	R^2 = 0.037 Adjusted R^2 = 0.027 R = 0.192		F = 3.562 Significance of F = 0.002			
Obstetrics	A	0.041	0.060	0.0032	5.359	0.021
	B	0.025	0.035	0.0011	1.904	0.168
	C	-	-	-	-	-
	D	0.047	0.106	0.0095	16.169	< 0.001
n = 1427	R^2 = 0.165 Adjusted R^2 = 0.161 R = 0.407		F = 40.168 Significance of F = < 0.001			
Ophthalmology	A	-0.059	-0.061	0.0036	3.957	0.047
	B	0.042	0.040	0.0016	1.725	0.189
	C	-0.022	-0.058	0.0033	3.571	0.059
	D	-	-	-	-	-
n = 1028	R^2 = 0.062 Adjusted R^2 = 0.056 R = 0.249		F = 9.634 Significance of F = < 0.001			
General surgery	A					
	B	0.061	0.091	0.0071	4.701	0.031
	C	-	-	-	-	-
	D	0.052	0.114	0.0110	7.330	0.007
n = 549	R^2 = 0.184 Adjusted R^2 = 0.175 R = 0.429		F = 20.413 Significance of F = < 0.001			
Trauma and orthopaedics	A	0.161	0.274	0.0676	54.172	< 0.001
	B	0.073	0.298	0.0021	1.705	0.192
	C	-	-	-	-	-
	D	0.109	0.246	0.0539	43.230	< 0.001
n = 645	R^2 = 0.206 Adjusted R^2 = 0.197 R = 0.453		F = 23.549 Significance of F = < 0.001			
Urology	A	0.128	0.173	0.0273	12.484	0.001
	B	0.012	0.020	0.0003	0.154	0.695
	C	-0.024	-0.061	0.0031	1.439	0.231
	D	-	-	-	-	-
n = 441	R^2 = 0.053 Adjusted R^2 = 0.037 R = 0.229		F = 3.435 Significance of F = 0.001			

Table 6.17. Standard multiple regression of all variables on log of number of adverse events, undertaken for patient admissions in each specialty separately.

It is immediately evident from table 6.17 that none of the multiple regression analyses produced a model which predicted the log of the number of adverse events particularly well. Although all of them returned multiple correlation coefficients that were statistically significant, the variables listed in table 6.15 which were included in each multiple regression analysis accounted for at most just under 20% of the variation in the log of the number of adverse events - and at worst, accounted for only 2.2% of the variation. This finding might be viewed quite positively, since it demonstrates that the project staff variables were generally not responsible for more than a relatively insignificant variation in the adverse event score. But it also suggests that other variables (such as length of stay and admission type) which were shown to be associated with variations in the adverse event score in chapter 5, were also not strongly associated with adverse event score in this analysis. One explanation might be that the data contained in the data set, despite the transformations performed to reduce skew, was sufficiently outside the various assumptions of the multiple regression technique discussed earlier to make the results of analysis overly conservative and reduce its power. It could also be argued that the relationship between adverse event score and the various independent variables included in the analysis has already been demonstrated in chapter 5 to be non-linear to some degree, and so the linear model underlying multiple regression is not appropriate. It seems that given the poor predictive value of these multiple regression models, the results should be interpreted with some caution and set alongside the findings from various bivariate analyses already presented and discussed.

Table 6.18 presents a rather more concise analysis of the results from the multiple regressions contained in table 6.17. For each specialty it shows the proportion of variation in the log of the number of adverse events which the regression model was able to explain; the proportion of variation which could be attributed to the project staff variables alone (calculated by summing the ($;sr^2$) , a process which has some limitations, described below); and the statistical significance of the variation attributable to each member of the project staff.

Specialty	Proportion of variation explained by all variables (Adjusted R^2)	Proportion of variation explained by project staff variables (;sr^2)	Significance of individual project staff variables's contribution to variation (P value associated with F statistic)			
			A	B	C	D
Accident and emergency	2.2%	3.2%	0.543	0.071	< 0.001	-
ENT	11.8%	7.5%	n/a	0.958	-	< 0.001
Gynaecology	2.7%	2.9%	0.087	0.008	0.011	-
Obstetrics	16.1%	1.4%	0.021	0.168	-	< 0.001
Ophthalmology	5.6%	0.85%	0.047	0.189	0.059	-
General surgery	17.5%	1.8%	n/a	0.031	-	0.007
Trauma and orthopaedics	19.7%	12.4%	< 0.001	0.192	-	< 0.001
Urology	3.7%	2.9%	0.001	0.695	0.231	-

Table 6.18. Summary of selected results from standard multiple regression of all variables on log of number of adverse events, undertaken for patient admissions in each specialty separately

One anomaly immediately evident in the table is that in two specialties, accident and emergency and gynaecology, ;sr^2 is greater than the value of adjusted R^2. It was noted above that the sum of the squared semipartial correlation coefficients is usually interpreted as the unique contribution of those variables to the multiple correlation (or, in other words, the amount by which adjusted R^2 would be reduced if the variables were removed from the regression), but that when the value of adjusted R^2 is very small, ;sr^2 can exceed it and this interpretation is no longer appropriate (Tabachnick and Fidell 1989, p151).

The multiple regression analyses suggest that somewhere between under 1% and just over 12% of the variation in adverse event scores could be attributed to the project staff, but they do not provide an explanation for much of the rest of the observed variation. When the results in table 6.18 are compared with the summary of findings from bivariate analyses in the same specialties in table 6.11, there are a number of similarities. In both analyses, the variations in adverse event score associated

with the project staff seem to be least in ophthalmology and general surgery, and greatest in trauma and orthopaedics and ENT.

The data in table 6.18 for individual project staff is more difficult to interpret, because the collinearity between the four variables caused one - either C or D - to be eliminated from the regression analysis for each specialty. In addition, it should be remembered that the p values reported in the table are in part a product of the varying numbers of cases screened by each screener in each specialty (see table 6.10) as well as of variations in their association with the adverse event score. However, it is notable that the results broadly echo those from bivariate analyses which were summarised in table 6.12, showing screener B to be the least associated with any variation, while screener D in particular appears to be associated with variation in every specialty for which results are available.

6.3.4 **Conclusions**

These studies of the data collected during the RSCH project seem to suggest that significant variations in the results of screening were associated with the identity of the project staff undertaking screening. It is not possible to identify the causes of these variations, but it might be argued that the non-random distribution of cases across project staff, the development of some degree of specialty specialisation among staff, the existence of differences in actual practice in screening, differences in staff members' interpretation of the written definitions of adverse-event measures, and differences in staff members' clinical knowledge and skills would all have played a part. It is not possible to attribute these variations solely to differences in screener performance or behaviour.

However, these results suggest that, for whatever reason, the reliability of these adverse-event measures used in a realistic clinical setting with the usual potential distractions and logistic problems is likely to be lower than their reliability when used in the more controlled and artificial setting of the interrater and intrarater reliability studies reported earlier in this chapter. This adds force to the argument, already made in section 6.2.4, that the continuing use of adverse-event measures such as these should incorporate some form of ongoing monitoring of reliability, through

periodic rescreening or data analysis linked to appropriate interventions aimed at maintaining and improving reliability, such as workshops, case presentations and refresher training for those undertaking screening.

Chapter 7

Conclusions

7.1 Introduction

The research reported in this thesis was undertaken in order to expand our knowledge and understanding of the validity and reliability of adverse-event measures of quality in healthcare, as the objectives set out in chapter 1 made clear. This final chapter is intended to draw together the results from the research reported in chapters 4, 5 and 6 with the evidence from previous studies which was presented in chapter 3, and to place those findings in context in order that their implications for those who might want to develop or use adverse-event measures of quality in healthcare are explored.

First, the chapter presents an overview of the key findings and conclusions from the research presented in chapters 4, 5 and 6, and compares those findings with those reported elsewhere. It then reviews a number of issues raised by the studies, and discusses their implications, and considers what specific conclusions the research offers to those developing or using adverse-event measures of quality in healthcare. Finally, a number of limitations concerning the research reported in this thesis are outlined, and an agenda for future research is then set out.

7.2 Research findings, conclusions and issues raised

7.2.1 The validity of adverse-event measures of quality

The questionnaire and interview study undertaken to explore the face and content validity of an adverse-event measure of quality produced broadly concordant results, which supported the use of these measures in measuring the quality of healthcare. The questionnaire study found that while a

few criteria within the measure being tested were not supported by respondents, and many useful suggestions for improvements to the measure were made, there was overall support for the validity of the measure, and for the practicality of using such a measure in the context of a British acute hospital. The interview study largely reiterated and supported the views expressed in the questionnaire study about the utility of adverse-event measures, and indicated that while there were some concerns about the potential for bias in quality measurement if adverse-event measures were the only source of information, there was considerable support for the use of such measures.

In considering what reliance we are able to place on the findings from the questionnaire study, we should take into account the evidence that in areas where respondents' views could be compared with empirical data (on the incidence of adverse events) there was good general agreement, which suggests that respondents' views in other areas may be similarly valid and reliable. While the questionnaire study did not set out to sample the views of a representative sample of clinicians or other groups with an interest in adverse-event measures of quality, it was possible to compare the views expressed by quite different groups of respondents - practising clinicians and public health physicians. Their opinions on the validity of the adverse-event measure were generally in agreement, with only minor differences of opinion in a few areas, which might suggest that the views expressed in the questionnaire study would command broader support from other groups too. Finally, the interview study, though much smaller in scale, produced similar results and conclusions from a different research method and so lends further support to the credibility and generalisability of the findings on the validity of adverse-event measures of quality.

Overall, the two studies described in chapter 4 offer strong support for the content and face validity of the adverse-event measure of quality they tested. It was noted in chapter 3 that no other studies of the face and content validity of adverse-event measures of quality were identified in the literature, so the studies reported in chapter 4 cannot be compared with evidence from elsewhere.

In chapter 5, the construct validity of some adverse-event measures of quality was assessed, using the data drawn from the RSCH occurrence screening project. Six constructs were tested, of which three had already been tested and supported in previous studies reviewed in chapter 3. A series of bivariate analyses and a multivariate analysis using loglinear or multiway frequency analysis

confirmed support for five of the six constructs across a number of specialties. It was, however, clear that other, sometimes complex associations existed in the data set which could confound comparisons made using measures of quality based on adverse events. Earlier studies reported in chapter 3 had already demonstrated in other settings and with other adverse-event measures of quality that three of the six constructs tested were supported.

Most of the previous research into the validity of adverse-event measures described in chapter 3 concerned the criterion-related validity of such measures. It was noted that the fundamental problem for such studies was the identification of an appropriate and meaningful criterion variable. Most of the studies used some form of implicit professional review, with one or more clinicians making judgements about whether or not an adverse event had occurred, and if so rating its effects, causation and so on. Even if the methodological flaws in a number of these studies (such as an absence of blinding, and the use of inappropriate measures of association to rate agreement) are ignored, the evidence from a number of sources that such implicit reviews have low validity and reliability in themselves makes it unlikely that this approach to testing validity can be relied upon. In this light, the evidence presented above on the face, content and construct validity of these measures may be given additional weight.

7.2.2 **The reliability of adverse-event measures of quality**

The experimental and observational studies undertaken to explore the reliability of adverse-event measures of quality produced rather more equivocal findings. The experimental studies broadly indicated that the measures tested had moderate to good reliability. The measures were certainly capable of improvement, since some criteria within each measure were particularly unreliable and their removal or modification would certainly result in improved reliability. However, these experimental studies highlighted the important contribution of rater training to reliability, with the poorest reliability found in a study which involved a screener with extensive clinical experience but limited training in using the measure. They also provided some data which suggested that more adverse events were found by the screeners in the experimental studies than during their use of the same measures in the RSCH occurrence screening project, which supports the hypothesis that experimental studies of reliability may have poor generalisability because raters behave differently

during the experimental study (are, for example, more careful or attentive in applying the measures) from the way they would behave when using the measure in non-experimental conditions. It was noted in chapter 3 that the results from earlier experimental studies of the reliability of adverse-event measures were rather mixed, and reached varying conclusions about their reliability. It might be argued, in the light of the findings reported above, that these heterogeneous findings reflect differences in the adverse-event measures tested and in the approaches to rater training employed.

The observational study of reliability, based on an analysis of data from the RSCH occurrence screening project, suggested that significant variations in the results of screening were associated with the identify of the screeners applying the adverse-event measures used. While it was not possible to attribute these variations solely to differences in screeners' performance or behaviour, and a number of other causes could be posited, it seemed likely that the substantial differences sometimes observed indicated that the reliability of the adverse-event measures being used was not as high as the experimental studies might have indicated. No other similar analyses were identified in the literature reviewed in chapter 3, so it is not possible to compare these findings with others from elsewhere.

Overall, the results of the studies of reliability give some cause for concern. While they suggest that the reliability of adverse-event measures can be quite high, they also indicate that reliability is rather dependent on the individual measure and the conditions in which it is used. They suggest that the reliability of adverse-event measures in actual practice may be lower that the theoretical estimates of reliability derived by developers of those measures might suggest.

7.2.3 **General issues and considerations raised by the research**

Testing the validity and reliability of measures is an essential part of their development. It should give cause for concern that many adverse-event measures of quality, and measures of other kinds, have become widely used despite an absence of empirical evidence to demonstrate that they offer valid and reliable information on the concepts that they purport to measure. This research raises four key questions concerning the testing of validity and reliability which deserve further discussion:

a) *How should information about different dimensions of validity and reliability be integrated in assessing measures of healthcare quality?*

It was evident from the research that the results of validity and reliability testing are unlikely to be unequivocal and unanimous, and may even point to quite different conclusions about the performance of the measure being tested. For example, within the concept of validity there are a number of dimensions - criterion-related, construct, face and content - which when tested may give divergent results. The developers and users of measures need to be able to balance such divergent results and reach decisions about whether to develop further or deploy such measures on the basis of this information. When a measure has apparently poor validity and poor reliability, those decisions are relatively easy to make. But if, for example, a measure combines good validity and poor reliability, it is less clear whether and how it should be used. It could be argued that the users of such measures should be encouraged to use only those measures with proven good validity and good reliability, and that other measures should remain the preserve of their developers until their performance has been improved.

Perhaps one general lesson which emerges from the research is that broadly based assessments of validity and reliability, which make assessment across a number of different dimensions rather than focusing on a single dimension (such as criterion-related validity), may be more useful and provide more assurance to the users of measures. Where, for example, there is research evidence from a range of settings, using both observational and experimental methods, of the criterion-related, construct, face and content validity and interrater and intrarater reliability of a measure - then, perhaps, its users can feel justifiably confident about how the measure will perform in actual practice.

b) *How meaningful are the results of experimental approaches to assessing reliability?*

The results of reliability testing reported in chapter 6 raise some concerns about the differences in apparent reliability when a measure is being formally tested and when it is simply being used in actual practice. The environment for experimental studies of reliability - with specially trained and well-motivated raters, more time and resources for applying the

measure, better feedback on performance, or whatever - may be quite atypical of the environment for actual practice. However, experimental studies of reliability eliminate the many potential confounding factors and biases which make the interpretation of observation studies of reliability so difficult. It might be argued either that the users of measures should seek both experimental and observational evidence of reliability from the developers of those measures, or that experimental studies of the reliability of such measures should be undertaken in conditions which as much as possible approximate the settings of actual practice in which the measure would be used. There is a clear parallel here with the debate over the use of experimental and observational methods in testing the effectiveness of healthcare interventions, and with the development of both explanatory and pragmatic trials of such interventions.

c) *How generalisable are the results of validity and reliability testing undertaken by the developers of measures of healthcare quality?*

The studies reported in chapters 4, 5 and 6 raise important questions about how much the users of measures can rely on the testing undertaken by their developers. Ideally, once a measure has been tested and its validity and reliability have been demonstrated, we should then be able to use it without any further debate or discussion about these issues. However, these studies suggest that the external validity or generalisability of the testing undertaken by the developers of measures might be limited. In other words, while the developers of measures might be able to show that they are valid and reliable in the conditions in which they are tested, those conditions may be quite atypical of the environment in which the measure is subsequently used and the differences may have profound implications for validity and reliability.

This means that it may be wise to treat data on validity and reliability from the developers of measures as maximal estimates of the validity and reliability which might be obtained rather than necessarily as indicators of the performance which can be expected in normal practice. It may be advisable for the potential users of any measure to seek and rely on reports of validity and reliability from other users of that measure, in circumstances as similar as possible to those in which it is to be used, and to place greater reliance on those data than on

those from the developers of the measure. When this kind of data on validity and reliability is not available, or if there is reason to believe that the generalisability of validity and reliability data for the measure is particularly low, it could be argued that some testing of reliability and validity should be incorporated into the actual use of the measure. Indeed, the importance of rater training and other conditions to the reliability of measurement means that the regular testing of reliability (through, for example, periodic intrarater or interrater rescreening linked to the feedback of results and comparative analyses) may be a necessary step in the routine use of such measures.

d) *What is the relationship between validity and reliability for measures of healthcare quality*?

The studies reported in chapters 4, 5 and 6 also suggest that there may be a balance to be struck between the validity and reliability of adverse-event measures of quality. It was noted earlier that the development of the measures used in the RSCH occurrence screening project and whose validity and reliability was tested in this research was led by clinicians within the project. Their priorities were not necessarily to maximise either validity or reliability, but it could be argued that the former rather than the latter was the more important consideration in their eyes. Certainly, the comments from respondents in the questionnaire and interview studies were more concerned with issues of validity than reliability. This may have resulted in the production of measures with high validity (reported in chapters 4 and 5) but rather less good reliability (reported in chapter 6). Some of the changes to the measures suggested by the examinations of reliability undertaken and reported in chapter 6 which would improve their reliability seem, on the face of it, likely to also reduce their validity. There is a known statistical relationship between validity and reliability for any measure, as was noted in chapter 2. The lower the reliability of a measure is, the lower the maximum achievable validity will be. But, beyond that, there may circumstances in which decisions about the content, structure or definition of a measure have to be made which may either increase validity at the expense of reliability, or vice versa.

7.2.4 Implications for the development and use of adverse-event measures of quality in healthcare

This research reinforces the conclusions of many earlier studies reviewed in chapter 3 which indicate that adverse events are an important phenomenon, worthy of study and measurement because of their implications for patients and healthcare organisations and because of the insight they offer into the quality of care and how it might be improved. Broadly, it supports the use of such measures in quality measurement in healthcare as being a meaningful and worthwhile component of wider quality assurance programmes, with two caveats. Firstly, measurement of quality should not focus solely on data about adverse events but should also take account of other important characteristics and dimensions of quality. Secondly, measures based on adverse events should be developed and used in ways that assure their validity and reliability.

This research suggests that the developers of adverse-event measures of quality should be more rigorous in testing the validity and reliability of their measures before recommending them to potential users. It is evident that high validity and reliability cannot be assumed, and the study of validity and reliability during the development of measures is likely to identify opportunities to make improvements to the measures.

In turn, the users of adverse-event measures of quality should, perhaps, be more discerning or demanding in their selection and application of measures, seeking better evidence from more extensive testing of the validity and reliability of the measures which they adopt. They should also be more mindful of the potential threats to validity and reliability inherent in the conditions in which measures are actually used, and should ensure not only that measures are used as they were intended to be used, with the appropriate level of rater training and support, but also that some element of ongoing measurement of the performance of the measure is incorporated into its routine use.

7.3 **Further research**

When this research is viewed in the context of the framework for evaluating measures of healthcare quality proposed by the Institute of Medicine (1990) which was discussed in chapter 2, it can be seen that this research has a worthy but narrow focus, and that many important dimensions of the performance of such measures have not been evaluated. This research has been primarily focused on what that framework labelled the scientific grounding of such measures - but there are equally important considerations to do with their general design, efficiency and latitude which have not been addressed. Future research should be aimed at broadening the scope of this evaluation, and exploring these other aspects of the performance of adverse-event measures.

Given that it was argued in chapter 2 that the limitations of quality assurance programmes in healthcare often lay more in the process of quality improvement than in their approaches to quality measurement, it is an important limitation of this research that it was confined to an examination of the use of adverse-event measures in quality measurement. There are a range of other issues concerning the application of the adverse event data produced by measurement in changing clinical practice and bringing about quality improvements which need to be addressed. For example, the utility of different approaches to presenting the data, and the nature of organisational arrangements, incentives and mechanisms for then using that data, deserve further examination.

For almost all the adverse-event measures of quality discussed in chapter 3, and for all the research reported in chapters 4, 5 and 6, the primary source of data has been the written clinical record. Many of the concerns about the validity and reliability of such measures arise, directly or indirectly, from the nature, structure and content of those records. For example, the reliability with which adverse events can be identified depends, among other things, on the quality of record keeping. With the growing use of information technology in healthcare, and the increasing availability of a wide range of clinical data on various computer systems, there may be opportunities to both reduce the cost of identifying adverse events and increase the validity and reliability with which they are identified by making greater use of automated approaches to screening for adverse events. The potential for such techniques merits further research.

Bibliography

Agency for Health Care Policy and Research (1995). Using clinical practice guidelines to evaluate quality of care [vols I & II]. Rockville, Maryland: AHCPR.

Allan EL, Barker KN (1990). Fundamentals of medication error research. American Journal of Hospital Pharmacy, 47: 555-571.

Allery LA, Owen PA, Robling MR (1997). Why general practitioners and consultants change their clinical practice: a critical incident study. British Medical Journal 314(7084):870-4.

Allsop J, Mulcahy L (1996). Regulating medical work: formal and informal controls. Buckingham: Open University Press.

American Medical Association (1986). Quality of care. Journal of the American Medical Association, 256(8):1032-1034.

Amess M, Walshe K, Shaw C et al (1995). The audit activities of the medical Royal Colleges and their Faculties in England. London: CASPE Research.

Angell M, Kassirer JP (1996). Quality and the medical marketplace - following elephants. New England Journal of Medicine, 335(12):883-885.

Appleby J, Walshe K, Ham C (1995). Acting on the evidence: a review of clinical effectiveness, sources of information, dissemination and implementation. NAHAT research paper 17. Birmingham: National Association of Health Authorities and Trusts.

Ash A, Shwartz M, Payne SMC, Restuccia JD (1990). The self-adapting focused review system: probability sampling of medical records to monitor utilisation and quality of care. Medical Care, 28(11):1025-1039.

Bagust A (1996). League tables. British Journal of Hospital Medicine, 55(6):369-370.

Bardsley M, Coles J (1992). Practical experiences in auditing patient outcomes. Quality in Health Care, 1:124-130.

Barnes C, Moynihan C (1988). Accuracy of generic screens in identifying quality problems: analysis of false positive and false negative occurrences. Topics in Health Record Management, 9(1):72-80.

Barr DP (1955). Hazards of modern diagnosis and therapy - the price we pay. Journal of the American Medical Association, 159(15):1452-1456.

Barrable B (1992). A survey of medical quality assurance programs in Ontario hospitals. Canadian Medical Association Journal, 146(2):153-160.

Bates DW, O'Neil AC, Boyle D et al (1994). Potential identifiability and preventability of adverse events using information systems. Journal of the American Medical Informatics Association 1: 404-411.

Bates DW, Cullen DJ, Laird N et al (1995). Incidence of adverse drug events and potential adverse drug events: implications for prevention. Journal of the American Medical Association, 274(1): 29-34.

Bennett J, Walshe K (1990). Occurrence screening as a method of audit. British Medical Journal, 300:1248-1251.

Berlin A, Spencer JA, Bhopal RS et al (1992). Audit of deaths in general practice: pilot study of the critical incident technique. Quality in Health Care 1992; 1:231-235.

Berwick DM (1988). Toward an applied technology for quality measurement in healthcare. Medical Decision Making, 8(4):253-258.

Berwick DM (1989). Continuous improvement as an ideal in health care. New England Journal of Medicine, 320(1):53-56.

Berwick DM (1996). A primer on leading the improvement of systems. British Medical Journal, 312:619-622.

Berwick DM, Knapp MG (1990). Theory and practice for measuring healthcare quality. In: Graham NO (ed). Quality assurance in hospitals: strategies for assessment and implementation. Rockville, Maryland: Aspen Publishers.

Berwick DM, Godfrey AB, Roessner J (1990). Curing healthcare: new strategies for quality improvement. San Francisco: Jossey Bass.

Berwick DM, Wald DL (1990). Hospital leaders opinions of the HCFA mortality data. Journal of the American Medical Association, 263(2):247-249.

Black N (1990). Quality assurance of medical care. Journal of Public Health Medicine, 12(2):97-104.

Bomhof J, Arends JW, van der Beek J (1993). Adverse patient occurrences in a university hospital: a comparison of screening results registered by specialists and by external review. Quality Assurance in Health Care, 5(2):157-165.

Bomhof J, Nieman FHM, Reerink E (1993). Registration of adverse patient occurrences in a university hospital: relations between adverse patient occurrences and characteristics of hospitalised patients. Quality Assurance in Health Care, 5(2):167-174.

Bowden D, Williams G, Stevens G (1986). Medical quality assurance in Brighton Health Authority - can American translate to English? In: Moores B (ed). Are they being served? Oxford: Philip Allan.

Bowden D (1990). The trouble with danger. Health Service Journal, (19 April):589.

Bradley CP (1992). Uncomfortable prescribing decisions: a critical incident study. British Medical Journal 304:294-296.

Brennan P, Silman A (1992). Statistical methods for assessing observer variability in clinical measures. British Medical Journal, 304:1491-1494.

Brennan TA (1995). Medical injuries: international perspectives. Medical Journal of Australia. 163(9):475-6.

Brennan TA, Localio RJ, Laird N (1989). Reliability and validity of judgements concerning adverse events suffered by hospitalised patients. Medical Care, 27(12):1148-1158.

Brennan TA, Hebert LE, Laird NM et al (1991a). Hospital characteristics associated with adverse events and substandard care. Journal of the American Medical Association, 265(24):3265-3269.

Brennan TA, Leape LL, Laird NM et al (1991b). Incidence of adverse events and negligence in hospitalised patients: results of the Harvard Medical Practice Study I. New England Journal of Medicine, 324(6):370-276.

British Medical Journal (1974). Editorial. Towards medical audit. British Medical Journal, i:255-257.

British Medical Journal (1976). Editorial. Audit again. British Medical Journal, ii:714-715.

British Standards Institute (1979). BS4778: Glossary of terms used in quality assurance. London: British Standards Institute.

British Standards Institute (1987). BS5750: Quality systems: part I. Specification for design/development, production, installation and servicing. London: British Standards Institute.

Brook RH (1977). Quality - can we measure it? New England Journal of Medicine, 296(3):170-172.

Brook RH, Appel FA (1973). Quality of care assessment: choosing a method for peer review. New England Journal of Medicine, 288:1323-1329.

Brook RH, Kosecoff J (1988). Competition and quality. Health Affairs, 7(3):150-161.

Burr M (1990). Feasibility trial of concurrent casenote screening for clinical indicators to assure quality of care at a large Australian teaching hospital. Australian Clinical Review, 10:114-116.

Buttery Y, Walshe K, Coles J et al (1994). The development of audit: findings of a national survey of healthcare provider units in England. London: CASPE Research.

California Medical Association (1977). Report on the medical insurance feasibility study. San Francisco: California Medical Association/Sutter Publications.

Caplan RA, Posner KL, Cheney FW (1991). Effect of outcome on physician judgements of appropriateness of care. Journal of the American Medical Association, 265(15):1957-1960.

Caranasos GJ, Stewart RB, Cluff LE (1974). Drug-induced illness leading to hospitalisation. Journal of the American Medical Association, 228(6):713-717.

Carlow D (1988). Occurrence screening can improve QA programs. Dimensions, xx(June):20-22.

Carmines EG, Zeller RA (1979). Reliability and validity assessment. London: Sage Publications.

Carr W, Szapiro N, Heisler T, Krasner M (1989). Sentinel health events as indicators of unmet need. Social Science in Medicine, 29(6):705-714.

Carr-Hill R, Dalley G (1992). Assessing the effectiveness of quality assurance. Journal of Management in Medicine, 6(1):10-18.

CASPE Research (1987). Use and validity of performance indicators - a national survey. London: CASPE Research.

CASPE Research (1989). The development of screening criteria as a practical tool for medical audit and peer review in a major district general hospital. London: CASPE Research.

CASPE Research (1990). Brighton quality assurance project - final report (volumes I-V). London: CASPE Research.

CASPE Research/Commission on Professional Hospital Activities (1992). Three hospitals study. London: CASPE Research.

CHKS Ltd (1995). Acute care 95: healthcare resource groups national statistics. Alcester: CHKS.

Charlton JRH, Hartley RM, Silver R, Holland WW (1983). Geographical variations in mortality from conditions amenable to medical intervention in England and Wales. Lancet, i:691-696.

Charlton JR. Velez R (1986). Some international comparisons of mortality amenable to medical intervention. British Medical Journal, 292(6516):295-301.

Chassin M (1996). Improving the quality of care. New England Journal of Medicine, 335:1060-1063.

Citro FC, Deneselya JA, Kopper MG et al (1988). Risky business: the first encounter with APOs. QRC Advisor, 5(2):1-7.

Clements R (1995). Essentials of clinical risk management. In: Vincent C (ed). Clinical risk management. London: BMJ Publishing.

Cluff LE, Thornton GF, Seidl LG (1964). Studies on the epidemiology of adverse drug reactions. Journal of the American Medical Association, 188(11):976-983.

Codman EA (1914). The product of a hospital. Surgery, Gynaecology and Obstetrics, 10:491-494.

Coles J (1990). Outcomes management and performance indicators. In: Hopkins A, Costain D (ed). Measuring the outcomes of medical care. London: Royal College of Physicians.

Commission on Professional Hospital Activities (1990). CPHA today. Ann Arbor, Michigan: CPHA.

Couch JB (1989). The Joint Commission on Accreditation of Healthcare Organisations. In: Goldfield N, Nash DB (ed). Providing quality care: the challenge to clinicians. Philadelphia: American College of Physicians.

Couch NP, Tilney NL, Rayner AA et al (1981). The high cost of low frequency events: the anatomy and economics of surgical mishaps. New England Journal of Medicine, 304:634-637.

Craddick JW (1979). The Medical Management Analysis system: a professional liability warning mechanism. Quality Review Bulletin, 5(4):2-8.

Craddick JW, Bader BS (1983). Medical Management Analysis: a systematic approach to quality assurance and risk management. Volume I: Introduction. Auburn, California: Joyce W Craddick.

Craddick JW (ed) (1984). Medical Management Analysis: a systematic approach to quality assurance and risk management. Volume II: Implementation Manual. Auburn, California: Joyce W Craddick.

Crombie IK, Davies HTO, Abraham SCS et al (1993). The audit handbook: improving health care through clinical audit. Chichester: John Wiley.

Deming WE (1986). Out of the crisis. Cambridge, MA: Massachusetts Institute of Technology.

Department of Health (1989a). Health service indicators: guidance - dictionary. London: HMSO.

Department of Health (1989b). Working for patients: working paper 6. London: HMSO.

Department of Health (1991a). Report on confidential enquiries into maternal deaths in the UK, 1985-87. London: HMSO.

Department of Health (1991b). HC(91)2. Medical audit in the hospital and community health services. London: Department of Health.

Department of Health (1993a). EL(93)115. Improving clinical effectiveness. London: Department of Health.

Department of Health (1993b). EL(93)59. Meeting and improving standards in healthcare - a policy statement on clinical audit. London: Department of Health.

Department of Trade and Industry (1995). Implementing BS EN ISO 9000: a guide for small firms. London: HMSO.

DesHarnais SI, McMahon LF, Wroblewski RT, Hogan AJ (1990). Measuring hospital performance: the development and validation of risk adjusted indexes of mortality, readmissions and complications. Medical Care, 28(12): 1127-1141.

Diamond MR, Kamien M, Sim MG et al (1995). A critical incident study of general practice trainees in their basic general practice term. Medical Journal of Australia. 162(6):321-4.

Dick W, Hegarty N (1971). Topics in measurement: reliability and validity. New York: McGraw Hill.

Dingwall R, Fenn P (1995). Risk management: financial implications. In: Vincent C (ed). Clinical risk management. London: BMJ Publishing.

Donabedian A (1966). Evaluating the quality of medical care. Milbank Memorial Fund Quarterly, 44(3):166-206.

Donabedian A (1980). Explorations in quality assessment and monitoring: volume I. The definition of quality and approaches to its assessment. Ann Arbor, Michigan: Health Administration Press.

Donabedian A (1982). Explorations in quality assessment and monitoring: volume II. The criteria and standards of quality. Ann Arbor, Michigan: Health Administration Press.

Donabedian (1990). The quality of care: how can it be assessed? In: Graham NO (ed). Quality assurance in hospitals: strategies for assessment and implementation. Rockville, Maryland: Aspen Publishers.

Dubois RW, Rogers WH, Moxley JH et al (1987). Hospital inpatient mortality: is it a predictor of quality? New England Journal of Medicine, 317(6):1674-1680.

Dubois RW, Brook RH, Rogers WH (1987b). Adjusted hospital death rates: a potential screen for quality of medical care. American Journal of Public Health, 77(9):1162-1167.

Dubois RW, Brook RH (1988). Preventable deaths: who, how often and why? Annals of Internal Medicine, 109:582-589.

Ellwood P (1988). Shattuck lecture. Outcomes management - a technology of patient experience. New England Journal of Medicine, 318(23):1549-1556.

Ennis M, Vincent CA (1990). Obstetric accidents: a review of 64 cases. British Medical Journal, 300:1365-1367.

Erdmann MD (1990). The value of an occurrence screening quality assurance program. Clinical Research, 38(3):911A.

Erdmann MD (1991). The value and cost of an occurrence screening quality assurance program. Clinical Research, 39(2):188A.

Escovitz GH, Burkett GL, Kuhn JC et al (1978). The effects of mandatory quality assurance: a review of hospital medical audit processes. Medical Care, 16(11): 941-949.

Field PA, Morse JM (1985). Nursing research: the application of qualitative approaches. London: Chapman and Hall.

Fitzpatrick R (1989). Measurement of patient satisfaction. In: Hopkins A, Costain D (ed). Measuring the outcomes of medical care. London: Royal College of Physicians.

Flanagan JC (1954). The critical incident technique. Psychological Bulletin, 51(4): 327-358.

Fleiss JL (1981). Statistical methods for rates and proportions. New York: John Wiley and Sons.

Fowkes FGR (1982). Medical audit cycle: a review of methods and research in clinical practice. Medical Education, 16:228-238.

Frater A (1992). Health outcomes: a challenge to the status quo? Quality in Health Care, 1:87-88.

Frater A. Costain D (1992). Any better? Outcome measures in medical audit. British Medical Journal 304:519-520.

Freeborn DK, Greenlick MR (1973). Evaluation of the performance of ambulatory care systems: research requirements and opportunities. Medical Care, 11 (Suppl):68-75.

Friedman M (1982). Iatrogenic disease: addressing a growing epidemic. Postgraduate Medicine 71(6): 123-129

Fuchs VR (1978). Public policy and the medical establishment: who's on first? Journal of Medical Education, 54(1):8-11.

Gaba DM (1989). Human error in anesthetic mishaps. International Anesthesiology Clinics, 27(3):137-47.

Gabbay J, McNicol M, Spiby J et al (1990). What did audit achieve? Lessons from preliminary evaluation of a year's medical audit. British Medical Journal, 301:526-529.

Ghiselli EE, Campbell JP, Zedeck S (1981). Measurement theory for the behavioural sciences. San Francisco: W H Freeman and Company.

Gill M (1993). Purchasing for quality: still on the starting blocks? Quality in Health Care 2:179-182.

Goldman RL (1989). Development of a Veterans Administration occurrence screening program. Quality Review Bulletin, 15(10): 315-319.

Goldman RL, Walder DJ (1992). An initial assessment of the Veteras Affairs occurrence screening program. Quality Review Bulletin, 18(10):327-332.

Goldman RL (1992). The reliability of peer assessments of quality of care. Journal of the American Medical Association, 267(7): 958-960.

Graham NO (1990). Historical perspective and regulations regarding quality assessment. In: Graham NO (ed). Quality assurance in hospitals: strategies for assessment and implementation. Rockville, Maryland: Aspen Publishers.

Grimshaw J, Freemantle N, Wallace S et al (1995). Developing and implementing clinical practice guidelines. Quality in Health Care, 4:55-64.

Griffiths R (1983). NHS management inquiry report. London: Department of Health and Social Security.

Groves EW (1908). A plea for a uniform registration of operation results. British Medical Journal, ii:1008-1009.

Gruer R, Gunn AA, Gordon DS et al (1986). Audit of surgical audit. Lancet, i:23-26.

Ham CH (1985). From efficiency monitoring to quality assurance: the development of monitoring in the NHS. Hospital and Health Service Review, 81(3):110-113.

Hartwig SC, Denger SD, Schneider PJ (1991). Severity indexed, incident report-based medication error reporting program. American Journal of Hospital Pharmacy 48:2611-2616.

Harvard Medical Practice Study (1990). Patients, doctors and lawyers: medical injury, malpractice litigation and patient compensation in New York. Harvard, New York: Harvard College.

Hays WL (1994). Statistics for the social sciences (5th edition). Forth Worth, Texas: Harcourt Brace.

Heywood AJ, Wilson IH, Sinclair JR (1989). Perioperative mortality in Zambia. Annals of the Royal College of Surgeons of England, 71:354-358.

Hiatt HH, Barnes BA, Brennan TA, Laird NM, Lawthers AG, Leape LL et al (1989). A study of medical injury and medical malpractice: an overview. New England Journal of Medicine, 321(7):480-484.

Hopkins A (1990). Measuring the quality of medical care. London: Royal College of Physicians.

Hopkins A, Gabbay J, Neuberger J (1994). Role of users in achieving a quality service. Quality in Health Care, 3:203-209.

Hopkins A (1995). Some reservations about clinical guidelines. Archives of Disease in Childhood, 72(1):70-75.

Hopkins A (1996). Clinical audit: time for a reappraisal? Journal of the Royal College of Physicians of London. 30(5):415-25.

Horn SD, Horn RA (1986). Reliability and validity of the severity of illness index. Medical Care, 24(2):159-178.

Hulka BS (1979). Peer review in ambulatory care. Medical Care, 17(3) Suppl:1-75.

Illich I (1975). Medical nemesis: the expropriation of health. London: Marion Boyes, 1975.

Institute of Medicine (Lohr K ed) (1990). Medicare: a strategy for quality assurance. Volume I. Washington: National Academy Press.

Jencks SF, Wilensky GR (1992). The Health Care Quality Improvement Initiative: a new approach to quality assurance in Medicare. Journal of the American Medical Association, 268(7): 900-903.

Jenkins D. Investigations: how to get from guidelines to protocols. British Medical Journal, 303:323-324.

Jessee WF (1977). Quality assurance systems: why aren't there any? Quality Review Bulletin, 3(11):16-18.

Joint Commission on Accreditation of Healthcare Organisations (1988). Accreditation Manual for Hospitals. Chicago: JCAHO.

Joss R, Kogan M (1995). Advancing quality: total quality management in the National Health Service. Buckingham: Open University Press.

Jost TS (1989). Medicare Peer Review Organisations. Quality Assurance in Health Care, 1(4):235-248.

Jost TS (1990). Assuring the quality of medical practice: an international comparative study. London: King's Fund.

Juran JM (1979). Quality control handbook. New York: McGraw-Hill.

Kahn KL, Brook RH, Draper D, Keeler EB, Rubenstein LV, Rogers WH, Kosecoff J (1988). Interpreting hospital mortality data: how can we proceed? Journal of the American Medical Association, 260(24):3625-3628.

Kane RL (1980). Iatrogenesis: just what the doctor ordered. Journal of Community Health, 5(3):149-158.

Kaplan SH, Ware JE (1989). The patient's role in healthcare and quality assessment. In: Goldfield N, Nash DB (ed). Providing quality care: the challenge to clinicians. Philadelphia: American College of Physicians.

Kelson M (1996). User involvement in clinical audit: a review of developments and issues of good practice. Journal of Evaluation in Clinical Practice, 2(2):97-109.

Kind P (1988). The development of health indices. In: Teeling Smith G (ed). Measuring health: a practical approach. London: John Wiley and Sons.

King TM, Jones JL (1989). The Johns Hopkins Health System quality review process. In: Spath PL (ed). Innovations in health care quality management. Chicago: American Hospital Publishing.

Kings Fund (1996). Hospital accreditation programme: organisational standards and criteria. London: Kings Fund.

Kitson A (1989). A framework for quality: a patient centred approach to quality assurance in healthcare. Harrow, Middlesex: Royal College of Nursing/Scutari Press.

Kleefield S, Churchill WW, Laffel G (1991). Quality improvement in a hospital pharmacy department. Quality Review Bulletin, 17(5): 138-143.

Koch H, Higgs A (1991). What does quality healthcare cost? International Journal of Health Care Quality Assurance, 4(4):4-7.

Komaroff A (1978). The PSRO quality assurance blues. New England Journal of Medicine, 298(21):1194-1196.

Kritchevsky SB, Simmons BP (1991). Continuous quality improvement: concepts and applications for physician care. Journal of the American Medical Association, 266(13):1817-1823.

Krukowski ZH, Matheson NA (1988). Ten year computerised audit of infection after abdominal surgery. British Journal of Surgery, 75(9):857-861.

Laffel G, Blumenthal D (1989). The case for using industrial quality management science in health care organisations. Journal of the American Medical Association, 262(20):2869-2873.

Laffel G, Berwick D (1992). Quality in health care. Journal of the American Medial Association, 268(3):407-409.

Lakshmanan MC, Hershey CO, Breslau D (1980). Hospital admissions caused by iatrogenic disease. Archives of Internal Medicine, 146:1931-1934.

Lancet (1984). NHS Health Advisory Service [editorial]. ii: 763-764

Lang NM, Clinton JF (1984). Assessment of quality of nursing care. Annual Review of Nursing Research, 2:135-163.

Lawrence-Brown MMD, Manning K (1989). Experience of a concurrent screening quality assurance programme on a vascular surgical unit. Australian Clinical Review, 9:17-26.

Leape LL (1994). Error in medicine. Journal of the American Medical Association 272(23): 1851-1857.

Leape LL, Brennan TA, Laird N et al (1991). The nature of adverse events in hospitalised patients: results of the Harvard Medical Practice Study II. New England Journal of Medicine, 324(6):377-384.

Lembcke PA (1956). Medical auditing by scientific methods illustrated by major female pelvic surgery. Journal of the American Medical Association, 162(7):646-655.

Lembcke PA (1967). Evolution of the medical audit. Journal of the American Medical Association, 199(8):111-118.

Lewis P, Charny M (1992). Early bath. British Journal of Healthcare Computing, 9(4):36.

Localio AR, Lawthers AG, Brennan TA et al (1991). Relation between malpractice claims and adverse events due to negligence: results of the Harvard Medical Practice Study III. New England Journal of Medicine, 325(4):245-251.

Localio AR, Weaver SL, Landis JR et al (1996). Identifying adverse events caused by medical care: degree of physician agreement in a retrospective chart review. Annals of Internal Medicine, 125: 457-464.

Lohr K (1990). A strategy for quality assurance in Medicare. New England Journal of Medicine, 322(10):707-712.

Longo DR, Wilt JE, Laubenthal RM (1986). Hospital compliance with Joint Commission standards: findings from 1984 surveys. Quality Review Bulletin, 12(11):388-94.

Longo DR, Ciccone KR, Lord JT (1989). Integrated quality assessment: a model for concurrent review. Chicago: American Hospital Publishing.

Lord J, Littlejohns P (1997). Evaluating healthcare policies: the case of clinical audit. British Medical Journal, 315:668-671.

Lyall J (1990). Switching on to medical audits. Health Service Journal, 100(5224):1596.

Main DS, Pace WD (1991). Measuring health: guidelines for reliability assessment. Family Medicine, 23(3):227-30.

Manning K, Lawrence-Brown MMD, Hirsch RL, Tibbett P, North J, Mackay L, Archibald J, Brown M (1990). Experience of a concurrent screening quality assurance programme in an emergency department. Australian Clinical Review, 10:117-125.

Mant J, Hicks N (1995). Detecting differences in quality of care: the sensitivity of measures of process and outcome in treating acute myocardial infarction. British Medical Journal, 311:793-796.

Massanari RM (1992). Quality improvement: controlling the risk of adverse events. In: Wenzel RP (ed). Assessing quality healthcare: perspectives for clinicians. Baltimore, Maryland: Williams and Wilkins.

Maxwell RJ (1984). Quality assessment in health. British Medical Journal, 288:1470-1472.

Mayer W, Clinton JJ, Newhall D (1988). A first report of the Department of Defense external civilian peer review of medical care. Journal of the American Medical Association, 260(18):2690-2693.

Maynard A (1991). Case for auditing audit. Health Service Journal, 101(18 July):26.

McAuliffe WE (1978). Studies of process-outcome correlations in medical care evaluations: a critique. Medical Care, 14(11):907-929.

McDonald I (1991). Coming up to standard. International Journal of Health Care Quality Assurance, 4(4):17-20.

McDonald PJ, Royle GT, Taylor I, Karran SJ (1991). Mortality in a university surgical unit: what is an avoidable death? Journal of the Royal Society of Medicine, 84:213-216.

McDowell I and Newell C (1991). Measuring health : a guide to rating scales and questionnaires. Oxford: Oxford University Press.

McIntyre N, Popper K (1983). The critical attitude in medicine: the need for a new ethics. British Medical Journal, 287:1919-1923.

McLamb JT, Huntley RR (1967). The hazards of hospitalisation. Southern Medical Journal, 60(5):469-472.

Mendenhall S (1987a). The use of billing data in quality assurance. Quality Review Bulletin, 13(1):31-33.

Mendenhall S (1987b). The ICCS code: a new development for an old problem. Proceedings of Symposium on Computer Applications in Medical Care, November 1987. Washington: IEEE Computer Society Press.

Merrison A (1979). Royal Commission on the National Health Service. London: HMSO.

Merry MD (1987). What is quality care? A model for measuring health care excellence. Quality Review Bulletin, 13(9):298-301.

Mills DH (1978). Medical insurance feasibility study - a technical summary. Western Journal of Medicine, 128(4):360-365.

Mills DH, Von Bolschwing GE (1995). Clinical risk management: experiences from the United States. In: Vincent C (ed). Clinical risk management. London: BMJ Publishing.

Mills I (1987). Outcome measures - getting there. Health Service Journal, xx(xx 16 Jul):822.

Milne R, Clarke A (1990). Can readmission rates be used as an outcome indicator? British Medical Journal, 301:1139-40

Mitchell MW, Fowkes FG (1985). Audit reviewed: does feedback on performance change clinical behaviour? Journal of the Royal College of Physicians of London. 19(4):251-254.

Morgan C, Murgatroyd S (1994). Total quality management in the public sector. Buckingham: Open University Press.

Morlock L, Lindgren OH, Mills DH (1989). Malpractice, clinical risk management and quality assessment. In: Goldfield N, Nash DB (ed). Providing quality care: the challenge to clinicians. Philadelphia: American College of Physicians.

Morlock L, Malitz FE (1991). Do hospital risk management programmes make a difference? Relationships between risk management program activities and hospital malpractice claims experience. Law and Contemporary Problems, 54(2):1-22.

Morse JM (1991). Strategies in sampling. In: Morse JM (ed). Qualitative nursing research: a contemporary dialogue. London: Sage Publications.

Moser RH (1956). Diseases of medical progress. New England Journal of Medicine, 255(13):606-614.

Moss F, Smith R (1991). From audit to quality and beyond. British Medical Journal, 303:199-200.

Mugford M, Banfield P, O'Hanlon M (1991). Effects of feedback of information on clinical practice: a review. British Medical Journal, 303: 398-402.

Murphy JG, Jacobson S (1984). Assessing the quality of emergency care: the medical record versus patient outcome. Annals of Emergency Medicine 13(3):158-165.

National Audit Office (1988). Quality of clinical care in NHS hospitals. London: HMSO.

National Committee For Quality Assurance (1996). Standards for the accreditation of managed care organizations. Washington: NCQA.

NCEPOD (1989). The report of the National Confidential Enquiry into Perioperative Deaths. London: NCEPOD.

Nelson AR (1976). Orphan data and the unclosed loop: a dilemma in PSRO and medical audit. New England Journal of Medicine, 295(11):617-619.

Nobrega FT, Morrow GW, Smoldt RK et al (1977). Quality assessment in hypertension: analysis of process and outcome methods. New England Journal of Medicine, 296(3):145-148.

Nordal CA, Ang J (1988). Occurrence screening in long term care. Dimensions in Health Services, 65(2):23-27.

Norman I, Redfern S (1995). What is audit? In: Kogan M, Redfern S (eds). Making use of clinical audit. Buckingham: Open University Press.

Norman IJ, Redfern SJ, Tomalin DA et al (1992). Developing Flanagan's critical incident technique to elicit indicators of high and low quality nursing care from patients and their nurses. Journal of Advanced Nursing. 17(5):590-600.

Oakland JS (1989). Total quality management. Oxford: Heinemann Professional Publishing.

O'Connor GT, Plume SK, Olmstead EM et al (1996). A regional intervention to improve the hospital mortality associated with coronary artery bypass graft surgery. Journal of the American Medical Association, 275(11):841-846.

O'Leary MR, Smith MS, O'Leary DS et al (1989). Application of clinical indicators in the emergency department. Journal of the American Medical Association, 262(24):3444-3447.

O'Leary DS (1991). Beyond generic occurrence screening. Journal of the American Medical Association, 265(15):1993-1994.

Øvretveit J (1994). All together now. Health Service Journal, 1 Dec: 24-26.

Øvretveit J (1997). Would it work for us? Learning from quality improvement in Europe and beyond. Joint Commission Journal on Quality Improvement, 23(1):7-22.

Palmer RH, Louis TA, Peterson HF et al (1996). What makes quality assurance effective? Results from a randomised controlled trial in 16 primary care group practices. Medical Care, 34(9):SS29-SS39.

Panniers TL, Newlander J (1986). The adverse patient occurrences inventory: validity, reliability and implications. Quality Review Bulletin, xx(Sept):311-315.

Panniers TL (1987). Severity of illness, quality of car and physician practice as determinants of hospital resource consumption. Quality Review Bulletin, xx(May):158-165.

Patton MQ (1990). Qualitative evaluation and research methods (2nd edition). London: Sage Publications.

Phelps CE (1976). Benefit/cost analysis of quality assurance programs. In: Egdahl RH, Gertman PM (ed). Quality assurance in health care. Germantown, Maryland: Aspen Systems Corporation.

Plsek P (1997). Systematic design of healthcare processes. Quality in Health Care, 6(1):40-48.

Pollitt C (1990). Doing business in the temple? Managers and quality assurance in the public services. Public Administration, 68: 435-452.

Pollitt C (1996). Business approaches to quality improvement: why are they hard for the NHS to swallow. Quality in Health Care, 5(2):104-110.

Practitioner (1980). Editorial. Down with audit. Practitioner, 223:427-8.

Prass RL (1996). Iatrogenic facial nerve injury: the role of facial nerve monitoring. Otolaryngologic Clinics of North America. 29(2):265-75.

Pringle M, Bradley CP, Carmichael CM et al (1995). Significant event auditing: a study of the feasibility and potential of case-based auditing in primary medical care. London: Royal College of General Practitioners.

Reason J (1995). Understanding adverse events: human factors. Quality in Health Care 4(2):80-89.

Reerink E (1990). Defining quality of care: mission impossible? Quality Assurance in Health Care, 2(3/4):197-202.

Relman A (1989). Confronting the crisis in healthcare. Technology Review, July 1989:31-40.

Reynolds JL (1995). Reducing the frequency of episiotomies through a continuous quality improvement program. Canadian Medical Association Journal, 153(3):275-282.

Rich MW, Shah AS, Vinson JM et al (1996). Iatrogenic congestive heart failure in older adults: clinical course and prognosis. Journal of the American Geriatrics Society. 44(6):638-43.

Richards T, Lurie N, Rogers WH, Brook RH (1988). Measuring differences between teaching and nonteaching hospitals. Medical Care, 25(5)Suppl:S1-S140.

Richardson FM (1972). Peer review of medical care. Medical Care, 10(1):29-39.

Robinson ML (1988). Sneak preview: JCAHO's quality indicators. Hospitals, 5 July:38-43.

Robinson LA, Stacy R, Spencer JA, et al (1995). Use of facilitated case discussions for significant event auditing. British Medical Journal, 311(7000):315-8.

Rooney E (1988). A proposed quality system specification for the National Health Service. Quality Assurance, 14(2):45-53.

Roos LL, Cageorge SM, Austen E, Lohr KN (1985). Using computers to identify complications after surgery. American Journal of Public Health, 75(11):1288-1295.

Roos NP, Roos LL, Mosset J, Havens B (1988). Using administrative data to predict important health outcomes: entry to hospital, nursing home, and death. Medical Care, 26(3):221-239.

Rosenberg SN (1990). Choosing the assessment method that meets your needs. In: Graham NO (ed). Quality assurance in hospitals: strategies for assessment and implementation. Rockville, Maryland: Aspen Publishers.

Royal College of Physicians (1989). Medical audit - a first report: what, why and how? London: Royal College of Physicians.

Royal College of Surgeons (1989). Guidelines to clinical audit in surgical practice. London: Royal College of Surgeons.

Runciman WB, Sellen A, Webb RK et al (1993). The Australian Incident Monitoring Study. Errors, incidents and accidents in anaesthetic practice. Anaesthesia & Intensive Care. 21(5):506-19.

Rutstein DD, Berenberg W, Chalmers TC, Child CG, Fishman AP, Perrin EB (1976). Measuring the quality of medical care: a clinical method. New England Journal of Medicine, 294(11):582-588.

Rutstein DD, Mullan RJ, Frazier TM, Halperin WE, Melius JM, Sestito JP (1983). The principle of the sentinel health event and its application to the occupational diseases. American Journal of Public Health, 73:1054-1062.

Sanazaro PJ (1974). Medical audit: experience in the USA. British Medical Journal, i: 271-282.

Sanazaro PJ, Worth RM (1978). Concurrent quality assurance in hospital care: report of a study by Private Initiative in PSRO. New England Journal of Medicine, 298(21):1171-1177.

Sanazaro PJ, Mills DH (1991). A critique of the use of generic screening in quality assessment. Journal of the American Medical Association, 265(15):1977-1981.

Sanderson, HF (1987). Performance indicators. British Journal of Hospital Medicine. 37(3):245, 248, 250-1.

Sartwell PE (1974). Iatrogenic disease: an epidemiologic perspective. International Journal of Health Services, 4(1):89-93.

Schimmel EM (1964). The hazards of hospitalisation. Annals of Internal Medicine, 60(1):100-110.

Schumacher DN, Parker B, Kofie V, Munns JM (1987). Severity of illness index and the adverse patient occurrence index: a reliability study and policy implications. Medical Care, 25(8):695-704.

Scrivens E (1995). Accreditation: protecting the professional or the consumer? Buckingham: Open University Press.

Sellu D (1986). Audit: its effect on the performance of a surgical unit in a district general hospital. Hospital and Health Service Review, 82:64-69.

Shanks J, Frater A (1993). Health status, outcome and attributability: is a red rose red in the dark? Quality in Health Care, 2:259-262.

Shaw CD (1980a). Aspects of audit: the background. British Medical Journal, i:1256-1258.

Shaw CD (1980b). Aspects of audit: acceptability of audit. British Medical Journal, i:1443-1446.

Shaw CD (1989). Medical audit: a hospital handbook. London: King's Fund.

Shaw CD (1990). Criterion based audit. British Medical Journal, 300:649-651.

Shaw CD (1992). Specialty medical audit. London: King's Fund Centre.

Sheldon TA (1994). Please bypass the PORT. British Medical Journal, 309(6948):142-3.

Siegel S, Castellan NJ (1988). Non-parametric statistics for the behavioural sciences (2nd edition). New York: McGraw Hill.

Smith C (1992). Occurrence screening in general practice. Medical Audit News, 2(2):41.

Smith R (1991). Where is the wisdom…? The poverty of medical evidence. British Medical Journal 303: 798-799.

Standing Medical Advisory Committee, Department of Health (1990). The quality of medical care. London: HMSO.

Steel K, Gertman PM, Crescenzi C, Anderson J (1981). Iatrogenic illness on a general medical service at a university hospital. New England Journal of Medicine, 304(11):638-642.

Steffen GE (1988). Quality medical care: a definition. Journal of the American Medical Association, 260(1):56-61.

Stevens G, Wickings HI, Bennett J (1988). Medical quality assurance: research in Brighton Health Authority. International Journal of Health Care Quality Assurance, 1(2):5-11.

Stevens G, Bennett J (1989). Clinical audit - occurrence screening for QA. Health Service Management, 85(Aug):178-181

St Leger AS, Schneiden H, Walsworth-Bell JP (1992). Evaluating health services' effectiveness. Milton Keynes: Open University Press.

Stuart D (1989). Beyond MMA - lessons for Australian healthcare. Sydney: Royal North Shore Hospital.

Tancredi LR, Bovbjerg RR (1992). Creating outcomes-based systems for quality and malpractice reform: methodology of accelerated compensation events. Milbank Quarterly, 70(1):183-216.

Tarlov AR, Ware JE, Greenfield S, Nelson EC, Perrin E, Zubkoff M (1989). The Medical Outcomes Study: an application of methods for monitoring the results of medical care. Journal of the American Medical Association, 262(7):925-930.

Thomson MA, Oxman AD, Haynes RB, Davis DA, Freemantle N, Harvey EL. Audit and feedback to improve health care professional practice and health care outcomes (Part I) [Protocol]. In: Bero L, Grilli R, Grimshaw J, Oxman A Collaboration on Effective Professional Practice Module of The Cochrane Database of Systematic Reviews , [updated 03 March 1997]. Available in The Cochrane Library [database on disk and CDROM]. The Cochrane Collaboration; Issue 2. Oxford: Update Software; 1997. Updated quarterly.

Thompson JS, Prior MA (1992). Quality assurance and mortality and morbidity conferences. Journal of Surgical Research 52:97-100.

Tribe DMR, Korgaonkar G (1989). Medical negligence - 1. Journal of Management in Medicine, 4(3):204-209.

Trunet P, Le Gall JR, Lhoste F, Regnier B, Saillard Y, Carlet J, Rapin M (1980). The role of iatrogenic disease in admissions to intensive care. Journal of the American Medical Association, 244(23):2617-2620.

UK Clearing House on Health Outcomes (1996). Outcomes briefing 7: outcomes within clinical audit. Leeds: Nuffield Institute for Health.

Vuori H (1989). Research needs in quality assurance. Quality Assurance in Health Care, 1(2/3):147-159.

Walshe K, Lyons C, Coles J, Bennett J (1991). Quality assurance in practice: research in Brighton Health Authority. International Journal of Health Care Quality Assurance, 4(2):27-35.

Walshe K, Coles J (1993). Evaluating audit in the NHS: developing a framework. London: CASPE Research.

Walshe K, Coles J (1993b). Medical audit: evaluation needed. Quality in Health Care, 2:189-190.

Walshe K (1995). Evaluating clinical audit: past lessons, future directions. London: Royal Society of Medicine, 1995.

Walshe K (1995b). Opportunities to improve the practice of clinical audit. Quality in Health Care, 4(4):231-232.

Walshe K (1997). Indicators won't turn the tables. Health Service Journal 1997; 107(5562): 24.

Walshe K, Bennett J, Ingram D (1995). Using adverse events in healthcare quality improvement: results from a British acute hospital. International Journal of Health Care Quality Assurance, 8(1):7-14.

Walshe K, Buttery Y (1995). Measuring the impact of audit and quality improvement activities. Journal of the Association for Quality in Healthcare, 2(4):138-147.

Welch CE, Grover PL (1991). An overview of quality assurance. Medical Care 29(9):AS8-AS28.

Whitehead TP, Woodford FP (1981). External quality assessment of clinical laboratories in the United Kingdom. Journal of Clinical Pathology, 34:947-957.

Wilkin A, McColl I (1987). Surgical audit: the clinician's view. Theoretical Surgery 1:195-206.

Williamson JA, Mackay P (1991). Incident reporting. Medical Journal of Australia, 155:340-344.

Williamson JW (1978). Formulating priorities for quality assurance activities. Description of a method and its application. Journal of the American Medical Association, 239:631-637.

Wilson RM, Runciman WB, Gibberd RW (1995). The quality in Australian healthcare study. Medical Journal of Australia 163:458-471.

Wolff AM (1992). Limited adverse occurrence screening: a medical quality control system for medium sized hospitals. Medical Journal of Australia, 156:449-452.

Wolff AM (1995). Limited adverse occurrence screening: an effective and efficient method of medical quality control. Journal of Quality in Clinical Practice 15:221-223.

Woodyard J (1990). Facing up to errors. Health Service Journal, 100(5194):468-469.

World Health Organisation (1985). Targets for health for all. Copenhagen: World Health Organisation Regional Office for Europe.

Yates JM, Davidge MG (1984). Can you measure performance? British Medical Journal, 288:1935-1936.

Yates JM, Vickerstaff L (1982). Interhospital comparisons in mental handicap. Mental Handicap, 10:45-47.

Yim SF, Lam SK, Haines CJ (1996). Iatrogenic cardiac tamponade during pregnancy. Australian & New Zealand Journal of Obstetrics &Gynaecology. 36(2):205-6.

Appendices

4.1 Generic adverse-event measure of quality developed for and used in the RSCH occurrence screening project.

G01 Unexpected admission following outpatient management.

Unexpected admission following treatment in the outpatient or accident and emergency department of the hospital, or in the community.

DEFINITION

The patient's admission was the direct result of the adverse results of or some complication of outpatient management, given by hospital staff in outpatients departments or Accident and Emergency, or by general practitioners or other community health services.

EXCEPTIONS

Patients whose admission is expected or planned as a result of a chronic condition such as unstable diabetes, carcinoma, glaucoma or cataracts, or progressive renal failure.

GUIDANCE NOTES

Look for evidence of delayed diagnosis or treatment, failures or breakdowns in service provision, complications of outpatient drug therapy or procedures.

If the admission resulted from care provided by another district, or by community services outside the Health Authority, this should be noted.

EXAMPLES

Delayed diagnosis; any condition attributed to outpatient procedures e.g. radiation burns; wound infections; delayed treatment.
In renal medicine, sudden worsening renal failure due to drug therapy e.g. ACE inhibitors, tetracycline, NSAI's, diuretics, infection when on immunosuppressive therapy.

G02 Unexpected readmission to hospital.

Unexpected readmission to hospital following a previous admission.

DEFINITION:

The reason for the patient's readmission was the development/diagnosis of complications arising from a previous admission, or the incomplete management of problems diagnosed or present during the previous admission.

EXCEPTIONS:

Patients whose re-admission was planned and documented at the time of the previous admission, or who are expected to have multiple admissions.

GUIDANCE NOTES:

Look for complications or problems unresolved during the previous admission requiring readmission and examine carefully all readmissions within 7 days of previous discharge.

For General Surgery record all re-admissions within 7 days of discharge.

If the previous admission was to a hospital outside this district, this should be noted.

EXAMPLES:

Following dilatation and curretage, patient readmitted for retained products.
Patient readmitted with unresolved infection following inpatient antibiotic treatment for chest infection.
In Renal patients: Readmission for bleeding following renal biopsy;
Readmission post renal transplant for surgical complication , i.e. not rejection;
Readmission post Tenchoff catheter placement for P.D. fluid leak, or AT ANY TIME for hernia;
Readmission post A-V fistula formation for infection, or revision of fistula. [See also RE 02]

G03 Error in obtaining consent to operative procedure.

Errors made in obtaining and documenting the patient's consent to operative procedure(s).

DEFINITION

The operation consent form(s) for the procedure(s) carried out on the patient cannot be found in the notes, or can be found but contain error(s).

EXCEPTIONS

Patients undergoing emergency surgery, where the patient and patient's family could not give their consent in advance of surgery. Reasons for this should be documented in the notes.
Omission of the hospital name at the top of the consent form should not be included under this criteria.
Life threatening problems found and dealt with during surgery.

GUIDANCE

Look for operation consent forms on which:
- the details of the procedure do not correspond with the operation notes.

- sections are not completed or are incorrect,inaccurate illegible or contain abbreviations. L. and R. for left and right is permissible if clear. Omission of "hospital":
- doctor's or patient's signature missing.

EXAMPLES of renal operations: Renal transplant, Transplant nephrectomy, A-V fistula formation, Ureteric stent formation, Tenchoff catheter placement or removal, renal biopsy, parathyroidectomy.

G04 Unplanned removal/injury/repair of structure during surgery.

Unplanned removal, injury, or repair of an organ or structure during surgery or invasive procedure.

DEFINITION:

During surgery or invasive procedure,the unplanned removal,injury or repair of an organ or structure occurred.

EXCEPTIONS:

There are no exceptions to this criteria.

GUIDANCE NOTES:

Compare operation proposed on consent form with operation performed as documented in the notes. Look for evidence of:
a) Departure from proposed procedure.
b) Injury to an organ or structure.
c) Removal of healthy organs or tissue.
d) Repair to structures/organs inadvertantly damaged.
e) Confirmation of above by reference to pathology reports.
f) Any injury to the patient undergoing anaesthesia and surgery.

Look particularly closely at procedures where the risk of such incidents is higher, such as: intubations, percutaneous aspirations, biopsies, catheterisations, endoscopic procedures, lumbar punctures.

EXAMPLES:

Perforation of bowel during during endoscopic procedure.
Corneal abrasion occurred while under anaesthetic for colporrhaphy.
Paralysis of vocal cord due to recurrent laryngeal nerve damage during parathyroidectomy.
Nephrectomy after renal biopsy because of bleeding but NOT because of severe irreversible rejection.

EXCEPTIONS:
Some renal operations e.g. Renal transplant may necessitate extra procedures, (once surgeon has opened the abdomen) to be successful, e.g. native nephrectomy and use of patient's own ureter.

G05 Unplanned return to theatre.

Unplanned return to theatre for complications of a previous procedure.

DEFINITION

An unplanned return to theatre for a second or subsequent procedure occurred.

EXCEPTIONS

Planned second procedures, or second stages of two stage procedures.

GUIDANCE

For indications of whether the return to theatre was planned, check the operation notes for the first procedure. Check for procedures repeated because of lack of success on first attempt.

Check for surgical correction of complications e.g. insertion of a drain for a wound infection.

EXAMPLES

Return to theatre following haemorrhoidectomy for haemostasis for post-operative bleeding.
Return to theatre for re-suturing following burst abdominal wound.

G06 Delay or error in diagnosis

An error or undue delay in diagnosing the patient's condition is documented.

DEFINITION

An error or undue delay in diagnosing the patient's condition is documented in the record, which has resulted in a significant difference in the patient's care or treatment. Errors or delays may come to light through:

a) Pathology/histology results following surgery
b) Postmortem report
c) Documentation of delayed diagnosis in later medical records
d) Documentation of error in diagnosis in later medical records

EXCEPTIONS

There are no exceptions to this criterion.

GUIDANCE

Check medical notes for evidence of delays in diagnosing conditions, or for erroneous diagnoses later corrected. Check for negative histology reports following surgery. For all deaths where post-mortems are carried out, check the post-mortem report opinion on cause of death against the ante-mortem diagnosis and treatment.

Renal patients: Level of Diagnosis.
An accurate diagnosis of renal failure may be made, but the patient not treated due to unsuitability for dialysis therapy. There may be an accompanying inaccurate guess as to the cause of renal failure, the true diagnosis later revealed at post-mortem. This is irrelevant to the treatment the patient received.

INFORMATION TO RECORD

Record details of the timescale on which decisions about diagnosis were made and changed, and evidence of the effect of any delay or error on the patient's care and treatment.

EXAMPLES
Fracture missed on original admission, but later diagnosed and treated.
Mastectomy performed and histological report shows tissue to be benign.
Polyps removed presumed benign, sent for histological examination and

G07 Transfusion problems:reactions,complications,usage.

Problems arising from transfusions of blood or blood derivatives, such as significant complications and reactions or misuse of the service or unavailability of blood or blood derivatives.

DEFINITION

Following transfusion of blood or blood derivatives, significant reaction(s) or complication(s) occurred, or the Blood Transfusion Service was misused or blood/blood derivatives were unavailable.

EXCEPTIONS

There are no exceptions to this criterion.

GUIDANCE

Look for mismatch of blood group.
Look for evidence of reaction/complication such as rash, rigor, breathing difficulties, temperature spike greater than 37.5C , tachycardia, loin pain, chest pain, hypotension, jaundice following transfusion (within 12 hours), renal failure. Check for evidence that the complication/reaction was addressed, such as slowing or stopping transfusion rate, drug therapy, etc.
Single unit transfusions are not indicated (except for some renal patients) therefore record details.

Renal patients: Dialysis patients often receive transfusions for severe symptomatic anaemia, usually two units per transfusion, but occasionally one.

EXAMPLES

Urticaria and pyrexia following transfusion of whole blood.
Mis-matched blood given in error followed by renal failure.

G08 Hospital acquired infection.

Infection acquired during the patient's admission.

DEFINITION

The patient developed an infection during their admission, which was acquired after they had been admitted.

EXCEPTIONS
There are no exceptions to this criterion.

GUIDANCE

Look for evidence of freedom from infection or presence of infection in medical or nursing notes on admission and check laboratory reports. Compare with later documentation of signs and symptoms and laboratory reports.
Urinary: Match up positive culture with reports of signs and symptoms, such as dysuria, haematuria, frequency, pyrexia.
Urinary tract infections are not a relevant event in the care of renal patients, and should NOT therefore be recorded.
Chest: Match up signs and symptoms such as coughing, pyrexia, purulent sputum and a positive culture. If negative result but signs and symptoms present check if patient is on antibiotics. If so, count as variation.
Wound: Match signs and symptoms e.g.pyrexia, redness, swelling, and drainage around surgical site, with positive culture.
Other infections: Match signs and symptoms with positive laboratory report.
Check date of diagnosis and onset of infection, and admission notes, to establish whether the infection was acquired pre or post admission.

EXAMPLES
Urinary tract infections in catheterised patients.

G09 Antibiotic/drug utilisation problems.

One (or more) of the specified drug/antibiotic usage problems occurred.

DEFINITION

One of more of the following specified drug/antibiotic usage problems occurred:

a) Medications omitted, prescribed or given in error, given at wrong rate,delayed with no reason given.
b) Medication reactions and anaphylaxis.

GUIDANCE

Check kardex for any documented medication errors. Note subsequent action.
Check for comments made by the pharmacist in notes or on prescription sheet.
Include in (a), error in prescribed dose in patients suffering from renal failure. Check prescription sheet for changed dosage.

EXAMPLES

Penicillin prescribed or given when known allergy documented.

G10-1 Cardiac or respiratory arrest,

Cardiac arrest,or respiratory arrest, occurring during the admission.

DEFINITION

The patient had a cardiac arrest or respiratory arrest, during their admission.

EXCEPTIONS

There are no exceptions to this criterion.

GUIDANCE

Look for evidence in records leading up to the event, such as major changes in vital signs, administration of medication or other treatments just prior to the event,which might suggest its cause.

Check records of treatment of the arrest/shock itself, such as calling of crash team, availability of resuscitation equipment,etc.
Check records for ECG prior to surgery and also availability of ECGs. See if result of ECG is in record.
Look for other clinical indicators-
- reports of tests for urea and electrolytes, (especially if K+ > 6.0)
- evidence of adequate hydration (fluid charts, medical notes) or overload

- if monitored, evidence of rhythm changes

G10-3 CVA or MI or PE following surgery.

A CVA or MI or PE occurred following a surgical procedure during the same admission.

DEFINITION:

A cerebrovascular accident or myocardial infarction occurred within the 48 hour period immediately following a surgical procedure, or pulmonary embolus occurred at any stage following surgery.

EXCEPTIONS:

There are no exceptions to this criterion.

GUIDANCE NOTES:

In Urology notes, also include patients who suffered a CVA, MI or PE at any time during their admission.
Please note whether there were any cardiovascular symptoms not addressed by admitting doctor and whether ECG was considered.

Pulmonary embolus: Look for documentation for clotting time pre-surgery and note whether anti-coagulant therapy was considered. Look at accounts of post-operative nursing care: deep breathing exercises, passive or active limb exercises and consideration for referral to physiotherapy and early ambulation.

EXAMPLES

Pulmonary embolus following bowel resection.
Cerebrovascular accident following hysterectomy.

G11 Unexpected transfer to higher dependency unit.

Unexpected transfer from general care to higher dependency unit.

DEFINITION:

Patient was transferred unexpectedly from general care to a higher dependency unit.

EXCEPTIONS:

Planned transfers.

GUIDANCE

Transfers to higher dependency units include transfers to coronary care units,intensive therapy units here or in other hospitals.

Check pre-operative medical records for evidence that a transfer was planned.
Look for reasons for transfer and examine events leading up to transfer.
Check pre-operative medical records for evidence that a transfer was planned.

EXAMPLES:

Patient transferred to intensive therapy unit following routine herniorrhaphy with breathing difficulties.

G12-1 Patient related clinical complications occurred.

One or more of the specified patient-related clinical complications occurred.

DEFINITION

One or more of the following patient-related clinical complications occurred during the admission:
b) IV problems: Overload. Phlebitis requiring treatment. Time Venflon tissued and times of subsequent action. Infected IV sites.
c) Pain control not addressed.
d) Development/worsening of skin problems resulting from pressure.
e) Faecal impaction developed during admission requiring physical evacuation.
g) Deep vein thrombosis not evident on admission.
h) Wound haematoma in breast surgery patients
i) Urinary retention occurred following surgery using epidural anaesthesia.
EXCEPTIONS
There are no exceptions to this criterion.

GUIDANCE
Check whole of medical and nursing records for complications specified. Note whether preventive measures taken and complications addressed.
b) Note time Venflon is recorded as tissued in nursing kardex, time reported to doctor and time Venflon replaced. Record descriptions of infected IV sites and action taken. c) Look for complaints of pain not addressed and the effect of analgesia not evaluated. Look for pain-control charts. In General Surgery, record for abdominal emergency admissions their arrival in A and E Dept, and time when first analgesia given.
d) For each patient who develops skin problems, record original nursing assessment of pressure risk, date problem first noticed, subsequent action.
h) Check Breast Operation form and nursing kardex.

G12-2 Patient-related non-clinical problems/incidents occurred

One or more of the specified patient-related non-clinical problems or incidents occurred.

DEFINITION

One or more of the following patient-related non-clinical problems or incidents occurred:

a) Theatre booking cancelled/delayed.
b) No ITU bed available when clinically needed
c) Patient transferred/admitted to/receiving care on an inappropriate ward because of bed shortages.
d) Delay in obtaining a second opinion from another specialty as recorded in the notes.
e) Casenotes, X-rays or other records or results missing/not available when needed.
g) Patient had slip or fall, or other accident.
h) Equipment failure.
i) Necessary equipment not available when needed.
j) Prescribed drugs not obtainable.
k) Delay in undergoing diagnostic or therapeutic procedures.
l) Portering service problems
m) Staff unavailable
n) Other

EXCEPTIONS

There are no exceptions to this criterion

GUIDANCE
Check medical and nursing records. Document the causes of the problem or incident if possible. Record if theatre instruments unavailable. Note if theatre delayed/cancelled due to TSSU problems or if ECG or other missing report delays operation. If notes are missing, record reason.
b) For General Surgery patients, record reason for non-availability of ITU bed if given. Record age of patient, with stated diagnosis and reasons for ITU admission.

EXAMPLES
Patient's X-rays are missing when they go to theatre.
Specimen results not available for review of treatment e.g. frozen section.
Patient staying nil by mouth while waiting long period due to theatre rescheduling.

G14 Neurological deficit on discharge not present on admission

Neurological deficit developed which was not present when the patient was admitted.

DEFINITION

The patient developed a neurological deficit during their admission which was not present when they were admitted.

EXCEPTIONS

Patients coming under criteria G10-3 and G12-1 and patients developing planned or expected neurological deficit.

GUIDANCE

Look for evidence of neurological damage or compromise throughout the medical and nursing records. Look particularly for neurological deficits resulting from surgery. Check assessments of orientation and sensory, circulatory and motor function. Look for evidence of seizures, urinary or faecal incontinence, or intractable pain which were not present or not documented on admission.

EXAMPLES
Facial palsy post-parotidectomy. Foot-drop following knee replacement or discectomy. Radial nerve palsy following humeral fracture plating. Laryngeal nerve palsy following parathyroidectomy. Median nerve palsy (or ulnar or radial) following formation of A-V fistula.

G15 Transfer to another hospital

Patient transferred to another hospital for one of the specified reasons.

DEFINITION

Patient transferred to another acute hospital for one of the following reasons:

- because of staff shortages, bed shortages, or ward closures

- for treatments normally available at this hospital

EXCEPTIONS

Transfers for treatment and tests not normally available at this hospital, and tertiary referrals. Transfers of patients to be nearer to home/relatives.

GUIDANCE

Look for the reasons for the transfer, and check the appropriateness of the original admission to this hospital.

G16 Unexpected death

Unexpected patient deaths.

DEFINITION

Patient died unexpectedly during admission

EXCEPTIONS

Patients admitted for terminal care or receiving terminal care, and patient recorded category C under the district resusitation policy.

GUIDANCE

Check nursing and medical records for evidence that the death was unexpected or any indications that it might have been prevented.

G17 Medical record review

Review of quality, consistency and completeness of medical records.

DEFINITION

One (or more) of the following specified deficiencies were found in the medical records:

a) Admission clerking not dated.
b) Notes not filled in every third day and at other important stages of clinical management.
e) Notes illegible or unsigned.
f) Incomplete or contradictory recording of information.
g) Breast Operation Form incomplete (Breast Surgery patients only).
h) Clinical information form not present in notes or incomplete.

EXCEPTIONS

There are no exceptions to this criterion

GUIDANCE

Check throughout medical records for the specified deficiencies.
See BHA guidance for the writing of the clinical record.
In b), if a patient is discharged within 3 days, ensure that an entry in the notes is made post-operatively by doctor. (Especially relevant in Ophthalmology.) g) Check completeness of form except for histology results which will not be available before patient discharge.

G18 Nursing record review

Review of quality, consistency and completeness of nursing records.

DEFINITION

One or more of the specified deficiencies were found in the nursing records:

a) No nursing assessment completed within 24 hours of admission
b) Patient's care plan missing or not updated
d) Incomplete, contradictory, or inadequate recording of information
e) Nurses' signatures missing on drug chart.
f) Property form not completed/no disclaimer.
g) Evidence of nursing insensitivity.

EXCEPTIONS

There are no exceptions to this criterion.

GUIDANCE

Check throughout the nursing records for the specified deficiencies
Compare TPR frequency with nursing orders in care plan. Ensure frequency of observations reflects the condition of the patient.

EXAMPLES

Patient's care plan giving post-operative nursing orders including care of intravenous infusion remains unchanged during remainder of patient's stay in hospital.
Quarter hourly observations following surgery become less frequent without corresonding adjustment in care plan.

G19 Evidence of patient and/or family dissatisfaction

Evidence of patient and/or family dissatisfaction in records

DEFINITION

There is evidence in the medical or nursing records, that the patient and /or the patient's relatives or friends expressed dissatisfaction with the care given to the patient.

EXCEPTIONS

There are no exceptions to this criterion.

GUIDANCE

Look for information on cause of dissatisfaction and how it was expressed in nursing and medical records. Look for evidence that the patient/family complaint was handled appropriately.

All discharges taken against medical advice should be carefully checked against this criterion.

Look for dissatisfaction with food, cleanliness, conduct of staff, clinical care, noise at night in ward, lack of privacy.

EXAMPLES

Patient's daughter complains to staff nurse that a window was left open overnight causing her mother to become chilled.
Patients complain about dirty washing facilities.
Patient took own discharge when operation postponed and incomplete explanation given by ward staff.

G20 Discharge related problems

Problems relating to the patient's discharge occurred.

DEFINITION

One (or more) of the following specified problems occurred in relation to the patient's discharge:
e) For General Surgery: Discharge checklist not completed.
c) Absence of copy of discharge summary to GP. (With renal patients also to referring hospital doctor.)
d) Discharge delayed for non-clinical (organisational or social) reasons
f) No evidence of patient information leaflet given to patient on discharge.
EXCEPTIONS
There are no exceptions to this criterion

GUIDANCE

Look through medical notes for copy of discharge prescription and letter to general practitioner. Look for evidence of assessment of patient for community services, provision of necessary dressings, drugs to take out, liaison with relatives and evidence of planned nursing discharge. Look for reasons for discharge delay.

EXAMPLES

Elderly patient remained in hospital for three weeks following medical

4.2 Pilot questionnaire used in the questionnaire study of clinician opinion.

Occurrence Screening Validity Study QUESTIONNAIRE

About this questionnaire

This questionnaire contains a set of screening criteria. it is intended to find out how good you think each screening criterion would be at selecting cases where there was some lapse in the standard of care. Clearly, none of the criteria can select every case where there has been some lapse, and every criterion will select some cases where there has been no lapse in the standard of care. The lower these "false negative" and "false positive" rates can be made, the better the screening criteria will be at selecting cases where the quality of care has been lower than it should have been. Therefore, both your opinion on the screening criteria as they are now, and your suggestions for improvements to the criteria will be very welcome.

Filling in this questionnaire

There is one sheet in this questionnaire for each screening criterion, and each sheet is laid out in exactly the same way.

The first part of the sheet gives you information about the screening criterion. It gives the title of the criterion, and its definition, and lists any known exceptions to the criterion. It also gives some guidance notes and example cases, which would be used by the person who was screening patients' records to guide them in applying the criterion.

The second part of the sheet contains a series of questions which we would like you to answer, about the screening criterion. Please answer these questions about the criterion exactly as it stands.

Having answered the questions, we would like you to give any suggestions for amendments or alterations to any part of the screening criterion which you feel would make it better at selecting cases where there has been some lapse in the standard of care.

Returning the questionnaire

When you have completed the questionnaire, please place it in the enclosed stamped addressed envelope and return it to me at CASPE Research.

Study report

If you would like to receive a copy of the study report based on the responses to these questionnaires, please tick the box on the right. [] tick box

G01 - Admission for the adverse results of or complications resulting from out-patient management.

DEFINITION
The patient's admission was the direct result of the adverse results of or some complication of out-patient management, given by hospital staff in outpatients departments or Accident and Emergency, or by general practitioners or other community health services.

EXCEPTIONS
Patients whose admission is expected or planned as a result of a chronic condition such as unstable diabetes, carcinoma, glaucoma or cataracts.

GUIDANCE NOTES
Look for evidence of delayed diagnosis or treatment, failures or breakdowns in service provision, complications of outpatient drug therapy or procedures.
If the admission resulted from care provided by another district, or by community services outside the Health Authority, this should be noted.

EXAMPLES
Delayed diagnosis; any condition attributed to out-patient procedures e.g. radiation burns; wound infections; delayed treatment.

---

1. How closely do you think this criterion is related to the quality of care that patients receive?

0% ¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦ 100%
Not at related — Very closely related

2. Do you think the information needed for this criterion could be found in patients' routine medical and nursing records?

0% ¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦ 100%
Never found — Always found

3. If all acute patients were screened using this criterion, what proportion of patients do you think would fit the criterion?

0% ¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦ 100%
No patients — All patients

4. Do you think that a detailed review of cases which fit this criterion would show that those cases had more lapses in the quality of care than those not fitting the criterion, or fewer lapses than those not fitting the criterion?

[] More lapses in quality of care than others
[] Same level of lapses as others
[] Fewer lapses than others

5. Do you think that a detailed review of cases which fit this criterion would show that those cases had, on average, a better quality of care than those not fitting the criterion, or a worse quality of care than those not fitting the criterion?

[] Better quality of care than others
[] Same quality of care as others
[] Worse quality of care than others

6. How serious, in terms of the effect on the patient's health, do you think the lapses in the quality of care identified in cases which fit this criterion would be?

0% ¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦ 100%
No effect — Very serious effect

7. Do you think this screening criterion could be altered to make it relate more closely to the quality of care that patients receive?

[] Yes [] No

If you answered YES to question 7, please give details of the amendments you would suggest below. Please continue overleaf if you need more space.

G02 - Re-admission for complications arising from a previous admission, or because of the incomplete management of problems on a previous admission.

DEFINITION
The reason for the patient's re-admission was the development/diagnosis of complications arising from a previous admission, or the incomplete management of problems diagnosed or present during the previous admission.

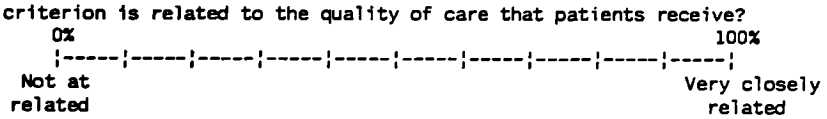
EXCEPTIONS
Patients whose re-admission was planned and documented at the time of the previous admission, or who are expected to have multiple admissions.

GUIDANCE NOTES
Look for complications or problems unresolved during the previous admission requiring re-admission and examine carefully all re-admissions within 7 days of previous discharge.
If the previous admission was to a hospital outside this district, this should be noted.

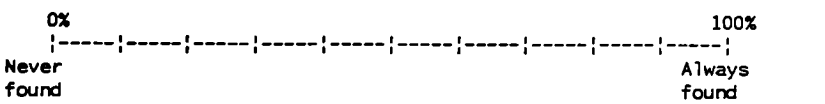
EXAMPLES
Following dilatation and curretage, patient re-admitted for retained products.
Patient re-admitted with unresolved infection following in-patient anti-biotic treatment for chest infection.

---

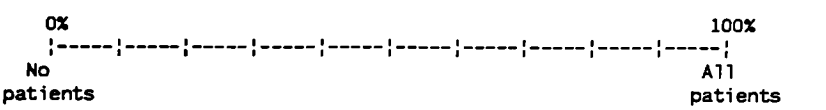
1. How closely do you think this criterion is related to the quality of care that patients receive?



2. Do you think the information needed for this criterion could be found in patients' routine medical and nursing records?



3. If all acute patients were screened using this criterion, what proportion of patients do you think would fit the criterion?



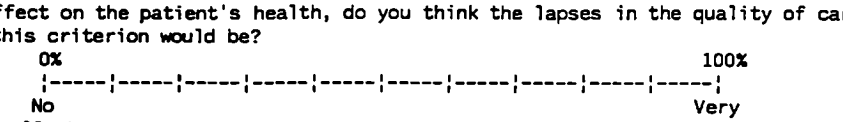
4. Do you think that a detailed review of cases which fit this criterion would show that those cases had more lapses in the quality of care than those not fitting the criterion, or fewer lapses than those not fitting the criterion?

[] More lapses in quality of care than others
[] Same level of lapses as others
[] Fewer lapses than others

5. Do you think that a detailed review of cases which fit this criterion would show that those cases had, on average, a better quality of care than those not fitting the criterion, or a worse quality of care than those not fitting the criterion?

[] Better quality of care than others
[] Same quality of care as others
[] Worse quality of care than others

6. How serious, in terms of the effect on the patient's health, do you think the lapses in the quality of care identified in cases which fit this criterion would be?



7. Do you think this screening criterion could be altered to make it relate more closely to the quality of care that patients receive?

[] Yes [] No

If you answered YES to question 7, please give details of the amendments you would suggest below. Please continue overleaf if you need more space.

G03 - Errors made in obtaining and documenting the patient's consent to operative procedure(s).

DEFINITION
The operation consent form(s) for the procedure(s) carried out on the patient cannot be found in the notes, or can be found but contain error(s).

EXCEPTIONS
Patients undergoing emergency surgery, where the patient and patient's family could not give their consent in advance of surgery. Reasons for this should be documented in the notes.
Life threatening problems found and dealt with during surgery.

GUIDANCE
Look for operation consent forms on which:
- the details of the procedure do not correspond with the operation notes.
- sections are not completed or are incorrect, inaccurate illegible or contain abbreviations. L. and R. for left and right is permissible if clear.
- doctor's or patient's signature missing.

---

1. How closely do you think this criterion is related to the quality of care that patients receive?

0% ¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦ 100%
Not at related — Very closely related

2. Do you think the information needed for this criterion could be found in patients' routine medical and nursing records?

0% ¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦ 100%
Never found — Always found

3. If all acute patients were screened using this criterion, what proportion of patients do you think would fit the criterion?

0% ¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦ 100%
No patients — All patients

4. Do you think that a detailed review of cases which fit this criterion would show that those cases had more lapses in the quality of care than those not fitting the criterion, or fewer lapses than those not fitting the criterion?

[] More lapses in quality of care than others
[] Same level of lapses as others
[] Fewer lapses than others

5. Do you think that a detailed review of cases which fit this criterion would show that those cases had, on average, a better quality of care than those not fitting the criterion, or a worse quality of care than those not fitting the criterion?

[] Better quality of care than others
[] Same quality of care as others
[] Worse quality of care than others

6. How serious, in terms of the effect on the patient's health, do you think the lapses in the quality of care identified in cases which fit this criterion would be?

0% ¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦ 100%
No effect — Very serious effect

7. Do you think this screening criterion could be altered to make it relate more closely to the quality of care that patients receive?

[] Yes [] No

If you answered YES to question 7, please give details of the amendments you would suggest below. Please continue overleaf if you need more space.

G04 - Unplanned removal, injury, or repair of an organ or structure during surgery or invasive procedure.

DEFINITION
During surgery or invasive procedure,the unplanned removal,injury or repair of an organ or structure occurred.

EXCEPTIONS
There are no exceptions to this criteria.

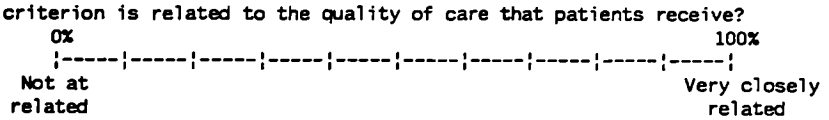
GUIDANCE NOTES
Compare operation proposed on consent form with operation performed as documented in the notes. Look for evidence of:
a) Departure from proposed procedure.
b) Injury to an organ or structure.
c) Removal of healthy organs or tissue.
d) Repair to structures/organs inadvertently damaged.
e) Confirmation of above by reference to pathology reports.
f) Any injury to the patient undergoing anaesthesia and surgery.

Look particularly closely at procedures where the risk of such incidents is higher, such as: intubations, percutaneous aspirations, biopsies, catheterisations, endoscopic procedures, lumbar punctures.

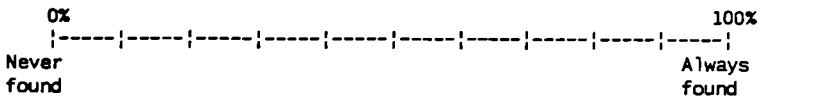
EXAMPLES
Perforation of bowel during during endoscopic procedure.
Corneal abrasion occurred while under anaesthetic for colporrhaphy.

---

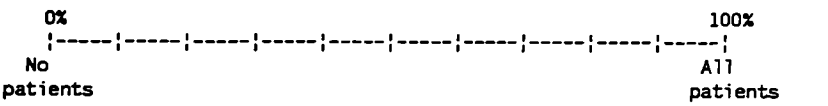
1. How closely do you think this criterion is related to the quality of care that patients receive?



2. Do you think the information needed for this criterion could be found in patients' routine medical and nursing records?



3. If all acute patients were screened using this criterion, what proportion of patients do you think would fit the criterion?



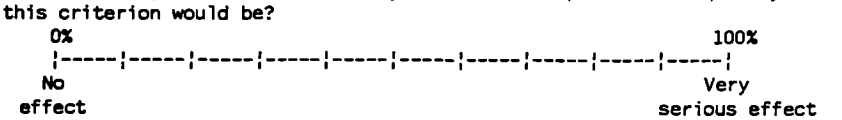
4. Do you think that a detailed review of cases which fit this criterion would show that those cases had more lapses in the quality of care than those not fitting the criterion, or fewer lapses than those not fitting the criterion?

[] More lapses in quality of care than others
[] Same level of lapses as others
[] Fewer lapses than others

5. Do you think that a detailed review of cases which fit this criterion would show that those cases had, on average, a better quality of care than those not fitting the criterion, or a worse quality of care than those not fitting the criterion?

[] Better quality of care than others
[] Same quality of care as others
[] Worse quality of care than others

6. How serious, in terms of the effect on the patient's health, do you think the lapses in the quality of care identified in cases which fit this criterion would be?



7. Do you think this screening criterion could be altered to make it relate more closely to the quality of care that patients receive?

[] Yes [] No

If you answered YES to question 7, please give details of the amendments you would suggest below. Please continue overleaf if you need more space.

G05 - Unplanned return to theatre for complications of a previous procedure.

DEFINITION
An unplanned return to theatre for a second or subsequent procedure occurred.

EXCEPTIONS
Planned second procedures, or second stages of two stage procedures.

GUIDANCE
For indications of whether the return to theatre was planned, check the operation notes for the first procedure. Check for procedures repeated because of lack of success on first attempt.
Check for surgical correction of complications e.g. insertion of a drain for a wound infection.

EXAMPLES
Return to theatre following haemorrhoidectomy for haemostasis for post-operative bleeding.
Return to theatre for re-suturing following burst abdominal wound.

---

1. How closely do you think this criterion is related to the quality of care that patients receive?

0% ... 100%
Not at related ... Very closely related

2. Do you think the information needed for this criterion could be found in patients' routine medical and nursing records?

0% ... 100%
Never found ... Always found

3. If all acute patients were screened using this criterion, what proportion of patients do you think would fit the criterion?

0% ... 100%
No patients ... All patients

4. Do you think that a detailed review of cases which fit this criterion would show that those cases had more lapses in the quality of care than those not fitting the criterion, or fewer lapses than those not fitting the criterion?

[] More lapses in quality of care than others
[] Same level of lapses as others
[] Fewer lapses than others

5. Do you think that a detailed review of cases which fit this criterion would show that those cases had, on average, a better quality of care than those not fitting the criterion, or a worse quality of care than those not fitting the criterion?

[] Better quality of care than others
[] Same quality of care as others
[] Worse quality of care than others

6. How serious, in terms of the effect on the patient's health, do you think the lapses in the quality of care identified in cases which fit this criterion would be?

0% ... 100%
No effect ... Very serious effect

7. Do you think this screening criterion could be altered to make it relate more closely to the quality of care that patients receive?

[] Yes [] No

If you answered YES to question 7, please give details of the amendments you would suggest below. Please continue overleaf if you need more space.

G06 - Pathology or histology report varies significantly from pre- operative diagnosis, or post-mortem report varies significantly from the ante-mortem diagnosis.

DEFINITION
Pathology or histology results following surgery are significantly at variance with pre-operative diagnosis, and this variance has resulted in a significant difference in a patient's care or treatment; or the post-mortem report is significantly at variance with the ante-mortem diagnosis and this variance has resulted in a significant difference in a patient's care or treatment.

EXCEPTIONS
There are no exceptions to this criterion.

GUIDANCE
Check for negative histology reports following surgery. For all deaths where post-mortems are carried out, check the post-mortem report opinion on cause of death against the ante-mortem diagnosis and treatment.
Note whether post-mortem has been carried out or not.

EXAMPLES
Mastectomy performed and histological report shows tissue to be benign. Polyps removed presumed benign, sent for histological examination and found to be malignant on report.

1. How closely do you think this criterion is related to the quality of care that patients receive?

0% ... 100%
Not at related ... Very closely related

2. Do you think the information needed for this criterion could be found in patients' routine medical and nursing records?

0% ... 100%
Never found ... Always found

3. If all acute patients were screened using this criterion, what proportion of patients do you think would fit the criterion?

0% ... 100%
No patients ... All patients

4. Do you think that a detailed review of cases which fit this criterion would show that those cases had more lapses in the quality of care than those not fitting the criterion, or fewer lapses than those not fitting the criterion?

[] More lapses in quality of care than others
[] Same level of lapses as others
[] Fewer lapses than others

5. Do you think that a detailed review of cases which fit this criterion would show that those cases had, on average, a better quality of care than those not fitting the criterion, or a worse quality of care than those not fitting the criterion?

[] Better quality of care than others
[] Same quality of care as others
[] Worse quality of care than others

6. How serious, in terms of the effect on the patient's health, do you think the lapses in the quality of care identified in cases which fit this criterion would be?

0% ... 100%
No effect ... Very serious effect

7. Do you think this screening criterion could be altered to make it relate more closely to the quality of care that patients receive?

[] Yes [] No

If you answered YES to question 7, please give details of the amendments you would suggest below. Please continue overleaf if you need more space.

G07 - Problems arising from transfusions of blood or blood derivatives, such as significant complications and reactions or misuse of service.

DEFINITION
Following transfusion of blood or blood derivatives, significant reaction(s) or complication(s) occurred, or the Blood Transfusion Service was misused.

EXCEPTIONS
There are no exceptions to this criterion.

GUIDANCE
Look for mismatch of blood group.
Look for evidence of reaction/complication such as rash, rigor, breathing difficulties, temperature spike greater than 37.5C , tachycardia, loin pain, chest pain, hypotension, jaundice following transfusion (within 12 hours), renal failure. Check for evidence that the complication/reaction was addressed, such as slowing or stopping transfusion rate, drug therapy, etc. Single unit transfusions are never indicated, therefore record details.

EXAMPLES
Urticaria and pyrexia following transfusion of whole blood.
Mis-matched blood given in error followed by renal failure.

---

1. How closely do you think this criterion is related to the quality of care that patients receive?

0% ... 100%
Not at related ... Very closely related

2. Do you think the information needed for this criterion could be found in patients' routine medical and nursing records?

0% ... 100%
Never found ... Always found

3. If all acute patients were screened using this criterion, what proportion of patients do you think would fit the criterion?

0% ... 100%
No patients ... All patients

4. Do you think that a detailed review of cases which fit this criterion would show that those cases had more lapses in the quality of care than those not fitting the criterion, or fewer lapses than those not fitting the criterion?

[] More lapses in quality of care than others
[] Same level of lapses as others
[] Fewer lapses than others

5. Do you think that a detailed review of cases which fit this criterion would show that those cases had, on average, a better quality of care than those not fitting the criterion, or a worse quality of care than those not fitting the criterion?

[] Better quality of care than others
[] Same quality of care as others
[] Worse quality of care than others

6. How serious, in terms of the effect on the patient's health, do you think the lapses in the quality of care identified in cases which fit this criterion would be?

0% ... 100%
No effect ... Very serious effect

7. Do you think this screening criterion could be altered to make it relate more closely to the quality of care that patients receive?

[] Yes [] No

If you answered YES to question 7, please give details of the amendments you would suggest below. Please continue overleaf if you need more space.

G08 - Infection acquired during the patient's admission.

DEFINITION
The patient developed an infection during their admission, which was acquired after they had been admitted.

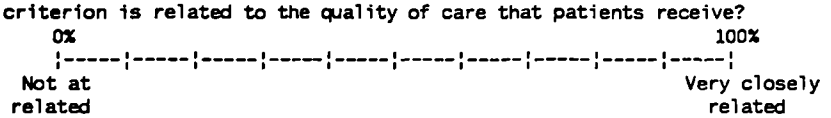
EXCEPTIONS
There are no exceptions to this criterion.

GUIDANCE
Look for evidence of freedom from infection or presence of infection in medical or nursing notes on admission and check laboratory reports. Compare with later documentation of signs and symptoms and laboratory reports.
Urinary: Match up positive culture with reports of signs and symptoms, such as dysuria, haematuria, frequency, pyrexia.
Chest: Match up signs and symptoms such as coughing, pyrexia, purulent sputum and a positive culture. If negative result but signs and symptoms present check if patient is on antibiotics. If so, count as variation.
Wound: Match signs and symptoms e.g.pyrexia, redness, swelling, and drainage around surgical site, with positive culture.
Other infections: Match signs and symptoms with positive laboratory report.
Check date of diagnosis and onset of infection, and admission notes, to establish whether the infection was acquired pre or post admission.

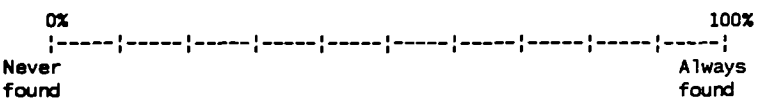
EXAMPLES
Urinary tract infections in catheterised patients.
Wound abscess post appendicectomy.

---

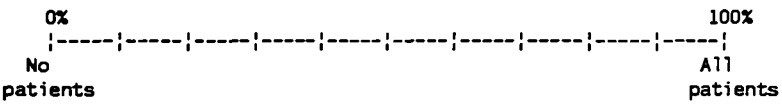
1. How closely do you think this criterion is related to the quality of care that patients receive?



2. Do you think the information needed for this criterion could be found in patients' routine medical and nursing records?



3. If all acute patients were screened using this criterion, what proportion of patients do you think would fit the criterion?



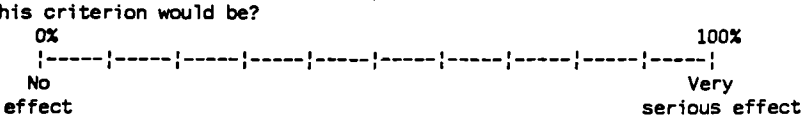
4. Do you think that a detailed review of cases which fit this criterion would show that those cases had more lapses in the quality of care than those not fitting the criterion, or fewer lapses than those not fitting the criterion?

[] More lapses in quality of care than others
[] Same level of lapses as others
[] Fewer lapses than others

5. Do you think that a detailed review of cases which fit this criterion would show that those cases had, on average, a better quality of care than those not fitting the criterion, or a worse quality of care than those not fitting the criterion?

[] Better quality of care than others
[] Same quality of care as others
[] Worse quality of care than others

6. How serious, in terms of the effect on the patient's health, do you think the lapses in the quality of care identified in cases which fit this criterion would be?



7. Do you think this screening criterion could be altered to make it relate more closely to the quality of care that patients receive?

[] Yes [] No

If you answered YES to question 7, please give details of the amendments you would suggest below. Please continue overleaf if you need more space.

G09 - One (or more) of the specified drug/antibiotic usage problems occurred.

DEFINITION
One of more of the following specified drug/antibiotic usage problems occurred:
a) Medications omitted, prescribed or given in error, given at wrong rate, delayed with no reason given.
b) Medication reactions and anaphylaxis.

GUIDANCE
Check kardex for any documented medication errors. Note subsequent action.
Check for comments made by the pharmacist in notes or on treatment card.

EXAMPLES
Penicillin prescribed or given when known allergy documented.
Anaphylactic shock in patient given intravenous antibiotics.
Omnopon infusion found to have been running at twice prescribed rate overnight.

---

1. How closely do you think this criterion is related to the quality of care that patients receive?

0% ¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦ 100%
Not at related — Very closely related

2. Do you think the information needed for this criterion could be found in patients' routine medical and nursing records?

0% ¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦ 100%
Never found — Always found

3. If all acute patients were screened using this criterion, what proportion of patients do you think would fit the criterion?

0% ¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦ 100%
No patients — All patients

4. Do you think that a detailed review of cases which fit this criterion would show that those cases had more lapses in the quality of care than those not fitting the criterion, or fewer lapses than those not fitting the criterion?

[] More lapses in quality of care than others
[] Same level of lapses as others
[] Fewer lapses than others

5. Do you think that a detailed review of cases which fit this criterion would show that those cases had, on average, a better quality of care than those not fitting the criterion, or a worse quality of care than those not fitting the criterion?

[] Better quality of care than others
[] Same quality of care as others
[] Worse quality of care than others

6. How serious, in terms of the effect on the patient's health, do you think the lapses in the quality of care identified in cases which fit this criterion would be?

0% ¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦ 100%
No effect — Very serious effect

7. Do you think this screening criterion could be altered to make it relate more closely to the quality of care that patients receive?

[] Yes [] No

If you answered YES to question 7, please give details of the amendments you would suggest below. Please continue overleaf if you need more space.

G10-1 - Cardiac arrest,or respiratory arrest, occurring during the admission.

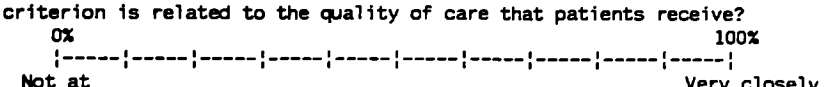
DEFINITION
The patient had a cardiac arrest or respiratory arrest, during their admission.

EXCEPTIONS
There are no exceptions to this criterion.

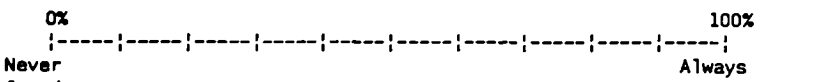
GUIDANCE
Look for evidence in records leading up to the event, such as major changes in vital signs, administration of medication or other treatments just prior to the event,which might suggest its cause. Check records of treatment of the arrest/shock itself, such as calling of crash team, availability of resuscitation equipment, etc.
Check records for ECG prior to surgery and also availability of ECGs. See if result of ECG is in record.
Look for other clinical indicators-
- reports of tests for Urea and Electrolytes
- evidence of adequate hydration (fluid charts, medical notes) or overload
- if monitored, evidence of rhythm changes
If unsuccessful, check for evidence of correct procedures followed after death.

---

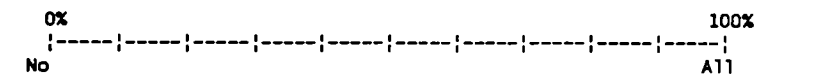
1. How closely do you think this criterion is related to the quality of care that patients receive?



2. Do you think the information needed for this criterion could be found in patients' routine medical and nursing records?



3. If all acute patients were screened using this criterion, what proportion of patients do you think would fit the criterion?



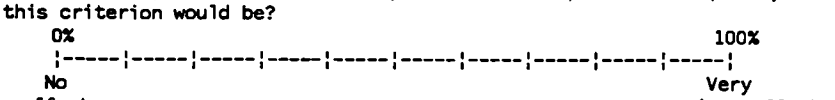
4. Do you think that a detailed review of cases which fit this criterion would show that those cases had more lapses in the quality of care than those not fitting the criterion, or fewer lapses than those not fitting the criterion?

[] More lapses in quality of care than others
[] Same level of lapses as others
[] Fewer lapses than others

5. Do you think that a detailed review of cases which fit this criterion would show that those cases had, on average, a better quality of care than those not fitting the criterion, or a worse quality of care than those not fitting the criterion?

[] Better quality of care than others
[] Same quality of care as others
[] Worse quality of care than others

6. How serious, in terms of the effect on the patient's health, do you think the lapses in the quality of care identified in cases which fit this criterion would be?



7. Do you think this screening criterion could be altered to make it relate more closely to the quality of care that patients receive?

[] Yes [] No

If you answered YES to question 7, please give details of the amendments you would suggest below. Please continue overleaf if you need more space.

G10-3 - A CVA or MI or PE occurred following a surgical procedure during the same admission.

DEFINITION
A cerebrovascular accident, myocardial infarction or pulmonary embolus occurred at any time during the admission following surgery.

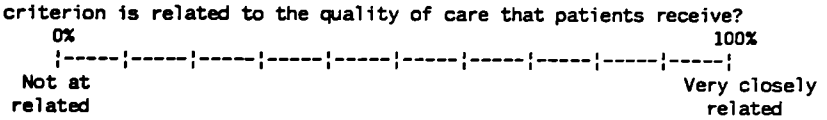
EXCEPTIONS
There are no exceptions to this criterion.

GUIDANCE NOTES
Note whether there were any cardiovascular symptoms not addressed by admitting doctor and whether ECG was considered.
Pulmonary embolus: Look for documentation for clotting time pre-surgery and note whether anti-coagulant therapy was considered. Look at accounts of post-operative nursing care: deep breathing exercises, passive or active limb exercises and consideration for referral to physiotherapy and early ambulation.

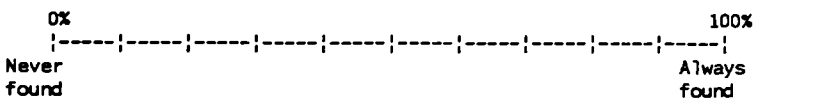
EXAMPLES
Pulmonary embolus following bowel resection.
Cerebrovascular accident following hysterectomy.

---

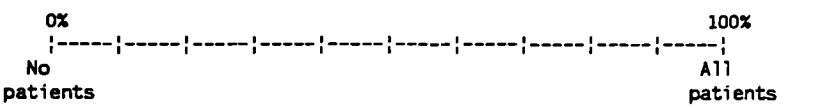
1. How closely do you think this criterion is related to the quality of care that patients receive?



2. Do you think the information needed for this criterion could be found in patients' routine medical and nursing records?



3. If all acute patients were screened using this criterion, what proportion of patients do you think would fit the criterion?



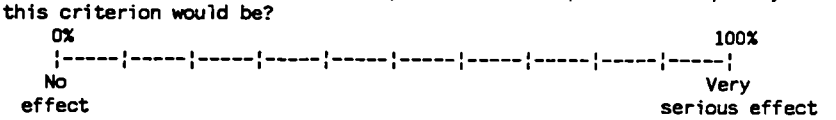
4. Do you think that a detailed review of cases which fit this criterion would show that those cases had more lapses in the quality of care than those not fitting the criterion, or fewer lapses than those not fitting the criterion?

[] More lapses in quality of care than others
[] Same level of lapses as others
[] Fewer lapses than others

5. Do you think that a detailed review of cases which fit this criterion would show that those cases had, on average, a better quality of care than those not fitting the criterion, or a worse quality of care than those not fitting the criterion?

[] Better quality of care than others
[] Same quality of care as others
[] Worse quality of care than others

6. How serious, in terms of the effect on the patient's health, do you think the lapses in the quality of care identified in cases which fit this criterion would be?



7. Do you think this screening criterion could be altered to make it relate more closely to the quality of care that patients receive?

[] Yes [] No

If you answered YES to question 7, please give details of the amendments you would suggest below. Please continue overleaf if you need more space.

G11 - Unexpected transfer from general care to higher dependency unit.

DEFINITION
Patient was transferred unexpectedly from general care to a higher dependency unit.

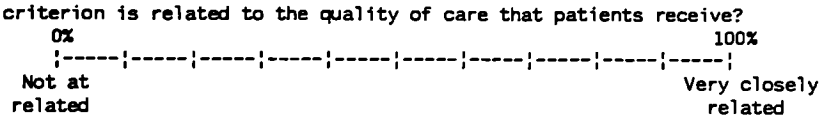
EXCEPTIONS
Planned transfers.

GUIDANCE
Transfers to higher dependency units include transfers to coronary care units, intensive therapy units here or in other hospitals.
Check pre-operative medical records for evidence that a transfer was planned.
Look for reasons for transfer and examine events leading up to transfer.
Check pre-operative medical records for evidence that a transfer was planned.

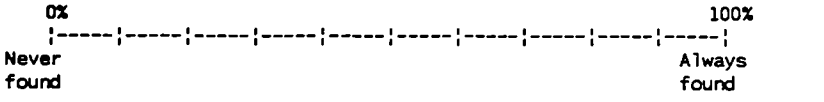
EXAMPLES
Patient transferred to intensive therapy unit following routine herniorrhaphy with breathing difficulties.
Patient transferred to coronary care following cardiac arrest.

---

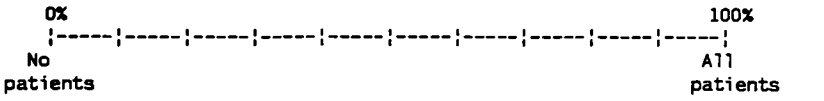
1. How closely do you think this criterion is related to the quality of care that patients receive?



2. Do you think the information needed for this criterion could be found in patients' routine medical and nursing records?



3. If all acute patients were screened using this criterion, what proportion of patients do you think would fit the criterion?



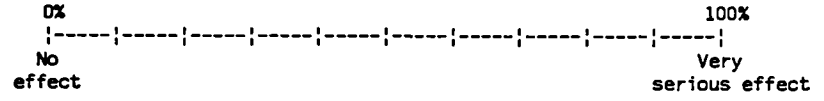
4. Do you think that a detailed review of cases which fit this criterion would show that those cases had more lapses in the quality of care than those not fitting the criterion, or fewer lapses than those not fitting the criterion?

[] More lapses in quality of care than others
[] Same level of lapses as others
[] Fewer lapses than others

5. Do you think that a detailed review of cases which fit this criterion would show that those cases had, on average, a better quality of care than those not fitting the criterion, or a worse quality of care than those not fitting the criterion?

[] Better quality of care than others
[] Same quality of care as others
[] Worse quality of care than others

6. How serious, in terms of the effect on the patient's health, do you think the lapses in the quality of care identified in cases which fit this criterion would be?



7. Do you think this screening criterion could be altered to make it relate more closely to the quality of care that patients receive?

[] Yes [] No

If you answered YES to question 7, please give details of the amendments you would suggest below. Please continue overleaf if you need more space.

G12-1 - One or more of the specified patient-related clinical complications occurred.

DEFINITION
One or more of the following patient-related clinical complications occurred during the admission:
b) IV problems: Overload. Phlebitis requiring treatment.
c) Pain control not addressed.
d) Development/worsening of skin problems resulting from pressure.
e) Faecal impaction developed during admission requiring physical evacuation.
g) Deep vein thrombosis not evident on admission.

EXCEPTIONS
There are no exceptions to this criterion.

GUIDANCE
Check whole of medical and nursing records for complications specified.
Note whether preventive measures taken and complications addressed.Check
for pain control chart.Look for complaints of pain not addressed and effect of analgesia not evaluated.

---

1. How closely do you think this criterion is related to the quality of care that patients receive?

0% 100%
Not at related — Very closely related

2. Do you think the information needed for this criterion could be found in patients' routine medical and nursing records?

0% 100%
Never found — Always found

3. If all acute patients were screened using this criterion, what proportion of patients do you think would fit the criterion?

0% 100%
No patients — All patients

4. Do you think that a detailed review of cases which fit this criterion would show that those cases had more lapses in the quality of care than those not fitting the criterion, or fewer lapses than those not fitting the criterion?

[] More lapses in quality of care than others
[] Same level of lapses as others
[] Fewer lapses than others

5. Do you think that a detailed review of cases which fit this criterion would show that those cases had, on average, a better quality of care than those not fitting the criterion, or a worse quality of care than those not fitting the criterion?

[] Better quality of care than others
[] Same quality of care as others
[] Worse quality of care than others

6. How serious, in terms of the effect on the patient's health, do you think the lapses in the quality of care identified in cases which fit this criterion would be?

0% 100%
No effect — Very serious effect

7. Do you think this screening criterion could be altered to make it relate more closely to the quality of care that patients receive?

[] Yes [] No

If you answered YES to question 7, please give details of the amendments you would suggest below. Please continue overleaf if you need more space.

G12-2 - One or more of the specified patient-related non-clinical problems or incidents occurred.

DEFINITION
One or more of the following patient-related non-clinical problems or incidents occurred:
a) Theatre booking cancelled/delayed.
b) No ITU bed available when clinically needed
c) Patient transferred/admitted to/receiving care on an inappropriate ward because of bed shortages.
d) Delay in obtaining a second opinion from another specialty.
e) Casenotes, X-rays or other records or results missing/not available when needed.
g) Patient had slip or fall, or other accident.
h) Equipment failure.
i) Necessary equipment not available when needed.
j) Prescribed drugs not obtainable.
k) Delay in undergoing diagnostic or therapeutic procedures.
l) Portering service problems
m) Staff unavailable

EXCEPTIONS
There are no exceptions to this criterion
GUIDANCE
Check medical and nursing records. Document the causes of the problem or incident if possible. Check for list of valuables recorded. Note if theatre instruments unavailable. Note if theatre delayed/cancelled due to TSSU problems or if ECG or other missing report delays operation. If notes are missing record reason.
EXAMPLES
Patient's X-rays are missing when they go to theatre.
Specimen results not available for review of treatment e.g. frozen section.
Patient staying nil by mouth while waiting long period due to theatre rescheduling.

1. How closely do you think this criterion is related to the quality of care that patients receive?

0% ... 100%
Not at related — Very closely related

2. Do you think the information needed for this criterion could be found in patients' routine medical and nursing records?

0% ... 100%
Never found — Always found

3. If all acute patients were screened using this criterion, what proportion of patients do you think would fit the criterion?

0% ... 100%
No patients — All patients

4. Do you think that a detailed review of cases which fit this criterion would show that those cases had more lapses in the quality of care than those not fitting the criterion, or fewer lapses than those not fitting the criterion?

[] More lapses in quality of care than others
[] Same level of lapses as others
[] Fewer lapses than others

5. Do you think that a detailed review of cases which fit this criterion would show that those cases had, on average, a better quality of care than those not fitting the criterion, or a worse quality of care than those not fitting the criterion?

[] Better quality of care than others
[] Same quality of care as others
[] Worse quality of care than others

6. How serious, in terms of the effect on the patient's health, do you think the lapses in the quality of care identified in cases which fit this criterion would be?

0% ... 100%
No effect — Very serious effect

7. Do you think this screening criterion could be altered to make it relate more closely to the quality of care that patients receive?

[] Yes [] No

If you answered YES to question 7, please give details of the amendments you would suggest below. Please continue overleaf if you need more space.

G14 - Neurological deficit developed which was not present when the patient was admitted.

DEFINITION
The patient developed a neurological deficit during their admission which was not present when they were admitted.

EXCEPTIONS
Patients coming under criteria G10/3 and G12/1 and patients developing planned or expected neurological deficit.

GUIDANCE
Look for evidence of neurological damage or compromise throughout the medical and nursing records. Look particularly for neurological deficits resulting from surgery. Check assessments of orientation and sensory, circulatory and motor function. Look for evidence of seizures, urinary or faecal incontinence, or intractable pain which were not present or not documented on admission.

EXAMPLES
Facial palsy post-parotidectomy. Frozen shoulder post-laryngectomy.

---

1. How closely do you think this criterion is related to the quality of care that patients receive?

0% 100%
Not at related — Very closely related

2. Do you think the information needed for this criterion could be found in patients' routine medical and nursing records?

0% 100%
Never found — Always found

3. If all acute patients were screened using this criterion, what proportion of patients do you think would fit the criterion?

0% 100%
No patients — All patients

4. Do you think that a detailed review of cases which fit this criterion would show that those cases had more lapses in the quality of care than those not fitting the criterion, or fewer lapses than those not fitting the criterion?

[] More lapses in quality of care than others
[] Same level of lapses as others
[] Fewer lapses than others

5. Do you think that a detailed review of cases which fit this criterion would show that those cases had, on average, a better quality of care than those not fitting the criterion, or a worse quality of care than those not fitting the criterion?

[] Better quality of care than others
[] Same quality of care as others
[] Worse quality of care than others

6. How serious, in terms of the effect on the patient's health, do you think the lapses in the quality of care identified in cases which fit this criterion would be?

0% 100%
No effect — Very serious effect

7. Do you think this screening criterion could be altered to make it relate more closely to the quality of care that patients receive?

[] Yes [] No

If you answered YES to question 7, please give details of the amendments you would suggest below. Please continue overleaf if you need more space.

G16 - Unexpected patient deaths.

DEFINITION
Patient died unexpectedly during admission

EXCEPTIONS
Patients admitted for terminal care or receiving terminal care, and patient recorded category C under the district resusitation policy.

GUIDANCE
Check nursing and medical records for evidence that the death was unexpected or any indications that it might have been prevented.

---

1. How closely do you think this criterion is related to the quality of care that patients receive?

0% ¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦ 100%
Not at related — Very closely related

2. Do you think the information needed for this criterion could be found in patients' routine medical and nursing records?

0% ¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦ 100%
Never found — Always found

3. If all acute patients were screened using this criterion, what proportion of patients do you think would fit the criterion?

0% ¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦ 100%
No patients — All patients

4. Do you think that a detailed review of cases which fit this criterion would show that those cases had more lapses in the quality of care than those not fitting the criterion, or fewer lapses than those not fitting the criterion?

[] More lapses in quality of care than others
[] Same level of lapses as others
[] Fewer lapses than others

5. Do you think that a detailed review of cases which fit this criterion would show that those cases had, on average, a better quality of care than those not fitting the criterion, or a worse quality of care than those not fitting the criterion?

[] Better quality of care than others
[] Same quality of care as others
[] Worse quality of care than others

6. How serious, in terms of the effect on the patient's health, do you think the lapses in the quality of care identified in cases which fit this criterion would be?

0% ¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦ 100%
No effect — Very serious effect

7. Do you think this screening criterion could be altered to make it relate more closely to the quality of care that patients receive?

[] Yes [] No

If you answered YES to question 7, please give details of the amendments you would suggest below. Please continue overleaf if you need more space.

G17 - Review of quality, consistency and completeness of medical records.

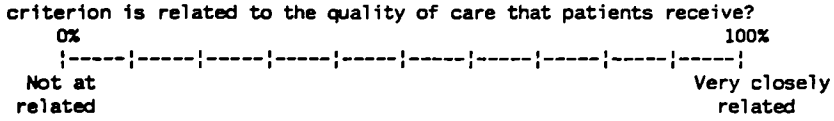
DEFINITION
One (or more) of the following specified deficiencies were found in the medical records:
a) Admission clerking not dated or not present.
b) Notes not filled in every third day and at other important stages of clinical management.
c) Notes illegible or unsigned.

EXCEPTIONS
There are no exceptions to this criterion

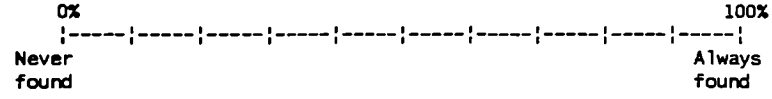
GUIDANCE
Check throughout medical records for the specified deficiencies. See BHA guidance for the writing of the clinical record. In b), if a patient is discharged within 3 days, ensure that an entry in the notes is made post-operatively by doctor. (Especially relevant in Ophthalmology.)

---

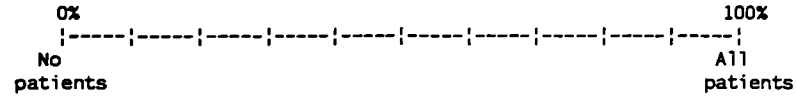
1. How closely do you think this criterion is related to the quality of care that patients receive?



2. Do you think the information needed for this criterion could be found in patients' routine medical and nursing records?



3. If all acute patients were screened using this criterion, what proportion of patients do you think would fit the criterion?



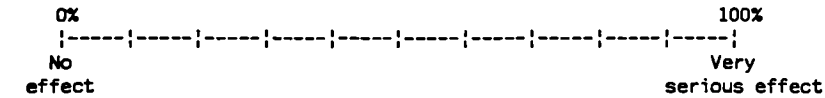
4. Do you think that a detailed review of cases which fit this criterion would show that those cases had more lapses in the quality of care than those not fitting the criterion, or fewer lapses than those not fitting the criterion?

[] More lapses in quality of care than others
[] Same level of lapses as others
[] Fewer lapses than others

5. Do you think that a detailed review of cases which fit this criterion would show that those cases had, on average, a better quality of care than those not fitting the criterion, or a worse quality of care than those not fitting the criterion?

[] Better quality of care than others
[] Same quality of care as others
[] Worse quality of care than others

6. How serious, in terms of the effect on the patient's health, do you think the lapses in the quality of care identified in cases which fit this criterion would be?



7. Do you think this screening criterion could be altered to make it relate more closely to the quality of care that patients receive?

[] Yes [] No

If you answered YES to question 7, please give details of the amendments you would suggest below. Please continue overleaf if you need more space.

G18 - Review of quality, consistency and completeness of nursing records.

DEFINITION
One or more of the specified deficiencies were found in the nursing records:
a) No nursing assessment completed within 24 hours of admission
b) Patient's care plan missing or not updated
d) Incomplete, contradictory, or inadequate recording of information
e) Nurses' signatures missing on drug chart.
f) Property form not completed/no disclaimer.

EXCEPTIONS
There are no exceptions to this criterion.

GUIDANCE
Check throughout the nursing records for the specified deficiencies
Compare TPR frequency with nursing orders in care plan. Ensure frequency of observations reflects the condition of the patient.

EXAMPLES
Patient's care plan giving post-operative nursing orders including care of intravenous infusion remains unchanged during remainder of patient's stay in hospital.
Quarter hourly observations following surgery become less frequent without corresonding adjustment in care plan.

---

1. How closely do you think this criterion is related to the quality of care that patients receive?

0% ... 100%
Not at related ... Very closely related

2. Do you think the information needed for this criterion could be found in patients' routine medical and nursing records?

0% ... 100%
Never found ... Always found

3. If all acute patients were screened using this criterion, what proportion of patients do you think would fit the criterion?

0% ... 100%
No patients ... All patients

4. Do you think that a detailed review of cases which fit this criterion would show that those cases had more lapses in the quality of care than those not fitting the criterion, or fewer lapses than those not fitting the criterion?

[] More lapses in quality of care than others
[] Same level of lapses as others
[] Fewer lapses than others

5. Do you think that a detailed review of cases which fit this criterion would show that those cases had, on average, a better quality of care than those not fitting the criterion, or a worse quality of care than those not fitting the criterion?

[] Better quality of care than others
[] Same quality of care as others
[] Worse quality of care than others

6. How serious, in terms of the effect on the patient's health, do you think the lapses in the quality of care identified in cases which fit this criterion would be?

0% ... 100%
No effect ... Very serious effect

7. Do you think this screening criterion could be altered to make it relate more closely to the quality of care that patients receive?

[] Yes [] No

If you answered YES to question 7, please give details of the amendments you would suggest below. Please continue overleaf if you need more space.

G19 - Evidence of patient and/or family dissatisfaction in records.

DEFINITION
There is evidence in the medical or nursing records, that the patient and /or the patient's relatives or friends expressed dissatisfaction with the care given to the patient.

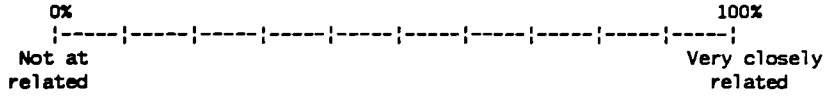
EXCEPTIONS
There are no exceptions to this criterion.

GUIDANCE
Look for information on cause of dissatisfaction and how it was expressed in nursing and medical records. Look for evidence that the patient/family complaint was handled appropriately.
All discharges taken against medical advice should be carefully checked against this criterion.
Look for dissatisfaction with food, cleanliness, conduct of staff, clinical care, noise at night in ward, lack of privacy.

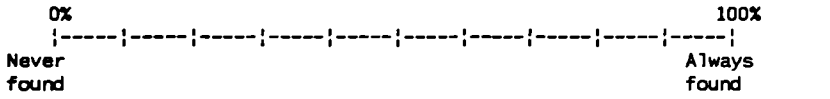
EXAMPLES
Patient's daughter complains to staff nurse that a window was left open overnight causing her mother to become chilled.
Patients complain about dirty washing facilities.
Patient took own discharge when operation postponed and incomplete explanation given by ward staff.

---

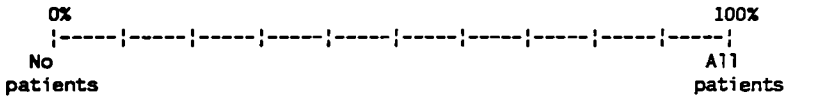
1. How closely do you think this criterion is related to the quality of care that patients receive?



2. Do you think the information needed for this criterion could be found in patients' routine medical and nursing records?



3. If all acute patients were screened using this criterion, what proportion of patients do you think would fit the criterion?



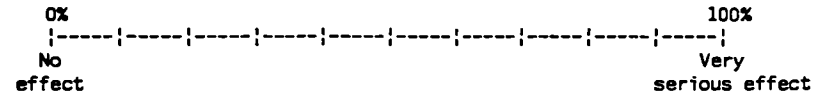
4. Do you think that a detailed review of cases which fit this criterion would show that those cases had more lapses in the quality of care than those not fitting the criterion, or fewer lapses than those not fitting the criterion?

[] More lapses in quality of care than others
[] Same level of lapses as others
[] Fewer lapses than others

5. Do you think that a detailed review of cases which fit this criterion would show that those cases had, on average, a better quality of care than those not fitting the criterion, or a worse quality of care than those not fitting the criterion?

[] Better quality of care than others
[] Same quality of care as others
[] Worse quality of care than others

6. How serious, in terms of the effect on the patient's health, do you think the lapses in the quality of care identified in cases which fit this criterion would be?



7. Do you think this screening criterion could be altered to make it relate more closely to the quality of care that patients receive?

[] Yes [] No

If you answered YES to question 7, please give details of the amendments you would suggest below. Please continue overleaf if you need more space.

G20 - Problems relating to the patient's discharge occurred.

DEFINITION
One (or more) of the following specified problems occurred in relation to the patient's discharge:
c) Absence of copy of discharge summary to GP
d) Discharge delayed for non-clinical (organisational or social) reasons

EXCEPTIONS
There are no exceptions to this criterion

GUIDANCE
Look through medical notes for copy of discharge prescription and letter to general practitioner. Look for evidence of assessment of patient for community services, liaison with relatives and evidence of planned nursing discharge. Look for reasons for discharge delay.

EXAMPLES
Elderly patient remained in hospital for three weeks following medical decision for discharge,awaiting placement in nursing home.

---

1. How closely do you think this criterion is related to the quality of care that patients receive?

0% ¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦ 100%
Not at related — Very closely related

2. Do you think the information needed for this criterion could be found in patients' routine medical and nursing records?

0% ¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦ 100%
Never found — Always found

3. If all acute patients were screened using this criterion, what proportion of patients do you think would fit the criterion?

0% ¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦ 100%
No patients — All patients

4. Do you think that a detailed review of cases which fit this criterion would show that those cases had more lapses in the quality of care than those not fitting the criterion, or fewer lapses than those not fitting the criterion?

[] More lapses in quality of care than others
[] Same level of lapses as others
[] Fewer lapses than others

5. Do you think that a detailed review of cases which fit this criterion would show that those cases had, on average, a better quality of care than those not fitting the criterion, or a worse quality of care than those not fitting the criterion?

[] Better quality of care than others
[] Same quality of care as others
[] Worse quality of care than others

6. How serious, in terms of the effect on the patient's health, do you think the lapses in the quality of care identified in cases which fit this criterion would be?

0% ¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦-----¦ 100%
No effect — Very serious effect

7. Do you think this screening criterion could be altered to make it relate more closely to the quality of care that patients receive?

[] Yes [] No

If you answered YES to question 7, please give details of the amendments you would suggest below. Please continue overleaf if you need more space.

4.3 Final questionnaire used in the questionnaire study of clinician opinion.

Medical Audit - Occurrence Screening
QUESTIONNAIRE

About this questionnaire

This questionnaire contains a set of 20 screening criteria. it is intended to find out how good you think each screening criterion would be at selecting cases where there has been some lapse in the quality of care. Clearly, none of the criteria can select every case where there has been such a lapse, and every criterion will select some cases where there has been no lapse in the quality of care. The lower these "false negative" and "false positive" rates can be made, the better the screening criteria will be at selecting cases where the quality of care has been lower than it should have been. Therefore, both your opinions on the screening criteria as they are now, and your suggestions for improvements to the criteria will be very welcome.

Filling in this questionnaire

The middle two pages of this questionnaire contain a table, in which each of the 20 screening criteria is listed. For each criterion, there are five questions for you to answer. The first four questions are answered by giving a rating from 0 to 10, and the fifth question is answered by ticking either YES or NO. There is also some space for you to write comments about the criterion.

On the back page of the questionnaire is some supplementary information about each of the screening criteria. It explains their meanings in more detail, and gives some examples of the sorts of cases which might be selected by each criterion.

Confidentiality

The answers you give and comments you make in this questionnaire will not be passed on to anyone apart from the research team, nor will they be used in an identifiable way in the research report. The questionnaire numbering will only be used to follow up unreturned questionnaires, and to examine differences in opinions amongst different clinical groupings. It will not be used to identify individual respondents.

Study report

If you would like to receive a copy of the study report based on the responses to these questionnaires, please tick the box on the right. [] tick box

Returning the questionnaire

When you have completed the questionnaire, please place it in the enclosed stamped addressed envelope and return it to me at:

CASPE Research, King Edward's Hospital Fund,
14 Palace Court, Bayswater, London W2 4HT.

Thank you again for your help with this study.

Kieran Walshe.
Research Co-ordinator,
CASPE Research.

	Screening criterion (for more information about each criterion, see the table on the back page of this questionnaire).	How closely do you think this criterion is related to the quality of care that patients receive? Please rate it from: **0 (not at all related)** to **10 (very closely related).**	Do you think the information needed for this criterion could be found in patients' routine medical and nursing records? Please give a rating from: **0 (never found)** to **10 (always found).**
1.	Admission for adverse results of or complications resulting from outpatient management.		
2.	Readmission for complications arising from a previous admission, or because of the incomplete management of problems on a previous admission.		
3.	Errors made in obtaining and documenting the patient's consent to operative procedures.		
4.	Unplanned removal, injury or repair of an organ or structure during surgery or invasive procedure.		
5.	Unplanned return to theatre for complications resulting from a previous procedure.		
6.	Pathology or histology report varies significantly from preoperative diagnosis, or postmortem report varies significantly from antemortem diagnosis.		
7.	Problems arising from transfusions of blood or blood derivatives, such as significant complications or reactions, or misuse of service.		
8.	Infection acquired during the patient's stay in hospital.		
9.	Medications omitted, prescribed or given in error; medication reactions and anaphylaxis.		
10.	Cardiac or respiratory arrest occurring during the patient's stay in hospital.		
11.	A CVA or MI or PE occurred during the patient's stay in hospital, following a surgical procedure during the same admission.		
12.	Unexpected transfer of patient from general care to higher dependency unit (ITU, CCU, etc).		
13.	One or more of the specified patient-related clinical complications occurred.		
14.	One or more of the specified patient-related non-clinical/organisational problems or incidents occurred.		
15.	Neurological deficit developed which was not present when the patient was admitted.		
16.	Unexpected patient death.		
17.	Inadequacies found in the quality, consistency and completeness of the medical records.		
18.	Inadequacies found in the quality, consistency and completeness of the nursing records.		
19.	Evidence of patient and/or family dissatisfaction.		
20.	Problems relating to the patient's discharge occurred.		

If all acute patients were screened using this criterion, what proportion of patients do you think would fit the criterion? Please rate the proportion, from: 0 (no patients) to 10 (all patients).	How serious, in terms of its effect on the patient's health, do you think the circumstances outlined in this criterion would be? Please rate the effect, from: 0 (no effect) to 10 (very serious effect)	Do you think this criterion could be altered to make it relate more closely to the quality of care that patients receive? Please tick YES or NO. If you answer YES, please give details of amendments in the comments column.	If you have any comments on any of the screening criteria, or any suggestions for ways in which they could be improved, please write them in this column. Please continue on a separate sheet if there is not enough space for your comments below.
		[ ] Yes [ ] No	
		[ ] Yes [ ] No	
		[ ] Yes [ ] No	
		[ ] Yes [ ] No	
		[ ] Yes [ ] No	
		[ ] Yes [ ] No	
		[ ] Yes [ ] No	
		[ ] Yes [ ] No	
		[ ] Yes [ ] No	
		[ ] Yes [ ] No	
		[ ] Yes [ ] No	
		[ ] Yes [ ] No	
		[ ] Yes [ ] No	
		[ ] Yes [ ] No	
		[ ] Yes [ ] No	
		[ ] Yes [ ] No	
		[ ] Yes [ ] No	
		[ ] Yes [ ] No	
		[ ] Yes [ ] No	

The information provided on this page is intended to assist you in completing the questionnaire, by explaining the meaning of each screening criterion, and providing some examples of cases which would be selected by that criterion.

	Screening criterion	Further details: detailed definition of criterion, known exceptions, and examples of cases which would be selected by the criterion.
1.	Admission for adverse results of or complications resulting from outpatient management.	Admission was direct result of complications of outpatient management, given by hospital staff in outpatients/A&E, or by community staff. For example, delayed diagnosis, complications of outpatient procedures, etc.
2.	Readmission for complications arising from a previous admission, or because of the incomplete management of problems on a previous admission.	Readmission was the direct result of complications of a previous admission, or of problems which were not resolved on a previous admission. Expected/planned readmissions are not included.
3.	Errors made in obtaining and documenting the patient's consent to operative procedures.	For example, the consent form is missing, or incomplete, or illegible. Emergencies, where consent cannot be obtained, are excepted.
4.	Unplanned removal, injury or repair of an organ or structure during surgery or invasive procedure.	Intended to pick up inadvertent or mistaken damage to or removal of tissue/organs. For example, perforation of the bowel during an endoscopic procedure, or corneal abrasion under general anaesthetic.
5.	Unplanned return to theatre for complications resulting from a previous procedure.	For example, return to theatre for resuturing of operation wound. Planned returns to theatre for two stage procedures, and expected returns to theatre (such as trauma cases) are excepted.
6.	Pathology or histology report varies significantly from preoperative diagnosis, or postmortem report varies significantly from antemortem diagnosis.	Cases where histology reports following surgery contradict presurgery findings, or cause of death cited in post-mortem at odds with antemortem clinical conclusions. For example, mastectomy performed, and histological examination shows tissue to be benign.
7.	Problems arising from transfusions of blood or blood derivatives, such as significant complications or reactions, misuse of service.	Such as severe reactions to blood transfusions, delay in transfusion due to unavailability of supply, or transfusions given when not clinically necessary (for example, single unit transfusions).
8.	Infection acquired during the patient's stay in hospital.	Match signs and symptoms with positive diagnosis and laboratory report. Check date of onset and date of admission to establish when infection acquired. For example, wound abcess post appendicectomy.
9.	Medications omitted, prescribed or given in error; medication reactions and anaphylaxis.	For example, penicillin prescribed when known allergy documented, or IV infusion found to be running at twice prescribed rate.
10.	Cardiac or respiratory arrest occurring during the patient's stay in hospital.	Check for evidence in records that arrest might have been expected. Check handling of arrest - availability of resuscitation equipment, crash team.
11.	A CVA or MI or PE occurred during the patient's stay in hospital, following a surgical procedure during the same admission.	A cerebrovascular accident, myocardial infarction or pulmonary embolus occurred postoperatively. Check for evidence of cardiovascular symptoms preoperatively, and of appropriate post-operative care.
12.	Unexpected transfer of patient from general care to higher dependency unit (ITU, CCU, etc).	Patient transferred from a general ward to a high dependency unit unexpectedly.
13.	One or more of the specified patient-related clinical complications occurred.	Complications such as: development/worsening of pressure areas; deep vein thrombosis; inadequate pain control; faecal impaction requiring physical evacuation developed during the admission.
14.	One or more of the specified patient-related non-clinical/organisational problems or incidents occurred.	Problems such as: patient slips and falls; theatre booking delayed/cancelled; patient on inappropriate ward due to bed shortages; records missing; equipment failures/unavailability; delay in undergoing diagnostic procedures.
15.	Neurological deficit developed which was not present when the patient was admitted.	Check for evidence of neurological damage occurring during the admission, especially from surgery. For example, facial palsy post parotidectomy.
16.	Unexpected patient death.	Patients receiving terminal care, or patients who are not for resuscitation, are excepted.
17.	Inadequacies found in the quality, consistency and completeness of the medical records.	Notes not filled in regularly, at important stages of clinical management. Notes illegible, undated, or unsigned. For example, no notes postoperatively, or no documentation of decision to discharge.
18.	Inadequacies found in the quality, consistency and completeness of the nursing records.	Nursing assessment not done within 24 hours of admission, or care plan missing/not updated, or Kardex illegible, undated, unsigned, incomplete.
19.	Evidence of patient and/or family dissatisfaction.	For example, patient's daughter complains that food is cold and inedible. Patient takes own discharge when operation postponed several times. Patient complains about attitude/manner of member of staff.
20.	Problems relating to the patient's discharge occurred.	Such as delays in discharge for non-clinical reasons (such as waiting for rest-home vacancy), or inadequate discharge planning.

4.4 Transcription of all textual comments made by respondents to the questionnaire study of clinician opinion.

In this listing, the comments from respondents have been grouped into three main areas: comment about the adverse-event measure in general; comment relating to specific screening criteria or adverse events within the measure; and those pertaining to the questionnaire study design and implementation. Within each area, comments have been grouped around subjects or themes.

All comments have been transcribed as they were written.

1. Comments on the adverse-event measure

Grouping of adverse events within criteria

2001	Difficult to assess - some criteria include a range of possible incidents/factors which may be serious at one end of the range to trivial at the other.
2041	Many of the proposed groups combine trivial with serious events
7111	Some of the questions cover too wide a field with incidents ranging from the trivial to the life-threatening - eg those on drugs and blood transfusions.

Precision of definition of adverse events

2001	It is always difficult to be precise and perhaps not vital if used as a screen for further evaluation. However subjectivity on individual cases cannot be avoided.
2003	A number of these criteria depend on appropriate record keeping and subjective decisions about what is acceptable medical performance. These need to be achieved/established before these criteria have validity.
2016	Found most categories too general ie could be of major impact/little impact within same category.
2017	Specify instances - standards needed
2026	Some effort made to define severity of complications [needed]

Denominator for adverse event rates

2003	What about patients for whom events might have been expected but did not happen - denominator problem.

General validity of measure

2016	Idea OK but would need to be finely tuned.
2019	As screening criteria I think most of the 20 are very useful, though I found difficulty grading quality relevance
2022	No of criteria need to be reduced
2028	I think many of the criteria have the potential to be highly sensitive - if the information was available in the notes (which it isn't) - but are of low specificity.

2031 I think the abysmal quality of clinical records will scupper other aspects of quality/audit and need to be first area addressed

2040 Medical audit by occurrence screening is a useful form of medical audit it is however unsatisfactory in that medical audit should compare normal practice against an optimum standard eg an old procedure against a modern procedure

2041 I would prefer an examination of a sample of notes to detect not only occurrences but also whether the patient's care conformed to agreed standards and resulted in the expected (anticipated) outcome.

2044 The effects on the patient are obviously related to the nature of the mistake - these criteria act as indicators only - review with medical and nursing staff is necessary to determine the impact of the mistake - also necessary to determine how inevitable the mistake (ie postop infection) is - eg food poisoning is avoidable, UTI may not be

2057 The vast majority of patient complaints dissatisfaction and litigation have their origins in failure of communication. None of these criteria addresses itself to this (eg 3 or 19). The occurrence is largely unmeasured because doctors seldom write what they tell their patients or even whether they did. Areas of communication failure are - about the disease, about the procedure, about the complications, about aftercare and about prevention. How can we audit all this? How can we detect occurrences of this failure? Food for thought here. Also what about cancellation of admission long duration of waiting, deterioration of condition while waiting? probably under 14 but I'm not entirely sure.

2070 the whole approach horrifies me - bad apples vs continuous quality improvement. Too much emphasis on outliers - it hasn't worked in US.

2072 This questionnaire convinces me even more that this type of audit is costly, has a low payback rate and is threatening and about blame. 90% of the clinical issues will have a perfectly rational explanation. let us learn from the American experience

2105 Can I suggest you consider criteria related to the frequency of medical attention ie ward rounds, time before first seen, frequency of consultant examination etc. Also inappropriate investigation - example repeat X-rays or Ba meal in 95 year olds when not clinically useful.

2117 Do you really mean all acute patients screened - complications are rare occurrences. It is usually helpful to focus on positive as well as negative aspects of quality of care.

2121 Nearly all the criteria are too loosely defined at the present. The need to be much more specific in connection with this the answers in column 4 (effect) will vary when the criteria are tightened.

2127 Time and cost are the issues that seem a problem. To go through all records would take a long time especially if all this information is being sought. Will the benefits be worth the costs of this approach?

2137 Estimates of value could depend on whether is truly medical audit as perceived by doctors or risk management as seen by lawyers - maybe you should try this on a sample of regional solicitors!

2155 Some criteria cover many situations - some of which may be serious - in health or quality terms - and some may be irrelevant. in other words they are quite general. I think this is an advantage. If the screening criteria are too specific then relevant cases may be missed.

2161 My chief difficulty with this approach is that it tries to establish a uniform set of criteria for all specialties. To take a simple example a cardiac/respiratory arrest would be unexpected and worthy of investigation if it occurred in a patient admitted say for cold surgery, if the patient was admitted for a suspected myocardial infection the picture that emerges is quite different. It also lumps together variations of a problem with quite different consequences. For example, medication errors and medication reactions are all included in one criterion.

2179 Most of the clinical indicators are valid in terms of quality of care but there will be great difficulty identifying blame even where an event is documented. I would hope that the worst case scenarios outlined on the back page are very rare indeed. However, there may be many patients where (eg patient is given antibiotic and bacteria is resistant) there are errors which lead to sub-optimal treatment. 0-10 means very variable outcome possible. I thought the measures were well derived.

7102 I am growing less enthusiastic about occurrence screening as it is negative in approach. it does not encourage new ideas and smacks of policing. Junior doctors may (some do) see the monthly review as at Bromley as a list of their mistakes.

2. Comments on specific screening criteria

1. Admission for adverse results of outpatient management

2003 This category is too broad - delayed diagnosis and acknowledged complications are two different issues.

2010 Relate to predicability of adverse event.

2019 Could be very powerful if only could be deduced from routine records

2020 Clinical judgements - not sure whether in reality screening criteria

2028 Need to be more specific

2030 Would depend on what events specified and ease of extraction from records

2031 Intended to audit GP management?

2032 Limit this to one or two diagnoses only

2040 More specific

2042 You have to relate it in some way to the patients' expressed wishes on admission since some reluctance to be admitted leads to longer o/p than might be usual practice

2045 Difficult to conceptualise criterion

2052 Depends on nature of adverse effect

2053 Specify a time period since last outpatients? I think it would need to be more diagnosis specific eg diabetes and admission for amputation

2065 Almost always impossible to detect

2067 Relative risk stated before preventative treatment results added

2072 Effect is not measurable - could be any from 1 to 10.

2082	use limiter for diagnosis to exclude well known complications
2087	Outpatient management is probably good even if goes wrong sometimes.
2088	GP information
2091	The relationship to the waiting list is important here, and the ranking of urgency.
2092	Difficult criterion to identify.
2094	Not as a screening criterion but was it an event that was predicted as a possibility.
2097	Urgency of admission?
2098	Define OPD management eg medical/surgical etc.
2099	Effect will vary enormously depending on adverse occurrence - cannot score for overall group.
2100	Separate various factors eg active treatment causing problems, from delay causing problems.
2112	Waiting times should be separated from clinical process/outcome problems.
2113	Add unexpected.
2124	Could be used effectively in fracture clinics, dermatology etc
2129	Readmissions within 28 days.
2135	Major problem with this and 2 are the false negatives - no readmission may reflect poor quality care. Effect on patient's health may be much more if they are not readmitted.
2137	Need to differentiate between iatrogenic as opposed to side effects.
2139	Add: which would not have occurred had optimum management in the opinion of peer group been given.
2144	Drug interactions specifically for side effects.
2146	By defining what is appropriate outpatient management for each specialty.
2147	The adverse results or complications themselves could be clearly defined and serious.
2150	Effect would vary from 0-10 depending on seriousness of complication.
2151	Target diagnoses.
2158	Some complications are an implicit risk of the treatment eg marrow down from chemotherapy.
2162	Not always preventable and therefore a quality issue?
2178	Some complications appear unavoidable. Seriousness not defined.
2178A	No sense of degree of seriousness of the complication ? avoidable.
2180	Specify severity of complications and appropriate treatment.
2197	More specific adverse reactions.
7100	Outpatient management is increasing brinkmanship with admission being a no go area.
7101	Omit delayed diagnosis.
2.	*Readmission for complications relating to previous admission*
2003	A judgement is needed on predictable and unpredictable events.
2019	Split readmission and incomplete management
2028	Need to be more specific - how do you define "complication" or "incomplete management"?
2041	Chronic incurable vs curable?

2042	Clarification of factors in readmission
2045	Compare to expected complication rate under good practice
2049	Some definition or grading of complications is needed to enable any useful comment/assessment
2052	Depends upon the complication
2062	Length of initial stay? Delivery of discharge information.
2072	Omit - any good geriatrician will tell you how inappropriate this is.
2081	Timing - ie 7 days etc.
2098	Too wide a set of possible circumstances included here.
2112	Separate things discharges may be risked knowing that readmission may be needed - this may be good quality of care for most.
2113	Add unexpected.
2120	Within 7 days of discharge.
2144	Deep vein thrombosis.
2158	Just stick to second half of statement.
2178	Difficult to define in such a general sense ?readmissions.
2178A	No sense of degree of seriousness.
2199	Specify complications eg wound infection.
7203	Formalised case notes with protocol and/or problem oriented medical records listing abnormal results.

3.	*Error in obtaining consent to operative procedure*
2001	Can only be evaluated with patient input
2019	Shows sloppy approach but not necessarily directly related to healthcare quality
2025	It is more important to document that the patient fully understands the procedure
2028	Need to be more specific - eg no consent obtained/documented
2031	Quality of expression etc is closest related to quality of care but never documented
2032	More of QA implications together with possible legal consequences
2036	Recording manner and way in which consent was given and received
2041	? adverse outcome
2053	Use as a tracer situation as it may imply a general laxity in care
2062	Review of procedures required
2065	More important to assess whether adequate explanation was given
2088	Are operative procedures clearly understood by all staff?
2092	Feedback from patient.
2094	Some attempt at how the patient understood - not a screening criterion though.
2099	Restrict to missing or inconsistent consent.
2113	Very peculiar question.
2115	Too vague.
2139	Meaning of proportion question not clear here as not everyone has a procedure.
2162	Quality ++
2178	Depends on condition and consent to what?

2190 relate consent to patient understanding of procedure.
2193 Answered in relation to "documenting consent" - obtaining consent is more directly related to quality of care but no information currently available particularly about the process of quality of the consent obtained.
5006 Not applicable to GU medicine.
5047 Revision of consent procedures as addressed at present.
7102 Leave out obtaining as impossible to find out circumstances at time.

4. *Unplanned removal/injury/repair of organ/structure during surgery*
2003 This criterion needs to address removal necessitated by poor quality care - unplanned removal of diseased organs is good quality care unless poor diagnostic performance.
2004 Definite complication
2019 Specify more to allow for whether in difficult/understandable clinical conditions or normal uncomplicated anatomy. Even so, sometimes just bad luck
2028 Need specialist knowledge to recognise from the records that it was unplanned
2036 Record whether negligence was involved and whether it was inevitable
2041 ? effect - eg wrong leg
2042 You have to distinguish unplanned due to unexpected extent of disease from unplanned due to operator error
2053 Delete unplanned removal and unplanned repair - leave surgical injury
2087 Could be bad or good for patient - competent or incompetent work.
2094 Not as a screening criterion but what were the consequences.
2100 State whether in planned procedure resulted from mistaken action or unexpected finding.
2103 Inadvertent rather than unplanned.
2113 Unplanned and not necessitated by findings at operation.
2124 Consider auditing referral to another specialty as a result of complication eg hysterectomy and then referral to urology.
2139 Care may be good despite removal or repair being unplanned.
2140 Should say unintended rather than unplanned.
2144 Small numbers.
2147 Unplanned mistakes could be defined and serious.
2154 Include assessment of patients comprehension of what they have consented to and associated risks.
2168 It is an accident which can occur even in expert hands - it isn't a quality issue.
2178 Useful question. Must be of highest importance.
2178A Some damage unavoidable.
2189 Word repair should be omitted.
5999 This is a very varied group and I would split this into serious and non-serious problems.
7203 Regular review of surgical "mishaps".
7102 There are acceptable risks to some organs in some difficult procedures.

5.	*Unplanned return to theatre*
2030	Very difficult to assess in some cases
2033	Some complications may not be related to quality of care
2040	Time interval
2053	A time period may need to be specified eg re-op within 1 week
2067	Possibly selected conditions with less risk if confined those to signs of age
2113	From a previous related procedure.
2120	Concentrate on routine procedures in first instance.
2135	As for 1,2 the return to theatre reflects good quality care the complications poor quality care so effect left blank.
2144	Burst wound, secondary haemorrhage.
2172	Specificity?
2178	Not all procedures can be perfectly excellent.
2178A	No sense of degree of seriousness ? avoidable.
2197	More specific reason.
7203	Regular review of "returns to theatre".
6.	*Pathology/histology varies from diagnosis*
2025	Depends on the exact circumstances
2028	Focus on a small number of specific diagnoses - eg appendicitis - aim to take a random sample of all deaths or ?just screen PM incidence
2031	Simply forcing doctors to record a diagnosis preop has strong bearing on quality of care
2035	Effect varies far too much to be used non-specifically
2039	Separate issues - histology/diagnosis, diagnosis/death
2041	? effect - eg wrong treatment
2042	PM reports differ in 20-90% of cases" often this may not matter if the disease is not treatable anyway. But pre-op histology is different
2053	Judgement needed on significance of difference in relation to alternative outcome for patient
2081	It all depends - need to subdivide
2082	Define significant? eg different ICD chapter?
2088	I differentiate between effect of postmortem and histology report.
2091	Links between path and clinical notes usually poor - eg records not updated when new evidence appears particularly PM.
2094	How was the management affected prediagnosis at PM/histology by the [illegible]
2097	Pathology report varies significantly from preop diagnosis such that procedure was inappropriate (this would exclude biopsy report not as expected but action appropriate).
2098	Your example (overleaf) is a very extreme occurrence of the criteria.
2099	Investigation protocols.
2100	Separate antemortem from postmortem pathology.
2113	Need to clarify severity of variance.
2137	10-20% of PMs differ substantially - what is "significant"?
2139	Effect question is impossible to answer as a generalisation - some patients are adversely affected others are not!

2147	Depends on diagnoses.
2158	Not a helpful criterion though.
2162	Quality ++ but ? effect on outcome.
2178	Well known for PM results to differ 20-40%. Depends on degree of seriousness.
2178A	What about the normal appendix that is removed?
7203	PM especially - have target uptakes by consultant?
7102	Perioperative not preoperative - true diagnosis often made at the operation.

7.	*Transfusion problems: reactions, complications, usage*
2001	Too varied - from anaphylaxis to delay
2005	What does misuse mean? other than single unit. Either define explicitly or leave out.
2019	Bipolar again - split complications reactions from misuse of service
2024	Misuse of service important
2028	Be more specific - what reactions. ?screen for blood transfusions in certain common conditions
2039	Reactions due to mismanagement/omission
2040	More details of situation
2041	Severity of response
2043	?factors in transfusion reactions
2054	Misuse of service is different from problems arising...
2065	Doesn't distinguish "inevitable" reactions from errors
2071	Omit misuse of service - this will have an overall impact on the service offered to all patients but is a very different matter to blood reactions
2082	Limit to mismatched transfusion or transmission of infection - some reactions are to be expected.
2091	Misuse is of a different magnitude to severe reactions - one may kill the other may be far less.
2094	Consequences.
2099	Restrict to avoidable.
2113	Are we assuming blood matching errors cause the reactions?
2121	Make criterion more specific eg just look for people receiving 1 unit.
2137	Ragbag - a mismatch is a rare disaster, overordering is almost universal.
2139	Some problems are avoidable, others unavoidable could put "avoidable problems arising etc"
2162	Not always preventable but clearly important in outcome terms and therefore valid as a criterion for medical audit.
2172	Separate out unnecessary transfusions from reactions etc.
2178	Depends on degree of risk involved.
2178A	What precisely does misuse of service mean?
2180	Specify misuse of service.
2187	Misuse of service - not understood. Leave out or redefine.
2196	Misuse is potentially serious. Reactions do not imply poor service unless they could be anticipated.
7102	Must be seen in context of clinical need.
7103	I would separate complications and misuse.

8.	*Hospital acquired infection*
2001	Precise definitions needed
2004	Monitor outbreaks - not individuals?
2020	Needs further refinement
2028	Difficult to detect from notes
2030	Need to measure avoidability of infection
2031	Getting them to record is an uphill course
2035	Too non-specific
2036	How was infection acquired, type eg chest, wound etc
2041	Severity
2047	Extremely difficult to define hosp acquired infection
2058	Site of infection - yes if wound infection but no if UTI/chest infection
2062	Infection acquired by clinical procedures or defective domestic arrangements
2065	Definition of infection
2072	Change to nosocomial infections.
2081	Varies - as in 9
2083	Can be difficult to conclusively identify source
2092	Examine clean ops subset.
2098	Scorings relate to severity.
2099	Restrict to nosocomial ? infections, surgical infections.
2137	What is infection?
2139	Try separating chest infection, wound infection, urine infection etc.
2146	Infections following elective procedures.
2158	Wound infections are important, recurrence of chest infection in chronic bronchitis is unimportant.
2178	Criteria don't embody any degree of seriousness.
2178A	? avoidable no sense of degree of seriousness.
7203	Burns and wound infections (theatre cases excluding open trauma).
7102	In orthopaedics very serious, in urology trivial. May be unavoidable.
9.	*Antibiotic/drug utilisation problems*
2001	Covers huge range of possibilities from severe to minor
2012	Division of indicator [needed] by severity of outcome
2019	Again helpful to split indicator - rare complications from known hypersensitive very different
2025	There is a big difference between drugs omitted and drugs given in error
2036	These are mostly due to negligence however anaphylaxis/reaction where allergy is not known is not negligence
2039	Error = quality of care not reactions not due to carelessness
2041	Could be trivial or life threatening
2053	Delete medication reactions and anaphylaxis as they may be less due to quality of care
2054	Medication reactions/anaphylaxis are different to medication omitted.. etc
2058	Anaphylaxis only if it was already known that pt allergic to drug otherwise it can't be predicted.
2062	Anaphylaxis and reactions should not be linked with errors

2065	Distinguish errors of prescribing, dispensing and delivery to patient
2071	Split the two parts - medications omitted prescribed or given in error are very closely related to quality, reactions and anaphylaxis may be.
2072	Separate errors of prescribing and administration from reactions
2081	Vague - some errors could be very minor, some life-threatening
2083	Omitted/given in error causes and effects very different - ?separate
2099	Omit reactions, anaphylaxis.
2100	Separate omissions etc where no adverse reaction happened from those where it did.
2124	Define criteria more precisely eg compare prescription orders with timing of doses given.
2126	Giving wrong medication and medication reaction different - unless you state reaction to known drug allergy.
2137	As above difference between human side effects and medical cockup
2139	Errors and omissions are one criterion (always avoidable); reactions are another (may or may not be avoidable).
2140	First part of question only.
2146	Medications given for which there are contraindications, medications given in incorrect dosage.
2148	A ragbag of problems.
2158	Anaphylaxis only important if known previous reaction not elicited.
2178	Potentially fatal easier to see criterion in this question.
2178A	A better question but again a range of effects is possible.
2183	Better two separate criteria.
2189	Too many parts to the question.
2196	As above, reactions need not result from poor practice.
2197	Needs splitting omissions from reactions.
2199	Range too wide from medication to anaphylaxis.
7209	The range of possibilities described is too great (trivial reactions to anaphylaxis).
7102	Leave out omitted or ask a separate question.

10.	*Cardiac or respiratory arrest in hospital*
2001	?age related
2003	Mixture of structure/process/outcome audit here
2004	Concentrate on outcome [of resuscitation] not occurrence
2028	Detailed info from records probably not available
2035	Omit
2036	Inevitable or not?
2040	Reasons for admission to hospital - more details of illness and age?
2041	Outcome
2065	Worth checking for avoidable factors
2091	Sensitive indicator of quality.
2094	Why - is it predictable/avoidable.
2099	Following maladministration of drugs?
2120	Criterion should be age-related.
2137	Depends on casemix/risk; eg acute MI or elective hernia?

2154 May not be quality of care but underlying condition.
2158 Not a helpful criterion.
2164 Little impact on patients health in general but dramatic effect on odd individual patients.
2178 Question strictly one of avoidability.
2178A If unexpected or avoidable.
7209 Checking for evidence of possible arrest could be difficult.

11. CVA or MI or PE following surgery
2028 Good criterion
2034 Difference between CVA/MI and PE due to potentially preventable cause eg non-use of prophylaxis during surgery
2035 Omit
2036 Preexisting predisposition to CVA or PE
2041 Embolus is more easily prevented
2053 Within ? time period of admission to include MI occurring shortly after discharge
2058 CVA/MI are mostly due to independent pathology - only precipitating event would be hypotension. PE - consider whether routine prophylaxis considered.
2065 Definitions required for each.
2091 But again CVA/MI are clinically different from PE. The latter should be totally avoidable, the former may not be.
2115 Does this question mean poor management of cardiac/respiratory arrest
2120 Criterion should be age-related.
2124 A more sensitive indicator if related to specific events following specific surgery eg DVT/PE following THR.
2131 CVA and MI not in same category as PE.
2146 Depends on screening test used, particularly for PE.
2148 PE - quality; MI/CVA may be.
2151 Stick to Pes in surgical patients.
2158 Stick to post-surgery with or without prophylaxis.
2178 Examples differ eg PE should be more easily preventable.
2178A If avoidable.
2183 Isn't PE more preventable than the other two.
7203 Pes in orthopaedic lower limb surgery.

12. Unexpected transfer to special care/higher dependency unit
2003 ? value
2028 How do you define unexpected transfer
2032 Depends on nature of problem - difficult to generalise
2035 Give specific circumstances
2036 Definition of unexpected
2039 Ability to transfer may indicate a high level of care
2041 Outcome
2043 What if could have been foreseen
2048 Reason for transfer should be probed - eg haemorrhage/MI

2051	Not clear whether this is considered a good thing or bad thing
2058	Better to be moved to ITU than not!
2067	Policies for transfer explicit
2124	Confine this to pts admitted for routine procedures.
2129	Pulmonary embolus and venous thromboembolism more preventable
2178	Decision to transfer not against patient's interests.
2178A	Depends if result of an avoidable incident.
7208	Could mean better care!

13.	*Patient related clinical complications occurred*
2001	Vague
2003	? take account of patient condition and interaction with services
2005	Inadequate pain control - how defined or recorded
2019	Separate inadequate pain control - sadly probably very rarely routinely recorded
2028	How do you define inadequate pain relief and detect from notes?
2030	Needs careful specification and recording of events
2032	Important quality of care criteria
2041	Outcome
2053	Confine the lists to say 3 or 4 easily definable and relevant situations
2065	Each complication must be defined.
2099	Depends entirely on what condition is chosen.
2100	Too many things in the one category.
2101	Explain - do you mean 1-12 above?
2112	Depends on the [?] available.
2137	Too miscellaneous; need to select out.
2147	Effect depends on complications.
2150	Separate out inadequate pain control.
2167	Too vague not useful.
2178	Wide range of different conditions - criterion difficult.
2178A	Again no one answer to all these scenarios.
2187	Not clear what is specified because it isn't.
7203	Uncertain about DVT - not as strong as pressure sores and infection.
7208	Could be more specific.

14.	*Patient-related non-clinical problems/incidents occurred*
2001	Vague
2003	Needs checklist
2005	? delay in undergoing diagnostic procedures
2019	Separate missing records from rest
2028	Most of this information is not recorded
2032	These are the quality issues on which patients judge their care
2040	More specific
2041	Could be trivial or life threatening
2052	Depends upon exactly what is meant - more precise please! too open-ended

2113	Question needs tightening up as some are clearly lifethreatening ie nonavailability of equipment others inconvenient ie bed on another ward, delays.
2137	better than 13 because reflects evident failures.
2172	By individual complication.
2178	Frequent, irritating, organisational, avoidable.
2178A	What is the alternative to some of these - no treatment?
7208	Falls, especially fractures, need separate heading.
15.	*Neurological deficit on discharge not present on admission*
2024	This assumes that no neurological deficit was present on initial examination.
2028	OK for well defined conditions
2040	Reasons for admission and procedures done
2041	Outcome
2042	You need to be more specific - ?surgical post-op, ?medical expected or unexpected or drug related etc
2053	Very difficult to separate from an underlying disease process.
2091	Effect depends on whether CVA or single nerve palsy.
2099	For specific procedure-related events.
2126	Depends on nature of deficit and whether it was part of natural history.
2137	Would score higher but may include deficit related to primary disease.
2150	Death would be recorded but possibly not enough details of surrounding circumstances.
2158	Avoid developing stroke inclusion.
2178	Depends on type of deficit and whether avoidable or careless.
2178A	Depends on whether avoidable or not.
2183	Depends on length of stay and age of patients.
7208	It depends on the procedure.
7100	Not in general form very relevant - specific circumstances maybe.
16.	*Unexpected death*
2004	Its very vague
2012	I felt a bit daft indicating this had serious effect on patients health
2028	Define unexpected!
2029	Omit
2040	Reason for death may be unrelated to condition treated
2043	Can learn from unexpected events - will have long term benefits
2049	This seems an unhelpful criterion - too vague to be of value, too vague to be easily assessed
2051	Coroner's job surely
2072	Omit or include with complications
2082	Define unexpected - is death following hip replacement in very old people expected or not?
2099	Exclude other occurrences included in other criteria.
2137	Unexpected = ??
2178	Again a question of avoidability.
2178A	Depends on whether avoidable or not ? some other unknown pathology.

2183	I presume the death would be recorded but not always the unexpectedness.
7209	Difficult to see how to categorise death (is 10-20% unexpected?)
7103	Even though unexpected it may have been unavoidable.
17.	*Medical record deficiency*
2019	Needs more specification of significant inadequacies
2028	Need to define a standard against which to compare
2032	Assumptions are made that sloppiness in recordkeeping infer sloppiness of care
2039	If records incomplete will be no record ? able to recognise incomplete - recognise routine omissions
2041	Could be trivial
2051	Inadequacies in records should be apparent from records, almost by definition - can the question be improved
2053	Far too vague - specify signal events eg USA has lots of experience with this - BP, smoking history, alcohol history, social history not recorded
2065	Clear guidelines required.
2099	Use specific omissions.
2124	Audited against a "gold standard" of medical/nursing records and scored accordingly.
2144	Duplicate records to enable discrepancy noted?
2146	Require standardised proformas.
2162	Part of wider concept of quality of care.
2167	Needs to be more specific.
2178	No excuse in quality sense acceptable.
2178A	It depends on circumstances - total picture.
2197	Split quality from other questions.
7203	Irrelevance is more of a problem.
7209	Maybe think about levels of severity of omission.
18.	*Nursing record deficiency*
2028	In contrast to medical records, nursing records are usually comprehensive and available.
2178A	Note does not precisely relate to the question here.
19.	*Evidence of patient and/or family dissatisfaction*
2001	Should involve surveys
2005	Specified complaints by type on scale
2028	Unlikely to find relevant info in notes
2040	More detail of cause of dissatisfaction
2041	Could be trivial
2053	Specify with clinical treatment as opposed to non-clinical parameters?
2065	Only possible to detect by talking to patients
2092	Written complaints only.
2099	Stratification needed.
2103	Previous history of complaints ie threshold for dissatisfaction.
2126	I have assumed dissatisfaction with care not the magazines available.

2137 Specify how defined, recorded etc
2154 If not recorded, not useful.
2172 Categorisation.
2178 Criteria not comparable.
2178A You cannot compare cold food with an operation cancelled several times.
2197 Split patient from family.
7209 I'm extending my answer to cover more than just organisational dissatisfaction.

20. Discharge related problems
2001 Needs patient input
2005 Inadequate discharge planning - what's that
2028 Unlikely to find relevant info in notes
2036 Needs to be defined more
2039 Information exchange to GP
2053 Probably best picked up by length of stay greater than say 28 days
2081 More detailed problems.
2096 Need to separate out hospital related poor quality ie poor discharge planning from external forces ie resthome bed not available.
2112 Social delays (DSS etc) should be excluded unless that is [?] poor case management.
2162 Quality ++
2172 Categorisation into difficulties due to hospital care and due to home arrangements
2178 Matter of record. No excuse.
2178 Depends on type of patient.
7209 Should disaggregate screwed up discharge procedures and waiting for nursing home etc.

3. Comments on the questionnaire study

General scaling comments
2004 Scale 1-10 is too wide
2183 I have assumed that 0 to 10 is a log scale rather than a linear 1 eg 1 means very rare not 10% and 9 means very common not 90%.

Relationship to quality questions
2019 Often difficult to quantify because of mixture of poor quality and bad clinical luck - bipolar causation
2028 Because of low specificity I found column 1 v difficult to complete.
2096 It is really all most important.

Effect on health question
2010 If there is a desire to relate the criteria to the effects on health these are too general for comment and require more specification.
2096 In some cases even the patient dissatisfaction could be an indication of serious and significant events.

2135 Difficult to score effect as for each adverse event the consequences may be trivial or very severe and it is difficult to generalise.

2178B General comment given the criteria in 6 to 20 I don't think it is possible to comment on effect. Effect has so many individual variables.

2179 I don't know and couldn't make a valid guess.

2191 In several places on this questionnaire the answers would be "could be important" type of reply ie inadequacy in patient records which are common but usually have little consequence yet occasionally can be important.

Scale for expected incidence/proportion question

2001 Proportion likely to be less than 10% for most.

2004 Estimating proportion anticipated seems unhelpful at this stage.

2024 Couldn't answer this column

2034 No idea but I suppose will vary between hospitals and specialties

2036 I didn't understand the proportion column. Do you mean the proportion of all acute admission who would fulfil the criteria - thats what I took it to mean although I'm not sure that's what you wanted.

2057 The proportion column is less meaningful than it might have been - if 0=none and 10=all then it stands to reason that each unit is 10% unless a complex scale is used. Most criteria would be prevalent between 0-10% which doesn't discriminate between them

2082 Column for proportion should be rate per 100 or rate per 1000 not 0-10

2084 We found the proportion column almost impossible to answer - but otherwise found the questionnaire useful.

2096 I have absolutely no data available on this important question.

2097 The proportion column is too condensed.

2999 I found the occurrence proportion difficult as many complications occur but not at 10% range!

Problems of completion

2015 Completed independently by medical and nursing personnel - both had difficulty completing especially columns 2,3,4

2019 Thought I'd find it very difficult, but thanks to your explanatory notes it wasn't so bad.

2028 I found questionnaire very difficult to complete and suspect its repeatability will be v low.

2048 I have addressed these questions from the viewpoint of medical audit and not the wider clinical audit. I have very little confidence in my replies to records/proportions/effects questions.

2068 I found this a very difficult questionnaire. very subjective. Mostly guesses. I very nearly didn't fill it in. i don't think that the results will be very useful.

2085 I am sorry my answers are so unsatisfactory - the adverse effects depend so much on the eventual outcome - however, they should not occur. The process items indicate efficiency but may have no effect on outcome

2091 Well thought out, some confusing questions eg 17 and 18. I bet clinicians would answer the questions differently? Why give it to non-clinical people to fill out?

2093 The proportion question is difficult to interpret.

2094 I understand what this is about in theory - answering the questions by valid indicators of rating is really not easy to achieve in practice. This took 30 mins and I do not feel I have done justice to everything. I wonder how many of those filling this in will understand what the questions really mean? some clinicians may do better - or have better data to go on.

2101 V difficult to answer especially "effect" - since various incidents would have very different effects.

2117 It is almost impossible to give general values for these criterion - they have differing significance with age, specialty etc.

2131 I find this very difficult, since I do not know what these criteria themselves mean especially improvement.

2137 A bold adventure - I have tried to be consistent (and it took 45 minutes!). Did MMA not do validation studies in USA?

2164 A bit too long - needed a big threshold to be overcome to face filling it in.

2172 I have great doubts over validity of my responses. it was extremely difficult to answer - for example, death diagnosed as due to cerebral haemorrhage instead of abscess wouldn't matter much but diagnosis of head cancer when benign cyst would matter a great deal.

2174 I always find this sort of questionnaire extraordinarily difficult to complete and I wonder if it has much value.

2178 There was so much difference between degrees of seriousness of examples given and of imprecision of questions as to render the questionnaire virtually useless. there appears to be lacking of what may be available and what is pure happenstance. I doubt whether clinicians would appreciate this type of enquiry.

2183 Most answers are guesstimates. About 25 minutes to fill in with 1 phone interruption. I don't feel I can complete column 5 (improvement) in most cases as I have not direct experience and little knowledge of studies which would validate these criteria.

2191 I found this questionnaire difficult to complete in places - partly due to lack of knowledge on my own part. I've done my best but some answers are little more than wild guesses.

2999 Only my dedication to audit caused me to persist!

Questionnaire layout

2016 Sorry did not see back page until filled in middle pages.

2053 The preamble and questionnaire are nicely designed - well done. I found the questions very difficult to answer

2057 Whoever completes this questionnaire accurately and without previous experience in 20 minutes should be a candidate for MENSA! - it is not complex, just thought and labour intensive.

2058 It was not clear that further information was available on the last page.

2062 Possibly design could be improved - it isn't user-friendly to turn to the back page quite so often

4.5 Interview schedule used in interview study of clinician opinion.

This appendix contains the proforma which was used to structure and record the interviews with study participants about the validity of adverse event measures.

What do you see as the advantages and disadvantages of occurrence screening, in measuring and assessing the quality of inpatient care? - What are its advantages? - What are its disadvantages?
How strong do you think the parallels are between occurrence screening and more traditional problem focused audits - such as CEPOD/NCEPOD, M&M meetings, etc - What are the similarities? - What are the differences?
How useful do you think occurrence information would be in: - measuring quality of care for individual patients? - measuring quality of care for groups of patients (eg in aggregate)? - creating and promoting changes in practice/quality?
Do you think occurrence screening is equally applicable to all areas or specialty of inpatient care? If not - What areas is it best suited for? - What areas is it least suited for?

What do you think would be important factors in getting an audit system based on occurrence screening to work well, and what would impede it? - Helps? - Hindrances?
What do you think are the most important/useful products, results or outcomes of occurrence screening - and what are the least important/useful products? - Most important/useful? - Least important/useful?
Would you use (or do you use) occurrence screening in your own specialty/hospital in medical audit/clinical audit? - By itself, in combination with other systems, not at all... - Why?

5.1 Obstetrics adverse-event measure of quality developed for and used in the RSCH occurrence screening project

OB01A Management of spontaneous rupture of membranes.

Management of spontaneous rupture of membranes in specified circumstances.

DEFINITION

Spontaneous rupture of membranes occurred in the following specified circumstances:
a) SROM occurred more than 24 hours prior to admission.
b) Mother presented to hospital following SROM and was discharged prior to delivery.

INFORMATION TO RECORD

For (a): Record length of time elapsed between SROM and admission to hospital and reason for delay if known and time and date of eventual delivery.
For (b): Record time of SROM, time of review by RSCH doctor, time of readmission and the mode of onset of labour e.g. spontaneous onset, stimulation.

OB02A Elective induction of labour.

The mother underwent elective induction of labour.

DEFINITION

The mother underwent elective induction of labour for one of the following specified reasons:
a) Post-maturity.
b) PIH.
c) Maternal medical reasons.
d) Fetal medical reasons.
e) Social reasons.
f) No reason given.
g) Other reasons.

GUIDANCE

Induction of labour means that the membranes are still intact and there are no contractions when intervention takes place.

OB03A Problems of labour/delivery.

Specified problems occurred during labour/delivery.

DEFINITION

One or more of the specified problems occurred during labour/delivery:
a) Failed trial of scar.
b) Failed forceps delivery.
c) Third degree tear.
d) Maternal injury.
e) More than three hours in active second stage of labour.
f) Undiagnosed breech presentation.
g) Undiagnosed multiple birth.
h) Delivery of infant following previous sterilisation.
i) Low lying placenta seen at booking scan.

GUIDANCE

Check mother's record for evidence of the above and record details.
e) Active 2nd stage means the cervix is fully dilated AND active pushing has commenced.

INFORMATION TO RECORD

For a) Trial of scar.
1) Parity with respect to previous caesarean section.
2) Gestation.
3) Reason for trial of scar (e.g. maternal preference or advice of obstetrician if discernible).
4) Method of onset of labour. If labour was induced, record type e.g. prostin or propess pessary, ARM, syntocinon and time from induction until start of labour i.e. when 3 cms.
5) Length of each stage. (1st stage: 3cm to fully; 2nd stage: fully to delivery).
6) Outcome.

For b) Failed forceps.
1) Maternal height
2) Evidence of the baby being large
3) Was labour augmented
4) Evidence of the mother having inadequate analgesia during labour
5) Length of time fully dilated
6) Position of head
7) Station
8) Evidence of maternal distress

9) Evidence of fetal distress
10) Reasons for failure
11) Apgars
12) Was this attempt a "trial of forceps" and was it in theatre

For Third Degree Tear:

Record:
1) Type of delivery - vaginal, Ventouse, forceps.
2) Position of mother at delivery.
3) Position of baby at birth.
4) Episiotomy, if performed.
5) Other relevant factors - e.g. hand presented with head.

i) For Low Lying Placenta: look at scan report at booking

Record:
1) Any bleeding pre-delivery and stage of gestation.
2) Mode of eventual delivery.
3) Whether admitted because of bleed from asymptomatic placenta praevia.

4) If Ultrasound located low lying placenta, whether USS repeated, at what gestation, and whether placenta still low lying at that time.

OB04A Caesarean section

Reasons for caesarean section

DEFINITION

The mother underwent caesarean section for specified reasons, classified as a) elective or b) emergency.

GUIDANCE

Check mother's record for required information.

INFORMATION TO RECORD

Record the reasons given for performing caesarean section and whether elective or emergency.

OB05A Problems of caesarean section

One or more of the specified problems occurred relating to caesarean section

DEFINITION

One or more of the following problems occurred relating to caesarean section:
a) Unplanned injury to, or repair of an organ or structure.
b) Excessive blood loss, e.g. more than 1 litre.
c) Deep vein thrombosis or pulmonary embolus.
d) Wound infection.
e) Other problems.
f) Baby needed resuscitation by intubation or bagging.

GUIDANCE:
For f) state whether Elective or Emergency Caesarean

OB06 Neo-natal problems

Neo-natal (up to 28 days following a live birth) problems occurred.

DEFINITION
One or more of the following neo-natal problems occurred:
a) Congenital defects (noted in first 10 days).
b) Apgar score less than 7 at 5 minutes.
c) Small for gestational age: Birthweight of less than the 10th percentile for that gestation.
d) Big baby: Birthweight greater than the 90th percentile for that gestation.
e) Apnoea: Episode of stopping breathing for longer than 20 seconds.
f) Transfer to SCBU. (Record reason).
g) Infection: record cot position on ward where possible.
h) Feeding problems including: poor effort at feeding, poor milk supply, obsessional feeding.
i) Hypothermia. Temperature less than 35 degrees Celsius or however defined.

GUIDANCE
Check mother's and baby's record for above problems and record details.

OB06A Perinatal problems.

Specified perinatal problems occurred.

DEFINITION

One or more of the following perinatal problems occurred:
a) Injury to the baby during delivery requiring follow-up.
b) Apgar score less than 7 at 5 minutes.
c) Infection.
d) Hypoglycaemia.
e) Death of baby.
f) Stillbirth.

GUIDANCE

Check mother and baby's record for evidence of the above problems.

OB07A Post - natal problems.

The mother experienced specifed post-natal problems.

DEFINITION

The mother experienced one or more of the specified post-natal problems:

a) Post-partum haemorrhage, blood loss of more than 500 mls.
b) Secondary haemorrhage, (after 24 hours and before 10 days post-partum)
c) Anti-D not given within 24 hours when indicated.
d) Infection not present on admission.

OB08A Problems with analgesia/anaesthesia

Problems with analgesia/anaesthesia occurred

DEFINITION

One or more of the following specified problems with analgesia or anaesthesia occurred:
a) Epidural or spinal anaesthetic problems e.g. spinal headache, prolonged block or urinary retention longer than 8 hours following administration.
b) Neurological deficit not present on admission.
c) Baby drowsy due to transmission of sedative effect of mother's analgesia.
d) Non-availability of analgesia/anaesthesia when required.

GUIDANCE

Check mother's and baby's record for evidence of the above problems.

OB09 Drug-related problems

One or more drug-related problems occurred.

DEFINITION
One or more of the following specified drug usage problems occurred:

a) Medications omitted, prescribed or given in error, given at wrong rate, delayed with no reason given.
b) Medication reactions and anaphylaxis.
c) Incomplete record e.g. omission of midwife's or doctor's signature.

GUIDANCE
Check mother's and baby's record for evidence of the above problems and record action taken.
All drugs administered should be recorded in the notes.

OB10 Mother/family dissatisfaction.

Evidence of mother and/or family dissatisfaction

DEFINITION
There is evidence in the mother's record that the mother or mother's relatives or friends expressed dissatisfaction with the care given including complaints of inadequate pain control.

GUIDANCE
Look for information on cause of dissatisfaction and how it was expressed in midwifery and medical records. Look for evidence that the mother/family complaint was handled appropriately;

All discharges taken against medical advice should be carefully checked against this criterion.

OB11 Mother or baby related non-clinical problems

Mother or baby related non-clinical problems/incidents occurred.

DEFINITION

One or more of the specified mother or baby related non-clinical problems or incidents occurred:

a) Theatre delay
b) Delay in obtaining a second opinion from another speciality
c) Casenotes, X-rays or other records or results missing/not available when needed
d) Mother or baby had slip or fall or other accident
e) Equipment failure
f) Necessary equipment not available when needed
g) Delay in undergoing diagnostic/therapeutic procedures e.g. ultrasound scan.
h) Inadequate referral notes with intrauterine transfers

GUIDANCE
Check medical and midwifery records. Document the causes of the problem or incident if possible. Check for list of valuable recorded. Note if theatre instruments unavailable. Note if theatre delayed/cancelled due to TSSU problems or missing report delays operation.

OB12 Obstetric record review

Review of quality, consistency and completeness of mother's and baby's records and related documentation

DEFINITION

One or more of the following specified deficiencies were found in the records or related documentation:
a) observations not recorded on partogram at half-hourly intervals - give reasons why if not.
b) lack of continuity or incomplete/contradictory recording of information.
c) documentation missing
d) notes illegible
e) doctor's or midwife's signature missing after entries in notes and after delivery.
f) signature not printed

GUIDANCE
Check mother's and baby's records and related documentation.

OB14 Problems of obstetric anaesthesia

Problems of obstetric anaesthesia

DEFINITION

The patient experienced one or more of the following problems relating to anaesthesia:

a) Delay.
b) Technical problems.
c) Intra-operative problems.
d) Post-operative problems.
e) Neonatal problems attributable to anaesthesia.

GUIDANCE

a) Delay - due to non-availability of anaesthetist, results of clotting screen (eg for pre-eclampsia) delayed, etc

b) Technical problems. General anaesthesia: difficulty with intubation,

oesophageal intubation, equipment malfunction, etc. Epidural or spinal anaesthesia: failure to produce block/inadequate block, proceeding to general anaesthesia, second anaesthetist needed to site block, dural tap, prolonged procedure, total spinal, etc.

c) Intra-operative problems - broken teeth, corneal abrasion, cardio-respiratory problems, difficulty in reversing anaesthetic, patient aware during procedure.

d) Post-operative problems - hypotension, respiratory depression, ITU admission, backache, neurological dysfunction, (including headache, weakness), bladder dysfunction, patient death within one week of anaesthesia.

e) Neonatal problems possibly attributable to anaesthesia - poor Apgar score,(< 6 at 1 minute, < 8 at 5 minutes), resuscitation required (eg. IVI inserted, ventilation for more than 5 minutes), transfer to SCBU, need to give Narcan to baby.

INFORMATION TO RECORD

Record category of problem(s) identified, and brief details of situation and

OB15 Problems of epidural/non-epidural pain relief in obstetrics

Problems of epidural and non-epidural pain relief in obstetric inpatients.

DEFINITION

The patient experienced one or more of the following problems while receiving epidural or non-epidural pain relief.

a) Delay.
b) Technical problems.
c) Post-epidural problems.
d) Neonatal problems attributable to analgesia.
e) Poor/inadequate analgesia of any type.

GUIDANCE

a) Delay - Epidural requested but not given. Non-availability of anaesthetist. Delayed results of clotting screen (eg in pre-eclampsia).

b) Technical epidural problems - unable to site epidural, second

anaesthetist required, prolonged procedure, dural tap, unilateral block, inadequate block, resiting required, etc.

c) Post-epidural problems - headache, bladder dysfunction, backache, neurological dysfunction.

d) Neonatal problems possibly attributable to analgesia - poor Apgar scores (< 6 at 1 minute, < 8 at 5 minutes),resuscitation required - IVI sited, ventilation for more than 5 minutes, transfer to SCBU, need to give Narcan to baby, etc.

e) Poor/inadequate analgesia of any type - as documented by medical/midwifery staff in notes, or complained of by mother.

f) Pethidine given to mother during labour - record 1) Amount given
2) Time given
3) Time of delivery

INFORMATION TO RECORD

Record category of problem found, and brief details of the situation and its management. Record details of Pethidine administration for assessment of collective use.

5.2 Results of multiway frequency analyses undertaken to examine construct validity of adverse-event measures used in the RSCH occurrence screening project

Tests that K-way effects are zero.

K	DF	L.R. Chisq	Prob	Pearson Chisq	Prob	Iteration
1	10	4256.920	.0000	11431.487	.0000	0
2	38	124.116	.0000	152.947	.0000	0
3	68	32.982	.9999	32.247	.9999	0
4	57	6.490	1.0000	6.028	1.0000	0
5	18	.000	1.0000	.000	1.0000	0

Tests of PARTIAL associations.

Effect Name	DF	Partial Chisq	Prob	Iter
LOSGRP*AGEGRP	6	24.075	.0005	3
AGEGRP*SEX	2	73.308	.0000	3
SEX*ADMTYPE	1	3.808	.0510	4
AEGRP	3	1051.909	.0000	2
LOSGRP	3	1397.994	.0000	2
AGEGRP	2	422.858	.0000	2
ADMTYPE	1	1384.099	.0000	2

Table 1. Results of multiway frequency analysis of 1,028 admissions in accident and emergency screened for adverse events.

Tests that K-way effects are zero.

K	DF	L.R. Chisq	Prob	Pearson Chisq	Prob	Iteration
1	10	2290.814	.0000	5420.326	.0000	0
2	38	141.290	.0000	258.183	.0000	0
3	68	56.194	.8460	80.487	.1428	0
4	57	34.537	.9919	27.964	.9996	0
5	18	5.768	.9971	4.633	.9993	0

Tests of PARTIAL associations.

Effect Name	DF	Partial Chisq	Prob	Iter
AEGRP*SEX*ADMTYPE	3	12.423	.0061	4
AEGRP*LOSGRP	9	27.989	.0010	4
LOSGRP*AGEGRP	6	19.051	.0041	3
AEGRP*ADMTYPE	3	11.907	.0077	4
LOSGRP*ADMTYPE	3	32.597	.0000	4
AGEGRP*ADMTYPE	2	13.555	.0011	4
AEGRP	3	149.307	.0000	2
LOSGRP	3	1520.861	.0000	2
AGEGRP	2	303.111	.0000	2
ADMTYPE	1	314.883	.0000	2

Table 2. Results of multiway frequency analysis of 798 admissions in ENT screened for adverse events.

Tests that K-way effects are zero.

K	DF	L.R. Chisq	Prob	Pearson Chisq	Prob	Iteration
1	9	3121.788	.0000	13144.549	.0000	0
2	29	32.216	.3105	106.971	.0000	0
3	39	13.800	.9999	11.908	1.0000	0
4	18	.000	1.0000	.000	1.0000	0

Tests of PARTIAL associations.

Effect Name	DF	Partial Chisq	Prob	Iter
AEGRP*ADMTYPE	2	7.306	.0259	3
AEGRP	3	273.992	.0000	2
LOSGRP	3	1068.590	.0000	2
AGEGRP	2	1133.079	.0000	2
ADMTYPE	1	646.127	.0000	2

Table 3. Results of multiway frequency analysis of 566 admissions in gynaecology screened for adverse events.

Tests that K-way effects are zero.

K	DF	L.R. Chisq	Prob	Pearson Chisq	Prob	Iteration
1	9	21521.229	.0000	85368.955	.0000	0
2	29	502.109	.0000	607.517	.0000	0
3	39	.909	1.0000	.499	1.0000	0
4	18	.000	1.0000	.000	1.0000	0

Tests of PARTIAL associations.

Effect Name	DF	Partial Chisq	Prob	Iter
AEGRP*LOSGRP	9	489.761	.0000	2
AEGRP	3	1464.418	.0000	2
LOSGRP	3	5936.417	.0000	2
AGEGRP	2	8696.615	.0000	2
ADMTYPE	1	5423.780	.0000	2

Table 4. Results of multiway frequency analysis of 3,958 admissions in obstetrics screened for adverse events.

Tests that K-way effects are zero.

K	DF	L.R. Chisq	Prob	Pearson Chisq	Prob	Iteration
1	10	9228.247	.0000	36677.981	.0000	0
2	38	534.665	.0000	1724.113	.0000	0
3	68	89.887	.0390	101.703	.0051	0
4	57	36.879	.9823	32.534	.9962	0
5	18	9.087	.9577	6.807	.9917	0

Tests of PARTIAL associations.

Effect Name	DF	Partial Chisq	Prob	Iter
LOSGRP*SEX*ADMTYPE	3	11.194	.0107	5
AGEGRP*SEX*ADMTYPE	2	12.234	.0022	4
AEGRP*LOSGRP	9	59.446	.0000	5
AEGRP*AGEGRP	6	14.177	.0277	6
LOSGRP*AGEGRP	6	61.055	.0000	5
AGEGRP*SEX	2	110.287	.0000	6
LOSGRP*ADMTYPE	3	138.942	.0000	4
AGEGRP*ADMTYPE	2	66.359	.0000	6
SEX*ADMTYPE	1	4.034	.0446	6
AEGRP	3	1991.975	.0000	2
LOSGRP	3	4952.723	.0000	2
AGEGRP	2	669.688	.0000	2
SEX	1	102.777	.0000	2
ADMTYPE	1	1511.087	.0000	2

Table 5. Results of multiway frequency analysis of 2,231 admissions in ophthalmology screened for adverse events.

Tests that K-way effects are zero.

K	DF	L.R. Chisq	Prob	Pearson Chisq	Prob	Iteration
1	10	3906.491	.0000	6204.237	.0000	0
2	38	1727.670	.0000	2286.209	.0000	0
3	68	110.408	.0009	108.715	.0013	0
4	57	49.996	.7330	50.350	.7210	0
5	18	.955	1.0000	.843	1.0000	0

Tests of PARTIAL associations.

Effect Name	DF	Partial Chisq	Prob	Iter
LOSGRP*AGEGRP*ADMTYPE	6	19.686	.0031	5
AGEGRP*SEX*ADMTYPE	2	19.645	.0001	5
AEGRP*LOSGRP	9	250.205	.0000	7
AEGRP*AGEGRP	6	18.433	.0052	7
LOSGRP*AGEGRP	6	350.884	.0000	7
LOSGRP*SEX	3	11.461	.0095	6
AGEGRP*SEX	2	343.451	.0000	7
LOSGRP*ADMTYPE	3	126.744	.0000	8
AGEGRP*ADMTYPE	2	71.516	.0000	8
SEX*ADMTYPE	1	5.495	.0191	8
AEGRP	3	698.386	.0000	2
LOSGRP	3	2790.238	.0000	2
AGEGRP	2	241.877	.0000	2
SEX	1	17.845	.0000	2
ADMTYPE	1	158.150	.0000	2

Table 6. Results of multiway frequency analysis of 2,916 admissions in orthopaedics screened for adverse events.

Tests that K-way effects are zero.

K	DF	L.R. Chisq	Prob	Pearson Chisq	Prob	Iteration
1	10	807.314	.0000	1073.543	.0000	0
2	38	209.020	.0000	216.161	.0000	0
3	68	41.771	.9949	38.573	.9985	0
4	57	41.597	.9375	37.806	.9765	0
5	18	3.994	.9998	3.168	1.0000	0

Tests of PARTIAL associations.

Effect Name	DF	Partial Chisq	Prob	Iter
AEGRP*LOSGRP	9	79.032	.0000	5
LOSGRP*AGEGRP	6	35.825	.0000	5
AEGRP*ADMTYPE	3	10.699	.0135	5
LOSGRP*ADMTYPE	3	14.888	.0019	5
AGEGRP*ADMTYPE	2	34.907	.0000	5
SEX*ADMTYPE	1	7.817	.0052	5
AEGRP	3	99.438	.0000	2
LOSGRP	3	687.844	.0000	2
AGEGRP	2	8.478	.0144	2
ADMTYPE	1	11.407	.0007	2

Table 7. Results of multiway frequency analysis of 549 admissions in general surgery screened for adverse events.

Tests that K-way effects are zero.

K	DF	L.R. Chisq	Prob	Pearson Chisq	Prob	Iteration
1	10	1233.396	.0000	2134.724	.0000	0
2	38	117.602	.0000	123.885	.0000	0
3	68	41.345	.9956	44.908	.9862	0
4	57	39.242	.9650	35.416	.9890	0
5	18	.646	1.0000	.390	1.0000	0

Tests of PARTIAL associations.

Effect Name	DF	Partial Chisq	Prob	Iter
AEGRP*SEX*ADMTYPE	3	10.637	.0139	3
LOSGRP*AGEGRP	6	71.936	.0000	3
AGEGRP*SEX	2	13.586	.0011	4
AGEGRP*ADMTYPE	2	9.917	.0070	4
AEGRP	3	320.418	.0000	2
LOSGRP	3	631.231	.0000	2
AGEGRP	2	14.601	.0007	2
SEX	1	257.693	.0000	2
ADMTYPE	1	9.452	.0021	2

Table 8. Results of multiway frequency analysis of 520 admissions in urology screened for adverse events.