THE INFLUENCE OF CENTRE SELECTION ON THE GENERALISABILITY OF ECONOMIC EVALUATIONS CONDUCTED ALONGSIDE RANDOMISED CONTROLLED TRIALS. A CASE STUDY FROM THE *ROSSINI* TRIAL

by

ADRIAN GHEORGHE

A thesis submitted to the University of Birmingham for the degree of

DOCTOR OF PHILOSOPHY

Primary Care Clinical Sciences
School of Health and Population Sciences
College of Medical and Dental Sciences
University of Birmingham

2013

Abstract

The thesis investigated the influence of centre selection on the generalisability across locations of trial-based economic evaluations. A novel methodology to assess and enhance the generalisability of trial findings was demonstrated using the comparison between wound-edge protection devices (WEPDs) and standard care to reduce surgical site infection (SSI) after open abdominal surgery as a case study.

A systematic review and a preliminary economic model suggested that WEPDs may be effective and cost-effective in reducing SSI compared to standard care, although the methodological quality of available studies was poor. ROSSINI was a high quality multi-centre randomised controlled trial (RCT) which demonstrated that WEPDs are unlikely to be effective or cost-effective, so their routine use cannot be recommended.

The impact of centre selection on trial results was then investigated using ROSSINI as a case study. Mixed methods research demonstrated that most RCTs do not enrol centres so as to ensure a representative sample at jurisdiction level. The Generalisability index (Gix) was introduced as the basis of a novel methodology to assess generalisability, which was demonstrated using simulation methods and ROSSINI data. The results suggested that the characteristics of the sample of participating centres can significantly affect RCT clinical and cost-effectiveness estimates.

Dedication

To my family, present and future

Acknowledgements

There are many people who made the completion of this research possible and to whom I would like to thank.

First and foremost I would like to thank my main supervisor, Dr. Melanie Calvert, for shaping a uniquely rewarding doctoral experience and for being a model in conducting research. I have always found in her encouragement, realism, trust and inspiration.

To my supervisor Professor Tracy Roberts for continuous support, extremely insightful contributions and unabated patience, but especially for being the healthiest challenger a PhD student could ever hope for.

To Professor Sue Wilson for her conceptual contributions to the initial stages of the research and for believing in this idea when it seemed only an improbable thought.

To Dr. Karla Hemming for her decisive contribution to the development of the Generalisability index in the most decisive of moments.

To Dr. Jonathan Ives for substantial contributions to the design and analysis of the mixed method study.

To my friend Benjamin Fletcher for his contributions to no less than two systematic reviews but, far more importantly, for endless and extremely enjoyable discussions on trial methodology, methodology in general and life.

To Thomas Pinkney, David Bartlett, Professor Dion Morton and the ROSSINI Trial Management Group for their trust and support.

To the National Institute for Health Research for funding this PhD studentship.

To Dr. Nicola Gale for her help in conducting one of the focus groups.

To my colleagues Val Redman and Richard Haslop for doing a fantastic job with running ROSSINI and for invaluable support.

To Dr. Sabi Redwood, my mentor, for her patience, thoughtfulness and for being the most gifted listener I have yet encountered.

To Dr. George Dowswell, Sally Warmington, Sue Clifford, Janet Jones, Melanie Guthrie, Laura Parker for being the best colleagues one could think of.

To Dr. Sarah Damery, Dr. Sarah Flanagan and Dr. Sue Neilson for offering me a warm academic home in times of need.

To my colleagues in the Health Economics Unit for their support and warmth.

To Dr. Ulla Griffiths, my manager at LSHTM, for strong support and encouragement during the write-up phase.

To Simon Jenkins and Dr. Greg Moorlock, my friends and partners in endless debates, pancakes and good curry.

To Mos Llangco and Benjamin Toro for being outstanding friends.

To Kiron Griffin, Naomi Adelson and Helena Lee for making 34 Quinton Road a home.

To my family at home, my mother Elena, my father Emil and my aunt Nina, who have always believed in me beyond words.

Contributorship statement

Chapters 1, 2, 3, 4, 6, and 9 are entirely the product of my own work with continuous guidance and inputs from my supervisors, Dr. Melanie Calvert and Prof. Tracy Roberts.

The ROSSINI trial, reported in Chapter 5, was designed by the West Midlands Research Collaborative and managed by the Birmingham Centre for Clinical Trials. I assisted Dr. Melanie Calvert in the main statistical analysis of the trial and conducted the entire trial-based economic evaluation, with continuous guidance and inputs from Dr. Melanie Calvert and Prof. Tracy Roberts.

The mixed methods study reported in Chapter 7 was designed primarily by me with advice from my two supervisors and Dr. Jonathan C. Ives, Senior Lecturer in Biomedical Ethics (Primary Care Clinical Sciences, University of Birmingham). One of the focus groups was moderated by Dr. Nicola K. Gale, Lecturer in the Sociology of Health Care (Health Services Management Centre, University of Birmingham).

The Generalisability index methodology reported in Chapter 8 is based on my design, developed under the guidance of my two supervisors and with substantial contributions from Dr. Karla Hemming, Senior Lecturer in Biostatistics (Public Health, Epidemiology and Biostatistics, University of Birmingham).

Mr. Benjamin R. Fletcher, Research Associate (Primary Care Clinical Sciences, University of Birmingham) contributed to the conduct and manuscript review of the systematic reviews reported in Chapter 3 and section 7.2.

Table of contents

List of illustrations

List of tables

List of abbreviations

BCa – Bootstrapped bias corrected and accelerated

CDC – US Centers for Disease Control and Prevention

CI – Confidence interval

CONSORT – CONsolidated Standards of Reporting Trials

GP – general practitioner

HCAI – Health-care associated infection

HPA – UK Health Protection Agency

HRQoL – Health-related quality of life

ICER – Incremental cost-effectiveness ratio

ISPOR – International Society for Pharmacoeconomics and Outcomes Research

MI – multiple imputation

MICE – multiple imputation using chained equations

MFF – Market Forces Factor (UK Department of Health)

MRSA – Methicillin-resistant *Staphylococcus aureus*

NHS – National Health Service (England and Wales)

NICE – UK National Institute for Health and Clinical Excellence

NIHR – UK National Institute for Health Research

NNIS – US National Nosocomial Infections Surveillance

OR – Odds ratio

PSA – Probabilistic sensitivity analysis

QALY – Quality-adjusted life year

RCI – Reference Cost Index (UK Department of Health)

RCT – Randomised controlled trial

RCT-EE – Randomised controlled trial with an embedded economic evaluation

ROSSINI – Reduction of Surgical Site Infection Using a Novel Intervention

RR – Risk ratio

SD – Standard deviation

SPIRIT – Standard Protocol Items: Recommendations for Interventional Trials

SSI – Surgical Site Infection

SWI – Surgical Wound Infection

US – United States

UK – United Kingdom

WEPD – Wound-edge protection device

WHO – World Health Organization

WTP – Willingness-to-pay

CHAPTER 1. INTRODUCTION

This research focuses on the generalisability across locations of the results of economic evaluations conducted alongside randomised controlled trials (RCTs). An intervention to reduce surgical site infection (SSI) rate after open abdominal surgery is used as a case study: after generating cost-effectiveness evidence by applying a range of standard methods, the thesis investigates the importance of centre selection for trial results and demonstrates a novel approach to evaluating generalisability.

This Chapter describes the methodological background of the research and presents the structure of the thesis. The first section outlines the principles of RCTs, followed by a brief introduction to economic evaluation methods and the current issues concerning the generalisability of RCTs and trial-based economic evaluation results. The Chapter ends by stating the objectives of the research and by presenting an outline of the thesis' Chapters.

1.1. Randomised controlled trials

Clinical trials can be defined broadly as experiments which test a medical intervention on human subjects (1). RCTs are a category of clinical trials with two important features: an explicit control group, which enables a direct comparison between the intervention(s) being tested and a comparator; and a random treatment allocation process, which ensures that participants differ only by chance and the intervention they are about to receive. RCTs are conducted in order to answer one or more meaningful research questions concerning the benefits and harms of a given intervention relative to the chosen comparator.

The study outcomes operationalise a RCT's research questions. The primary outcome is the most important outcome in a trial and was defined by the International Conference for Harmonisation - Statistical Principles for Clinical Trials (ICH E9, p.5) as "*the variable capable of providing the most clinically relevant and convincing evidence directly related to* *the primary objective of the trial*" (2). The choice of the primary outcome should be informed by the available clinical evidence and key stakeholders such as patients, investigators and clinicians. Additional outcomes may include other clinically important variables, safety markers, health-related quality of life (HRQoL) and cost-effectiveness; the last two are important to policy makers to inform market authorization and reimbursement decisions (3). In order to improve comparability, it is recommended that all trials conducted in a particular therapeutic area adopt a core set of outcomes (4).

The principal research question of a given RCT can be formulated, by means of the primary outcome, as a testable hypothesis which is usually labelled the 'null hypothesis' (H_0). For example, in a trial of an antihypertensive drug X compared to placebo, a suitable primary outcome may be the difference from baseline in systolic blood pressure after 90 days of treatment. The null hypothesis may, thus, be that 'Drug X is not more effective than placebo in controlling systolic blood pressure'. In order to test the null hypothesis, primary data are collected from an appropriate sample of participants, the relevant sample statistic is calculated (e.g. mean difference in systolic blood pressure across treatment groups) and a decision rule based on the sample statistic is used to decide whether sample data support the null hypothesis or not i.e. H_0 can be rejected or not. Upon making such a decision, two types of errors can be made: type I error refers to the case when H_0 is rejected when it is in fact true; the notation for the probability of committing a type I error is α. Conversely, type II error refers to the case when H_0 cannot be rejected, but it is in fact false; the notation for the probability of committing a type II error is β. The power of a statistical hypothesis test measures its capacity to reject the null hypothesis when it is indeed false i.e. the capacity to make a correct decision $(1 - \beta)$ (5).

The primary outcome is also important because it informs the calculation of the trial sample size i.e. the number of patients who need to be recruited in order to maximise statistical power. The following generic types of data inputs are necessary to calculate the sample size: the minimally important difference in the primary outcome between the trial's arms that investigators expect to observe; the level of statistical significance α (usually 0.05); the required power (usually 0.80); and, for continuous outcomes, the standard deviation of the measurements (6).

Bias can be understood as any systematic error in results and is a major concern in any experiment (2). RCTs are regarded as the gold standard in clinical research because of their potential to minimise the effect of several important biases. For example, randomisation can minimise selection bias by ensuring that patients are allocated to the intervention or control arm purely by chance and not subject to patient or clinician preferences. This can be achieved by using a treatment allocation sequence generation method which is unpredictable and cannot be easily tampered with. An acceptable example is a computer-generated sequence communicated to researchers via a secure Internet or phone connection, while poor methods include the use of sealed envelopes and allocation according to the day of the week. More sophisticated randomisation procedures include balancing the trial arms across known risk factors (stratification) and randomising sequentially within blocks of patients of random size (blocking) to maintain the desired intervention to control allocation ratio. Blinding refers to keeping the study personnel and participants unaware of treatment assignment and, if implemented appropriately, minimises performance bias i.e. uneven medical attention across trial arms. Blinded outcome assessment i.e. the professionals who are conducting the assessments are unaware of the treatment allocation, reduces the risk of detection bias, thereby ensuring the study outcomes are evaluated objectively. A comprehensive discussion of sources of bias and available options to minimise and assess them is given in the Cochrane Handbook of Systematic Reviews of Interventions (7).

A number of trial designs are available, their appropriateness depending primarily on the therapeutic area, the trial intervention and the trial objectives. The most straightforward configuration is the parallel design, where patients are randomised to either the intervention or control arm and are subject only to the corresponding regimen. By contrast, in cross-over designs patients are randomised to either of the arms and after a specified time interval they switch to the other arm (8, 9). The main advantages of cross-over designs are that each patient acts as their own control and smaller sample sizes are required to observe a significant effect. However, these designs are only applicable to chronic, reversible conditions and there are issues associated with carry-over effects i.e. the effect of a treatment may be such that by the time patients switch to an alternative regimen they are not is the state they would have been had they not received the initial treatment. One way to deal with carry-over effects is the introduction of a wash-out period, after which all measurements are assumed to be unaffected by the previous treatment.

In factorial designs two or more treatments are evaluated simultaneously; such designs are particularly useful if the objective is to understand interactions or to describe dose-response characteristics (10). Parallel, cross-over and factorial designs are conventional trial configurations where the unit of randomisation is the individual (arguably in cross-over trials the unit of randomisation is the sequence of interventions that the individual undergoes). However, the unit of randomisation can be more complex, such as a health care institution or a geographical area, where participants within that unit undergo only the allocated treatment. This is the case of cluster randomised trials, which are particularly useful in evaluating interventions where randomisation at individual level is problematic (the case of contamination effects, such as in the evaluation of health care professionals training programmes) or impossible (the case of environmental factors such as air quality) (11). More recently, adaptive designs allow updating trial characteristics based on accumulating information without jeopardising the integrity of the analysis (12, 13).

The trial protocol describes the objectives, design, methodology, statistical considerations and organisation of a trial (14). The protocol fulfils several roles: it documents how data should be collected, managed and analysed; it presents the trial to funding, regulatory and ethics bodies when applying for grants and approvals; it demonstrates the trial's compliance with official regulations, norms and guidelines; and it acts as a reference document throughout trial conduct. The SPIRIT Initiative (Standard Protocol Items for Randomized Trials) have recently published a list of standard items to be included in RCT protocols (15). In addition to data collection and analysis methods, the trial protocol must include procedures for issues such as confounding and handling missing data. In the case of missing data, it is important to investigate the reasons for which the data are missing in order to estimate the degree of bias likely to be incurred and to inform the methods for dealing with it (16).

Missing data are a sensitive topic in RCTs (17, 18). Despite the best efforts to ensure complete data collection, small amounts of missing data are inevitable. This is even more the case with patient self-completed case-report forms and patient-reported outcomes (PRO) assessments, which may be returned incomplete or not returned at all (19). The International Conference for Harmonisation - Statistical Principles for Clinical Trials (2) does not specify precise guidelines with respect to the volume of missing data, but only require the analyses to be “sensible”. Of paramount importance is, however, the mechanism responsible for data missingness and researchers are strongly encouraged to investigate this mechanism prior to making definitive decisions. The main types of missingness mechanisms were conceptualised by Little and Rubin (20):

a) 'missing completely at random' (MCAR): the probability of an observation to be missing is independent of both observable and unobservable variables;

b) 'missing at random' (MAR): the probability of an observation to be missing is dependent on observable variables and independent of unobservable variables. MAR is the weakest assumption based on which valid inferences can be produced using only the observed data and without having any other information regarding the missingness mechanism; and

c) 'missing not at random' (MNAR): the probability of an observation to be missing depends on both observable and unobservable variables. Valid inferences can only be obtained by considering a joint model of the observed data and the missingness mechanism.

The nature of the missingness mechanism can never be known with certainty, although a distinction can be made between MCAR and MAR in the sense that close inspection of the data can rule out MCAR. It is always the case that a number of assumptions have to be made before proceeding to handling missing data. Two types of approaches to missing data can be distinguished: traditional (*ad hoc*) methods and likelihood methods. *Ad hoc* methods (listwise deletion, casewise deletion, mean marginal imputation, last value carried forward) make strong assumptions about the data and have been strongly critiqued (17, 21).

An attractive modern method is multiple imputation (MI) (22). The underlying principle is the following: instead of imputing a single value for a missing observation, MI imputes m>1 values, generating m alternative complete datasets which can be analysed using standard statistical techniques. MI operates under the MAR assumption and the imputed values for each observation are conditional on the joint distribution of the missing variables and other observed variables for that observation. A multivariate normal distribution is assumed, which raises questions about the method's suitability for non-normally distributed data. However, Graham and Schafer (23) showed that MI performs well even for extremely non-normal variables. The estimates of the m analyses are ultimately combined using a set of rules formulated by Rubin (22). Although Rubin demonstrated that more than five imputations bring negligible gains in efficiency, more recent accounts recommend a larger number of imputations (24, 25).

Multiple imputation using chained equations (MICE) is a method for generating imputed values based on imputation model constructed for each variable with missing data (26). The underlying principle is that, following an initial filling of all missing values using random sampling with replacement from the observed values, each variable in turn is regressed against all the others and the missing values are replaced with values drawn randomly from its posterior predictive distribution. The process is repeated for a number of k cycles (usually 10-20) and m datasets are produced, similar to MI. The estimates are then combined using Rubin's rules. The important strength of MICE over MI is that it can easily handle variables with different distributions. Moreover, each variable can have its own imputation model, as opposed to MI which did not distinguish between independent and dependent variables. Nevertheless, it does not yet have firm theoretical grounds and is sensitive to model (mis)specification.

Adequate RCT reporting is of utmost importance for assessing the value of the findings and for planning future research. The CONSORT (CONsolidated Standards of Reporting Trials) Statement aims to provide a framework for the appropriate reporting of RCT methods and results (27). CONSORT extensions are also available for various types of designs (28, 29) and outcomes (30).

1.2. Economic evaluation

Drummond *et al.* (31) defined economic evaluation as "*the comparative analysis of alternative courses of action in terms of both their costs and consequences*" (p. 9). Given the resource scarcity in the health care sector, economic evaluations can inform choices between existing alternatives by making explicit the criteria underlying the decision. There are two principal economic paradigms from which the evaluation can be conducted: welfarist and extra-welfarist (32). The differences between these two perspectives are substantive in what concerns the relevant outcomes, the sources of outcome valuation, the weighting of the outcomes and the extent to which interpersonal comparisons are possible. Welfarism assumes that individuals make rational choices by selecting the options which maximise their welfare; individuals are the best judges of their welfare; utility derives from outcomes or behaviours rather than from processes; and utility information is the only argument used to assess the merit of a given state. Central to the welfarist paradigm is the concept of 'utility', which has received a range of interpretations across history (33), but can be understood as an individual's preference ordering over bundles of goods or states of the world (32). By contrast, extra-welfarism allows the use of other relevant outcomes than utility, does not consider individuals as the only source of valuation, explicitly allows outcome weighting based on non-preference principles and explicitly allows interpersonal outcome comparisons (34, 35). Although extra-welfarism allows the incorporation of relevant outcomes other than utility, such as equity, its practical applications have been criticised for focusing solely on health (36).

Several techniques of economic evaluation can be distinguished based on their approach to the valuation of consequences of health care interventions. The main types of economic evaluation are cost-benefit analysis (CBA), cost-effectiveness analysis (CEA) and cost-utility analysis (CUA) (31). CBA (37) is rooted in the welfare economic theory and evaluates the net social benefit of an intervention by comparing the costs and benefits of a given alternative, both valued in monetary units. An intervention is judged to be worth implementing if the net social benefit is positive i.e. net benefits outweigh net costs. Economists noted the methodological and ethical difficulties associated with assigning monetary values to health outcomes, a key step in CBA (38).

Both CEA and CUA assess a given alternative's value by comparing it to an external standard and assume that the decision makers' objective is to maximise health outcomes, but they do not measure benefits using the same unit: CEA uses natural units (e.g. cases averted, deaths), while CUA employs HRQoL measures. The two methods are similar in both application and interpretation, to the point where no formal distinction is made between them: for instance, in the US literature the term 'cost-effectiveness analysis' comprises both CEA and CUA, and increasingly so in the UK as well (39). Although CEA/CUA avoid a direct monetary valuation of health outcomes, in contrast to CBA, an external criterion of value is necessary to inform decision-making. An example of such a criterion is an accepted willingness-to-pay (WTP) threshold value (40).

It can be argued that CBA has a broader scope than CUA/CEA. First, by assigning monetary values to outcomes, CBA is suitable to compare programmes across different sectors of the economy, while CEA and CUA are restricted to comparing interventions which produce similar outcomes. Second, CUA/CEA often focus solely on health benefits and thus mainly address questions of production efficiency, while CBA can easily inform allocative efficiency decisions because it assigns relative values to both health and non-health outcomes.

Third, CEA and CUA are less equipped to capture health externalities because they usually focus on health outcomes, while CBA can quantify a wider range of effects.

Nevertheless, CUA is particularly useful because it allows comparability between largely different programmes and provides a means to integrate patients' preferences in the decision process (31). The costing exercise involves accounting for the monetary value of the resources associated with the programme's implementation. The choice of the considered costs is a delicate issue and a balance must be struck among several factors e.g. the perspective of the evaluation, the costs' relevance and the resources available for the evaluation itself.

Benefits in CUA are expressed in quality-adjusted life years (QALYs), a measure which combines morbidity and mortality such that it reflects an intervention's implications on both quality and quantity of life (41, 42). QALYs are generated by weighting the life expectancy with health utility weights informed by patients' preferences. Utility weights are anchored on death and perfect health and are measured on an interval scale – usually 0 to 1, where 0 corresponds to death and 1 to perfect health, although negative values are possible to indicate health states perceived as worse than death. A multitude of instruments are available for assessing the preference-based utility weights, both general (e.g. EQ-5D (43), SF-6D (44), HUI2 (45)) and disease specific (e.g. EORTC QLQ-C30 for cancer patients (46)).

The outputs produced by CUA are the costs and QALYs for each of the alternatives under scrutiny. In practice, one of the alternatives is usually the current standard of care, be it an intervention or simply no intervention. The metric of interest for decision-making purposes in CUA and CEA is the incremental cost-effectiveness ratio (ICER), defined as:

$$\mathrm{ICER} = \frac{C_i - C_o}{E_i - E_o} = \frac{\Delta C}{\Delta E} \quad (1.1),$$

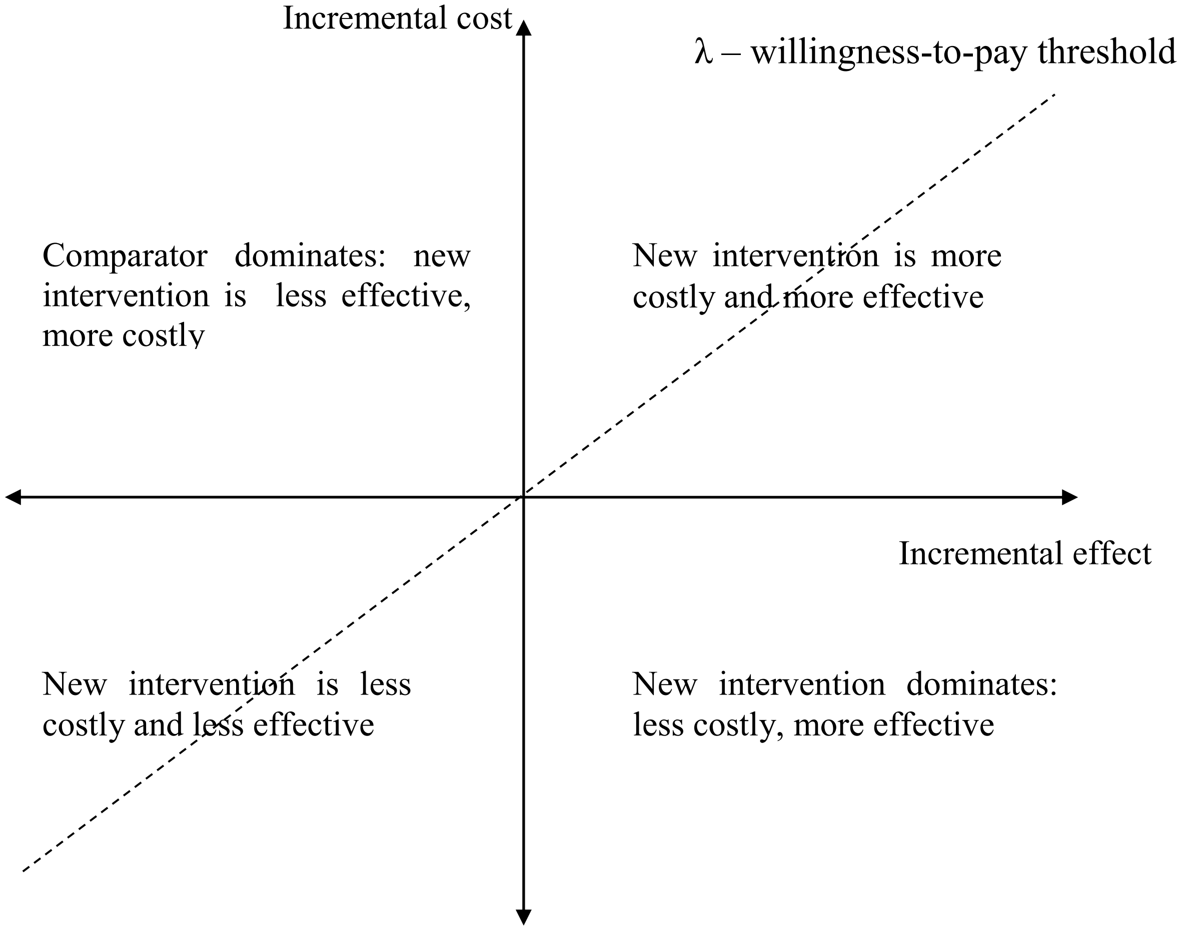
where:

C_i, E_i – costs (monetary units) and effects (QALYs) of the intervention under study; and

C_0, E_0 – costs (monetary units) and effects (QALYs) of the comparator (standard care).

The ICER represents the additional spending on a medical intervention compared to another in order to gain one extra QALY. There are instances where the value of the ICER does not communicate much about the relative implications of the two alternatives – for example, the ICER is positive both when the intervention is less costly and less effective, but also more costly and more effective than the comparator. The cost-effectiveness plane (47) is a graphic tool that clarifies such instances, allowing the straightforward visualisation of the incremental costs and effects (Figure 1.1).

Figure 1.1 The cost-effectiveness plane


Incremental cost
λ – willingness-to-pay threshold
Comparator dominates: new intervention is less effective, more costly
New intervention is more costly and more effective
Incremental effect
New intervention is less costly and less effective
New intervention dominates: less costly, more effective

The decision rule based on the ICER is that an intervention can be judged to be cost-effective if the ICER is below a set WTP threshold favoured by the decision maker. The UK decision body, the National Institute for Health and Clinical Excellence (NICE), currently favours interventions with an ICER between £20,000 and £30,000 per additional QALY (39), although the legitimacy of this interval is controversial (48, 49).

The ICER is constructed as a ratio of two differences between means (equation 1.1). While the differences can be assumed to asymptotically normal (if the sample size is large enough via the central limit theorem or if costs and effects are normally distributed), the sampling distribution of the ratio itself cannot be known. This raises serious difficulties in specifying confidence intervals around the ICER. Two types of methods have been suggested: a) parametric approaches, including the confidence box method, Fieller's theorem and Taylor series (50); and b) bootstrapping approaches, which include several variations such as the normal approximation, the percentile method, the bias-corrected and accelerated method (BCa) and parametric bootstrapping (51).

The objective of bootstrapping (52) is to make inferences about a population parameter based on a sample drawn from that population. The underpinning principle is that, given a sample of size n, repeatedly sampling with replacement from this sample and calculating the statistic of interest for each of the resulting samples of size n will construct an empirical distribution of the sampling distribution of the statistic of interest. While the process of obtaining the random samples and the statistic for each of them is straightforward, various methods of constructing the confidence intervals based on the empirical sampling distribution have been proposed. The normal approximation employs the traditional formulation of the variance and assumes that the sampling distribution of the statistic is normal. The percentile method involves ranking the statistics obtained from the replicated samples and selecting the i*th* percentile values and the bounds of the confidence interval. The bias corrected and accelerated method (BCa) is a modification of the percentile method which corrects for the estimator bias i.e. unequal proportion of bootstrap replicates above and below the sample statistic, and for the skew of the sampling distribution (53). The Fieller's theorem approach and BCa have been shown to outperform other methods (54).

Due to the statistical difficulties in expressing uncertainty around the ICER using parametric methods, after rearranging equation 1.1 the incremental net benefit (INB) was proposed as an alternative statistic of interest for cost-effectiveness (55):

$$\text{INMB} = \Delta E * \lambda - \Delta C \ (1.2)$$

$$\text{INHB} = \Delta E - \Delta C/\lambda \ (1.3),$$

where:

INMB – incremental net monetary benefit;

INHB – incremental net health benefit;

ΔE – incremental effect;

ΔC – incremental cost;

λ – willingness-to-pay threshold (£/QALY).

The net benefit (NB) framework has several advantages compared to incremental cost-effectiveness ratios. First, its interpretation is unambiguous and does not require information about the joint distribution of (ΔC; ΔE) pairs: positive values favour the intervention under scrutiny, while negative values do not. Second, net benefits are generally asymptotically normal, which allows obtaining unbiased estimates of variance.

Cost-effectiveness acceptability curves (56) describe the probability of an intervention to be cost-effective at a given willingness-to-pay threshold λ. The rules can be formally different depending on the chosen cost-effectiveness estimator: for the ICER, the CEAC is specified by the probability that the ICER<λ, if ΔE>0 and ICER>λ, if ΔE<0. In terms of the net benefit framework (55), the CEAC is given by the probability of the NB(λ)>0, where NB(λ) is the net benefit estimator. A thorough account of the definition, calculation and interpretation of CEACs is given by Löthgren and Zethraeus (57) and Fenwick *et al.* (58).

It must be noted that the CEAC only refers to a single intervention at a time. When multiple alternatives are compared simultaneously, the cost-effectiveness acceptability frontier (CEAF) extends the concept of CEAC by depicting the probability of the *optimal* option at each λ to be cost-effective. This may or may not be the alternative with the highest probability of being cost-effective, as indicated by the CEAC.

Although CEACs bring a more straightforward interpretation to the uncertainty around the cost-effectiveness estimator compared, for example, with confidence intervals around the ICER, they have been criticised on a number of grounds. Koerkamp *et al.* (59) pointed out that CEACs are insensitive to changes in the joint distribution of costs and effects differences, thereby masking potentially significant differences or exaggerating existing differences. Barton *et al.* (60) made a compelling case for not relying solely on CEACs when recommending the cost-effective option from a panel of interventions and advocate the mandatory representation of the CEAF as well. Jakubczyk and Kaminski (61) demonstrated that the properties of the CEAC are strongly influenced by factors such as the skewness of the NB estimator and correlation between ΔC and ΔE, and advise their use only for illustration purposes.

There are two principal types of economic evaluation: trial-based and model-based evaluations. The former entails collecting individual patient data on costs and outcomes alongside a RCT which compares two or more alternatives (62-64). This is often done using case report forms which record resource utilisation and outcome information (such as

HRQoL) for every enrolled patient. The quantities of interest for economic evaluation are the differences in mean cost and effect between the trial arms. Cost analysis may be particularly challenging because of inherent right skewness of cost data, potential difficulties in identifying the unit costs (as opposed to prices of health care provider charges) and censoring (missing data due to inappropriate data collection processes, patient drop-out or other reasons) (65).

Model-based economic evaluations predict under uncertainty the costs and outcomes associated with each alternative by means of a decision-analytic model, which "*uses mathematical relationships to define a series of possible consequences that would flow from a set of alternative options being evaluated*" (66, p.6). The key conceptual elements of a decision model are its structure and data inputs. Choosing the appropriate model type and associated structure are of utmost importance; to that end, categorisations and decision charts have been proposed to guide researchers (67, 68). Probabilities and expected values are the fundamental types of data inputs. Probabilities can be thought of as the likelihood of each possible consequence to occur; in a clinical setting they reflect the fact that clinically identical patients who are subject to the same intervention may respond differently. In relation to CUA, expected values concern the costs and outcomes (QALYs) associated with each alternative: these are calculated as the sum of costs and outcomes, respectively, of each possible consequence, weighted by the probability of each consequence.

Trial-based economic evaluations are now common, but several important shortcomings have been highlighted: evaluating a limited number of relevant interventions, providing information on restricted patient sub-groups and for a limited time-horizon (69). As such, the optimal approach to generating cost-effectiveness evidence entails multiple cycles of decision modelling and primary data analyses.

1.3. Generalisability of trials and trial-based economic evaluations

The generalisability of trial findings is a legitimate concern, both for clinical and economic outcomes. This section provides an overview of generalisability issues for RCTs, followed by an in-depth look at the generalisability of trial-based economic evaluations. A number of gaps in the current body of knowledge are identified and discussed. The section is informed by a pragmatic search of the relevant literature.

1.3.1. Clinical trials and external validity

RCTs have been considered the gold standard research design because of their potential to offer unbiased estimates of interventions' effectiveness. The strength of the RCT rests on three fundamental features: comparability of effects (through a placebo or control arm); comparability of populations in trial arms (through randomisation); and comparability of information (through blinding) (70). The extent to which a trial's results can be trusted is reflected in the study's quality. Quality itself is a complex, multidimensional concept which integrates elements of design, conduct, statistical analysis and reporting (71, 72). A definition of trial quality was proposed by Verhagen *et al.* (72) as a result of their Delphi study (p.1239): *"Quality is a set of parameters in the design and conduct of a study that reflects the validity of the outcome, related to the external and internal validity and the statistical model used"*. Validity is, thus, recognised as an important and conceptually rich dimension of quality. Moreover, validity is a fundamental pre-requisite for ethical and valuable research (73). A further distinction between internal and external validity was proposed by Campbell in 1957 (74); although it originated in psychology, this dichotomisation was adopted in social sciences and experimental design in general.

1.3.1.1 Internal validity

Internal validity refers to whether the results of the study are correct for the original study population. A study has internal validity when there are no suspicions that the differences in outcomes between the patient groups are due to other factors apart from chance and the intervention(s) that were administered. By contrast, external validity refers to whether the results of the study are applicable to other circumstances, such as a given patient population, a particular health care organisation or a geographical setting. External validity as a concept is meaningless without specifying the descriptive parameters of the setting where results are to be applied. Furthermore, internal validity is a pre-requisite of external validity, as misleading results cannot form a reliable basis for any further generalisation. This reality has been acknowledged by Campbell himself (p.310): "*If one is in a situation where either internal validity or representativeness must be sacrificed, which should it be? The answer is clear. Internal validity is the prior and indispensable consideration*" (74).

Assessing internal validity involves identifying the extent to which a study is vulnerable to a range of sources of bias. Bias is understood here as a systematic error in results or inferences. A formal definition of bias has been proposed by Murphy (p.345): "*any process at any stage of inference tending to produce results that differ systematically from the true values*" (75). Detailed lists of possible biases that can occur in experimental research have been proposed, for example by Murphy (75) and Sackett (76) in the late 1970s. The Cochrane Collaboration currently distinguishes between several major types of bias in relation to RCTs (7): selection bias – systematic differences between the patient groups being compared; performance bias – differential exposure in health care provision or other treatment outside the intervention under scrutiny; attrition bias – systematic variation in withdrawals or exclusions; detection bias – systematic differences in outcome assessment; and reporting bias

– preferential reporting of study's findings. It must be acknowledged that a methodological flaw falling into one of the categories outlined above may or may not actually introduce bias, therefore the term 'risk of bias' is more appropriate. The methods and procedures to avoid or minimise each of these biases have been extensively addressed in the literature (7, 71).

1.3.1.2 External validity

The conceptual content of 'external validity' in the context of RCTs is extremely rich, which explains the heterogeneity of its accounts. For example, Dekkers *et al.* (77) suggested a checklist of 11 individual items, grouped in four domains: eligibility criteria for participants and centres; temporal, ethnical, socio-economic and geographical aspects; patient characteristics going beyond eligibility criteria, such as age and comorbidities; and the applicability of study results. A comprehensive account was proposed by Rothwell (78), who indicated 39 relevant issues that should be considered and reported, grouped under six categories: the setting of the trial; patient selection; characteristics of randomised patients; differences between trial protocol and routine practice; outcome measures and follow-up; and adverse effects of treatment. Other checklists or frameworks for assessing external validity are also available (79, 80). The distinction between the study population and the population from which it has been sampled and is thought to represent (the target population) has often been the focus of generalisability research in trials (81). The example checklists cited above, however, show that there is more to context than the patient population: for example, the type of health care setting and the nature of clinical protocols are also important.

Enhancing the external validity of trials involves creating an experimental environment which is as close as possible to real-life settings i.e. pragmatic trials (82), for example by relaxing the inclusion/exclusion criteria, selecting a representative sample of clinicians and centres, devising protocols that are in accordance with clinical practice and evaluating relevant and meaningful outcomes. It must be acknowledged that it is challenging for a given RCT to produce results that are widely generalisable. Nevertheless, honest, transparent and detailed reporting of the trial's conduct would allow the readers to make their own opinion as to the findings' generalisability.

1.3.1.3 Enhancing the external validity of RCTs

The RCT as a research design is particularly valued in the scientific community for high internal validity, in other words for the potential to offer unbiased results. However, trials' potential for external validity has often been questioned (78, 83-85). Indeed, RCTs feature several strong limitations: they usually evaluate specific interventions one at a time, thus leaving potentially important questions unanswered; they focus on optimizing the conditions for obtaining a positive finding by minimising heterogeneity, for example by adhering to strict clinical protocols or over-selecting patients; and are bounded by logistical, financial and ethical constraints in choosing the questions they can answer (86). These limitations, especially the drive for positive findings, hinder the applicability of trial findings to real world practice. A wealth of empirical evidence supports this claim. Studies across a wide range of therapeutic areas have suggested that trial participants are often unrepresentative of the target population (87-94), which can introduce bias in the measures of effect (95). For example, Steg *et al.* reported that eligible patients with acute myocardial infarction enrolled in RCTs had lower baseline risk and lower mortality than non-enrolled ones (96).

The choice of participating centres can also influence the generalisability of trial results (78), especially in non-pharmacologic trials, as outcomes may be affected by factors like hospital volume (97) and practitioners' expertise (98). For example, the systematic review of Halm *et al.* (97) found that patients treated in higher volume hospitals have better clinical outcomes across a wide range of therapeutic areas. In surgical RCTs, restricting participation to centres where surgeons have a proven record of success may lead to results which depart greatly from real-life estimates (78). Practice guidelines can also differ from one hospital to another. For instance, negative pressure wound therapy (NPWT) is a technology currently used in the UK for the open abdomen at the discretion of UK National Health Service (NHS) trusts in the absence of a nationwide recommendation towards its implementation (99). Limited evidence suggests that RCTs are predominantly carried out in university and teaching centres, while non-teaching centres are somewhat better represented in non-randomised studies (100). The influence of centre-specific characteristics on treatment outcomes has been equally recognized in observational research (101).

Two types of strategies are available for enhancing the generalisability of clinical trials. One of them is, obviously, conducting RCTs which emulate closely ‘real’ clinical practice. This approach is based on a more than 40-year old conceptual distinction between explanatory and pragmatic clinical trials (102). Explanatory trials are usually conducted in tightly controlled, ‘laboratory’ conditions, with the aim of answering a scientific question. On the other hand, pragmatic trials would be run in ‘normal’ conditions in order to answer an applicability question, such as a policy decision. In accordance with the latter approach, pragmatic or practical clinical trial designs have been proposed (103-105) so as to maximise the value of trial findings to decision makers. The distinction between explanatory and pragmatic trial designs has been commented in more detail by MacPherson (106) and Treweek and Zwarenstein (82). Recommendations include comparing clinically relevant alternatives (placebo-controlled trials often have little relevance when alternative interventions are already available), enrolling a diverse study population, recruiting from a variety of settings and measuring a broad range of relevant outcomes. The issue of relevant outcomes is particularly important in at least two aspects: first, outcomes beyond health must also be considered, such as economic and quality of life consequences, as more and more decision makers include such considerations in their decisions (3). Second, generalisability is also linked to between-study comparability, therefore the need for trials to report a core, common outcome set for evidence synthesis purposes is more stringent than ever (4, 107).

Pragmatic clinical trials may appear to solve most of the problems associated with external validity, but Karanicolas *et al.* (108) pointed out that even pragmatism can be evaluated from at least three relevant standpoints: the policymaker, the clinician and the patient. The corollary is that there is no 'one size fits all' pragmatic approach, which can only underline even more RCT's major limitation of not being able to answer multiple questions at once. The PRECIS tool, developed by Thorpe *et al.* (109), introduced a summary measure of the pragmatic-explanatory continuum in order to assist both researchers who design trials and those who assess trials. While noting that 'pragmatism' has become more and more fashionable in research during the past two decades, Kent and Kitsios (110) warned against over-reliance on the results of pragmatic trials by pointing out that generalising the findings of an over-inclusive experiment may be equally as (or even more) flawed as doing the same with a severely restrictive one.

The second strategy involves stimulating the complete and transparent reporting of trial conduct and results in order to allow readers to make their own judgement on the general quality and, specifically in this case, the external validity of trial results. The CONSORT statement (27) provides a minimum set of recommendations for trialists in that respect. Initially developed for parallel group RCTs, further CONSORT extensions have become available, for example for pragmatic trials (29) and PRO data collected alongside RCTs (30). Since the introduction of the CONSORT statement in 1994 (111), cumulative evidence has suggested that RCT reporting has improved, but remains suboptimal (112-115).

Generalisability is a stand-alone item in the CONSORT 2010 checklist, where it is included as '*Generalisability (external validity) of trial findings*' and invites discussion on how the trial's results can be interpreted in light of the participants, setting, interventions and outcomes. The CONSORT extension for pragmatic trials (29) is more specific in guiding the discussion of contextual effects, as it requires to "*describe key aspects of the setting which determined the trial results. Discuss possible differences in other settings where clinical traditions, health service organisation, staffing, or resources may vary from those of the trial*" (p.6). Similarly, the CONSORT extension to RCTs of non-pharmacologic treatments (116) requires discussing generalisability in relation to the care providers and centres involved in the trial: *"Generalizability (external validity) of the trial findings according to the intervention, comparators, patients, and care providers and centers involved in the trial"* (p.W-63). The CONSORT extension to patient-reported outcomes (30) also refers to participating centres in the explanation of the generalisability item (p.820): "*In addition to the design and conduct issues relevant to the generalizability of the RCT overall, several PRO–specific limitations (including both patient- and center-level characteristics) may affect generalizability of the PRO results*".

However, it is often difficult to ascertain the generalisability of RCT results since reporting external validity in trial publications remains poor (117-119). One potential reason for this state of affairs is the focus of most guidelines and textbooks on internal rather than external validity (120, 121). For instance, the CONSORT Statement has only one item explicitly addressing generalisability out of 25 in total. This focus on internal validity has also been recognised by journal editors and the emerging picture is that more effort needs to go into improving external validity (122). In addition, the lack of specific information in trial reports has been identified as restricting applicability and interpretation (104, 123). Transparency is key not only to make an informed qualitative judgement on the transferability of trial findings to other settings, but analytical methods are now available to allow a quantitative adjustment of trial results to an appropriately specified target population (124).

1.3.2. Generalisability of trial-based economic evaluations

Trials consume enormous amounts of resources and have become increasingly expensive to run (125, 126). Reasons for this include: longer follow-up periods, increasing regulatory requirements and the need for ever larger sample sizes as the therapeutic benefit of new technologies are more and more marginal. Furthermore, the ethics of reproducing research is at least questionable. Under these auspices, the pressure to maximise the output of every research endeavour has increased continually and RCTs often recruit across jurisdictions (in this context, 'jurisdiction' refers to an administrative space where a medical intervention will be implemented e.g. a health care system, a local authority). For example, multinational trials recruit internationally in order to achieve the required sample size, to demonstrate that the clinical findings can be extrapolated to other populations or, in the case of industry-driven studies, to obtain the data required by local regulatory authorities for market authorisation purposes (127).

Ensuring the generalisability of RCT results may be particularly challenging for economic outcomes, which inform health policy decisions. This is because the relative clinical effect of an intervention has been historically assumed constant across settings, albeit not without challenges (78, 128, 129); however, this assumption may not hold for economic outcomes. Therefore economic evaluation results should not routinely be assumed to be completely transferable between jurisdictions.

In one of the earliest pieces of research on this topic, Drummond *et al.* (130) compared a decision model between four countries while allowing for slight adjustments of the model as well as for local input data. Their conclusion was that cost-effectiveness results were significantly different between countries; and the main drivers of variability were the cost variations and the patterns of care. O'Brien (131) further strengthened this case by identifying six generic 'threats to transferability' in economic evaluation studies: demography and epidemiology of disease; clinical practice and conventions; incentives and regulations for health care providers; relative price levels; consumer preferences; and the opportunity cost of resources. These issues are equally applicable to decision modelling and trial-based economic evaluations and, furthermore, to "*all levels of geographical grouping*" (p. S39). As a result, it is perfectly possible for the same medical technology to be cost-effective in one setting and cost-ineffective in another. Such a reality is likely to be of utmost concern for decision makers, who are interested in knowing whether results collected in other jurisdictions can inform decisions in their own. The focus group study of Hoffmann *et al.* (6) pointed out, indeed, that the generalisability of economic evaluation findings are of great interest UK policy makers. The major emergent issue was that economic evaluations too often ask narrowly focused questions which do not allow the portability of their results to other contexts.

Before going any further, a terminology note should be made. 'Generalisability', 'transferability', 'portability' have all been used to describe the extent to which economic evaluation results are applicable from one geographical setting to another. Boulenger *et al.* (132) suggested that 'transferability' may be a broader concept than 'generalisability' as it encapsulates both the intrinsic value of the results and the methods available to assess their applicability in various settings. Barbieri *et al.* (p.1028) defined *generalisability* and *transferability* as follows: "*Studies may be considered generalisable if they can be applied to a range of jurisdictions without any adjustment needed for interpretation. In addition, some studies may be transferable if they can be adapted to apply to other settings*" (133). This interpretation has been endorsed by the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Good Research Practice Task Force (134). No formal distinction will be made between these terms throughout the thesis for simplicity. Two principal research directions have been intensively explored in the generalisability literature in relation to economic evaluations: the factors which influence transferability; and the methods that address the transferability of cost-effectiveness results.

1.3.2.1 Factors influencing transferability

The factors linked with transferability of economic evaluations received close scrutiny in the literature and this sub-section gives an overview of the nature and content of these factors by drawing on several comprehensive papers which investigated them in detail. A summary of the most relevant factors, as identified in the literature, in presented in Table 7.1. Welte *et al.* (135) published in 2004 a systematic review of 44 studies which aimed to identify potential transferability factors i.e. any parameter which may influence economic evaluation results and may differ between countries. They identified 14 factors, grouped under three broad categories: methodological characteristics of the economic evaluation; health care system characteristics; and patient characteristics. In addition, the authors also made a judgement on the effort required to check for each factor the correspondence between the study and the decision country; eight out of 14 factors were rated as requiring 'medium-very high' such effort.

Table 1.1 Factors which influence the generalisability of economic evaluation results

Study	Factors
Welte *et al.*(135)	**Methodological characteristics**: Perspective; Discount rate; Medical cost approach; Productivity cost approach. **Health care system characteristics**: Absolute and relative prices in health care; Practice variation; Technology availability. **Population characteristics**: Disease incidence/prevalence; Case-mix; Life expectancy; Health-status preference; Acceptance, compliance, incentives to patients; Productivity and work-loss time; Disease spread.
Sculpher *et al.*(127)	**Patient factors**: demographics; epidemiology; case-mix; baseline risk; compliance. **Clinician factors**: skill/experience; practice style; incentives. **Health care system factors:** absolute/relative prices; exchange rates; clinical practice; resource utilisation; historical differences. **Wider socio-economic factors**: cultural attitudes; health-state preferences.
Goeree *et al.*(136)	**Patient characteristics**: demographics, education, socio-economic status; risk factors, medical history, genetic factors; lifestyle, environmental factors; mortality rates, life expectancy; attitudes toward treatment, culture, religion, hygiene, nutrition; compliance and adherence rates, ethical standards; population values (utilities); population density, immigration, emigration, travelling patterns; income, employment rates, productivity, work loss time, friction time; type of insurance coverage, user fees, co-payments, deductibles; incentives for patients. **Disease characteristics**: epidemiology; disease severity, case mix; disease interaction, co-morbidity, concurrent medications; mortality due to disease. **Provider characteristics**: clinical practice, conventions, guidelines, norms; experience, education, training, skills, learning curve position; quality of care provided; method of remuneration (supplier-induced demand); patient identification; cultural attitudes; incentives for providers, liability. **Health care system characteristics**: absolute or relative prices; available resources, programs, services; organization of delivery system, structure, level of competition; level of technology advancement, innovation and availability; available treatment options; capacity utilization, economies of scale, technical efficiency; input mix, specialization of labor, joint production; access to programs and services, gatekeepers, historical differences; waiting lists, referral patterns; regulatory and organizational infrastructure, licensing of products; availability of generics or substitutes; market forms of suppliers, payment of suppliers, supplier incentives; incentives for institutions. **Methodological characteristics**: costing methodology, estimation procedures; study perspective; study factors; timing of the economic evaluation; clinical endpoints/outcome measures; discount rates; exchange rates, purchasing power parities; opportunity cost; affordability.

In a similar but more targeted exercise published in 2005, Barbieri *et al.* (137) systematically reviewed European-wide cross-country comparisons of economic evaluations (both model- and trial-based) of pharmaceuticals in order to identify the factors which account for variations of cost-effectiveness results between countries. They included 46 inter-country comparisons and concluded that differences were not likely to be systematic in the sense there was no stable enough pattern to infer that if a given intervention was found cost-effective in country A it would automatically be, say, more cost-effective in country B. The principal finding of the study was that resource use (when it was allowed to vary[1]) and the local willingness to pay threshold were the main determinants of variation.

Sculpher *et al.* (127) published in 2004 a comprehensive account of the determinants of generalisability of economic evaluations in health care. They undertook a series of systematic reviews to identify, on one hand, the factors associated with variability in economic evaluations and, on the other hand, the methods used to assess variability and enhance generalisability. Their review was very broad as it referred to both trial-based and model-based economic evaluations and, furthermore, to variability across locations and time. For their systematic review on factors influencing generalisability, the authors reviewed 36 conceptual papers and identified 26 factors affecting the geographical variability of economic evaluation results, grouped under four categories: patient factors; clinician factors; health care system factors; and wider socio-economic factors. The authors highlighted that, at the time, generalisability appeared to be a particularly relevant issue in a multinational context as most of the included studies investigated cross-national comparisons, with only two studies (from the UK and US, respectively) looking at within-country variations.

[1] Specifically in trial-based economic evaluations where the analyst does not decide to pool resource use data across countries. The authors' opinion was that resource use should not be pooled whenever possible.

In their discussion of the potential effect of local factors on cost-effectiveness results, Sculpher *et al.* referred explicitly to several issues: first, they acknowledged the difference between centres which usually participate in clinical trials and those who do not, further implying that differences in outcomes between intervention and controls may not be transferable across these types of settings; second, they acknowledged the 'clinician effect', whereby the training, experience and habits of health care professionals are an integral part of the intervention that is being delivered and, therefore, the resulting costs and patient outcomes. As a result, local variations may have obvious implications on cost-effectiveness. Third, and in relation to the previous issue, the incentives that health care staff have across locations, such as payment and reimbursement schemes, also impact performance. Even more importantly, different centres are exposed to different population profiles (reflected by demographic characteristics such as age, sex and ethnicity), epidemiological profiles (reflected by different disease burdens) and, subsequently, to differences in patient case-mix. The latter are also correlated with institutional factors, as well - for example, teaching hospitals tend to see the more complicated and thus more resource intensive cases, but also offer better care than nonteaching hospitals (138).

Goeree *et al.* (139) published in 2007 the results of a systematic review where they looked at 102 papers (conceptual, empirical and review articles) and derived no less than 77 factors affecting transferability, grouped into five categories inspired by the earlier review of Welte *et al.* (135), as characteristics of the: patients; diseases; providers; health care systems; and methodologies. It has to be acknowledged that this review had a strong focus on international comparisons. The provider-specific factors identified in the review were similar to the ones pointed out by Sculpher *et al.* (127): clinical practice; staff experience and skills; the quality of care provided; method of remuneration; cultural attitudes; and provider incentives. Most importantly, the authors' final conclusive point drew attention to the importance of directing research efforts towards quantifying the relative impact of these factors when transferring economic evaluation data.

While the demographical, epidemiological and health care system characteristics can often be considered fixed in a particular context, the centre-specific variability factors are arguably of more interest to researchers and decision makers because they are (at least at an intuitive level) the most readily amenable to change. The selection of centres and health care professionals for inclusion in trials to ensure generalisability has received relatively little attention (78), but the limited available evidence suggests that most evaluative research takes place in university hospitals, while ‘common centres’ are slightly better represented in observational studies (100).

Unit costs are also expected to vary across locations and Sculpher *et al.* (127) pointed out that such variations exist not only in between-country comparisons, but also in within-country ones. In the context of trial-based economic evaluations, the issue thus becomes the suitability of using average unit costs across all the centres in the trial: in the absence of unit costs missing completely at random (MCAR), the average unit cost will most likely misrepresent the centre-specific cost (140, 141). This is a legitimate concern: a systematic review of economic evaluations conducted alongside trials funded by the UK Health Technology Assessment Programme revealed that only 52 of 95 reviewed studies used unit costs that were sourced locally (142). Of course, it may be impractical or even impossible to collect unit costs from all centres involved in a study, so a number of alternative solutions have been suggested: for example, Goeree *et al.* (23) reported a framework allowing the selection of the number of hospitals (one or more) from which unit costs should be used to perform economic evaluation calculations across multicentre economic evaluations; and

Grieve *et al.* (140) used multiple imputation to account for missing centre-specific unit cost data. However, in the UK the use of nationally averaged ('off the shelf') unit costs is accepted for the reference case of an economic evaluation (39).

Another relevant local factor relates to the quality of care provided. The differences in quality of care between providers have been documented thoroughly at multiple levels. For example, extensive literature reviews have shown that larger health care providers (both hospitals and physicians) seem to be associated with better outcomes (97) and that teaching hospitals are generally associated with superior health outcomes when compared with non-teaching hospitals (138, 143). In the case of the UK, there is evidence of variation in quality of care across settings both in primary care (144) and hospital care (145).

The empirical evidence on the variations of patient preferences across settings is more controversial. This has relevance for the results of cost-utility analyses i.e. cost per QALY: theoretically, if patients in different settings value the same health states differently, the results of an economic evaluation will subsequently vary irrespective of other contextual factors. For example, several national tariffs as well as a European tariff are available for the EQ-5D instrument (146) and there is evidence that valuations differ substantially between countries, mostly due to methodological differences in elicitation and cultural attitudes (147). However, the evidence on whether such variations affect the economic evaluation results is scarce and inconclusive: several studies have shown that using different tariffs to calculate QALYs has little impact on the overall cost-effectiveness findings (148, 149), while others have suggested that these differences may be relevant (150).

In summary, the current knowledge on factors influencing the generalisability of economic evaluations depicts a complex picture. First, while a plethora of potentially relevant factors have been proposed, accounting for the majority of them may be challenging in practice. Second, not all of these factors can be measured straightforwardly e.g. the impact of clinician incentives. Third, the impact of these factors on economic evaluation results is difficult to measure. Finally, even if the impact can be measured, empirical results on the magnitude and direction of these effects have not always been consistent. Some factors are less prone to such difficulties than others, such as the cost of the intervention itself, although even in this case the cost assumed in a research environment can be different from the one that the manufacturer (e.g. the drug company) may eventually agree with local decision makers.

1.3.2.2 Methods addressing transferability

Two broad categories of methods have been suggested to aid decision makers in addressing the transferability of economic evaluation results. The first category refers to methods which aim to assess the extent to which the results of economic evaluation studies as a whole can be transferred across settings. The second category includes methods that address transferability involves *adjusting* the results of an economic evaluation to obtain local cost-effectiveness estimates.

Methods which assess the transferability of economic evaluations

Such methods primarily target decision makers and aim to assist them in evaluating the extent to which the results of an economic evaluation conducted elsewhere are applicable in their own setting. A recent synthesis of these methods was given by Goeree *et al.* (139), whose systematic review of transferability approaches identified seven strategies: five aimed to offer a qualitative verdict on transferability and two proposed indices to quantify it. The main characteristics of the identified approaches are presented in Table 7.2.

Table 1.2 Approaches towards the generalisability of economic evaluation studies

Study	Type of approach	Preliminary criteria	Generalisability factors to be considered	Comments
Heyland *et al.*(151)	Checklist	Comprehensive description of competing alternatives; Sufficient evidence of clinical effectiveness or, as second best, clinical efficacy; Important costs were identified, measured and valuated appropriately; Appropriate sensitivity analysis taking into account all estimates of uncertainty.	Patient characteristics. Perspective of the analysis Intervention Costing methods Outcomes Discount rate	No clear decision rule Method piloted on 29 Canadian economic evaluations in the field of critical care, out of which four got past the first stage. Overall generalisability verdict interpretable.
Späth *et al.*(152)	Checklist	The study perspective is clear Two or more competing options are compared The evaluated therapies are described The therapies are applicable in the local setting	Perspective of the analysis Patient characteristics Health outcome data Resource utilisation Unit prices and discount rates	Decision rule: a study must comply with all generalisability criteria in order to be considered transferable. Method piloted on 26 economic evaluations (in the area of breast cancer) for transfer to the French health care system. Six studies met the methodological criteria, but none was judged to be transferable mainly due to insufficient reporting of resource use and unit prices.
Welte *et al.*(135)	Transferability chart	Relevant technology is relevant to local setting Comparator is relevant to local setting Study has acceptable quality	Methodological characteristics: (4 factors) Health care system characteristics (3 factors) Population characteristics (7 factors)	Decision chart guides the reader towards a generalisability assessment Method piloted on three case studies

Study	Type of approach	Preliminary criteria	Generalisability factors to be considered	Comments
Boulenger *et al.*(132)	Checklist and quantitative transferability index (0% not transferable to 100% completely transferable)	None (see Comments)	Intervention and comparator Countries Perspective Study population (2 factors) Effectiveness (2 factors) Benefit Costs (5 factors) Sensitivity analyses	Methodological quality of each study is assessed as an integral part of the evaluation, not a pre-requisite. Piloted on 25 economic evaluations, average transferability index 68.8%.
Drummond *et al.*(134)	Four-step application algorithm	Relevant technology is relevant to local setting Comparator is relevant to local setting Study has acceptable quality	Methodological characteristics: (4 factors) Health care system characteristics (3 factors) Population characteristics (7 factors)	Based on the Welte *et al.* criteria, the authors discuss practical approaches to adapting cost-effectiveness results to local settings.
Chase *et al.*(153)	HTA adaptation toolkit	Relevant policy and research questions Translation is possible Technology is described Scope is specified Report is peer-reviewed Conflict of interest Report is not outdated Methods are accurately described	Perspective; Preferences; Relative costs; Indirect costs; Discount rate; Technological context; Personnel characteristics; Epidemiological context; Factors that influence incidence and prevalence; Demographic context; Life expectancy; Reproduction; Pre- and post-intervention care; Integration of technology in health-care system; Incentives	Very comprehensive checklist, aimed at HTA reports. Checklist generated as part of a wide European consensus involving 28 HTA agencies. Out of five domains, one refers to economic evaluation: 26 questions in total, out of which 3 refer to transferability.

Study	Type of approach	Preliminary criteria	Generalisability factors to be considered	Comments
Antonanzas *et al.*(154)	Transferability index (0 not transferable to 1 completely transferable)	The relevant parameters needed to estimate cost-effectiveness are given in the study The quality of the study is acceptable	Perspective; Intervention and comparator; Clinical practice; Life expectancy; Health status preferences; Productivity measures; Epidemiology; Discount rate; Costs and health effects.	Global Transferability Index (IT) results from aggregating a general transferability index (IT1) and a specific transferability index (IT2) Method piloted on 27 economic evaluations on infectious diseases conducted in Spain, obtaining IT in the range 0.534 to 0.543, denoting low to moderate transferability.

Five of the seven strategies developed checklists or guidelines to inform a qualitative judgement of the extent to which the results of a given study are transferable (115, 134, 135, 151, 152). Most strategies comprised two steps: the first step was a methodological assessment of the study; if judged appropriate, an in-depth assessment of transferability then ensued. The criteria for the preliminary methodological assessment are largely similar across the checklists, but vary in focus: for example, Heyland *et al.* (151) emphasised the validity and quality of reporting by requiring a comprehensive description of the alternatives under scrutiny, evidence of effectiveness and efficacy, appropriate costing and appropriate sensitivity analyses. On the other hand, Welte *et al.* (135) included relevance as well by requiring the relevant technology and comparator to be comparable to the one that will be used in the decision country.

In terms of the generalisability assessment itself, a wide range of criteria were proposed. Heyland *et al.* (151) proposed a list of ten questions related to clinical and system generalisability. Späth *et al.* (152) suggested an assessment against five indicators, namely: potential users of the economic evaluation, characteristics of the patient population in the 'receiving' setting, the transferability of outcome data, the transferability of resource use and the transferability of unit prices. Welte *et al.* (135) described 14 specific knock-out criteria and suggested a flowchart along which the user is guided either towards a clear transferability verdict i.e. 'study results full/qualitatively transferable' or towards an assessment of whether modelling adjustments are needed and how they can be made, followed by a similar transferability verdict. Drummond *et al.* (134) also produced a decision chart which guides the reader through an assessment of transferability, with or without adjustment for 'specific knock-out criteria' such as unit costs, discount rate, time horizon and perspective. Finally, Chase *et al.*[2] (153) developed a comprehensive toolkit to focus the appraisers' efforts in extracting the relevant information from health technology assessments reports conducted elsewhere and making an informed judgement on the transferability to their own setting. The actual economic evaluation component of the toolkit contains 26 questions that assess relevance and reliability and three questions addressing transferability.

Two of the seven strategies used the generalisability criteria published in previous checklists to calculate numerical indices quantifying the measure of transferability (132, 154). Boulenger *et al.* (132) proposed their own checklist of relevant criteria and used it to construct a study-level transferability index: a score was assigned to each item in the checklist (1 for 'yes', 0.5 for 'partially' and 0 for 'no/no information') and an overall score was obtained by summation and then division by the maximum number of points, thereby obtaining the transferability index as a percentage. When they piloted it on a sample of 25 economic evaluations to assess the transferability of results between the UK and France, the authors found a mean transferability index of 66.9% for the entire checklist and 68.8% for the transferability sub-checklist. Antonanzas *et al.* (154) proposed a general index (IT_1) and a specific index (IT_2) applicable to economic evaluation studies. IT_1 assesses two critical and 16 non-critical *objective* factors in order to produce an index that evaluates the methodological quality of the economic evaluation. IT_2 assesses four critical and eight non-critical *subjective* factors to evaluate the extent to which a study is transferable to a different setting. For each of the factors, a score of 1 is given if the factor is completely addressed, 0.5 if partially addressed and 0 if not addressed at all. Ultimately, IT_1 and IT_2 are combined in a global transferability index using a number of alternative formulae such that a maximum value of 1

[2] In the Goeree *et al.* systematic review, this document is cited as 'Turner *et al.'* A more comprehensive publication of the same project and with the same authors has become available in the meantime with Chase D as the first author, therefore it is referred to and referenced in this Chapter as Chase *et al.* The content of the toolkit is identical in Turner *et al.* and Chase *et al.*

denotes a completely transferable study and 0 denotes non-transferability or insufficient information to make such a judgement. The authors tested their method by evaluating the transferability of 27 economic evaluations on infectious diseases conducted in Spain and found a mean value of the index in the range 0.534 to 0.543, denoting low to moderate transferability.

In summary, the methods proposed to assess the transferability across settings of economic evaluation results share a number of fundamental characteristics: first, they are predicated on the necessity of evaluating internal validity as a pre-requisite for external validity. All the proposed methods start with a preliminary phase where the methodological quality of the study is assessed; if deemed acceptable, a thorough investigation of transferability then becomes appropriate. Second, they recognise the difficulty of accounting for the plethora of factors that are thought to be relevant for the generalisability of economic evaluation results and attempt to integrate them in meaningful tools aimed at facilitating the decision-making process. Through the use of flowcharts, algorithms and scores, the reader (e.g. decision maker) is guided towards a rational and informed decision. Finally, it is acknowledged that transferability is a matter of judgement. Most of the reviewed methods offered clear-cut verdicts (e.g. findings are transferable/not transferable) only in the extreme cases where either all the information is available and appropriate or essential information is missing. In real-life policy making, most situations are likely to be mapped somewhere between these two extremes, where the decision becomes much more nuanced. Two of the seven proposed methods (132, 154) attempted to quantify transferability using indices, but no meaningful cut-off points were suggested.

Methods which adjust economic evaluation findings

This category includes analytical methods, aimed primarily at researchers, which can be used either independently or in the final stages of the previous category of methods. A distinction can be made between decision-modelling studies and trial-based economic evaluations: while adapting the results of a decision model to a local context usually involved populating the model with local input data (with or without adapting the model structure to reflect the local clinical pathways), the methods that obtain local adjustments based on individual patient data (i.e. from trials) are more complex. Manca *et al.* (155) conducted a comprehensive critical review of the proposed methods in the context of multinational RCTs and identified three broad categories of approaches.

The first type of approach uses tests for heterogeneity to *establish* whether the cost, effectiveness and cost-effectiveness results of multinational RCTs can be pooled or should undergo a stratified analysis. Cook and colleagues (156) proposed this approach and, citing the work of Gail and Simon (157), distinguished between qualitative interactions i.e. the treatment effect is positive in some countries and negative in others, and quantitative interactions i.e. only the magnitude of the treatment effect, but not its direction, varies across countries. The authors used a five-country RCT as a case study, calculated country-specific measures of effectiveness (mortality and hospitalization rate) and cost-effectiveness (incremental cost-effectiveness ratio and net monetary benefit) based solely on patients recruited from those countries, and then applied statistical tests for qualitative (157) and quantitative interactions (157, 158). This approach has several limitations: heterogeneity tests are often underpowered (159, 160); non-statistically significant differences may mask different cost-effectiveness recommendations; and, most importantly, this method cannot offer context-specific estimates.

The second type of approach aims to *estimate* local (country-specific) cost-effectiveness results without accounting for the hierarchical structure of the trial data i.e. by using centre characteristics and a centre-level dummy as regressors. The method involves applying a simple regression model of costs and outcomes against a number of patient-level and centre-level variables, as well as a centre dummy variable; the parameter of interest is thus the treatment coefficient estimate. Coyle and Drummond (161) applied this approach by using simple ordinary least squares regression to explain cost variation using data from two UK RCTs investigating interventions for head and neck cancer patients. However, such a framework does not incorporate the correlation between costs and outcomes (162) and Willan *et al.* (163) later addressed this limitation by regressing costs and effects simultaneously using seemingly unrelated regressions.

Finally, the third type of approach *estimates* local (country-specific) cost-effectiveness results while accounting for the hierarchical structure of the data. The key concept here is to account for the fact that individual patients are sampled within higher-level units (such as centres or countries) and thus individual effects are assumed to be drawn randomly from a distribution of higher-level effects. The advantages of using such a multilevel structure are clear: the correlation between individuals and countries can be modelled explicitly; and the analysis uses all the information in the trial as opposed to information only from a country-specific subset. As a result, adjusted country-specific cost-effectiveness estimates can be obtained. The application of multilevel modelling in the cost-effectiveness analysis of multinational trials was introduced by Manca *et al.* (164), who regressed net-benefits and then estimated centre-specific values using a fully Bayesian procedure (Markov chain Monte Carlo shrinkage estimation). Of note, regression on net benefits had been proposed earlier by Hoch *et al.* (165). Further developments of the method allowed for the correlation between costs and outcomes to be explicitly modelled using bivariate hierarchical modelling (166, 167), which is currently the method recommended by ISPOR in conducting cost-effectiveness analyses of multinational trials (134).

In summary, the third type of approach to generalisability uses regression methods to account for individual and centre-specific variables in order to obtain appropriately adjusted local cost-effectiveness estimates. The methods have varied in complexity as they incorporated the multi-level structure of the data and they allowed modelling the correlation between costs and outcomes. While bivariate hierarchical modelling is the current norm, it relies on the fundamental assumption of *exchangeability* (168, 169): given a collection of independent and identically distributed random variables, the property of exchangeability means that the joint distribution of the variables is symmetric or, equivalently, that the joint distribution of any permutation of the variables remains constant. In the context of multicentre RCT analyses, this means that prior to examining the data there is no reason to expect differences between the outcomes of interest at centre- (or country-) level. For example, in the simple case of a two-centre RCT, exchangeability holds if the probability of observing an incremental cost below £300 in centre A *and* below £500 in centre B is identical to the probability of observing an incremental cost below £500 in centre A *and* below £300 in centre B.

Obviously, systematic variations between centres/countries do exist in practice; if they could be completely explained by several factors (such as the health expenditure per capita for a given country or the proportion of qualified staff employed in each hospital), the assumption of exchangeability would still hold as *conditional exchangeability* - which is to say that exchangeability applies for a given set of values of the identified systemic factors. However, Manca *et al.* acknowledged that, in practice, it is rarely known whether the centres (countries) included in RCTs comply with the assumption of exchangeability. They reiterated the earlier appeal of Drummond *et al.* (134) to ensure that the sample of centres (countries) included in the RCT satisfies this assumption and that sufficient country- and centre-specific data are collected to allow relevant analyses.

Two observations can be made in relation to the categories of methods summarised above. First, the majority of studies approached transferability from an international perspective by referring to multinational RCTs. There remains the question whether existing evidence is sufficient to warrant the use of generalisability techniques in refining centre-specific economic evaluation results of multi-centre single-country RCTs.

Second, a series of general recommendations for further research in the area of generalisability have been made (127, 134, 155, 170). Despite repeated calls for addressing generalisability at the trial design stage, no practical guidance has been offered to date and most of the existing research contributions are to be employed in retrospective analyses using trial-wide results. The question still remains as to the role and scope for a prospective methodology, applicable at the trial design stage, to support generalisability. This issue will be addressed in the following sub-section.

1.3.3. Identifying knowledge gaps

Can economic evaluation estimates vary?

Extrapolating the results of a trial-based economic evaluation is of interest to decision makers, who want to know whether and the extent to which a particular intervention is cost-effective in their jurisdiction. The first question that arises is: *are there reasons to believe that economic evaluation results vary systematically across centres in a given jurisdiction?* At present it is difficult to answer. There are indications in the literature that within-country variations in economic evaluation results are possible. For instance, in their article which introduced multilevel modelling in economic evaluations, Manca *et al.* (164) obtained centre-specific net monetary benefits and cost-effectiveness acceptability curves for each of the 20 centres in an English RCT and the results clearly suggested that the intervention under scrutiny was cost-effective in some settings and cost-ineffective in others. In another study, which focused on using MI methods to obtain centre-specific unit costs as opposed to average unit costs for a trial-based economic evaluation set in the NHS, Grieve *et al.* (140) found that, for the particular comparison under scrutiny, the intervention was more cost-effective in teaching hospitals than in district general hospitals. More specifically, the intervention was 15% less likely to be cost-effective at £30,000/QALY in non-teaching hospitals compared to teaching hospitals when MI methods were used, and 40% less likely to be cost-effective when mean reference costs were used. Sculpher *et al.* (12) had mentioned in their review several studies reporting differences in cost-effectiveness estimates between centres as a result of differences in unit costs and practice variation. Only one of these studies was UK-based. Nevertheless, they suggested that obtaining centre-specific cost-effectiveness results required further exploration to establish their usefulness for local policy makers. Goeree *et al.* (136)

acknowledged that *"the little evidence that does exist suggests that hospital cost variation may be as large within countries as it is between countries"* (p. 565). Consequently, variables intrinsic to the patient (e.g. age, comorbidities), the health care inputs (e.g. qualification of surgeons, availability of particular medical technologies) and the health care system (e.g. financing streams) may explain reasonably well the source of these variations. Even in health care systems like the NHS where hospital reimbursement relies on largely fixed tariffs (Payment by Results), hospital-specific costs are expected to vary (171, 172). Furthermore, observational data from the English NHS suggested that between-hospital variation in cost of care for all obstetrics patients can be as high as 19% after controlling for patient characteristics (173), while between-hospital variation in length of stay for elective hip replacement was in the region of 5% (174). Coding inaccuracies, apportioning shared costs and managerial inefficiency were all indicated as potential explanations for the observed differences. In the light of this evidence and given that the interdependence between costs and outcomes is often difficult to quantify, there are reasons to expect a potentially significant systematic variation in cost-effectiveness between centres at the very onset of the RCT. This should lead to a proportionate interest from the part of local decision makers of accounting for as much of this variation as possible in economic analyses of interventions.

Do economic evaluation results vary?

Nevertheless, how can it be ascertained that economic evaluation results actually vary across locations? Within the constraints of an experimental design one can only enrol in research a sample of the potentially relevant centres, therefore the issue quickly becomes whether results vary across the centres involved in the RCT. This question can only be reliably answered in retrospect, once the trial results have been analysed. Gail and Simon

(157) described such tests for heterogeneity to test the influence of centre on any parameter of interest and their methods were applied by Cook *et al.* (156), as discussed in the previous sub-section. However, heterogeneity tests are usually underpowered and have limited informative value. Notwithstanding, once heterogeneity has been ascertained and the need for adjustment acknowledged, the methods outlined in the previous sub-section can be used to refine the cost-effectiveness estimates for each of the participating centres. Refining should be understood in this context as adjusting the centre-specific cost-effectiveness estimate based on the trial-wide results. It must be made clear once more that the existing methods refer to data analysis and are retrospective in nature because they are only applicable when the trial data have been collected.

Are adjusted cost-effectiveness estimates valid?

This leads to a further question: *are there limitations inherent in the retrospective approach of these methods that may lead us to question the validity of the adjusted cost-effectiveness estimates?* Two observations must be made. First, none of the existing methods makes any verifiable assumption regarding the sample of centres included in the analysis. In other words, pooling data from centres from within the same jurisdiction is assumed to reliably lead to a representative cost-effectiveness estimate for the entire jurisdiction. For example, if a multi-national RCT recruited patients from four centres in country A, adjusted cost-effectiveness estimates for country A will be based on information collected from those four centres and on trial-wide information. This is what Manca *et al.* demonstrated in their example of the ATLAS trial (167). However, is it correct to assume that the cost-effectiveness estimate for country A is valid without knowing how representative those four centres are for country A itself?

Intuitively, at least, this assumption should hold if the centres enrolled in the RCT were representative for the jurisdiction they represent. This could be achieved in two ways: either centres were deliberately chosen based on a number of covariates which recommended them as representative at jurisdiction level; or the centres were randomly selected from the pool of available centres in the jurisdiction. There is no evidence in the literature to date that either condition has been satisfied. Moreover, given the host of factors influencing cost levels, it is expected that the definition of 'representative' is both complex and difficult to specify. Purposive selection of recruiting centres/sites has been previously suggested without further details (101, 127, 170). Drummond *et al.* (170) suggested possible centre-level covariates and introduced the concept of minimum patients recruited from each centre, but no consistent method to address this suggestion has yet been developed.

The issue of randomly selecting centres has been touched upon in the literature rather as a limitation and an area where more research should be conducted (155, 164). Furthermore, choosing an insufficient number of centres and corresponding sample sizes can only lead to biased mean estimates and large variances. As discussed in sub-section 1.3.2, the issue applies to centre-specific unit costs, as well.

The second observation is that the existing generalisability methods still leave decision makers from jurisdictions that were not involved in the trial with difficulties in transferring the economic evaluation results. Building on the limitation outlined above, this equally applies to centres that belong to a jurisdiction included in the trial but have not contributed with primary data. Manca *et al.* (175) advised towards great caution when considering such extrapolations. If the generalisability refinements have not incorporated centre-specific covariates (e.g. patient case-mix), one potential approach would be to find a similar recruiting centre in terms of the covariates considered in the model and then simply use the economic evaluation result. However, there is no guarantee that such a centre exists and, more importantly and in relation with 'representativeness', there is no straightforward indication as to what exactly constitutes 'similarity'.

Manca *et al.* (175) also offered a comprehensive account of the analytical strategies available depending on the availability of individual-patient data (IPD) and participation in the trial. The proposed framework was designed to address multinational studies, but the authors suggested that it may be useful for within-country jurisdictions. In the absence of IPD and if the jurisdiction of interest did not participate in the trial, decision-modelling was the indicated option. Decision models usually offer cost-effectiveness estimates with confidence intervals around them according to the uncertainty and sensitivity analyses incorporated; if the uncertainty around the point cost-effectiveness estimate is large (i.e. the confidence interval is wide), the result is of little use and its applicability is restricted to jurisdictions which are assumed to have identical budgets and identical reimbursement priorities. Of course, constructing a decision model for each centre would be impractical. An alternative solution would be to make the decision model available to all interested decision makers, who may adjust the parameters to their own needs. This would involve specifying a transparent and user-friendly decision model and circulating it to decision makers.

Another approach involves the use of a preliminary decision model (69). For example, Glasziou *et al.* (176) used a preliminary cost-utility analysis to inform data selection and the required sample size. The question at hand is the following: when an estimate of cost-effectiveness robust to sensitivity analyses is already available, under what circumstances is it worth collecting prospectively additional centre-specific data (as required by multilevel modelling and bivariate hierarchical modelling, for instance, or for selecting centres based on centre-specific covariates)? Decision modelling is not ideal because the issue at hand is not an entirely statistical consideration and should not be treated as such. If the sample of participating centres is unrepresentative (e.g. a random sample) of the entire population of centres, any estimator based on the sample of centres will be biased from the nationwide estimator to an unknown degree and in an unknown direction. First, the preliminary decision model may often be based on effectiveness and resource use estimates from outside the jurisdiction (e.g. another country) and the impact of the differences would be difficult to assess, especially in the presence of structural uncertainty. Second, in relation to the concerns expressed in the previous paragraph, even if jurisdiction-specific cost-effectiveness estimates exist, there is no guarantee that they reflect national practice if the process of centre selection (centres from which primary data were collected and the estimates have been calculated) has not been justified. Third, the uncertainty around the decision model's cost-effectiveness output may be significant enough to prevent any yes/no recommendation to be formulated. Finally, let us assume that a jurisdiction-specific decision model exists and wide sensitivity analyses around the base-case estimates have proved virtually every scenario to be cost-effective. It would thus be expected that the intervention is cost-effective in any given centre within the jurisdiction and a yes/no decision can be made. However, a yes/no decision is simply not enough, as this would require local decision makers to have the same reimbursement priorities. Taking the example of the UK health reforms at hand, where increased decentralisation is about to be implemented and commissioning devolved to local clinical commissioning groups, this assumption is unlikely to hold (177).

Summary

The discussion above attempted to identify the knowledge gaps associated with retrospective methods concerned with patient-level data analysis from RCTs. Two main themes emerged, both giving reason for concern: there usually is no explicit method of selecting centres and their corresponding sample sizes, although this has been suggested in the literature at a conceptual level. This limitation may hamper the validity of a series of computations, from heterogeneity tests to adjusted cost-effectiveness estimates. Furthermore, there is no reliable tool available for decision makers representing centres and/or jurisdictions which did not participate in an RCT to relate to the trial-wide results when having to make decisions in their own settings. Some of the limitations of modelling methods in addressing these concerns have also been discussed. These themes suggest that centre selection has not been addressed in the literature and it may matter in deriving centre-specific cost-effectiveness estimates. However, its impact has not yet been established.

1.4. Thesis objectives and structures

The objectives of the thesis are as follows: first, to evaluate the implications of the current practice of centre selection in RCTs in the UK for the generalisability of trial results; second, to identify any discrepancies between the current and optimal practice of centre selection; and third, to propose and demonstrate a novel methodology i.e. the Generalisability index (Gix), as a tool to explore the influence of centre selection on RCT results and to allow the selection of representative centres at the trial design stage. The research will consider a given intervention as a case study i.e. wound-edge protection devices (WEPDs) to reduce surgical site infection (SSI) after open abdominal surgery, and will follow the standard steps in generating evidence on the clinical and economic benefits of medical interventions (69).

The novel generalisability methodology will be demonstrated using data collected alongside the ROSSINI trial (Reduction of Surgical Site Infection using a Novel Intervention), a UK-based RCT which evaluated the benefits of WEPD against standard care.

The thesis is structured as follows: the first part (Chapter 2 to Chapter 6) presents the clinical and economic evidence related to the chosen case study i.e. the benefits of WEPDs compared to standard care (no WEPDs) in reducing SSI. Chapter 2 introduces the main concepts and issues surrounding SSI together with the strategies available to reduce it, including WEPDs. Chapter 3 appraises and summarises the existing evidence on the clinical effectiveness of WEPDs by means of a systematic review and meta-analysis. Chapter 4 produces preliminary evidence on the cost-effectiveness of WEPDs compared to standard care using an original decision tree informed by secondary data relevant to the UK setting. Chapter 5 describes the rationale and principal clinical findings of the ROSSINI (Reduction of Surgical Site Infection Using a Novel Intervention) trial and presents the results of the trial-based economic evaluation of WEPDs compared to standard care. Chapter 6 provides an integrative discussion of the clinical and economic evidence on the benefits of WEPDs in reducing SSI, based on the findings presented in Chapters 3 to 5.

The second part of the thesis (Chapter 7 and Chapter 8) discusses in depth the generalisability of trial results from the perspective of centre selection and proposes an approach to evaluate generalisability, which will be demonstrated using ROSSINI data. Chapter 7 presents the methods and findings of a mixed methods study describing the current and optimal practice of centre selection for RCTs in the UK. Chapter 8 describes in detail the Generalisability index as a tool to explore the influence of centre selection on RCT results and demonstrates its utilisation using the ROSSINI trial as a case study. Ultimately, Chapter 9 offers an integrative discussion of the previous Chapters' findings.

CHAPTER 2. BACKGROUND TO SURGICAL SITE INFECTION AND WOUND-EDGE PROTECTION DEVICES

The aim of this Chapter is to introduce the clinical context which serves as a case study for the generalisability investigation i.e. surgical site infection (SSI) and the use of wound-edge protection devices (WEPDs). The Chapter starts with an exposition of the relevant concepts for SSI – definition, classification, epidemiology and consequences – and then discusses the types of strategies available to minimise SSI risk, with a focus on WEPDs.

2.1. Background to surgical site infection

Health care-associated infection (HCAI) can be defined as "*an infection occurring in a patient in a hospital or other health-care facility in whom the infection was not present or incubating at the time of admission. This includes infections acquired in the hospital, but appearing after discharge, and also occupational infections among staff of the facility*" (178, p.1). In a recent systematic review on the worldwide burden of HCAI (179), the World Health Organization (WHO) estimated that 7.1% of hospitalized patients acquire a HCAI, of which approximately 20% are SSIs. Surgical infections are postoperative complications with an overall average incidence among surgical patients in the range of 1-5% (180, 181). The burden of SSI is particularly high in developing countries: a recent systematic review (182) suggested a pooled cumulative incidence of SSI of 5.6 cases per 100 surgical procedures, almost twice the average value in the US (183) and Europe (184), thus making it the most prevalent type of HCAI in such settings. Examples include SSI rates of 12% in Bolivia (185), up to 17% in Egypt (186, 187), 24% in Brazil (188) and 26% in Tanzania (189).

Data from the US, UK and continental Europe indicate substantial variation in SSI incidence for different surgical sites, with hip replacement among the interventions with the lowest risk and large bowel surgery at the opposite end of the spectrum (183, 184, 190, 191). For example, the cumulative SSI incidence in English hospitals between 2006 and 2011 (Table 2.1) was 0.6% and 0.8% for knee and hip prosthesis, respectively, and 10.1% for large bowel surgery. These were the lowest and highest SSI rates, respectively, among all the 17 monitored surgical interventions (192). These concur with European-wide estimates (193), where the highest SSI rates were observed for colon surgery (9.2%) and the lowest for knee prosthesis (0.7%). In particular colorectal surgery is typically associated with average SSI incidence rates of 4-10%, but rates as high as 27% have been reported (194-198), especially in studies with intensive patient follow-up i.e. outside the inpatient setting.

Table 2.1 Cumulative SSI incidence by surgical category in England (2006-2011)

Type of surgery	Operations reported	SSI - inpatient & readmission	SSI rate (%) - inpatient & readmission	95% CIs
Abdominal hysterectomy	5,388	80	1.5	1.2-1.8
Bile duct, liver and pancreatic surgery	1,559	126	8.1	6.8-9.6
Breast	1,484	17	1.2	0.7-1.8
Cardiac (non-CABG)	1,286	13	1.0	0.5-1.7
Cholecystectomy	619	11	1.8	0.9-3.2
Coronary artery bypass graft (CABG)	26,468	1,172	4.4	4.2-4.7
Cranial	557	5	0.9	0.3-2.1
Gastric	1,093	48	4.4	3.3-5.8
Hip prosthesis	150,149	1,169	0.8	0.7-0.8
Knee prosthesis	162,728	895	0.6	0.5-0.6
Limb amputation	2,538	126	5.0	4.2-5.9
Large bowel	13,534	1,370	10.1	9.6-10.6
Reduction of long bone fracture	7,580	104	1.4	1.1-1.7
Repair of neck of femur	39,830	647	1.6	1.5-1.8
Small bowel	2,902	196	6.8	5.9-7.7
Spinal	13,166	126	1.0	0.8-1.1
Vascular	7,798	221	2.8	2.5-3.2

Source: Health Protection Agency (2011)

2.1.1. SSI definitions

The best known definition of SSI is the one elaborated by Horan *et al.* in 1992 (199), endorsed by the US Centers for Disease Control and Prevention (CDC) (200). This definition replaced the term 'surgical wound infection' (SWI) (201), which referred to incision infections, with 'surgical site infection' in order to comprise infections both at the organ and the incision level. Thus surgical infections are categorised according to their site in superficial SSIs, deep SSIs and organ/space SSIs (Table 2.2).

Nevertheless, a host of SSI definitions are available in clinical practice and research. Bruce *et al.* (202) conducted a comprehensive systematic review of the definition, measurement and monitoring of surgical wound infection and three other surgical adverse events. They reviewed 82 studies and identified 41 different definitions of surgical wound infection. Of the 41 definitions, five were nationally proposed definitions coming from US (199, 201) and UK (203-205) collaborative groups, respectively. Other studies used definitions largely based on the presence of purulent discharge with or without bacterial culture in combination with other criteria. The CDC definition was used in 29 studies from 12 countries, while the UK definitions were used in three UK-based studies. The plethora of SSI definitions and the apparent predominance of the CDC criteria informed the recommendation to consider the implementation of the CDC definition in the UK in the interest of consistency and comparability. The recommendation was later translated in practice and the current SSI definition employed by the UK Health Protection Agency (HPA) is in line with the CDC definition (206).

Table 2.2 CDC definition of SSI

Superficial Incisional SSI

Infection occurs within 30 days after the operation *and*
infection involves only skin or subcutaneous tissue of the incision *and*
at least *one* of the following:
1. Purulent drainage, with or without laboratory confirmation, from the superficial incision.
2. Organisms isolated from an aseptically obtained culture of fluid or tissue from the superficial incision.
3. At least one of the following signs or symptoms of infection: pain or tenderness, localized swelling, redness, or heat *and* superficial incision is deliberately opened by surgeon, *unless* incision is culture-negative.
4. Diagnosis of superficial incisional SSI by the surgeon or attending physician.

Do *not* report the following conditions as SSI:
1. Stitch abscess (minimal inflammation and discharge confined to the points of suture penetration).
2. Infection of an episiotomy or newborn circumcision site.
3. Infected burn wound.
4. Incisional SSI that extends into the fascial and muscle layers (see deep incisional SSI).

Note: Specific criteria are used for identifying infected episiotomy and circumcision sites and burn wounds.

Deep Incisional SSI

Infection occurs within 30 days after the operation if no implant is left in place or within 1 year if implant is in place and the infection appears to be related to the operation *and*
infection involves deep soft tissues (e.g., fascial and muscle layers) of the incision *and*
at least *one* of the following:
1. Purulent drainage from the deep incision but not from the organ/space component of the surgical site.
2. A deep incision spontaneously dehisces or is deliberately opened by a surgeon when the patient has at least one of the following signs or symptoms: fever (>38ºC), localized pain, or tenderness, unless site is culture-negative.
3. An abscess or other evidence of infection involving the deep incision is found on direct examination, during reoperation, or by histopathologic or radiologic examination.
4. Diagnosis of a deep incisional SSI by a surgeon or attending physician.

Notes:
1. Report infection that involves both superficial and deep incision sites as deep incisional SSI.
2. Report an organ/space SSI that drains through the incision as a deep incisional SSI.

Organ/Space SSI

Infection occurs within 30 days after the operation if no implant is left in place or within 1 year if implant is in place and the infection appears to be related to the operation *and*
infection involves any part of the anatomy (e.g., organs or spaces), other than the incision, which was opened or manipulated during an operation *and*
at least *one* of the following:
1. Purulent drainage from a drain that is placed through a stab wound‡ into the organ/space.
2. Organisms isolated from an aseptically obtained culture of fluid or tissue in the organ/space.
3. An abscess or other evidence of infection involving the organ/space that is found on direct examination, during reoperation, or by histopathologic or radiologic examination.
4. Diagnosis of an organ/space SSI by a surgeon or attending physician.

Source: Centers for Disease Control and Prevention (1999)

A comprehensive systematic review concluded that even small differences between SSI definitions can account for large variations in reported SSI rates both across institutions and across countries; therefore comparing estimates in the literature should be exercised with great caution (207). Furthermore, several important shortcomings have been pointed out in relation with the CDC definition: first, it relies on a relatively complex algorithm, which makes it difficult to implement and open to interpretation; second, it doesn't consider SSIs which occur beyond 30-days post-operatively and thus cannot account for the long-term impact of SSI. The authors of the review advocated the need for a more reliable and easy to implement definition of SSI before formally using SSI rates as a proxy for quality of health care services with a view to comparing hospitals.

A further aspect in SSI assessment pertains to the grading of wound infection, a useful instrument in SSI diagnosis. The same systematic review of Bruce *et al.* (202) identified 13 grading scales for surgical wound infection. The most prominent ones are the ASEPSIS scale (208) and the Southampton Wound Assessment Scale (209). The former was developed with the aim of evaluating wound healing after cardiac surgery and involves a point-based system relying on both clinical signs (serous discharge, erythema, purulent exudate and separation of deep tissue) and objective criteria such as antibiotic treatment and inpatient stay. A total score greater than 20 points indicates a SSI. The Southampton Wound Assessment Scale was developed for the assessment of hernia wounds and comprises five grades from 0 (normal healing) to 5 (deep or severe wound infection). Both grading scales were validated (210), but their practical implementation was judged to be cumbersome. A more recent comparison between the CDC and ASEPSIS definitions pointed out that ASEPSIS is more sensitive than CDC and the agreement between them is moderate at best; however, somewhat paradoxically, the two scales performed comparably (and also modestly) in predicting outcomes such as postoperative length of stay and the prescription of antibiotics (207). This finding only highlights the need for developing a more robust SSI definition in the future.

2.1.2. SSI microbiology

SSI can only develop if the surgical site is contaminated with microorganisms, which can originate either from the patient or from the environment in the operating room. When the skin is incised, the tissue is exposed to the flora on patient's skin, mucous membranes and hollow viscera, which constitute the causative agents of SSI in most cases (211). The pathogens responsible for SSI have been known for years and *Staphylococcus aureus* has long been indicated as the leading cause for SSI (200), but the microorganisms responsible for infection may differ across countries. For example, the HPA reported that *Enterobacter spp* were the predominant causes of SSI in 2011 (31% of isolated pathogens), followed by *Staphylococcus aureus* (27%), *Enterococcus spp* (8%), *Pseudomonas spp* (8%) and *Coagulase-negative staphylococci* (8%). Methicillin-resistant *S. Aureus* (MRSA) contributed 6% of all identified pathogens (192). The situation appears to be markedly different in the US, where *Staphylococcus aureus* is the leading causative pathogen (40%), followed by *Coagulase-negative staphylococci* (10%), *Streptococcus spp* (3.5%) and *Enterococcus spp* (2.6%), with MRSA accounting for 13.7% of infections and rising across time (212). *Staphylococcus aureus* has also been indicated as the major causative pathogen in countries such as Switzerland (213) and Egypt (186). The rise of MRSA as a cause for SSIs is an indication of the high proportion of immunocompromised individuals, potentially a consequence of the widespread use of antibiotics.

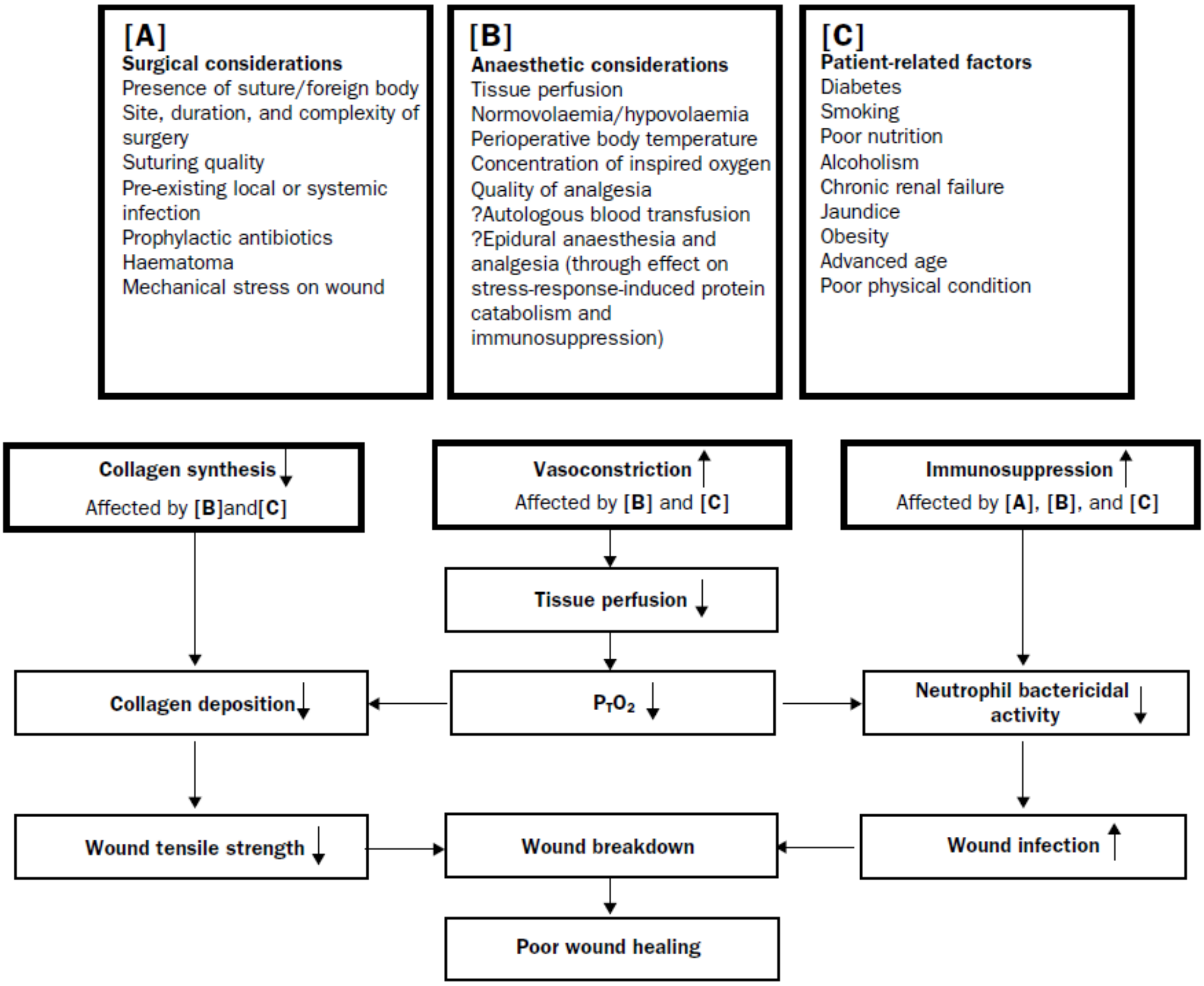
2.1.3. Risk factors for SSIs

The risk factors for SSI have been traditionally classified in two categories: patient characteristics and operative characteristics (200). Patient characteristics refer to factors such as age extremes, diabetes, smoking, obesity, malnutrition and the presence of infections at other sites; operative characteristics include skin antisepsis, the duration of the operation, preoperative shaving and preoperative skin preparation (Table 2.3). A further argument has been made for a distinct influence of anaesthetic considerations as a separate class of determinants, in addition to patient and operative characteristics (214). This is based on the influence that variables such as tissue perfusion, the perioperative body temperature and the concentration of inspired oxygen have on the wound healing process (Figure 2.1).

Table 2.3 Risk factors for SSI

Patient characteristics	Peri-operative characteristics
Age	Duration of surgical scrub
Diabetes	Skin antisepsis
Smoking	Preoperative shaving
Malnutrition (hypoalbuminemia)	Preoperative skin preparation
Obesity	Duration of operation
Coexistent infections at a remote body site	Antimicrobial prophylaxis
Colonisation with microorganisms	Operating room ventilation
Altered immune response	Inadequate sterilization of instruments
Length of preoperative hospital stay	Foreign material in the surgical site
	Surgical drains
	Surgical technique: poor haemostasis, failure to obliterate dead space, tissue trauma

Source: Adapted from CDC (1999)



P_TO_2 = Partial pressure of oxygen in tissues

Figure 2.1 Factors affecting surgical-wound healing

Source: Buggy (2000) reproduced with permission[3]

[3] Reprinted from The Lancet, 357(9227), Donal Buggy, Can anaesthetic management influence surgical-wound healing?, Copyright (2000), with permission from Elsevier.

Recent evidence suggested additional several SSI risk factors that hadn’t been accounted for in previous studies. For example, it has been shown that a history of skin infection is associated with enhanced susceptibility to SSI (215). Additionally, the surgeon himself has been found to be an independent risk factor, after controlling for adherence to guidelines and experience (216).

2.1.4. SSI risk categories

Three types of variables have been suggested as reliable predictors of SSI: 1) the intrinsic degree of microbial contamination of the surgical site; 2) the duration of an operation; and 3) markers for patient susceptibility (200). In relation to the degree of contamination of the surgical site, the widely accepted classification of surgical wounds distinguishes between four categories (classes) of wounds: clean, clean-contaminated, contaminated and dirty wounds (Table 2.4). At the end of the surgical procedure, a member of the surgical team assesses the type of wound according to the agreed criteria; the risk of SSI is increasingly higher from clean to dirty wounds.

Table 2.4 Classification of surgical wounds

Category	Description
Clean/ Class I	An uninfected operative wound in which no inflammation is encountered and the respiratory, alimentary, genital, or uninfected urinary tract is not entered. In addition, clean wounds are primarily closed and, if necessary, drained with closed drainage. Operative incisional wounds that follow non-penetrating (blunt) trauma should be included in this category if they meet the criteria.
Clean-contaminated/ Class II	An operative wound in which the respiratory, alimentary, genital, or urinary tracts are entered under controlled conditions and without unusual contamination. Specifically, operations involving the biliary tract, appendix, vagina, and oropharynx are included in this category, provided no evidence of infection or major break in technique is encountered.
Contaminated/ Class III	Open, fresh, accidental wounds. In addition, operations with major breaks in sterile technique (e.g. open cardiac massage) or gross spillage from the gastrointestinal tract, and incisions in which acute, non-purulent inflammation is encountered are included in this category.
Dirty-infected/ Class IV	Old traumatic wounds with retained devitalized tissue and those that involve existing clinical infection or perforated viscera. This definition suggests that the organisms causing postoperative infection were present in the operative field before the operation.

Source: Centers for Disease Control and Prevention (1999)

However, estimating SSI risk solely based on the type of incision is insufficient because other variables may also play a role and there is also the risk of an incorrect classification. A more comprehensive estimator is the National Nosocomial Infections Surveillance (NNIS) SSI risk index (217), which accounts for three independent risk factors and takes values between 0 (no risk factor present) and 3 points (all risk factors present), such that 1 point is awarded for each of the following instances: a) American Society of Anaesthesiologists (ASA) Physical Status Classification greater than 2, signifying a patient with severe systematic disease which may threaten his life (218); b) either contaminated or dirty/infected wound classification, as defined above; and c) length of operation greater than T hours, where T is approximately the 75th percentile of the duration of the specific intervention performed (219). Thus the NNIS risk can be regarded as a more reliable and objective risk estimator because it not only incorporates the wound classification system, but also the ASA class as a surrogate for patient susceptibility and the intervention-specific duration of surgery. Moreover, it allows surveillance authorities in each country to calibrate individual patient risk based on local data on length of surgery (220).

Although widely used, further research is needed towards the reliability of the NNIS risk index, as noted in the CDC guidelines (200). Indeed, the evidence around its performance is controversial: while there are indications that the NNIS risk index is highly correlated with SSI rates for some common operations (221), other results indicated that the NNIS index may actually have too low a sensitivity to be used as a prognostic tool (222).

2.1.5. Patient-level consequences of SSIs

SSIs are associated not only with considerable morbidity (223, 224), but also with excess mortality; it has been suggested that over one-third of postoperative deaths are related, at least in part, to SSI (225). Evidence suggests an increase of up to ten-fold in mortality rates in SSI patients compared to uninfected controls (226) and an increased likelihood of hospital readmission (227). Other clinical outcomes of SSIs include scars that are cosmetically unacceptable, such as those that are hypertrophic or keloid, persistent pain and itching (228).

There have been relatively few studies investigating the effects of SSI on health-related quality of life. The little available evidence suggests that SSI patients reported reduced quality of life in comparison with uninfected controls, both in relation to physical functioning and mental health (229, 230). A qualitative interview study on Swedish patients having deep SSIs revealed experiences of pain, insecurity and isolation extended over several months or even years (231).

2.1.6. Costs associated with SSIs

SSIs are associated with additional length of stay in hospital in the range of 6 to 17 days (181, 191, 232, 233) and even up to 23 days in the case of SSI due to Methicillin-resistant *S. aureus* (MRSA) (234). Moreover, SSI patients receive more intensive care after discharge compared to uninfected patients (226, 230), translated in higher number of home visits from a health care professional, ambulatory visits, emergency room visits and medication. These factors are responsible for an additional cost of care due to SSI of up to £10,500 (191, 195, 224, 235). Most cost studies in the US cited an additional cost due to SSI at approximately $3,000, but there is also evidence of differences in excess of $20,000 (229). The largest part of the health care costs associated with SSI are due to prolonged inpatient cost, but costs in an outpatient setting have been shown to amount to as much as 15% of total health care costs (195).

While the fact that SSIs are associated with supplementary costs is undisputed and the differences in magnitude between various estimates are most likely due to study design and setting, further aspects need to be considered. First, incurred costs are proportional with the depth of the SSI i.e. costs are lower for superficial SSI and higher for organ/space SSI (236, 237). Second, health care costs are only a fraction of the total costs associated with SSI. While only few studies have taken a societal perspective in cost-analysis to date (237, 238), the little available evidence suggests that health care costs may amount to as little as 10% of total costs when lost productivity is considered (239).

Evidence suggests that SSIs due to MRSA may be associated with prolonged hospital stay and even larger additional costs (240, 241). For example, Anderson *et al.* (234) collected resource use data from 509 patients along a 90-day time horizon and concluded that SSI-MRSA were associated with a $61,000 increase in total hospital charges compared to uninfected controls, as well as with a $24,000 increase compared to SSI cause by Methicillin-sensitive *Staphylococcus aureus*.

2.1.7. Strategies for SSI prevention

Evidence suggests that both HCAI and SSI are largely preventable, although it has been acknowledged that the currently available technologies do not allow 100% prevention (242). The effort aimed at reducing the burden of SSIs during the past decades has concentrated on three main directions: 1) mitigating the known SSI risk factors; 2) improving SSI prediction; and 3) improving SSI detection. In terms of addressing the known SSI risk factors, perioperative care factors have been comprehensively addressed in clinical guidelines issued by the US CDC (200) and the UK's National Institute for Health and Clinical Excellence (NICE) (243), respectively. The guidelines address in great detail practical considerations pertaining to the preoperative (e.g. hair removal, antibiotic prophylaxis, bowel preparation), intraoperative (e.g. hand decontamination, skin preparation) and postoperative phases (e.g. changing dressings, wound debridement) of the surgical intervention. Moreover, NICE clinical guidelines identified key priorities for further investigation, which include the benefit of various types of wound dressings, the benefit of nasal decontamination using mupirocin against *S. aureus*, the potential benefit of various techniques for maintaining patient homeostasis and the effect of the closure methods on SSI risk (243).

Second, the need for better prediction of SSI led to the development of more complex prognostic models to advance the understanding on protective and contributing factors (244-246). Prognostic models allow a more precise decomposition of the influence of various factors on the SSI risk as opposed to aggregating them in a single risk measure. For example, it has been noted that in some countries the NNIS risk index may discriminate poorly between high risk and low risk patients or that it may not be correlated with SSI rates at all, while locally constructed prognostic models or indices performed better (185, 188, 247). These findings highlight the need for constructing local prognostic models with a view to a better prediction of SSI occurrence. The same applies to subpopulations with particular co-morbidities: for example, Anaya *et al.* have recently developed a cancer specific SSI risk-stratification tool where preoperative chemotherapy emerged as an independent risk factor (248).

Third, the importance of SSI surveillance programmes for SSI detection has been acknowledged. This concerns both inpatient and, most importantly, post-discharge and outpatient surveillance. Although SSIs were traditionally believed to present up to six days postoperatively, more recent evidence indicates a later development, in the region of nine to 13 days postoperatively on the average (194, 195). Given the continuous pressures that hospitals face to decrease length of stay, this implies that many SSIs may manifest in an outpatient setting. Indeed, a large proportion of SSIs are detected post-discharge: European-wide data indicated that 48% of SSIs were detected post-discharge (193), in accordance with previous estimates from smaller studies in the UK (195, 249) and US (194). Estimates as high as 86% have also been cited (250).

There is still conflicting evidence on the most appropriate methods for post-discharge surveillance (PDS). A systematic review identified direct observation, telephone interviews with patients and patient questionnaires as the most common methods employed for data collection in PDS, but eventually concluded that information on the validity and reliability of the existing methods was insufficient to recommend either of them (251). National surveillance systems have been implemented in countries such as England (206), Scotland (252), France (253), the Netherlands (254), Germany (255) and Australia (256); most reports cited an overall reduction of SSI rates along the years of surveillance, thus suggesting the demonstrable beneficial effect of PDS in reducing SSI rates. While such a finding is in line with local reports (257-259), the difficulty of disentangling the effect of public reporting from other infection control measures has to be acknowledged, especially considering that most national surveillance schemes are still based on voluntary hospital reporting. Moreover, the decline in SSI rates over time only applies to hospitals with a history of SSI surveillance of several years and cannot be generalised to all types of surgical procedures (260).

2.2. Background to wound-edge protection devices

WEPDs are an intervention aimed at reducing SSI rates in patients undergoing open surgery (non-laparoscopic). Also known as 'wound guards', they have been used in abdominal surgery for more than 40 years, having been firstly mentioned at the end of the 1960s (261). There are several types of devices available on the market, but they all have the same basic design: a semi-rigid plastic ring placed into the abdomen via the laparotomy wound to which an impervious drape is circumferentially attached (262, 263). This plastic drape comes up and out of the wound onto the skin surface, thus protecting the incised wound edges from contact with contaminated media. The device is inserted in the abdominal cavity by the surgeon as soon as the incision has been made and is removed just before wound closure (Figure 2.2). The device also has retraction properties i.e. keeping the wound edges apart, which explains why it may also be marketed as a ‘wound retractor’.

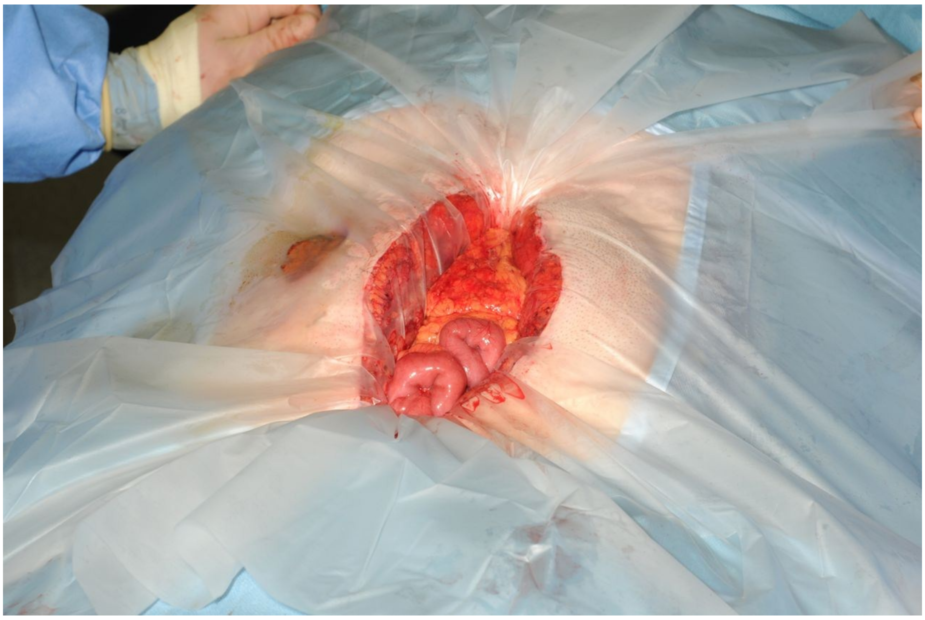


Figure 2.2 Wound-edge protection device used during open abdominal surgery
Source: Pinkney *et al.* (2013), reproduced with permission[4]

[4] Reprinted from BMJ, 347:f4305, Pinkney TD *et al.*, Impact of wound edge protection devices on surgical site infection after laparotomy: multicentre randomised controlled trial (ROSSINI Trial), Copyright (2013), with permission from BMJ Publishing Group Ltd.

Although WEPDs have been used for decades in the interest of reducing SSI rates, their mechanism of action is still unclear. Several explanations have been postulated: firstly, WEPDs create a physical barrier between the abdominal wound edges and viscera, visceral contents, contaminated instruments and gloves; this reduces the accumulation of endogenous and exogenous bacteria on the wound edges. In support of this mechanism, Raahave *et al.* (264) found different bacterial densities on and under such a device when used in laparotomy wounds. Their findings were replicated more recently in an observational study focusing on gastrointestinal surgery (265). WEPDs may also reduce necrosis from long procedure exposure of the incised tissue. Moreover, due to intrinsic retraction properties they reduce the need for handheld mechanical retraction and thus the associated tissue damage. However, it has been hypothesised that, conversely, contaminated intra-peritoneal fluid may advance through capillarity along the impervious wound guard and reach the wound edges thereby causing infection.

A distinction should be made between WEPDs and 'adhesive drapes': the latter are plastic drapes adherent to the skin and they do not come into direct contact with the wound margins. A Cochrane systematic review summarised the evidence on adhesive drapes and concluded that they do not show any benefit in reducing SSI rate (266). This distinction is particularly important because, on occasions, WEPDs have been referred to as 'ring drapes' or 'impervious drapes', such that confusion between the two types of devices may arise. The major difference in design between them is that adhesive drapes lack a plastic ring and remain entirely on the surface of the patient's skin without coming in contact with the abdominal cavity or with the wound edges (266).

Despite their potential for reducing SSIs when used intra-operatively by protecting the wound margins from contact with any contaminated materials, they have never come to widespread use and they are not even mentioned in the current UK clinical guidelines (243). Although there have been RCTs looking at the effectiveness of WEPDs versus that of various comparators (267-271), most of these trials are single-centre and the quality of their reporting appears to be questionable. Furthermore, there is no meta-analysis available on this topic. These factors may explain the limited uptake of WEPDs in current practice, as they are currently used solely at the surgeons' discretion. The next Chapter will formally synthesise the available evidence on the clinical effectiveness of WEPDs in reducing SSI.

2.3. Conclusion

SSI is a serious postoperative complication which affects HRQoL and is associated with significant costs. WEPDs have been used informally by surgeons for more than 40 years to prevent SSI, but the evidence of their effectiveness is controversial and has never been systematically reviewed. The following Chapter addresses this gap by presenting the results of a systematic review of the clinical effectiveness of WEPDs vs. standard care in reducing SSI.

CHAPTER 3. SYSTEMATIC REVIEW OF THE CLINICAL EFFECTIVENESS OF WOUND-EDGE PROTECTION DEVICES IN REDUCING SURGICAL SITE INFECTION IN ADULTS UNDERGOING OPEN ABDOMINAL SURGERY

A paper based on the work presented in this Chapter has been published as:

Gheorghe A, Calvert M, Pinkney TD, Fletcher BR, Bartlett DC, Hawkins WJ, Mak T, Youssef H and Wilson S on behalf of the West Midlands Research Collaborative and ROSSINI Trial Management Group. (2012) Systematic Review of the Clinical Effectiveness of Wound-edge Protection Devices in Reducing Surgical Site Infection in Patients Undergoing Open Abdominal Surgery. *Annals of Surgery* 255(6): 1017-1029.

As noted in Chapter 2, WEPDs have been used for decades based on anecdotal evidence regarding their effectiveness, but current UK clinical guidelines on SSI management do not mention them (243). The question arises whether there is sufficient evidence available to make a definitive decision on the appropriateness of their use as a means to reduce the rate of SSI. The aim of this Chapter is to appraise the available evidence on the clinical effectiveness of WEPDs compared to standard care in reducing the rate of SSIs in patients undergoing open abdominal surgery.

3.1. Methods

A systematic review was conducted according to a pre-specified protocol based on guidance from the Centre for Reviews and Dissemination (272) and the Cochrane Handbook of Systematic Reviews of Interventions (7). The review is reported in line with the PRISMA statement (273) (Appendix 1).

The elements of the research question addressed by the review are reported below in the PICOS format (274):

Population: human patients of any age undergoing open abdominal surgery, both elective and emergency.

Intervention: use of a WEPD (for the purpose of this review, a device was considered eligible if it covered the wound's cut edges with an impervious plastic sheet).

Comparator: standard care, as defined in each included study. The use of a protective device different to the WEPD was accepted if no other control arm was present in the study.

Outcome: SSI rate was a pre-specified study outcome.

Study design: acceptable study designs were RCTs, prospective controlled trials (CTs), prospective cohort studies and case-control studies.

The areas covered by the study protocol are reported below.

3.1.1. Eligibility criteria

Study eligibility was judged against the pre-specified inclusion criteria presented above in PICOS format. Neither publishing year nor language restrictions were applied. Any potentially relevant paper in a language other than English was translated into English. The following pre-specified exclusion criteria were applied: studies looking at a different device (e.g. adhesive drapes), unless a WEPD was also used in the study; definitions of SSI based solely on bacteriological information; study designs with a high risk of bias including case reports and retrospective studies. Reviews were not accepted for lack of primary data.

Purely bacteriological definitions of SSI were excluded for two reasons. First, the current clinical guidelines (200, 243) specify definitions of SSI based predominantly on clinical signs (e.g. discharge, pus, localised swelling, erythema). Bruce *et al.* (202) identified 41 separate definitions of wound infections in their systematic review looking at the validity and reliability of postoperative wound infection assessment. Only five of these were 'standard' definitions (issued by the CDC or by UK expert groups) and all of them were substantially based on clinical signs. The second argument builds upon the distinction between contamination, which is a bacteriological outcome, and infection, which is a clinical outcome. Evidence has indicated the difficulty to differentiate infection from contamination when interpreting the positive results of swab cultures, therefore ascertaining the presence of SSI based on bacteriological results is not recommended (205, 275).

3.1.2. Information sources

The following sources were searched in November 2010:

- Online bibliographic databases: OVID MEDLINE, OVID EMBASE, EBSCO CINAHL, ISI Web of Science (including Science Citation Index and Conference Proceedings) and The Cochrane Library;
- Proceedings of the annual conferences of the Association of Coloproctology of Great Britain and Ireland (ACPGBI) and of Association of Surgeons of Great Britain and Ireland (ASGBI);
- The identified manufacturers of WEPDs (3M^{TM}, Applied Medical^{TM} and MCD^{TM}) were contacted and were asked to provide details of any relevant studies they were aware of;
- The references of the included articles (see below in section *Study selection*) were hand-searched for further relevant studies and for articles citing them;
- The authors of the selected articles (see below) were contacted and asked to provide details about any other relevant studies.

3.1.3. Search strategy

A sensitive search strategy was devised in order to capture all the relevant studies. Given the variety of names under which WEPDs have been marketed, their four-decade history of utilisation and the broad range of surgical interventions considered such an approach i.e. a sensitive, encompassing search strategy, was judged appropriate given the review's objective. The pre-specified search terms were grouped in two thematic areas: the

WEPD terms and the SSI terms. Truncation was used where appropriate. The search strategy was applied to all the online databases, with slight adjustments inherent to the specific vocabulary of each database. All terms were searched as keywords. The search was performed independently by two researchers (AG and BF[5]). The search strategies for all databases are presented in Appendix 2.

3.1.4. Study selection

The study selection process took place in two consecutive steps. In phase 1 potentially relevant articles were selected by scanning their title and abstract in relation to the inclusion/exclusion criteria, as described in section *Eligibility criteria* above. In phase 2 the full-text versions of the articles selected in phase 1 were assessed in relation to the eligibility criteria. When a decision about eligibility could not be made in phase 1 based on the title and abstract, the full-text article was obtained. Only studies that fulfilled all the eligibility criteria were included.

The selection was performed independently by two reviewers (AG and BF). In case of disagreement, a consensual decision was made with the help of a third reviewer (MC[6]). The selection process was tested and piloted on a random sample of 20 papers during phase 1 of the selection process.

3.1.5. Data extraction

For each study having entered phase 2 of the selection, the following information items were extracted: study design; total number of participants and stratified by arm; type of surgery performed; intervention (including description of WEPD); description of the control

[5] Benjamin R. Fletcher – Research Associate, Primary Care Clinical Sciences, University of Birmingham

[6] Melanie Calvert – Reader in Epidemiology, Primary Care Clinical Sciences, University of Birmingham

group; pre-specified and reported outcomes; length of follow-up; effect estimates (i.e. effect on SSI rate); funding and other competing interests.

One reviewer (AG) extracted data for all selected studies in The Cochrane Collaboration's RevMan software 5.0 (276). The accuracy of the extracted data was verified for all the included studies by a second reviewer (BF).

3.1.6. Risk of bias in individual studies

The 'Risk of bias' tool presented in the Cochrane Handbook of Systematic Reviews of Interventions (7) was used to ascertain the suitability of each study selected after phase 2 for inclusion in a meta-analysis. Two reviewers (AG and BF) performed the assessment independently. In case of disagreement, a consensual decision was made with the help of a third reviewer (MC). It was acknowledged that blinding was impossible for surgeons, so patient and assessor blinding, respectively, were considered.

3.1.7. Synthesis of results

The outcome of interest was dichotomous - presence or absence of an SSI. Given the significant variation in the types of surgery considered and in the definitions of SSI applied, it was judged that a distribution of effects would realistically describe the influence of WEPDs on the SSI rate. Consequently a random-effects model (277) (Mantel-Haenszel method) meta-analysis was pre-specified. Nevertheless, random-effects models do not produce reliable estimates when few studies are included in the meta-analysis (278); therefore the results of the meta-analysis are presented both under fixed-effects and random-effects models, and the differences are discussed. The degree of heterogeneity between studies was explored using the I^2 statistic (279). In accordance with recommended practice (280), sources of heterogeneity were explored and subgroup analyses were conducted. A subgroup analysis was pre-specified in the protocol to investigate the influence of the degree of contamination (clean/clean contaminated/contaminated/dirty) on the SSI rate. This was based on evidence suggesting that higher degree of contamination is a risk factor for SSI (200).

RevMan 5.0 software (276) was used to perform the quantitative analyses. The main outcome measure is the risk ratio (RR), reported with 95% confidence intervals (CI), of developing an SSI in the intervention arm compared to the control arm.

3.1.8. Publication bias

A pre-specified publication bias assessment was performed by means of a funnel plot. Two formal tests for publication bias, Begg's test (281) and Egger's test (282), were also carried out using STATA 10 software (Stata Corp, College Station TX, US), as they are not supported in RevMan 5.0.

3.2. Results

3.2.1. Study selection

Following the two phases of the study selection process, 12 studies were included in the review (267-271, 283-289). Figure 3.1 summarises the stages of the selection process. The paper identified from other sources (i.e. not through searches of bibliographic databases) was a study by Harrower *et al.* (290) cited by some of the older studies (267, 268, 270, 283). It was subsequently excluded because it used bacteriological count as an outcome, in accordance with the pre-specified exclusion criteria. A further two excluded studies (264, 291) did not use a clinical definition of SSI. Another study was excluded because the WEPD

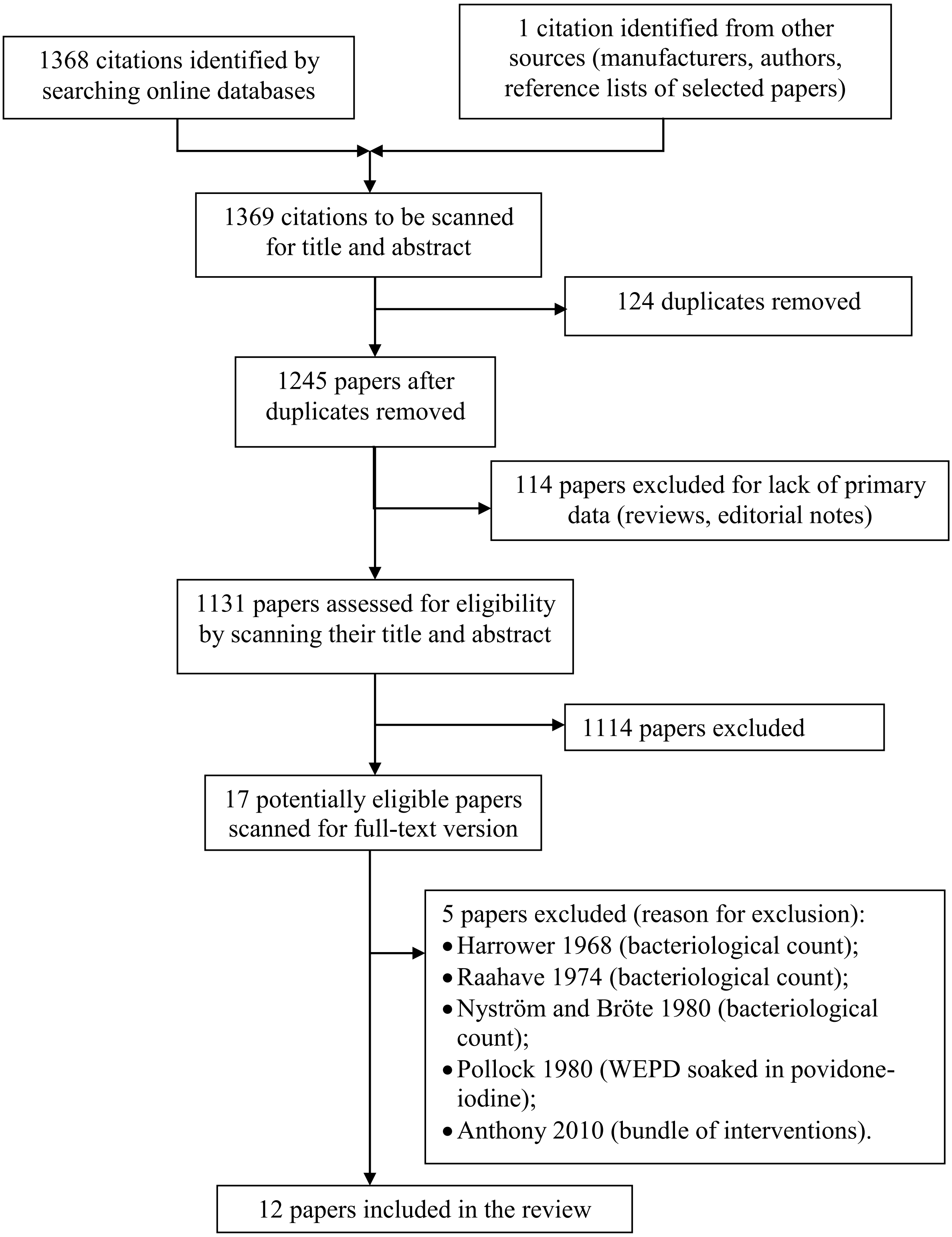

1368 citations identified by searching online databases
1 citation identified from other sources (manufacturers, authors, reference lists of selected papers)
1369 citations to be scanned for title and abstract
124 duplicates removed
1245 papers after duplicates removed
114 papers excluded for lack of primary data (reviews, editorial notes)
1131 papers assessed for eligibility by scanning their title and abstract
1114 papers excluded
17 potentially eligible papers scanned for full-text version
5 papers excluded (reason for exclusion):
• Harrower 1968 (bacteriological count);
• Raahave 1974 (bacteriological count);
• Nyström and Bröte 1980 (bacteriological count);
• Pollock 1980 (WEPD soaked in povidone-iodine);
• Anthony 2010 (bundle of interventions).
12 papers included in the review

Figure 3.1. PRISMA flowchart for systematic review of WEPD clinical effectiveness

was soaked in povidone-iodine (292), an antibacterial solution which would confound the effect of the WEPD according to its postulated mechanisms of action. Finally, one study was excluded because the intervention arm used a bundle of five interventions including a WEPD, therefore the study design did not allow an assessment of the individual effect of WEPDs (293).

Two included articles were not available in English: the German study of Batz *et al.* (285) and the French study of Brunet *et al.* (287), respectively. The full-text versions of these papers were analysed following translation into English. The authors of the most recently published studies (271, 288, 289) were contacted and asked whether they were aware of any other relevant published or unpublished studies. Only one of the authors (Horiuchi) responded and no further studies were identified. There were no disagreements between the two reviewers (AG and BF) with respect to the included studies and the extracted data.

3.2.2. Study characteristics

Table 3.1 summarises the characteristics of the 12 included studies and Table 3.2 presents the outcomes and the effect on SSI incidence reported in each study. Three studies were conducted in the UK, two in US and Ireland, respectively, and one each in Sweden, Germany, France, Japan and Australia.

Study design

Ten of the 12 included studies were RCTs and two were controlled trials (CTs) (267, 287). Two studies (267, 268) divided patients into three groups (two intervention groups and one control group) and the remaining ten studies compared two patient groups.

Patients

The 12 included studies reported data for a total of 1,933 patients. One paper (289) specified enrolment of patients over the age of 18 and another study (270) enrolled 'adults', but in the remaining studies patients' age was not reported as an inclusion/exclusion criterion.

Intervention and Control

In ten studies the WEPD used was identifiable by means of the description provided or by indicating the manufacturer or both. In the studies of Batz *et al.* (285) and Brunet *et al.* (287) no description was offered; in both these papers the WEPD was referred to as a 'ring drape'.

Only two of the 12 studies were multi-centre: one study (289) recruited from four hospitals and another study (270) recruited from two hospitals. Three studies (267, 268, 287) examined generic abdominal operations, one study (288) focused on appendicectomy and the remaining studies looked at gastrointestinal interventions, mostly colorectal surgery. In terms of the control group, two studies (288, 289) compared the WEPD against standard retraction and one study (285) compared the ring drape against incise drapes. As detailed in section 2.2, incise drapes are very similar to adhesive drapes (but different from 'ring drapes') in the sense that they do not come in contact with the abdominal cavity and with the wound edges. In the remaining studies the control group was described as the group where the WEPD was not used.

Table 3.1 Characteristics of the studies included in the systematic review

Study	Study type	Type of surgery	Number of patients	Inclusion/exclusion criteria	Intervention group	Control group
Maxwell 1969(267)	CT	elective or emergency major abdominal surgery	202: 82 intervention A 88 intervention R 32 control	Inclusion: based on type of surgery. No exclusion criteria reported.	Intervention A: plastic drape adherent to the skin; Intervention R: adherent plastic (as above) PLUS a circular plastic ring protector	"Towels were applied directly on the skin [...] and no plastic of any kind was used"
Alexander-Williams 1972(283)	RCT	midline or paramedian laparotomy associated with the opening of some part of the bowel or biliary tract	167: 84 intervention 83 control	Inclusion: based on type of surgery. No exclusion criteria reported.	"The impervious wound drapes used were vi-Drape (Parke-Davis). These are transparent plastic sheets having a central hole of 18, 23, or 28 cm diameter, around which is fixed a semi-rigid circular collar. This collar can be squeezed flat so that the central hole in the drape is introduced into the abdominal wound. When released the collar springs back to a circle beneath the abdominal wall, holding the plastic sheet in close proximity to the wound edge"	"either no wound protection or standard permeable cloth wound guards"
Psaila 1977(268)	RCT	abdominal surgery	144: 51 intervention A 46 intervention R 47 control	Inclusion: based on type of surgery. Exclusion: "Patients receiving preoperative antibiotics (with the exception of non-absorbable sulphonamides used for bowel preparation) were not included in the trial."	A: "The adhesive skin drape was Steri-Drape (Minnesota Mining and Manufacturing Co.)" PLUS standard linen towels; R: "Vi-Drape (Parke, Davis & Co.) was the plastic ring wound drape tested; this was placed through the wound itself, the ring being permitted to expand against the inner aspect of the abdominal wall and the drape being drawn over the wound surfaces" PLUS standard linen towels	"linen towels alone were used"
Gamble and Hopton 1984(284)	RCT	elective colonic surgery on one general surgical firm	56: 27 intervention 29 control	Inclusion: based on type of surgery. No exclusion criteria reported.	"The plastic ring drape consists of a flexible, semi-rigid plastic ring to the outer rim of which is welded a plastic sheet. The ring is compressed, inserted into the abdominal cavity and positioned under the abdominal wall inside the peritoneum. The plastic drape is smoothed out round the wound and clipped to the surrounding drapes, thus covering the edges of the incised abdominal wall, providing a barrier which should, in theory at least, reduce the risk of wound contamination"	"The ring drape was not used"

Study	Study type	Type of surgery	Number of patients	Inclusion/exclusion criteria	Intervention group	Control group

Nyström 1984(270)	RCT	elective colorectal surgery involving opening the bowel	140:
			70 intervention
			70 control
				Inclusion: adults;
				Preoperative exclusion: deferred surgery;
				Intraoperative exclusion: change of operative plans or an unforeseen therapeutic situation.
					"The drape is made of a polyvinyl plastic sheet with a central hole which is fitted with a plastic frame that can be adjusted to match the size of the incision (Op-drape, Triplus, Sweden). [...] The wound ring drape was adjusted to appropriate size and inserted into the abdomen before opening the bowel"	"without ring drape"

Batz 1987(285)	RCT	patients undergoing tumour resection for colorectal cancer	50: 25 intervention A 25 intervention B	Inclusion: based on the type of surgery. No exclusion criteria reported.	Ring drape	Incise drape
Redmond 1994(286)	RCT	gastrointestinal surgery	213: 102 intervention 111 control	Inclusion: based on the type of surgery. No exclusion criteria reported.	"wound edge protector"	"received no protection"
Brunet 1994(287)	CT	all interventions of abdominal surgery, elective and emergency	149: 73 intervention 76 control	Inclusion: based on the type of surgery. No exclusion criteria reported.	"champ à anneau", translated by the authors as "ring drape"	"no protection"
Sookhai 1999(269)	RCT	trans-abdominal surgery for GI disease	352: 170 intervention 182 control	Inclusion: based on the type of surgery. No exclusion criteria reported.	"This protector consists of an impermeable plastic drape with four adhesive patches that fits onto the abdomen. There is a hole in the middle with a semi-rigid plastic ring that fits into the abdominal wound and protects the wound edge from contact with viscera, visceral contents, contaminated instruments, and gloves"	"no wound-edge protector"

Study	Study type	Type of surgery	Number of patients	Inclusion/exclusion criteria	Intervention group	Control group
Horiuchi 2007(271)	RCT	non traumatic gastrointestinal surgery; laparoscopic surgery and minor surgery excluded	221: 111 intervention 110 control	No inclusion criteria reported. Exclusion: - patients who had severe adhesion with a history of laparotomy; - long-term use of steroids; - laparoscopic surgery or minor surgery such as appendectomy; - probable colon perforation.	"The Alexis retractor, a polyurethane wound retractor manufactured by Applied Medical"	"in the Without Alexis retractor group, a wound margin was left untreated"
Lee 2009(288)	RCT	open appendicectomy	109: 61 intervention 48 control	Inclusion: - clinical diagnosis of appendicitis; - planned open appendectomy; - and informed consent. Exclusion: - history of insulin-dependent diabetes; - and inability to follow-up owing to geographic location.	"Patients were than randomized [...] to receive intra-operative retraction with either standard retractors or the small (2.5-6cm) Alexis wound-protector system (Applied Medical, CA, USA). The Alexis wound retractor is a disposable plastic surgical retractor that provides 360 degrees retraction and wound protection for open procedures."	Standard retractor - see cell to the left
Reid 2010(289)	RCT	open elective colorectal resection	130: 64 intervention 66 control	Inclusion: patients older than 18 years. Exclusion: - patients who were cognitively impaired or otherwise unable to give informed consent; - and patients undergoing laparoscopic colorectal resection.	"This wound protector [Alexis - Applied Medical, CA, USA] is made up of 2 stiff rings with a cylinder between the 2 rings. The inner ring is placed in the peritoneal cavity, and the outer ring is placed outside of the abdomen. The outer ring is then rolled over the cylinder of impervious plastic until the plastic becomes taut circumferentially around the wound."	"in the control group wound retraction was achieved by retractors routinely used by the treating surgeon"

Abbreviations: CT - controlled trial; RCT - randomised controlled trial; SSI - surgical site infection; WEPD - wound-edge protection device.

Table 3.2 SSI incidence in the included studies

Study	Outcomes	SSI definition used	Time of assessment/ Length of follow-up	Effect on surgical site infection (SSI) incidence

Maxwell 1969(267)	Surgical site infection;
	Wound contamination;
	Various bacteriological outcomes
		"abnormal appearance of classical signs of inflammation in some area of the wound by the resident staff and a positive culture of wound exudate"	Unclear	Incidence of SSI:
				14.6% intervention A;
				18.2% intervention R;
				21.8% control.

Alexander-Williams 1972(283)	Surgical site infection; Wound complications	"...classifying the wound as showing either no infection, mild wound infection (erythema), moderate wound infection (exudate), or severe wound infection (pus)."	Initially at 3 and 7 days; then at 7 and 10 days	Incidence of SSI: 11.9% intervention; 12.0% control.

Psaila 1977(268)	Surgical site infection;
	Various bacteriological results
		"...at least one of the following criteria was used to identify the presence of infection: 1. Erythema around the sutures or along the wound edge with accompanying pyrexia. 2. Discharge of exudate or pus from the wound. 3. Wound breakdown."	Unclear	Incidence of SSI:
				16% intervention A;
				17% intervention R;
				21% control.

Gamble and Hopton 1984(284)	Surgical site infection; Various bacteriological results	"A wound was recorded as infected if a discharge occurred from it."	Unclear	Incidence of SSI: 32% intervention; 28% control.
Nyström 1984(270)	Surgical site infection; Bacteriological results	"Wound sepsis was defined as pus emptying spontaneously or upon incision."	Up to 30 days post-operatively	Incidence of SSI: 10% intervention; 9% control.
Batz 1987(285)	Surgical site infection; Bacteriological results	"A wound healing incident was defined as a spontaneous opening of the surgical abdominal wound with pus discharge."	Unclear	Incidence of SSI: 4% ring drape; 28% incise drape
Redmond 1994(286)	Surgical site infection;	"Wounds were deemed infected when there was overt pus or a culture-positive discharge."	At 5, 10 and 30 days post-operatively	Incidence of SSI: 10.8% intervention; 24.3% control.
Brunet 1994(287)	Surgical site infection; Bacteriological results	"Parietal infection was defined by the presence of pus at the wound level within a month following surgery"	One month post-operatively	Incidence of SSI: 8.2% intervention; 23.7% control
Sookhai 1999(269)	Surgical site infection;	"Postoperative wound infection was defined as the presence of a purulent discharge, a culture-positive discharge, pain/tenderness, localised swelling, erythema, or cellulitis which occurred within 30 days of surgery."	Up to 30 days post-operatively	Incidence of SSI: 13.5% intervention; 29.7% control.

Study	Outcomes	SSI definition used	Time of assessment/ Length of follow-up	Effect on surgical site infection (SSI) incidence

Horiuchi 2007(271)	Surgical site infection;
	Bacteriological results;
	Length of stay in hospital
		"SSI frequency and properties were analyzed according to the criteria of the United States Centres for Disease Control and Prevention (CDC)." (no reference provided)	Unclear	Incidence of SSI:
				7.2% intervention;
				14.5% control.

Lee 2009(288)	Surgical site infection;	"This [wound infection] was defined as any significant subcutaneous SSI necessitating wound opening or treatment with antibiotics. This also included any subject who was prescribed a separate course of antibiotics after discharge from hospital. All such events were coded as SSI."	Up to 3 weeks	Incidence of SSI: 1.6% intervention; 14.6% control.

Reid 2010(289)	Surgical site infection;
	Surgeons' satisfaction with the WEPD;
	Antibiotic usage;
	Length of stay
		"The principal outcome measure was the incidence of superficial or deep SSI occurring within 30 days of surgery, as defined by the Centres for Disease Control and Prevention."	At days 3 and 5 post-op and at discharge. Minimum follow-up 30 days post-operatively.	Incidence of SSI:
				4.7% intervention;
				22.7% control.

Outcomes

All the included trials pre-specified SSI as an outcome. However, there was considerable variation in how SSIs were defined. Only two papers (271, 289) referred to an internationally recognised definition of surgical infections, namely the CDC definition (200); in nine studies the authors used definitions of their own formulation (Table 3.2).

Most studies reported outcomes that had not been pre-specified in their Methods sections. With respect to this, seven studies reported various bacteriological outcomes, two studies (271, 289) reported the hospital length of stay associated with SSI and two studies (269, 288) estimated SSI-related costs. No studies reported patient quality of life as an outcome.

3.2.3. Risk of bias within the included studies

This section describes in detail the main sources of bias identified in the selected papers, discussed in the order of their publishing year. Table 3.3 summarises the risk of bias for the 12 included studies, assessed under the risk of bias tool presented in the Cochrane Handbook of Systematic Reviews of Interventions (7). Most of the studies were found to exhibit a high risk of bias.

Maxwell *et al.* (267) reported a trial where two interventions were evaluated at the study onset: a plastic drape vs. a plastic drape and a plastic ring protector used simultaneously. Patients were allocated alternately to these two initial study arms, which rendered the sequence generation clearly inadequate and allocation concealment unclear at best. The authors reported that a control group was introduced later, approximately half way through the trial, which explains the smaller number of patients compared to the two intervention groups. No information was given on how the allocation was done after the

Table 3.3 Risk of bias in the studies included in the systematic review (Cochrane 'Risk of bias' tool)

	Risk of bias category					
Study	**Adequate sequence generation**	**Allocation concealment**	**Blinding**	**Addressed incomplete outcome data**	**Free of selective reporting**	**Free of other bias**
Maxwell 1969(267)	No	Unclear	Unclear	No	Unclear	No
Alexander-Williams 1972(283)	Unclear	No	Unclear	No	Unclear	No
Psaila 1977(268)	Unclear	Unclear	Unclear	Yes	Unclear	No
Gamble and Hopton 1984(284)	Unclear	Unclear	Unclear	Unclear	Unclear	Unclear
Nyström 1984(270)	Unclear	Unclear	Unclear	No	Unclear	Unclear
Batz 1987(285)	Yes	Unclear	Unclear	No	Unclear	Unclear
Redmond 1994(286)	Unclear	Unclear	Unclear	Yes	No	No
Brunet 1994(287)	No	Unclear	Unclear	Unclear	Unclear	Unclear
Sookhai 1999(269)	Unclear	Unclear	Unclear	Yes	Unclear	No
Horiuchi 2007(271)	Unclear	Unclear	Unclear	Yes	Unclear	No
Lee 2009(288)	Yes	Unclear	Yes	No	Unclear	No
Reid 2010(289)	Yes	Unclear	Yes	No	Unclear	No

introduction of the control arm. Moreover, there was no mention regarding the blinding of the wound assessors. 16 patients were excluded from data analysis for death within ten days of surgery, but no indication is given on their allocation arm or their cause of death. The exclusion of this group of patients may potentially be significant because it accounted for approximately 6% of the total trial sample of studied cases (n=260). 'Wound contamination' and 'wound infection rate' were pre-specified outcomes, but other outcomes were reported as well, including various microbiological results and the influence of prophylactic antibiotic therapy on infection rate. The study is not free of other sources of bias: the time, frequency or length of follow-up are not specified, which makes it impossible to tell whether patients were reviewed at the same time intervals post-surgery and increases the concern over the influence of the 16 patients' exclusion discussed above.

In the trial reported by Alexander-Williams *et al.* (283) the study personnel apparently had unrestricted access to the patient allocation scheme, given that treatment allocation was recorded on a form (reproduced in the original paper) together with other patient identification data. Patients were randomised to either the intervention or control arm in blocks of ten, but no details were given as to how the randomisation sequence was generated. Measures intended to ensure the blinding of the outcome assessor (the bacteriologist) were presented, but no information was provided as to whether the patients themselves were aware of the allocation arm. Three patients were excluded from the analysis for having died within 24 hours of surgery, but no information was given about the treatment arm these patients belonged to or the cause of death. Incidence of wound infection was the only pre-specified outcome, but non-infectious wound complications were also reported.

The study is not free of other sources of bias: firstly, the first 96 patients were reviewed at days three and seven post-operatively and the remainder of the patients were reviewed at seven and ten days post-operatively. This was justified by the fact that no infection was ever recorded at day three. The short follow-up in general and this alteration in particular may have led to missing SSIs that occurred after the seven days post-surgery. Secondly, the presence of SSI could not be assessed with certainty in ten patients (five in each group), and a brief description of each case was reported as a 'wound complication'. With respect to the ten wound complications mentioned above, the authors claimed that *"even if some or all of them were included with the wound-infection patients the results would not materially be affected"* (p.145). While this statement is correct, it raises the question of whether the SSI definition was accurate enough. This hypothesis is further supported by the authors' initial intention to perform a stratified analysis by severity of SSI, which was ultimately abandoned and all wound-infections were pooled together. This may be explained by the low incidence of SSI reported in the study and may suggest data-driven analysis. Finally, a potential source of bias in this study comes from the fact that the manufacturer of the WEPD provided assistance in the design and sequence generation of the trial.

In the three-arm study of Psaila *et al.* (268) patients were reportedly randomised to the two intervention arms, but no information was given regarding sequence generation or the randomisation in the control group. Neither allocation concealment nor blinding could be ascertained based on the available information. No patients were lost to follow-up. Although wound infection was the only pre-specified outcome, microbiological results investigating the correlation between skin and drape contamination, respectively, and wound infection are presented. Another source of bias emerges from the failure to report the length of follow-up:

wounds were reviewed daily starting with the third day post-operatively, but it is not clear whether patients left the study after the same number of days following surgery or not.

In the RCT reported by Gamble and Hopton (284) sequence generation and allocation concealment could not be established due to lack of available information. No indication was given about the blinding of patients or of the outcome assessors, although the wound review process was appropriately described. Outcome data were reported for all patients enrolled in the study. Wound infection rate was the only pre-specified outcome; microbiological results were also reported.

Nyström *et al.* (270) reported the results of an RCT comparing a wound ring drape with standard care, but provided no information concerning sequence generation or allocation concealment. Furthermore, the authors specified the possibility of excluding patients after randomisation and stated that *"the postoperative course of the remaining patients is accounted for"* (p. 451). The number of patients excluded as such was not reported. This suggestion of per-protocol analysis is indeed confirmed: *"one hundred forty patients were treated according to the protocol"* (p.452). No indication of blinded outcome assessments or patient blinding was given. Although the exact number of patients in the intervention arm for whom this was the case was not specified, based on the authors' discussion this may amount to seven; given that the overall results reported seven SSI cases in the intervention arm (n=70) compared to six SSI cases in the control group (n=70), it can be argued that exclusion of these patients from data analysis due to protocol violation may well influence the trial's result.

Batz *et al.* (285) appropriately reported to have used a computer-generated randomisation sequence, but making a judgement on allocation concealment could not be made based on the published manuscript. The blinding of wound reviewers could not be evaluated. Patients were excluded from the study if they died after surgery but no supplementary information was given on the infection status of these excluded patients, therefore incomplete outcome data was not properly addressed. The authors apparently used the terms 'wound infection' and 'wound healing incident' interchangeably. No details were given on the number and frequency of wound assessments, or on the length of follow-up.

The RCT reported by Redmond *et al.* (286) did not accurately describe either sequence generation or allocation concealment measures. Blinded outcome assessment was ensured, but no mention was made on patient blinding. Data were reported for all the patients enrolled in the study. The number of SSI cases per treatment arm, stratified by degree of contamination, was reported but the wound infection rate (the pre-specified study outcome) was not calculated. While the definition for the SSI was explicit, when a subgroup analysis by degree of contamination was conducted no definitions of the three considered categories were given. A particular concern arose with respect to this study: the authorship list, the intervention and outcomes considered and the results' format were strikingly similar with the study of Sookhai *et al.* (269), published five years later (discussed below). This study's cohort is 139 patients smaller than reported by the latter. Two competing hypotheses were generated: either Redmond *et al.* (286) published interim results of the study reported by Sookhai and colleagues; or Sookhai *et al.* (269) conducted an original, larger study based on the smaller cohort study of Redmond *et al.*. The authors of both studies were contacted in order to elude this controversy, but no response was received. The study reported by Redmond *et al.* (286) was ultimately included in the analysis as independent research, but it must be noted that its originality can be questioned.

In the study reported by Brunet *et al.* (287) patients were allocated to the study arms based on an odd day/even day scheme, which renders sequence generation inadequate. No information was provided concerning allocation concealment or blinded outcome assessment.

Two patients in the intervention arm were excluded from the study because the WEPD could not be used, which suggests per-protocol analysis. The definition used for an SSI was explicit, but the statistic employed to examine the difference between groups was not clearly defined and thus result reporting was ambiguous. Bacteriological results, the influence of surgeon qualification on SSI incidence and inpatient length of stay were also presented, although not pre-specified.

Sookhai *et al.* (269) gave no information on sequence generation or allocation concealment. Wound reviewers were said to have been 'independent', which makes blinding unclear. Data were reported for all the patients enrolled in the study. The length of follow-up was indicated, but no mention was made regarding the timing and frequency of wound reviews.

Horiuchi *et al.* (271) did not give any information on sequence generation or allocation concealment. Outcome assessment was blinded, but no reference was made regarding patient blinding. The pre-specified outcome (SSI rate) was reported appropriately. A subgroup analysis of SSI rate by surgery site (colorectal and gastric surgery, respectively) was presented, but complete results with significance levels were presented only for colorectal surgery. Another source of bias stems from the failure to specify the length of follow-up as well as the timing and frequency of the wound reviews; as such, it is not clear whether patients in the two trial arms were assessed at comparable time points.

In the RCT reported by Lee *et al.*(288) patients were randomised using a computer-generated allocation sequence and were unaware of their treatment arm. Allocation concealment, however, was not made explicit. Four patients were lost to follow-up and they were excluded from the analysis without further details. No information was given on these patients' allocation arm. Most importantly, the trial was discontinued for early evidence of benefit, although the interim analysis which triggered this decision did not appear to be pre-specified.

Reid *et al.* (289) reported a RCT where patients were randomised in blocks of 20 by a computer-generated sequence. Allocation concealment was unclear because *"opaque envelopes opened at surgery by a third party"* were used, but there is no indication whether the envelopes were sealed or how 'third party' should be interpreted. Five randomised patients were excluded from the analysis: two of these exclusions were due to patient death, but no mention was made on the time of death relative to the surgery or the cause of death. Another source of bias resides in the unequal balance between the two patient groups in terms of mean body mass index (BMI), a known risk factor for SSI: patients in the control group had a significantly higher BMI compared to controls; subsequently, SSI rate was unexpectedly high in the control group.

Based on the analysis presented above, several key points can be made regarding the risk of bias in the selected papers. First, most categories in the risk of bias tool were judged as 'unclear' for lack of relevant information available in the full-text versions of the articles. This was mainly due to reporting failures therefore a straightforward quality assessment verdict was often impossible to reach. Second, when information was available the categories subject to bias most often were 'sequence generation' and 'incomplete outcome data'. Third, eight out of 12 studies were susceptible to biases falling in the 'Other sources of bias' category. Common shortcomings included not specifying either the length of follow-up or the timing and frequencies of the wound assessments and not reporting funding sources and competing interests.

Publication bias

The risk of publication bias was examined using a funnel plot (Figure 3.2). The slight asymmetry of the plot is due to the two studies of Batz *et al.* (285) and Lee *et al.* (288), which clearly favoured WEPDs and had some of the lowest sample sizes among all included studies (n=50 and n=109, respectively). Neither Begg's test (p=0.631, continuity corrected) nor Egger's test (p=0.242) were statistically significant in suggesting publication bias.

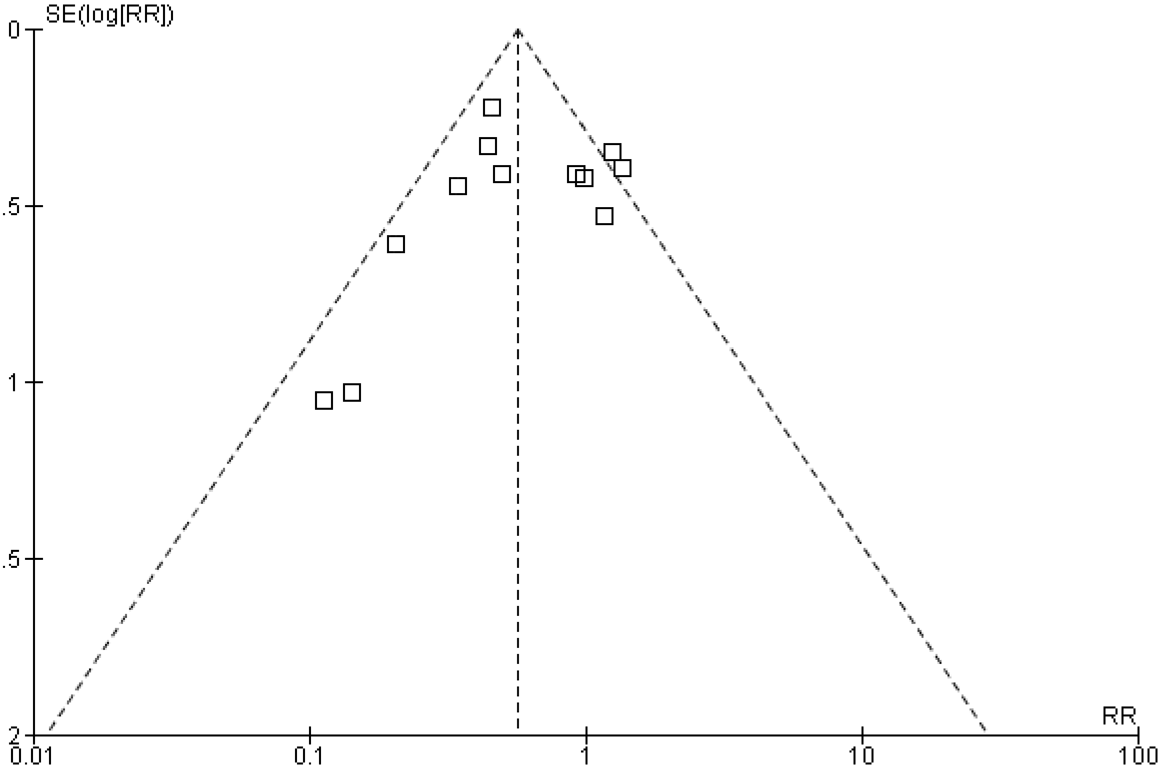


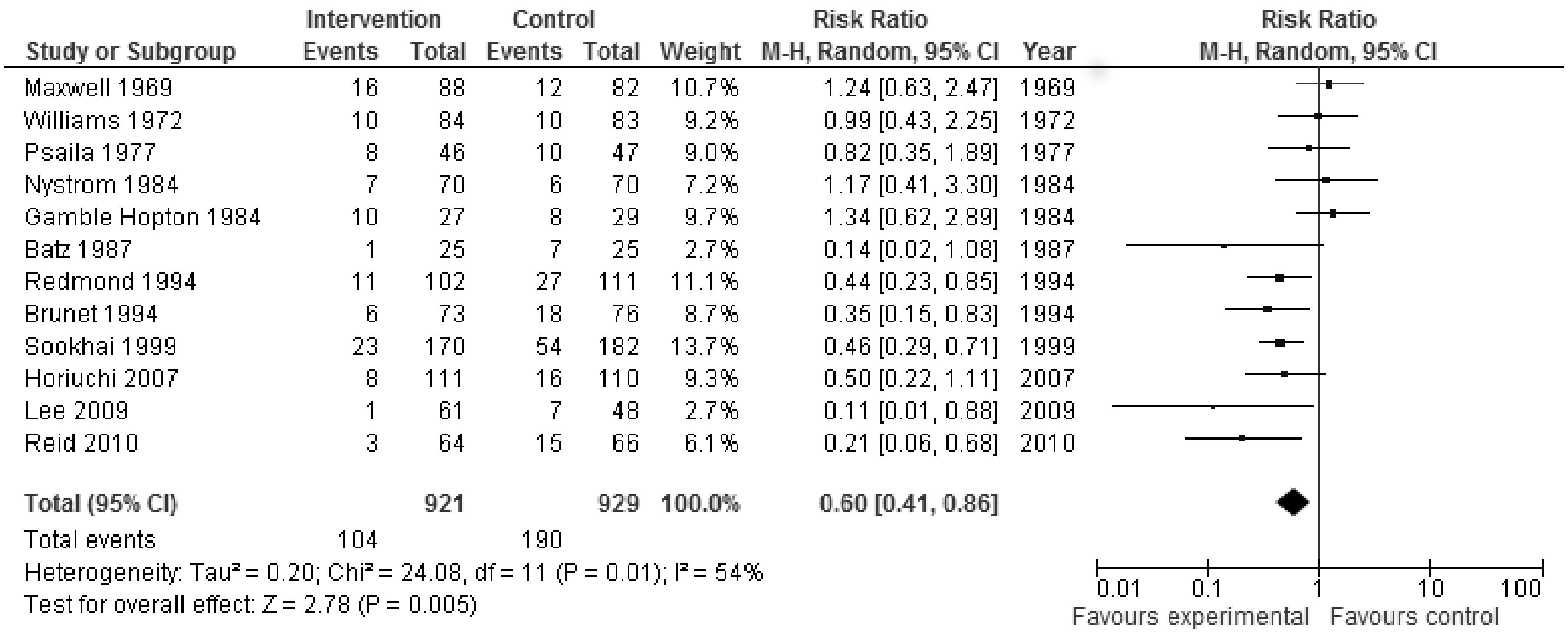
Figure 3.2 Funnel plot of the studies included in the WEPD systematic review

3.2.4. Results of individual studies and pooled results

All 12 studies had a medium or high risk of bias and none of them was judged to be of sufficient quality to be formally included in a meta-analysis. Given the lack of robust evidence and the contradictory results of the included studies, an exploratory meta-analysis was performed based on primary data from the 12 studies. The objective of this quantitative analysis was to provide an indication on the estimate of the effectiveness of WEPDs in reducing SSI rates.

The individual risk ratios and the 95% confidence intervals for the random-effects model meta-analysis are presented in Figure 3.3. The five included studies published prior to 1990 did not show a statistically significant benefit associated with the use of WEPDs. The remaining studies clearly favoured WEPDs, reporting a statistically significant benefit with the exception Horiuchi *et al.* (271) (RR 0.50, 95% CI 0.22 to 1.11). Lee *et al.* (288) reported the most favourable result for the use of WEPDs (RR 0.11, 95% CI 0.01 to 0.88).

The pooled risk ratio was 0.60 (95% CI 0.41 to 0.86) under a random-effects model. When a fixed-effects model was used, the pooled risk ratio was 0.56 (95% CI 0.45 to 0.70). These results suggest that the WEPDs appear to reduce the incidence of SSI when compared with standard care.



	Intervention		Control			Risk Ratio	
Study or Subgroup	Events	Total	Events	Total	Weight	M-H, Random, 95% CI	Year
Maxwell 1969	16	88	12	82	10.7%	1.24 [0.63, 2.47]	1969
Williams 1972	10	84	10	83	9.2%	0.99 [0.43, 2.25]	1972
Psaila 1977	8	46	10	47	9.0%	0.82 [0.35, 1.89]	1977
Nystrom 1984	7	70	6	70	7.2%	1.17 [0.41, 3.30]	1984
Gamble Hopton 1984	10	27	8	29	9.7%	1.34 [0.62, 2.89]	1984
Batz 1987	1	25	7	25	2.7%	0.14 [0.02, 1.08]	1987
Redmond 1994	11	102	27	111	11.1%	0.44 [0.23, 0.85]	1994
Brunet 1994	6	73	18	76	8.7%	0.35 [0.15, 0.83]	1994
Sookhai 1999	23	170	54	182	13.7%	0.46 [0.29, 0.71]	1999
Horiuchi 2007	8	111	16	110	9.3%	0.50 [0.22, 1.11]	2007
Lee 2009	1	61	7	48	2.7%	0.11 [0.01, 0.88]	2009
Reid 2010	3	64	15	66	6.1%	0.21 [0.06, 0.68]	2010
Total (95% CI)		**921**		**929**	**100.0%**	**0.60 [0.41, 0.86]**	
Total events	104		190				

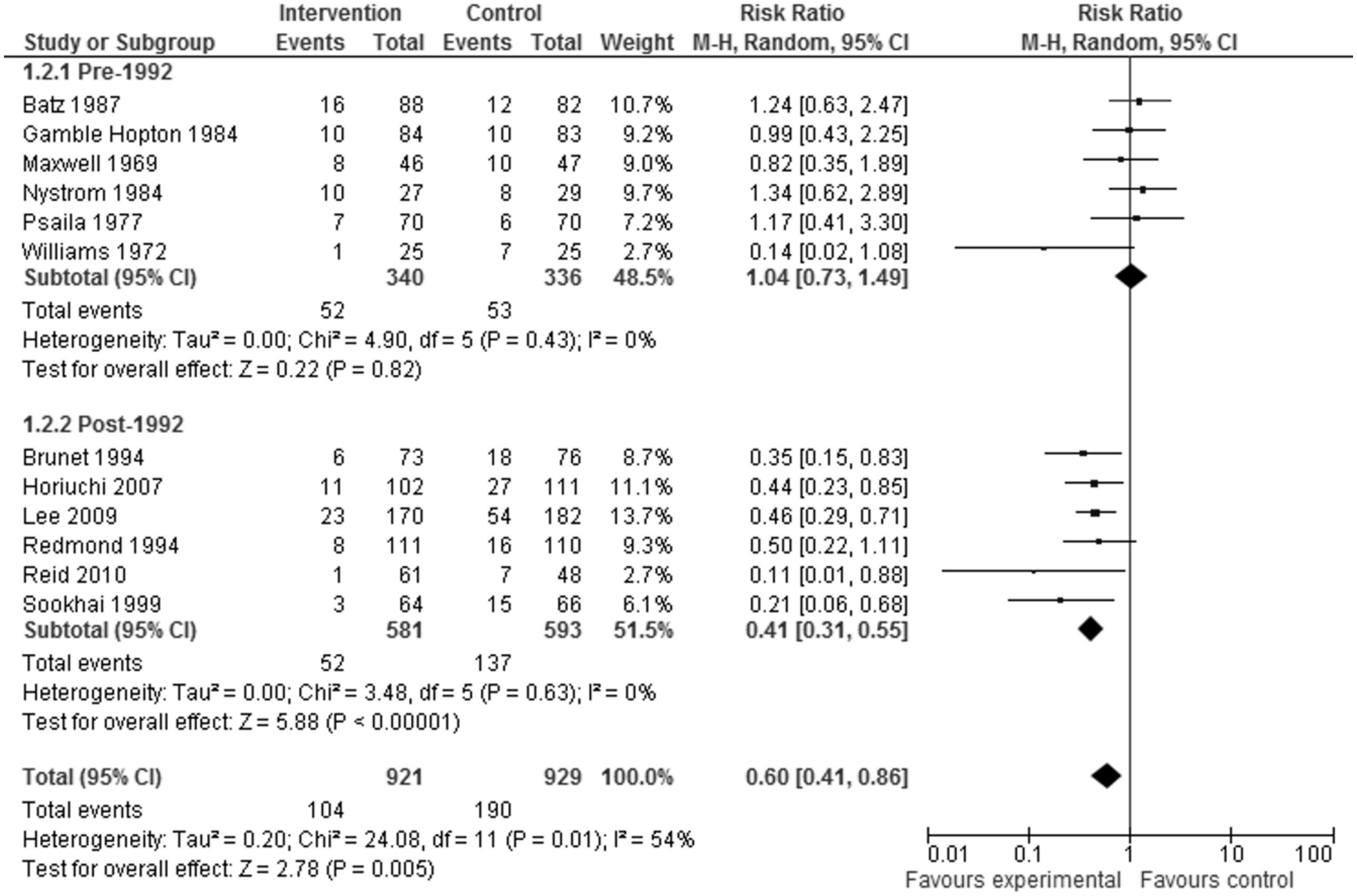
Heterogeneity: Tau² = 0.20; Chi² = 24.08, df = 11 (P = 0.01); I² = 54%

Test for overall effect: Z = 2.78 (P = 0.005)

Figure 3.3 Summary data, individual and pooled effect estimates for the studies included in the WEPD meta-analysis

3.2.5. Heterogeneity

Between-study heterogeneity was also assessed: the value of the I^2 statistic was 54% (P=0.01), suggesting statistically significant moderate heterogeneity (7). The source of heterogeneity was further explored by conducting two subgroup analyses: the first analysis grouped the studies in two categories according to their year of publishing i.e. pre-1992 and post-1992 studies. This analysis investigates the potential influence of the investigators' awareness of the CDC definition of SSI (published in 1992) on specifying SSI definitions. Of course, other factors may also have changed over time. Figure 3.4 presents the forest plots for the two subgroups together with an evaluation of between-study heterogeneity. The low values of the I^2 statistics suggest that both pre-1992 and post-1992 studies are highly homogeneous. The pooled RR suggest that pre-1992 studies are consistent in showing no benefit associated with WEPDs (RR 1.04, 95% CI 0.73 to 1.49), while post-1992 studies are largely consistent in demonstrating strong benefit (RR 0.41, 95% CI 0.31 to 0.55).



	Intervention		Control			Risk Ratio	Risk Ratio
Study or Subgroup	Events	Total	Events	Total	Weight	M-H, Random, 95% CI	M-H, Random, 95% CI
1.2.1 Pre-1992							
Batz 1987	16	88	12	82	10.7%	1.24 [0.63, 2.47]	
Gamble Hopton 1984	10	84	10	83	9.2%	0.99 [0.43, 2.25]	
Maxwell 1969	8	46	10	47	9.0%	0.82 [0.35, 1.89]	
Nystrom 1984	10	27	8	29	9.7%	1.34 [0.62, 2.89]	
Psaila 1977	7	70	6	70	7.2%	1.17 [0.41, 3.30]	
Williams 1972	1	25	7	25	2.7%	0.14 [0.02, 1.08]	
Subtotal (95% CI)		**340**		**336**	**48.5%**	**1.04 [0.73, 1.49]**	
Total events	52		53				
Heterogeneity: Tau² = 0.00; Chi² = 4.90, df = 5 (P = 0.43); I² = 0%							
Test for overall effect: Z = 0.22 (P = 0.82)							
1.2.2 Post-1992							
Brunet 1994	6	73	18	76	8.7%	0.35 [0.15, 0.83]	
Horiuchi 2007	11	102	27	111	11.1%	0.44 [0.23, 0.85]	
Lee 2009	23	170	54	182	13.7%	0.46 [0.29, 0.71]	
Redmond 1994	8	111	16	110	9.3%	0.50 [0.22, 1.11]	
Reid 2010	1	61	7	48	2.7%	0.11 [0.01, 0.88]	
Sookhai 1999	3	64	15	66	6.1%	0.21 [0.06, 0.68]	
Subtotal (95% CI)		**581**		**593**	**51.5%**	**0.41 [0.31, 0.55]**	
Total events	52		137				
Heterogeneity: Tau² = 0.00; Chi² = 3.48, df = 5 (P = 0.63); I² = 0%							
Test for overall effect: Z = 5.88 (P < 0.00001)							
Total (95% CI)		**921**		**929**	**100.0%**	**0.60 [0.41, 0.86]**	
Total events	104		190				
Heterogeneity: Tau² = 0.20; Chi² = 24.08, df = 11 (P = 0.01); I² = 54%							
Test for overall effect: Z = 2.78 (P = 0.005)							

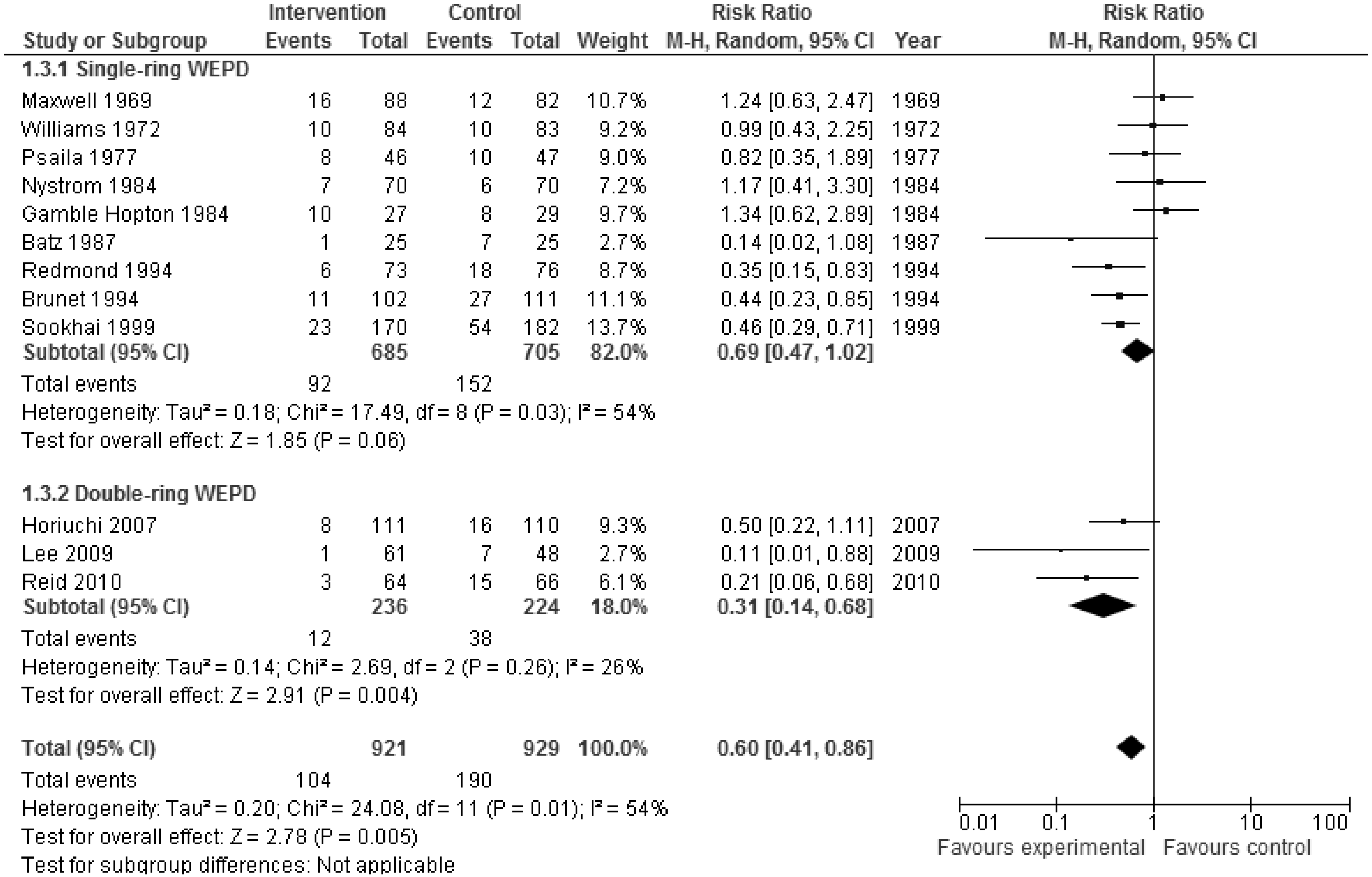
Figure 3.4 Exploratory subgroup analysis in WEPD systematic review - pooled effect estimates by year of publication

The second subgroup analysis grouped the included studies according to the design of the WEPD: most studies used the one ring design, while only the three most recent studies reportedly used the two-ring design. This analysis thus accounts for the potential influence of variations within the intervention under investigation. Figure 3.5 depicts the two forest plots and the values of the I^2 statistic suggest that both subgroups have some degree of heterogeneity, especially the trials using the one-ring design (I^2=54%). Although the two-ring design appears to have a stronger beneficial effect (RR 0.31, 95% CI 0.14 to 0.68) than the single-ring design (RR 0.69, 95% CI 0.47 to 1.02), other unexplored factors may contribute to between-study heterogeneity and are discussed below.

3.2.6. Pre-specified subgroup analyses

A pre-specified subgroup analysis was conducted to investigate the influence of the degree of contamination on SSI incidence. This analysis included only the studies where an explicit differentiation of different degrees of contamination was made by the authors: the 12 included papers described with variable amount of detail the type of surgery performed and assigning a degree of contamination based on partial descriptions would have led to unreliable results.

Four studies reported data on the relationship between SSI incidence and the surgical degree of contamination (268, 269, 286, 287). The descriptions of the degrees of contamination were neither identical nor completely consistent in these four papers: Psaila *et al.* (268) and Brunet *et al.* (287) used their own categorisations, while Redmond *et al.* (286) did not indicate any definition at all. Nevertheless, it is clear that all four papers referred to increasing levels of wound contamination.



Study or Subgroup	Intervention Events	Intervention Total	Control Events	Control Total	Weight	Risk Ratio M-H, Random, 95% CI	Year
1.3.1 Single-ring WEPD							
Maxwell 1969	16	88	12	82	10.7%	1.24 [0.63, 2.47]	1969
Williams 1972	10	84	10	83	9.2%	0.99 [0.43, 2.25]	1972
Psaila 1977	8	46	10	47	9.0%	0.82 [0.35, 1.89]	1977
Nystrom 1984	7	70	6	70	7.2%	1.17 [0.41, 3.30]	1984
Gamble Hopton 1984	10	27	8	29	9.7%	1.34 [0.62, 2.89]	1984
Batz 1987	1	25	7	25	2.7%	0.14 [0.02, 1.08]	1987
Redmond 1994	6	73	18	76	8.7%	0.35 [0.15, 0.83]	1994
Brunet 1994	11	102	27	111	11.1%	0.44 [0.23, 0.85]	1994
Sookhai 1999	23	170	54	182	13.7%	0.46 [0.29, 0.71]	1999
Subtotal (95% CI)		**685**		**705**	**82.0%**	**0.69 [0.47, 1.02]**	
Total events	92		152				
Heterogeneity: Tau² = 0.18; Chi² = 17.49, df = 8 (P = 0.03); I² = 54%							
Test for overall effect: Z = 1.85 (P = 0.06)							
1.3.2 Double-ring WEPD							
Horiuchi 2007	8	111	16	110	9.3%	0.50 [0.22, 1.11]	2007
Lee 2009	1	61	7	48	2.7%	0.11 [0.01, 0.88]	2009
Reid 2010	3	64	15	66	6.1%	0.21 [0.06, 0.68]	2010
Subtotal (95% CI)		**236**		**224**	**18.0%**	**0.31 [0.14, 0.68]**	
Total events	12		38				
Heterogeneity: Tau² = 0.14; Chi² = 2.69, df = 2 (P = 0.26); I² = 26%							
Test for overall effect: Z = 2.91 (P = 0.004)							
Total (95% CI)		**921**		**929**	**100.0%**	**0.60 [0.41, 0.86]**	
Total events	104		190				
Heterogeneity: Tau² = 0.20; Chi² = 24.08, df = 11 (P = 0.01); I² = 54%							
Test for overall effect: Z = 2.78 (P = 0.005)							
Test for subgroup differences: Not applicable							

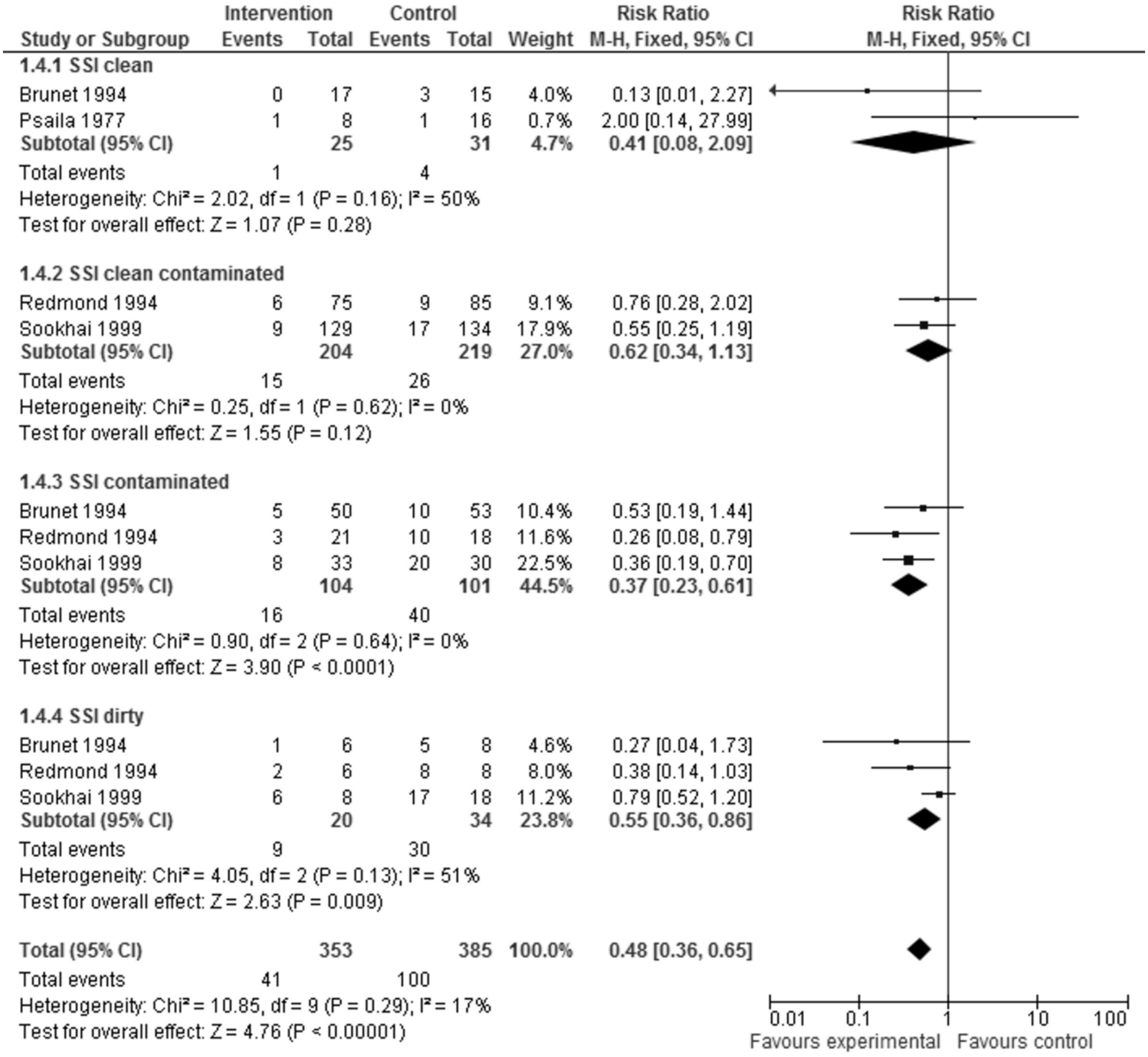
Figure 3.5 Exploratory subgroup analysis in WEPD systematic review - pooled effect estimates by type of WEPD design

An exploratory subgroup analysis was conducted in order to investigate the effect of WEPDs adjusted for the type of wound (clean/clean contaminated/contaminated/dirty). Results of this analysis are summarised in Figure 3.6. It is apparent that the use of WEPDs is associated with a statistically significant reduction in the risk of developing an SSI in patients undergoing contaminated (RR 0.37, 95% CI 0.23 to 0.61) and dirty surgery (RR 0.55, 95% CI 0.36 to 0.86). The point estimates also suggest a beneficial effect in clean (RR 0.41, 95% CI 0.08 to 2.09) and clean contaminated surgery (RR 0.62, 95% CI 0.34 to 1.13), but the CIs are wide.

3.3. Discussion

Summary of findings

12 prospective studies reporting primary data from 1,933 patients were included in the review. The quality assessment of these studies found them all to be at a significant risk of bias and six of them failed to address appropriately more than one risk category. Additionally, there was little consistency among studies with respect to the intervention used or the definition of an SSI. The type of WEPD used was not always accurately described. It appears that the device has displayed two different designs over time: the 'traditional design', identified up to the 1999 trial of Sookhai *et al.*(269), featured a plastic ring inserted in the abdomen and a large plastic drape emerging from it which covers the wound margins and extends out over the surrounding area; this device was also referred to as a 'ring drape' (268, 270, 285, 287). An alternative design was described in detail by Lee *et al.* (288) and appears to have been used in the three post-2005 studies (271, 288, 289): it had two plastic rings and also had retraction properties, therefore it was marketed under the name 'wound retractor'. These two designs have yet to be compared against each other.



	Intervention		Control			Risk Ratio
Study or Subgroup	Events	Total	Events	Total	Weight	M-H, Fixed, 95% CI
1.4.1 SSI clean						
Brunet 1994	0	17	3	15	4.0%	0.13 [0.01, 2.27]
Psaila 1977	1	8	1	16	0.7%	2.00 [0.14, 27.99]
Subtotal (95% CI)		**25**		**31**	**4.7%**	**0.41 [0.08, 2.09]**
Total events	1		4			
Heterogeneity: Chi² = 2.02, df = 1 (P = 0.16); I² = 50%						
Test for overall effect: Z = 1.07 (P = 0.28)						
1.4.2 SSI clean contaminated						
Redmond 1994	6	75	9	85	9.1%	0.76 [0.28, 2.02]
Sookhai 1999	9	129	17	134	17.9%	0.55 [0.25, 1.19]
Subtotal (95% CI)		**204**		**219**	**27.0%**	**0.62 [0.34, 1.13]**
Total events	15		26			
Heterogeneity: Chi² = 0.25, df = 1 (P = 0.62); I² = 0%						
Test for overall effect: Z = 1.55 (P = 0.12)						
1.4.3 SSI contaminated						
Brunet 1994	5	50	10	53	10.4%	0.53 [0.19, 1.44]
Redmond 1994	3	21	10	18	11.6%	0.26 [0.08, 0.79]
Sookhai 1999	8	33	20	30	22.5%	0.36 [0.19, 0.70]
Subtotal (95% CI)		**104**		**101**	**44.5%**	**0.37 [0.23, 0.61]**
Total events	16		40			
Heterogeneity: Chi² = 0.90, df = 2 (P = 0.64); I² = 0%						
Test for overall effect: Z = 3.90 (P < 0.0001)						
1.4.4 SSI dirty						
Brunet 1994	1	6	5	8	4.6%	0.27 [0.04, 1.73]
Redmond 1994	2	6	8	8	8.0%	0.38 [0.14, 1.03]
Sookhai 1999	6	8	17	18	11.2%	0.79 [0.52, 1.20]
Subtotal (95% CI)		**20**		**34**	**23.8%**	**0.55 [0.36, 0.86]**
Total events	9		30			
Heterogeneity: Chi² = 4.05, df = 2 (P = 0.13); I² = 51%						
Test for overall effect: Z = 2.63 (P = 0.009)						
Total (95% CI)		**353**		**385**	**100.0%**	**0.48 [0.36, 0.65]**
Total events	41		100			
Heterogeneity: Chi² = 10.85, df = 9 (P = 0.29); I² = 17%						
Test for overall effect: Z = 4.76 (P < 0.00001)						

Figure 3.6 Exploratory subgroup analysis in WEPD systematic review - pooled effect estimates by type of surgery

The inclusion/exclusion criteria were often not specified (Table 3.1): only two studies (288, 289) clearly described inclusion and exclusion criteria for the recruited patients; three other studies (268, 270, 271) specified exclusion criteria only; and the seven remaining studies did not give any such information. This makes it impossible to accurately assess the risk of SSI in recruited patients in most of the included studies.

Definitions of SSI varied greatly and belonged to the authors themselves with two exceptions (271, 289). This is in line with the finding of Bruce *et al. (202)*, whose systematic review of prospective studies on postoperative wound infection published between 1993 and 1997 revealed 41 distinct definitions of surgical wound infection. In addition to the diversity of SSI definitions, the number of SSI assessments and the follow-up period were either unclear or inconsistent (Table 3.2). When indicated, the length of follow-up was generally within 30 days post-operatively (269, 270, 286, 289).

The comparator was not consistent throughout the included studies: the control group received either no protection (269, 270, 284, 286), standard retraction (288, 289) or towels applied to the skin (267, 268). Moreover, the studies looked at different types of abdominal surgery and information about the degree of contamination, an accepted intra-operative risk factor for SSI, was available only in four studies (268, 269, 286, 287).

A large number of risk factors for SSI are known. Patient related risk factors include age, obesity, smoking status, diabetes, and underlying illnesses (243). Perioperative risk factors include appropriate preoperative antibiotic prophylaxis, mechanical bowel preparation with oral antimicrobials or not, duration of the procedure, and intraoperative blood transfusion (200). As previously described (sub-section 2.1.4), the NNIS SSI risk index is an internationally recognised predictor for SSI which incorporates simultaneously three important risk factors: the American Society of Anaesthesiologists score, the wound class and the duration of surgery (217). Table 3.4 summarises the extent to which such risk factors were assessed in the included studies.

Three methods of accounting for variability in the risk of SSI were considered: perioperative measures and inclusion/exclusion criteria; stratified randomisation and trial arm comparability; and stratified analysis of SSI incidence. Few studies adequately addressed risk factors when reporting the results. The most commonly mentioned topics under perioperative measures were skin preparation and antibiotic prophylaxis. No study reported patient randomisation stratified by risk. Four studies did not report any measure of comparability between trial arms (267, 268, 270, 283) and thus it cannot be ascertained whether randomisation was effective, while more recent studies presented results so as to allow comparing the trial arms with respect to an increasing number of risk factors. One study adjusted the results of SSI incidence for average length of inpatient stay (289), two studies adjusted for the site of surgery (271, 283) and four studies adjusted for the degree of wound contamination (268, 269, 286, 287). The remaining five studies did not stratify for any risk factor in their analyses. While reporting of risk factors appears to have improved with time, the lack of information on older studies makes it difficult to assess the validity of the results. This limitation was not due exclusively to the lack of available evidence at the time of publishing: for instance, Lee *et al.* (288) reported to have recorded operative time for each patient, but did not report whether patients in the two trial arms were comparable with respect to this parameter. Body weight was first acknowledged as a risk factor in the included studies in 1987 (285), but was neglected in the 1990's studies (269, 286, 287), only to be considered again (as body mass index) in the three most recent papers. As with the inclusion/exclusion criteria discussed previously, not controlling appropriately for risk factors makes it difficult to assess the real effect of WEPDs.

Table 3.4 Reporting and controlling for surgical site infection (SSI) risk factors in the WEPD systematic review studies

Study	Type of surgery	Reported information on SSI risk factors		
		Risk factors addressed via peri-operative measures or inclusion/exclusion criteria	**Risk factors addressed via stratified randomisation or reporting comparability between trial arms**	**Risk factors adjusted for in the analysis of SSI incidence**
Maxwell 1969(267)	elective or emergency major abdominal surgery	- the operative site was shaved during the morning of surgery; - standard skin disinfection;	None reported	None reported
Alexander-Williams 1972(283)	midline or paramedian laparotomy associated with the opening of some part of the bowel or biliary tract	- surgery type showed a high risk of wound contamination; - pre-operative antibiotic regimen was not standard;	None reported	Site of surgery
Psaila 1977(268)	abdominal surgery	- standard skin disinfection; - "a standard two-layer method of wound closure, using continuous chromic catgut and monofilament nylon, was employed in the majority of cases"; - an adhesive dressing was applied over each wound; - patients who received any other pre-operative antibiotics apart from sulphonamides for bowel preparation were excluded.	None reported	Degree of contamination (clean)
Gamble and Hopton 1984(284)	elective colonic surgery on one general surgical firm	- standard bowel preparation (metronidazole, ampicillin and neomycin); - standard skin disinfection;	Trial arms were reported to be comparable with respect to age and sex	None reported
Nyström 1984(270)	elective colorectal surgery involving opening the bowel	- no bowel preparation with antimicrobials; - antibiotic prophylactic regimen (either doxycycline or tinidazole); - all wounds were closed;	None reported	None reported

Study	Type of surgery	Reported information on SSI risk factors		
		Risk factors addressed via peri-operative measures or inclusion/exclusion criteria	Risk factors addressed via stratified randomisation or reporting comparability between trial arms	Risk factors adjusted for in the analysis of SSI incidence
Batz 1987(285)	patients undergoing tumour resection for colorectal cancer	- antibiotic prophylactic regimen (cephalosporin); - all wounds were closed primarily;	Trial arms were reported to be comparable with respect to: age, body weight and tumour stage	None reported
Redmond 1994(286)	gastrointestinal surgery	- standardised antibiotic prophylaxis; - standardised skin preparation;	Trial arms were reported to be comparable with respect to: age, sex, anaesthesia and operating time	Degree of wound contamination (clean contaminated, contaminated, dirty)
Brunet 1994(287)	all interventions of abdominal surgery, elective and emergency	- standard skin preparation (site shaving and skin disinfection); - antibiotic prophylactic regimen (cephalosporin) was given only to patients undergoing colorectal surgery;	Trial arms were reported to be comparable with respect to: age, length of surgery, elective/emergency surgery and degree of wound contamination	Degree of wound contamination (clean, contaminated, dirty)
Sookhai 1999(269)	trans-abdominal surgery for GI disease	- standardised antibiotic prophylaxis; - standardised skin preparation;	Trial arms were reported to be comparable with respect to: smoking status, pre-operative hospital stay, mean operation time, intraoperative temperature and number of blood units transfused	Degree of wound contamination (clean contaminated, contaminated, dirty)
Horiuchi 2007(271)	non traumatic gastrointestinal surgery; laparoscopic surgery and minor surgery excluded	- antibiotic prophylactic regimen different for upper-gastrointestinal surgery (ampicillin and cefazolin or flomoxef) and colorectal surgery (cefotiam, flomoxef or cefmetazol); - excluded patients with long-term steroid use;	Trial arms were reported to be comparable with respect to: sex, age, preoperative albumin level, body mass index, operative time, amount of blood loss during the operation, the lowest body temperature during the operation, amount of blood transfusion and the highest postoperative blood sugar level	Site of surgery (gastric surgery, colorectal surgery)

		Reported information on SSI risk factors		
Study	**Type of surgery**	**Risk factors addressed via peri-operative measures or inclusion/exclusion criteria**	**Risk factors addressed via stratified randomisation or reporting comparability between trial arms**	**Risk factors adjusted for in the analysis of SSI incidence**
Lee 2009(288)	open appendicectomy	- standard antibiotic prophylactic regimen (piperacilin-tazobactam or moxifloxacin); - all wounds were closed primarily; - excluded patients with insulin-dependent diabetes;	Trial arms were reported to be comparable with respect to: age, sex, body mass index, smoking status, history of diabetes and severity of appendicitis	None reported
Reid 2010(289)	open elective colorectal resection	- standardised antibiotic prophylaxis; - standardised skin disinfection; - use of oxygen and patient warming devices both intra- and post-operatively; - standard wound closure and wound dressing;	Trial arms were reported to be comparable with respect to: age, sex, body mass index, mechanical bowel preparation, immunosuppressant use, diabetes, preoperative chemoradiotherapy, anaemia, malnutrition, alcohol abuse, smoking history, skin disease, hypertension, ASA score, type of intervention and mean length of stay.	Total length of inpatient stay

Inadequate reporting together with the lack of stratification and adjustment for SSI risk factors are major limitations of the included studies.

Two studies reported advantages and disadvantages of using the WEPD as perceived by the operating surgeons, but these accounts are contradictory. Psaila *et al.* (268) noted that the ring drape may be associated with *"difficulty of access and even damage to intra-abdominal viscera"* (p.732). Reid *et al.* (289) included surgeons' satisfaction as an outcome in their study by means of a visual analogue scale (0 - WEPD offers no assistance; 10 - WEPD offers best possible assistance): the average score elicited from the eight participating surgeons was 7 (range 5-10). While it is clear that the two cited studies were conducted more than 30 years apart and refer to different WEPD designs, the result reported by Reid *et al.* (289) should be interpreted with caution due to the small number of respondents (n=8).

The quality of the studies was generally poor. Their methodological drawbacks primarily concerned inadequate sequence generation, inadequate blinding and questionable outcome reporting. In addition, failure to adequately specify SSI definitions, the wound assessment frequency and length of follow-up seriously hindered the reliable interpretation of individual study results. Moreover, the studies' sample sizes were generally low and the majority were single centre.

Given these limitations, the quantitative analysis presented in this review can only have exploratory value. Under a random-effects model, the pooled risk ratio was 0.60 (95% CI 0.41 to 0.86), indicating that WEPDs may reduce the average risk of developing an SSI in open abdominal surgery by approximately 40% compared to standard care. The results did not differ greatly when a fixed-effects model was used (pooled RR 0.56, 95% CI 0.45 to 0.70). The proposed biological mechanism responsible for the device's effect is based on the physical separation between wound margins and contamination sources at the surgical site.

Horiuchi *et al.* (294) have found that use of a WEPD protects wound margins from bacterial invasion.

Although there was moderate heterogeneity between studies (I^2=54%, P=0.01), the results of the fixed and random effects models were comparable. This suggests that heterogeneity did not greatly affect the pooled estimates and that the smaller trials had little effect on the pooled estimate. The exploration of heterogeneity sources by means of subgroup analysis led to two findings: first, the SSI definition and WEPD design could be major sources of between-study heterogeneity; and second, other unexplored factors are also likely to cause systematic variation between the studies' results. These could relate to patient characteristics, the type of surgical intervention and perioperative measures. Unfortunately, information on these characteristics is both incomplete and highly variable across the included studies; therefore a further investigation of these factors is fraught with difficulties. The same applies to trial design characteristics, which were often reported incompletely.

Most of the included studies were single-centre, with two exceptions (270, 289). In these two studies patients were recruited from four hospitals and two hospitals, respectively. Furthermore, in the study of Reid *et al.* (289) it appears that surgeons from only one hospital operated in all the four recruiting hospitals. This undermines the multi-centre character of the trial since practitioner expertise is an important dimension of centre-specific characteristics (98). This seriously limits the external validity of the individual results. Despite fairly consistent risk ratios, the variation in patient characteristics, surgical technique and hospital may have significantly altered the effect estimates.

Publication bias was investigated by means of a funnel plot. Two studies contributed to the slight asymmetry of the plot (285, 288). Apart from their small sample sizes, these particular papers are at high risk of bias due to the methodological issues previously discussed, thus it is difficult to ascertain that the funnel plot is actually indicating publication bias. This is in accordance with the non-significant results of both Begg's test and Egger's test, thereby suggesting that overall results are unlikely to have been influenced by publication bias.

The subgroup analysis by degree of contamination (Figure 3.6) revealed that WEPDs may be efficient in reducing SSI rates following surgery of various contamination degrees, although in this exploratory analysis statistical significance was reached only for contaminated and dirty surgery, respectively. This finding should be interpreted as merely an indication based on the existing evidence available to date. Considerable caution is needed when observing the results of this exploratory quantitative analysis: the poor quality of the studies, their small sample sizes, and the relative closeness of the risk ratio point estimates across the contamination groups and the inconsistent definitions of the contamination categories give reasons for concern. It is likely that only a larger, good quality RCT addressing these methodological drawbacks can provide a reliable answer as to which type of surgery by degree of contamination mostly benefits from the use of WEPDs.

It is apparent from the forest plot (Figure 3.3) that the clinical effectiveness of WEPDs improved with time: the recent trials reported a greater benefit from the intervention compared to earlier trials. Table 3.5 presents the SSI rates in the both the intervention and control arms of the included studies. While no trend is readily noticeable in the control groups, it appears that SSI rates have gradually declined over time in the intervention arms. This observation should be interpreted cautiously with respect to the protective effect of the device itself because the differences between these two categories of studies are significant. First, older studies are more susceptible to methodological limitations compared to more recent studies as the guidelines for conducting clinical trials have evolved substantially since the publication year of the oldest included study. The quality assessment (Table 3.3) confirms this hypothesis. Second, the two largest trials in the review belong to the 'recent trials' group. Third, an issue worth discussing is balancing group characteristics between treatment arms. The risk factors of SSI are now widely known, with evidence around them starting to gather during the 1980s (200). The majority of the included studies did not use stratified randomisation by risk factors; additionally, the older studies did not produce descriptive statistics to demonstrate the comparability between trial arms with respect to such risk factors. There is, thus, a potential for bias in an unknown direction for these studies, especially given their generally small sample sizes.

Table 3.5 WEPD systematic review: surgical site infection (SSI) rates over time in the intervention and control groups

Study	Control			Intervention		
	SSI cases	Total patients	SSI rate (%)	SSI cases	Total patients	SSI rate (%)
Maxwell 1969(267)	12	82	14.63	16	88	18.18
Alexander-Williams 1972(283)	10	83	12.05	10	84	11.90
Psaila 1977(268)	10	47	21.28	9	46	19.57
Gamble and Hopton 1984(284)	6	70	8.57	7	70	10.00
Nyström 1984(270)	8	29	27.59	10	27	37.04
Batz 1987(285)	7	25	28.00	1	25	4.00
Redmond 1994(286)	27	111	24.32	11	102	10.78
Brunet 1994(287)	18	76	23.68	6	73	8.22
Sookhai 1999(269)	54	182	29.67	23	170	13.53
Horiuchi 2007(271)	16	110	14.55	8	111	7.21
Lee 2009(288)	7	48	14.58	1	61	1.64
Reid 2010(289)	15	66	22.73	3	64	4.69

Strengths and limitations

The strength of this review lies in the comprehensive assessment of the relevant evidence regarding the use of WEPDs in open abdominal surgery. To my knowledge, this is the first review looking at the reduction in SSI incidence associated with this type of device. The review identified several studies that were seldom or never cited in the widely known papers belonging to this therapeutic field (284, 285, 287).

The findings have several limitations. First, it is possible that the search strategy failed to identify some unpublished studies or trials that are published in journals not included in the bibliographic databases. Only clinically based SSI definitions were accepted in order to increase relevance for the present clinical context. The review is limited to open abdominal surgery. Studies have been published on the use of WEPDs of a similar design in laparoscopic interventions (295, 296). However, SSI rates are much lower in laparoscopic surgery compared to open surgery across a range of interventions (197, 297-300) and it would be inappropriate to combine data for open and laparoscopic surgery. Finally, poor reporting of inclusion/exclusion criteria and inappropriate accounting for SSI risk factors in the included studies may limit the validity of the results.

Following the completion and submission for publication of this systematic review in April 2011, new evidence regarding the clinical effectiveness of WEPDs has emerged in the form of a parallel systematic review and two RCTs. Edwards *et al.* (301) systematically reviewed RCTs where WEPDs were evaluated in reducing SSI rate after gastrointestinal and biliary tract surgery. Their review included 6 studies (1008 patients in total), all of which had been included in the systematic review (269-271, 284, 288, 289). The pooled risk ratio estimated under a random-effects model was 0.55 (95% CI 0.31 to 0.98), thus largely comparable with the finding of this review. Edwards and colleagues also acknowledged the effect of the WEPD design as a potential source of between-study heterogeneity and conducted subgroup analyses based on structural design. The modern dual-ring design was associated with a larger reduction in SSI rate compared to the traditional one-ring design (RR 0.31, 95%CI 0.14 to 0.67 vs. RR 0.83, 95%CI 0.38 to 1.83).

The results of two further single-centre RCTs have been published. Theodoridis *et al.* (302) enrolled 231 women undergoing caesarean section at a general hospital in Thessaloniki (Greece) and used the Alexis wound retractor in the intervention group and a conventional Doyen retractor in the control group. The authors reported 3/116 (3%) SSI cases in the control group compared with 0/115 (0%) SSI cases in the intervention group. Due to insufficient reporting the methodological quality of the study could not be adequately assessed. Moreover, the surveillance period was not specified, although it was mentioned that patients were monitored only during hospitalization. Cheng *et al.* (303) enrolled patients undergoing colorectal resection at a university hospital in Kuala Lumpur (Malaysia) and investigated the effect of WEPD in preventing SSI. They reported 6/30 (20%) infections in the control group compared with 0/34 (0%) infections in the study group. While patients were followed-up for 30 days post-operatively and blinding appears to have been appropriately ensured, their study featured a potential risk of selection bias due to allocation concealment using sealed envelopes. Furthermore, their sample size is strikingly small because it relied on a very high effectiveness of the WEPD i.e. 1% SSI rate in the intervention group vs. 20% SSI rate in the control group, suggesting a RR of 0.05. This assumption was not supported by references and, in the light of any known published study, can be regarded as very optimistic: the most favourable studies for WEPDs (285, 288) estimated a RR in excess of 0.10.

Furthermore, new research is currently under preparation. A protocol for a Cochrane systematic review on the effectiveness of WEPDs has been recently published (304).

Furthermore, Mihaljevic *et al.* (305) published a protocol for a RCT (the BaFO trial) investigating the effectiveness of WEPDs (the one-ring design) in reducing SSI in adult patients undergoing midline and transverse laparotomy on the occasion of general and visceral surgery. The study aims to recruit 600 patients from 15 German hospitals and a 50% reduction in SSI informed the sample size calculation, which resonates with the result of Horiuchi *et al.*(271). The CDC definition of SSI will be used and patients will be monitored for 45 days post-operatively, the longest surveillance period in all trials known to date. BaFO initiated recruitment in September 2010 and is expected to finish in summer 2013.

3.4. Conclusion

The body of evidence surrounding the use of WEPDs in reducing SSI in patients undergoing open abdominal surgery is relatively rich and this review identified 12 relevant articles. The results of the exploratory meta-analysis suggested that WEPDs may significantly reduce the incidence of SSI post-operatively in patients undergoing open abdominal surgery when compared with standard care. However, the quality of the available evidence is generally very poor due to methodological flaws and reporting failures. All the included studies were single-centre with two exceptions and their sample sizes were generally low. This may explain why WEPDs have not yet been widely adopted in current practice.

Given the potential clinical benefit of WEPDs in reducing SSI, it is of interest to explore their potential economic benefits. Chapter 4 presents the methods and findings of an original decision analytic model which estimates the cost-effectiveness of WEPDs compared to standard care in the UK context.

CHAPTER 4. PRELIMINARY EVIDENCE ON THE COST-EFFECTIVENESS OF WOUND-EDGE PROTECTION DEVICES: A DECISION-ANALYTIC MODEL

The principal conclusion of Chapter 3 was that, based on the existing evidence, WEPDs are likely to be effective in reducing SSI rate following open abdominal surgery. However, there is currently no available evidence on the cost-effectiveness of WEPDs, against any comparator, in any setting and for any patient population. The aim of this Chapter was to produce preliminary evidence, based on the best available information, on the cost-effectiveness of WEPDs compared to standard care when used in adults undergoing open abdominal surgery in the UK context.

A cost-utility analysis was conducted such that patient outcomes were measured in QALYs and the result was expressed in incremental costs (£) per QALY gained. A preliminary literature search revealed there is only limited evidence on utility values associated with SSI and this evidence has not yet been reviewed systematically. This type of evidence is necessary to inform the decision model. The Chapter is, therefore, structured in two sections: a systematic review of SSI utility values; and the actual decision model, informed by the SSI utility systematic review and the evidence on the clinical effectiveness of WEPDs presented in Chapter 3.

4.1. Systematic review of SSI utility values

The aim of this systematic review was to identify utility values associated with SSI in order to inform the outcomes of the SSI health states in the decision model. The review was conducted according to a pre-specified protocol based on guidance from the Centre for Reviews and Dissemination (272).

4.1.1. Methods

The elements of the question addressed by the systematic review are reported below in PICOS format (274):

Population: human patients undergoing open surgery;

Intervention: if applicable, any type of intervention aimed at improving surgical wound outcomes;

Comparator: if applicable, any comparator;

Outcomes: utility information was collected from or used to model a cohort of patients experiencing a SSI;

Study design: any type of study was accepted, including studies reporting primary data, reviews and model-based economic evaluations.

Eligibility criteria

The review included studies of any design where utility values for SSI were invoked (e.g. decision models) or elicited (e.g. valuation exercises) or at least one generic or specific non-preference based instrument was applied to a cohort of SSI patients (e.g. clinical trials, burden of illness studies). All definitions of SSI were accepted as long as they were explicit and 'surgical wound outcomes' were one of the main outcomes of the study. Only studies investigating outcomes of patients after open surgery were accepted. No language restrictions applied.

The following categories of studies were excluded: studies that did not explicitly report utility values or HRQoL data for a cohort of SSI patients; studies that did not explicitly investigate surgical wound outcomes; studies where HRQoL data were elicited at more than 6 months after surgery without an explicit mention that patients still had a SSI at the time of elicitation; studies that had as a primary outcome a composite of multiple surgical outcomes, even if it included SSI or wound healing; and studies looking at non-surgical wounds (e.g. burns, diabetic ulcers, radiation wounds) or infections (e.g. systemic infections). Study protocols and conference abstracts were also excluded.

Information sources

The following databases were searched from the starting date until October 2011: OVID MEDLINE and MEDLINE-In-Process, OVID EMBASE, ISI Web of Knowledge (Science Citation Index) and NHS Economic Evaluation Database (NHS EED). The information sources were selected as such with the intention to include economic evaluations (both model- and trial-based) of interventions aimed at reducing SSI and standalone HRQoL studies on relevant cohorts of surgical patients.

Search strategy

The devised search strategy included two categories of search terms: terms associated with wound infection, largely inspired by the systematic review of clinical effectiveness of WEPDs presented in the previous Chapter; and a range of terms relevant for HRQoL studies, thus capturing both widely used generic preference-based multi-attribute utility instruments, such as EQ-5D (43), HUI2 (306) and HUI3 (307), QWB (308) and SF-6D (44), as well as non-preference-based generic health status measures, such as SF-12 (309) and SF-36 (310, 311) health surveys. The latter are of interest because their scores can be mapped through statistical algorithms to preference-based measures and thus generate utility values (312). The detailed search strategies are presented in Appendix 3. The search was performed in October 2011.

Study selection

The study selection process comprised three phases: in phase 1 the titles and abstracts of returned papers were scanned against the inclusion/exclusion criteria. Articles demonstrating any of the exclusion criteria were eliminated. Where a decision could not be made based on the title and abstract, the article was entered into phase 2. In phase 2 the full-text versions of the papers resulting from phase 1 were obtained and scanned against the inclusion/exclusion criteria. Only studies fulfilling all the inclusion criteria were accepted. In phase 3 backward and forward reference searches were conducted for the studies kept in the review at the end of phase 2 in order to identify other potentially relevant articles.

Data extraction

The following data items were extracted from the papers included after phase 3:
study type; setting; type of surgery; sample size (of the cohort/subgroup where SSI values were elicited from); HRQoL instrument(s) used (e.g. EQ-5D); time of elicitation (e.g. 4 weeks after surgery); and utility values/HRQoL scores.

Data analysis

The characteristics of the included studies and the utility values/HRQoL data relevant to SSI patients were tabulated and summarised in a narrative review. The main objective of the review was to inform the decision model, so a formal quality assessment of the included papers was not performed. One researcher (I) performed the searches, screening, study inclusion and data extraction.

4.1.2. Results of the systematic review of SSI utility values

4,427 papers were identified through the database search: 957 papers were retrieved from MEDLINE and MEDLINE in Process, 1,239 from EMBASE, 1,580 from ISI Web of Knowledge and 651 papers were retrieved from NHS EED. After removing 807 duplicates, 3,620 papers were scanned for title/abstract and 3,572 were excluded (phase 1). In phase 2, 48 full-text papers were read and further 37 studies were excluded for not complying with the inclusion/exclusion criteria. The list of excluded full-text articles and accompanying justifications are presented in Appendix 4. Eleven studies entered phase 3 and no relevant further studies were identified through the reference list search. The study of Elliott *et al.* (313) was excluded as it duplicated the previous publication of the same research team (314), leaving a total of ten studies included in the systematic review. The study selection process is presented in Figure 4.1.

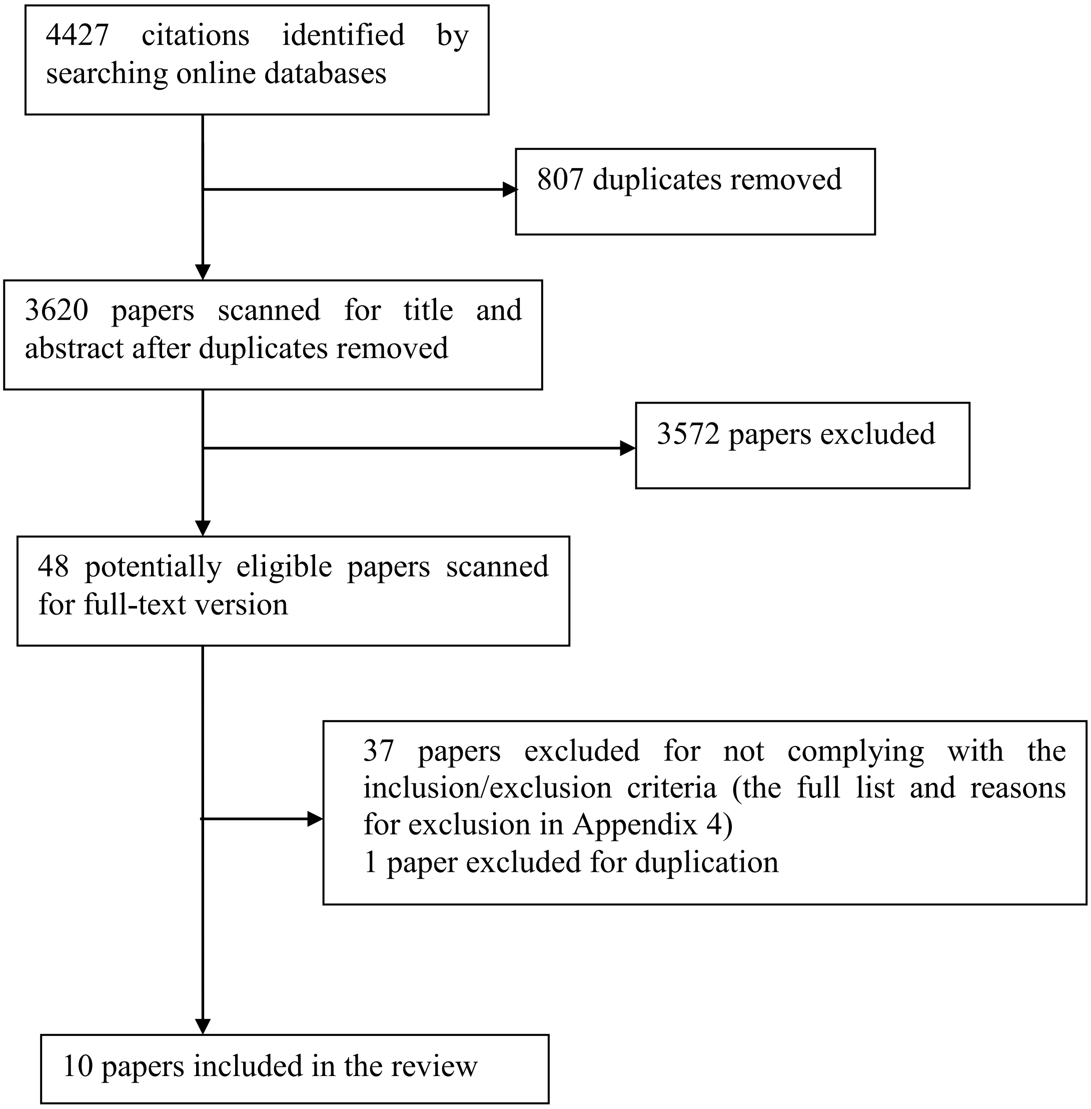

4427 citations identified by searching online databases
807 duplicates removed
3620 papers scanned for title and abstract after duplicates removed
3572 papers excluded
48 potentially eligible papers scanned for full-text version
37 papers excluded for not complying with the inclusion/exclusion criteria (the full list and reasons for exclusion in Appendix 4)
1 paper excluded for duplication
10 papers included in the review

Figure 4.1 Summary of the study selection process in the SSI utility systematic review

The characteristics of the ten papers included in the review are summarised in Table 4.1. Seven studies were conducted in the US (230, 315-320), while the remaining three were specific to the UK (314), Canada (321) and Denmark (322). Three studies considered orthopaedic surgery (314, 316, 318) and one study each considered cosmetic surgery (321), cardiac surgery (323), vascular surgery (320), caesarean delivery (319) and abdominal surgery (315). Furthermore, two studies considered a mix of surgical patients (230, 322). Eight articles were decision modelling studies (314-316, 318-321, 323) that cited utility values informing cost-utility analyses and two papers elicited SSI patients' own valuation of their health states using standardised questionnaires (230, 322). The utility data from the included studies are summarised in Table 4.2.

All eight modelling studies used decision trees to produce cost-effectiveness estimates for interventions aimed to reduce the risk of postoperative infection. The utility decrements associated with SSI varied from 0 i.e. no disutility (318), to 0.4 (315). However, the study with a null SSI utility decrement did not rely on published data for this decision, but rather assumed that SSI utility was equal to the utility of hospital confinement (318). A number of studies across several types of surgery cited utility decrements in the range 0.1 to 0.2 (314, 316, 321, 323).

Table 4.1 Characteristics of the studies included in the SSI utility systematic review

Study ID	Study type	Type of surgery	Country	Population characteristics	Intervention	Control
Brasel 1997(315)	Decision model	Appendicectomy	US	Hypothetical cohort of men and women with contaminated wounds	Primary wound closure; delayed primary wound closure	Secondary wound closure
Cranny 2008(314)	Decision model	General surgery	UK	Unclear, but data inputs compatible with a hypothetical cohort of 65-year old UK men	Glycopeptide prophylaxis: cephalosporin; vancomycin; cephalosporin and vancomycin	Unclear
Lee 2010(323)	Decision model	Cardiac surgery	US	Hypothetical cohort of 1,000 patients with median age 65	Surveillance strategy i.e. preoperative MRSA screening and decolonization	No MRSA surveillance strategy
Perencevich 2003(230)	Primary Study	General surgery	US	267 patients: SSI group - mean age 55.7, 48.3% male; control group - mean age 57.5, 52.8% male	SF-12	Individual domain scores MCS, PCS
Slobogean 2010(318)	Decision model	Surgical treatment of closed fractures	US	Hypothetical cohort of 52-year old men	Single-dose antibiotic prophylaxis	Multiple-dose antibiotic prophylaxis
Thoma 2003(321)	Decision model	Breast reconstruction	Canada	Unclear, but utility input data are compatible with a hypothetical cohort of 45-year old women	Free transverse rectus abdominis myocutaneous (TRAM) for breast reconstruction	Unipedicled TRAM for breast reconstruction

Study ID	Study type	Type of surgery	Country	Population characteristics	Intervention	Control
Bailey 2011(316)	Decision model	Orthopaedic surgery	US	Hypothetical cohort of 1,000 patients of age 63	Preoperative home-based chlorhexidine bathing cloth kits	No bathing cloth kits
Lee 2009(320)	Decision model	Vascular surgery	US	Hypothetical cohort of patients with median age 73	Surveillance strategy i.e. preoperative MRSA screening and decolonization	No MRSA surveillance strategy
Lee 2011(319)	Decision model	Caesarean delivery	US	Hypothetical cohort of 27-year old women	Preoperative *S. aureus* screening and decolonization	No *S. aureus* surveillance
Poulsen 1997(322)	Primary study	General, gynaecologic and orthopaedic surgery	Denmark	1301 patients: 47% over age 50, 52% male	GHQ and IADL	Infected minus uninfected differences Hospital cohort: -0.47 on GHQ scale 0.23 on IADL scale Patient cohort: 0.45 on GHQ scale -0.04 on IADL scale

Table 4.2 SSI utility data in the studies included in the SSI utility systematic review

Study ID	Study type	HRQOL instrument /source	Time of elicitation	HRQoL mean values for SSI
Brasel 1997(315)	Decision model	Clinical opinion	n/a	0.6 utility (0.4 utility decrement)
Cranny 2008(314)	Decision model	Clinical opinion: cited from Tengs and Wallace (2000), in its turn cited from Tsevat (1989)	n/a	0.9 utility (0.1 utility decrement)
Lee 2010(323)	Decision model	Cited from Selai and Rosser (1995)	Unclear	0.642 utility (0.198 utility decrement)
Perencevich 2003(230)	Primary Study	SF-12	8 weeks after surgery	Individual domain scores MCS, PCS
Slobogean 2010(318)	Decision model	Time trade-off, cited from Kuntz et al (2000), in its turn cited from Torrance (1987)	n/a	0.34 utility TTO (0 utility decrement)
Thoma 2003(321)	Decision model	Clinical opinion: a sample of 33 plastic surgeons	n/a	0.73 with drainage (0.14 utility decrement)
Bailey 2011(316)	Decision model	Clinical opinion: cited from Tengs and Wallace (2000), in its turn cited from Tsevat (1989)	n/a	0.9 utility (0.1 utility decrement)

Study ID	Study type	HRQOL instrument /source	Time of elicitation	HRQoL mean values for SSI
Lee 2009(320)	Decision model	Cited from Sackett and Torrance (1978)	Unclear	0.642 utility (utility decrement unclear)
Lee 2011(319)	Decision model	Clinical opinion: cited from Brasel et al (1997)	n/a	0.6 utility (0.32 utility decrement)
Poulsen 1997(322)	Primary study	GHQ and IADL	5.5 and 10 months after surgery	Infected minus uninfected differences Hospital cohort: -0.47 on GHQ scale 0.23 on IADL scale Patient cohort: 0.45 on GHQ scale -0.04 on IADL scale

The references for utility values invoked in the modelling studies were rarely primary studies themselves and cited other studies in their turn, sometimes of ambiguous relevance. For example, in the orthopaedic infection prophylaxis study of Brasel *et al.* a 0.1 SSI disutility was assumed, informed by the paper of Tengs and Wallace (324), which had reported a 0.9 utility for an infection of an artificial joint. The 0.9 value in the Tengs and Wallace review was in its turn informed by the 1989 decision modelling study of Tsevat *et al.* (325), who had assumed, based on their own judgement, 0.9 QALYs for a patient hospitalized for a year due to an infected artificial joint. In another example, Lee *et al.* used a utility value of 0.84 for an otherwise healthy patient following cardiac surgery and a utility of 0.642 for an infected surgical wound; the latter estimate was based on the 1995 study of Selai and Rosser (326). The study of Selai and Rosser was a pilot micro study on a sample of 40 patients in a UK general hospital whose aim was to compare the EQ-5D utility values of a sample of inpatients with those of the general population. It is not clear how many of the patients in this micro study actually experienced a SSI. The authors (Selai and Rosser) were contacted by email in an attempt to obtain a report of the original study: they responded and initially agreed to assist upon retrieving the document from their own archive, but eventually failed to provide the data.

There were also instances of unclear reporting: for example, Lee *et al.* used a 0.642 utility for SSI informed by the 1978 paper of Torrance and Sackett (327), but the utility decrement itself is unclear because the utility associated with uncomplicated surgery was not reported. Moreover, their paper appears to contain a referencing error: the paper of Torrance and Sackett did present utility values for hospital and home confinement due to dialysis and some contagious diseases, but not specifically for SSI and the numerical value of 0.642 did not even appear anywhere in their paper. However, a 0.642 utility associated with SSI appears in the Selai and Rosser study (326), discussed above, which was also referenced in a study of the same team (320) as a source for line infection utility.

Two included studies explicitly derived utility values based on expert opinion or clinical judgement (315, 321). Furthermore, the primary sources for three further studies (314, 316, 319) also relied on expert opinion to derive utility values. The methods of eliciting expert opinion also varied from the authors' own judgement (315) to conducting a survey among practicing surgeons (321). Thoma *et al.* acknowledged that utility values elicited from patients themselves are generally preferable, but argued that expert generated values are recommended when evaluating novel surgical interventions, as was their case (321).

Only one decision modelling study used a systematic review of the literature to inform its utility input data. Cranny *et al.* conducted systematic reviews of the effectiveness and cost-effectiveness of glycopeptide antibiotics and identified one economic evaluation study which reported HRQoL information for SSI using the SF-36 questionnaire (229). However, the authors of that study did not respond to their request for access to individual patient data, which would have allowed the derivation of utility scores. Eventually, Cranny *et al.* used a utility decrement of 0.1 for SSI based on Tengs and Wallace's (324) estimate of 0.9 utility for an infection of an artificial joint, which was described above.

Two primary studies employed standardised HRQoL instruments. Perencevich *et al.* (230) employed the SF-12 questionnaire and compared the health status of 50 patients with SSI at 8 weeks after surgery to that of 123 matched uninfected controls. Case-patients reported significantly lower scores than controls on the mental health component score of SF-12 (MCS-12); the difference between the groups was small, but not statistically significant on the physical health component (PCS-12). The authors were contacted by email and asked whether the individual patient scores for the SF-12 instrument were still available and could be shared. This would have allowed mapping the SF-12 scores onto the EQ-5D instrument and thus generate utility values (328). The authors promptly replied and reported that the original patient dataset had been deleted since the termination of their study and there was no backup copy. Given this situation, the mapping exercise could not be performed and utility values could not be calculated.

Poulsen *et al.* (322) enrolled 1301 Danish patients and compared the HRQoL between patients with and without a surgical wound infection (SWI). The authors looked at two cohorts: in the hospital cohort, the SWI was diagnosed while inpatient by a surgeon; this group included 58 cases and 648 controls. In the patient cohort, only SWIs diagnosed after discharge were included, either by the antibiotics prescription or a reopening of the wound because of purulent discharge. Patient outcomes were assessed using the 12-question edition of the General Health Questionnaire (GHQ) and the Instrumental Activity of Daily Living (IADL) questionnaire, which were mailed to patients twice (median 5.5 months and 10 months postoperatively). The differences between groups (infected vs. uninfected) were small, not statistically significant and inconsistent across cohorts (Table 4.2). A recent systematic review identified no mapping studies of GHQ or IADL to preference-based measures (312), therefore utilities cannot be calculated from these data.

4.1.3. Discussion

This systematic review identified ten studies which investigated interventions meant to reduce SSI following a range of surgical procedures, including general surgery (230), orthopaedic surgery (316) and caesarean delivery (319). The primary sources of utility values in the eight modelling studies were often informed by authors' own judgement or expert opinion. There also appear to be very few primary studies eliciting patient preferences on SSI health states. Circular referencing across the identified publications was common, thus indicating that the available literature on SSI utility values is scarce.

Given that the use of WEPDs lends itself mostly to open abdominal surgery, the appendicectomy study of Brasel *et al.* (315) appears to be the most relevant in this instance, especially given the 30-day time horizon which is in line with the SSI definition applicable in the UK. However, the utility value cited in this study was based solely on authors' own judgements and so its validity can be easily questioned. In this situation, I looked at the overall utility decrement associated with SSI in all the studies identified by the review in order to use (with due caution) all the available information on the impact of SSI on surgical patients' HRQoL.

In the eight modelling studies the utility decrement associated with a SSI was in the range of 0 i.e. no difference from uninfected surgical patients, to 0.4. Unfortunately, the utility scores could not be calculated from the two studies that elicited SSI patients' scores using validated questionnaires due to the lack of the individual patient data or absence of mapping algorithms. Two studies (314, 321) used a utility decrement of 0.1 for SSI and it was decided to use this estimate in the base-case analysis of the economic model for WEPDs. The impact of this value on the cost-effectiveness of WEPDs was explored in sensitivity analyses (see below sub-section 4.2.3).

4.2. SSI decision model

The aim of the decision model was to use the best available evidence to produce preliminary evidence on the cost-effectiveness of WEPDs compared to standard care in reducing SSI in order to inform decision makers on the wider benefits of using WEPDs in current surgical practice and on the need to gather additional evidence on the topic. The methods and results are reported below in line with the recommendations of the CHEERS Statement (329) (Appendix 5).

4.2.1. Methods

A model-based cost-utility analysis was conducted using TreeAge Pro 2011 software (330). The patient population considered in the base-case is represented by adult patients aged 60 undergoing open large bowel surgery in the UK. Large bowel surgery was chosen because it is one of the surgical procedures with the highest SSI incidence rates in the UK (192).

Setting and perspective

The model setting is the English NHS, where there is considerable interest in the surveillance of hospital-acquired infections in general and of SSI in particular (206, 331). The majority of the surveillance efforts refer to NHS hospitals, although the pathway of care of SSI patients also continues in primary care after discharge (195). There is less evidence about patterns of care in the primary setting than for inpatient care.

The model perspective was that of the NHS. Only resource utilisation relevant to the NHS was considered for costing purposes.

The base-case considered the use of a WEPD in adults undergoing open large bowel surgery compared with standard care (i.e. not using the WEPD). The comparator was chosen as such since the current clinical guidelines do not recommend any other intervention for the purpose of wound-edge protection, thus there is no obvious competitor for WEPD apart from the bundle of prophylactic measures used in surgery.

Time horizon

The model time horizon was 30 days post-operatively. Most SSI surveillance programmes as well as the SSI definitions used by the HPA (206) and the CDC (200) cite a 30-day interval post-surgery during which wound infections are being monitored and classed as an SSI, respectively. The minimum time horizon would, thus, be one month post-operatively. Moreover, the majority of SSIs do heal and patients recover full functionality. In the absence of published data on the average healing time of a SSI, a group of health care professionals were informally approached on this matter (two surgical registrars on rotation at University Hospitals Birmingham, one general practitioner (GP) with academic tenure in Primary Care Clinical Sciences – University of Birmingham, two district nurses affiliated with Sandwell Primary Care Trust and one practice nurse at University Hospitals Birmingham). In addition, the GP and the nurses were consulted about the likely resource utilisation and patient pathways in primary care. Their views were that SSIs can heal from as soon as several days to as long as several months, depending on the gravity of the infection and on patient co-morbidities. Given the under-reporting of SSIs, the paucity of data regarding SSI progression in time and the fact that most of the evidence concerning resource utilisation comes from studies with a 30-day period, the time horizon for the model was selected as 30 days after surgery. The implications of the time horizon on the cost-effectiveness findings are discussed later in the Chapter under *Strengths and limitations*. No discount rate for costs and outcomes was applicable due to the short time horizon.

Outcomes

Outcomes were measured in QALYs. Prior to the inclusion in the decision model, all utility values were adjusted accordingly for the one month time horizon. The utility associated with uninfected open abdominal surgery was informed by the study of Janson *et al.* (332), who elicited EQ-5D values from patients undergoing colon resection, one of the most common intervention in the 'large bowel surgery' category. A 0.1 utility decrement was assumed for SSI patients based on the systematic review presented in the previous section (section 4.1). The value of the utility decrement was varied in sensitivity analyses.

Clinical effectiveness

The clinical effectiveness estimates for the WEPD were informed by the findings of the systematic review of WEPD clinical effectiveness studies, which was presented in detail in Chapter 3. The base-case value for the relative risk of SSI associated with using the WEPD compared to standard care was 0.60 (95%CI 0.41 to 0.86).

Resource use and costs

All costs in the model are given in UK £ (2010 value). The price of the WEPD was sourced from the manufacturer 3M^{TM} (Steri-Drape©). Four WEPD sizes are available with differing prices and the medium sized WEPD was considered in the base-case. The Hospital and Community Health Services combined pay and price inflation index (HCHS) (333) was used to inflate all relevant costs obtained from the literature.

For inpatient care, the length of stay for uninfected patients and the additional length of stay for patients with superficial and with deep/organ-space SSI were informed by the study of Coello *et al.* (191). No conclusive evidence has yet suggested that MRSA-SSI affects the length of stay for patients undergoing large bowel surgery in the UK setting, although there is evidence from US hospitals that MRSA-SSI is associated with increased length of stay (234, 334). Since discharge practices are not transferable between countries, for the purpose of this model patients with MRSA and non-MRSA SSI were assumed to spend the same number of inpatient days and the impact of this assumption is discussed below under *Strengths and limitations*. The difference between the two types of infection was reflected through additional costs due to MRSA (i.e. barrier nursing), assumed to start being incurred half-way through the inpatient stay. The unit costs for an inpatient day, with or without SSI, were also informed by the study of Coello *et al.* and updated to the 2010 value (191). The unit costs for MRSA care, applicable to patients experiencing MRSA-SSI, were taken from the modelling study of Elliott *et al.* (313) and were added to the usual inpatient day cost. Patients who die in hospital as a result of a SSI have been assigned the cost for three organ support critical care (335).

Unit costs for care received in a primary setting (GP visit, district nurse and practice nurse time) were informed by Curtis (336). Costs for medication and painkillers prescribed by the GP (i.e. Co-fluampicil 250/250 and Co-codamol 8/500 4 times daily for 7 days, informed by discussion with one GP) were informed by Prescription Cost Analysis England 2010 (337).

Model structure and assumptions

The chosen model structure was a decision tree. This decision is supported by decision modelling methodological guidelines (338, 339) which recommend the use of decision trees to model interventions with relatively short duration outcomes beyond which the patient is expected to fully recover.

The model structure can be summarised as follows (Figure 4.2): following surgery, a patient may or may not develop a SSI. The SSI can be diagnosed during the initial inpatient phase or after discharge. If the SSI is diagnosed while in hospital, three main alternatives were explored: 1) the patient remains in hospital until the infection is healed; 2) the patient is discharged with a SSI and continues treatment in primary care; or 3) the patient dies in hospital as a result of the SSI or other complications. If discharged with a SSI (option 2), the infection may continue to heal or not in a primary care setting. If the infection does not heal, it has been assumed that patients will visit the GP, who may either refer them back to hospital or prescribe antibiotics and send the patient home. If the SSI develops after discharge, it was assumed that patients would visit the GP, who may refer them to hospital or not, as above.

The model distinguished between MRSA and non-MRSA SSIs because evidence suggests that MRSA-SSIs are associated with higher costs and higher mortality compared to non-MRSA SSIs (234). The model also differentiated between superficial and deep/organ-space SSIs because the inpatient length of stay has been shown to vary between the two categories (191). Superficial, deep and organ-space SSI can be distinguished according to severity and site (199).

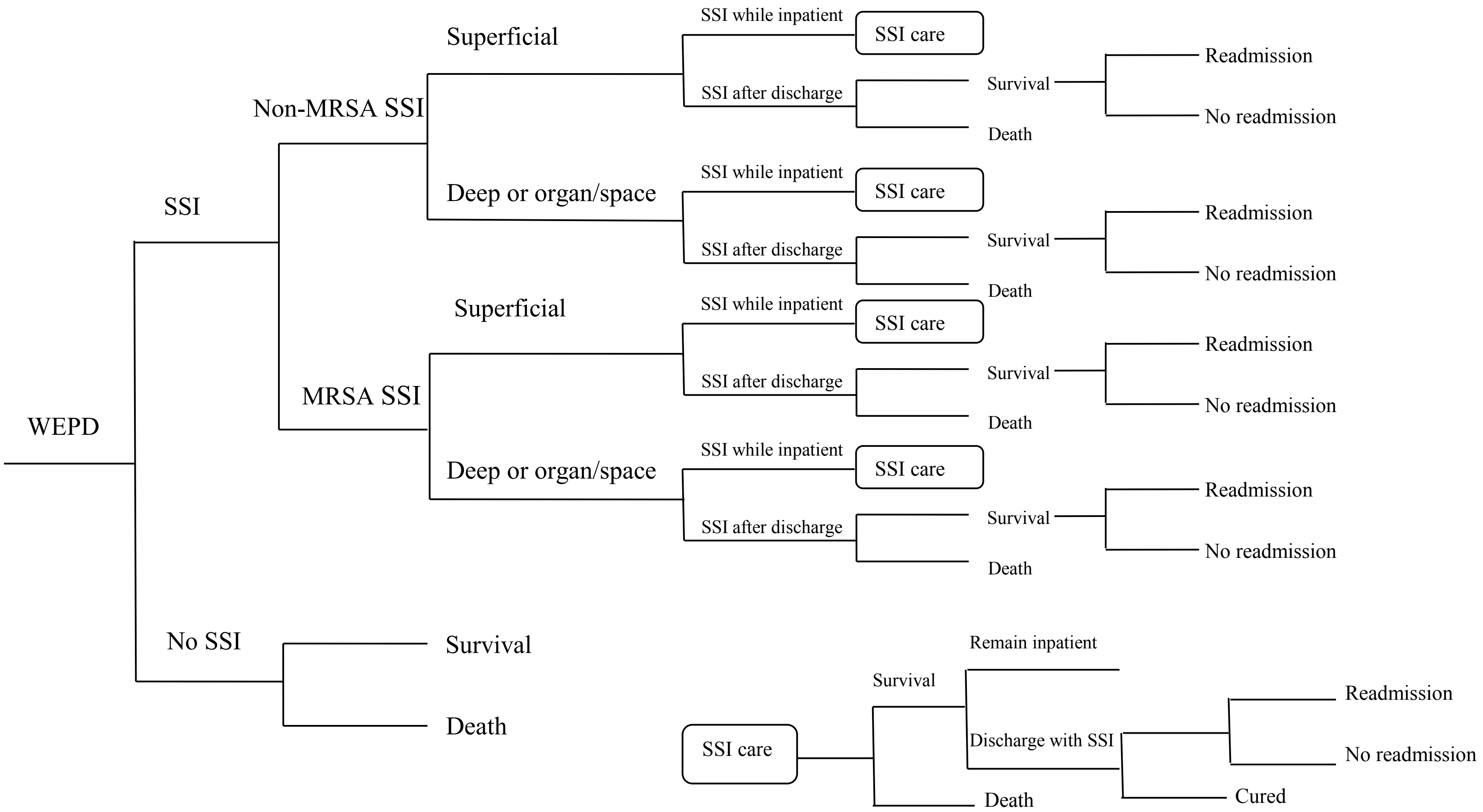


Figure 4.2 The structure of the decision model (WEPD arm)

It was assumed that patients with a SSI diagnosed whilst inpatients visit their GP once and receive seven visits from the district nurse after discharge. There is one exception: patients who are diagnosed with a SSI whilst an inpatient and remain in hospital until the SSI is cured will receive two district nurse visits upon discharge. In addition, if they have a recurrent SSI and the GP does not refer them to hospital, two practice nurse visits were considered. Patients for whom the SSI becomes apparent only after discharge visit their GP once and do not receive district nurse visits. Patients developing a MRSA-SSI after discharge visit their GP twice, undergo two practice nurse visits and they are referred back to hospital. These assumptions were informed by discussions with health care professionals, as described above.

A proportion of SSI inpatients were assumed to be discharged with a SSI and continue antibiotic therapy at home and in primary care, while the rest would remain as inpatients. A significant proportion of SSIs are diagnosed after discharge (250, 251, 340) and the clinical reality suggests that only a fraction of GPs will refer patients with a SSI back to hospital - in most cases wound care will continue in a primary setting under antibiotic treatment. Moreover, if the GP prescribes antibiotics for a non-MRSA-SSI, it was assumed that it would heal in primary care. On the other hand, it was assumed that MRSA-SSIs would not heal in primary care and would require hospital readmission.

It was also assumed that the use of a WEPD affects mortality and the probability of developing a SSI, but does not influence the probability of acquiring a particular type of SSI (i.e. MRSA/non-MRSA, superficial/deep/organ-space) or any other process variable (e.g. probability of detecting an SSI while inpatient, GP referral rate) compared to patients in the control group. For three probabilities (the probability of being discharged with a SSI, the probability of GP referral to hospital and the probability of having a recurrent SSI) there was no literature information available and the point estimates were informed by consultations with health care professionals, as described above.

Analytical methods

A probabilistic analysis was conducted in the base-case to reflect the uncertainty of the model input parameters, namely probability values, costs and utility values. Each model parameter was assigned a distribution reflecting the amount and pattern of its expected variation. Cost-utility results were calculated by simultaneously selecting random values from these distributions over 10,000 replications in a Monte Carlo simulation. The results of the simulations were depicted graphically using cost-effectiveness scatter plots and cost-effectiveness acceptability curves (CEACs) (58). The latter reflect the probability of either alternative being cost-effective at varying WTP thresholds, currently considered by NICE in the range of £20,000 to £30,000 per QALY gained (39).

Deterministic sensitivity analyses

A range of deterministic sensitivity analyses were carried out by varying the base-case values for the following parameters: the probability of discharging patients with a SSI; the probability of being referred to the hospital by the GP when developing a SSI; the utility decrement for SSI patients compared to uninfected patients; the length of stay for uninfected patients; the clinical effectiveness of WEPDs, reflected in the model by the relative risk of SSI in the WEPD arm; and the cost of the WEPD. These parameters were subject to sensitivity analyses because their base-case values were associated with the greatest uncertainty as there were no literature sources available to inform their estimates. For the first two probabilities, the intervals were chosen arbitrarily in order to investigate their influence on the ICER. In the sensitivity analysis of the utility decrement, the lower and upper bound have been set at 0 (no difference) and 0.4, respectively, according to the extreme values identified in the systematic review of SSI utility values (section 4.1). The lower bound of the length of stay analysis was informed by an average estimate for lower digestive tract surgery cited in Hospital Episode Statistics for England 2010 (341) and the upper bound was arbitrarily set at 20 days. The length of stay for uninfected patients influences that of SSI patients because the additional inpatient days due to SSI have been added to this core value. The relative risk of SSI in the intervention arm was varied across the entire possible range (0 to 1) to identify the threshold value at which the cost-effectiveness recommendation changes. The cost of the WEPD was varied from 0 to £100 – a conservative range given that the highest price for a WEPD, as communicated by the manufacturer 3M^{TM}, was £25.

Scenario analyses

In addition to the base-case, two alternative scenarios were analysed using probabilistic sensitivity analysis (PSA). The first scenario looked at the cost-effectiveness of WEPDs in small bowel surgery, as patients undergoing this type of intervention have a lower SSI risk compared to large bowel surgery. The average length of stay and extra length of stay due to SSI were modified accordingly (191). The second scenario referred to large bowel surgery, as in the base-case analysis, but assumed receiving more care in the primary setting; the relevant resource utilisation parameters (i.e. district nurse visits and medication) were informed by the study of Tanner *et al.* (195).

Structural uncertainty

Structural uncertainty in decision models refers to a wide range of sources of uncertainty, which cannot be classed as parameter or methodological uncertainty. These sources can be grouped under four main categories: inclusion/exclusion of relevant comparators; inclusion/exclusion of relevant events; statistical models to estimate specific parameters; and clinical uncertainty (342). In this particular case, there was some degree of uncertainty regarding the care pathway, given that SSI management is highly individualised and reported incompletely in the literature.

The base-case model attempted to reflect accurately the clinical reality underpinning SSI care, but it also relied on a large number of assumptions and had a complex structure. Consequently, an alternative decision model (model 2) was developed to explore the impact on the cost-effectiveness estimates of modelling a different patient pathway (Figure 4.3). The alternative model differed from the main decision model in two important aspects: it had a much simpler structure, thus making fewer assumptions about the pathway of care; and it used a bulk cost for SSI care as reported by Tanner *et al.* (195) as opposed to summing the individual cost elements. The total cost for SSI care included additional resource use due to SSI: inpatient days; district nurse, practice nurse and outpatient visits; medication and consumables (wound dressings, wound swabs); and readmission costs. The patient pathway can be summarised as follows: after undergoing open abdominal surgery, patients may or may not develop a SSI. In either case, they may survive or not. After discharge, all patients were assumed to receive two district nurse visits. The cost of death was assimilated with that of critical care for three organs, as in the main model. The cost of inpatient care was calculated by multiplying the average inpatient length of stay for large bowel surgery with the average unit cost of an inpatient day; in SSI patients, the cost attributable to SSI care was added to the total cost of inpatient care. All relevant probabilities, unit costs and utility values were the same as in the main model. The alternative model did not differentiate between severities of SSI (superficial vs. deep/organ) or causative agents (non-MRSA vs. MRSA).

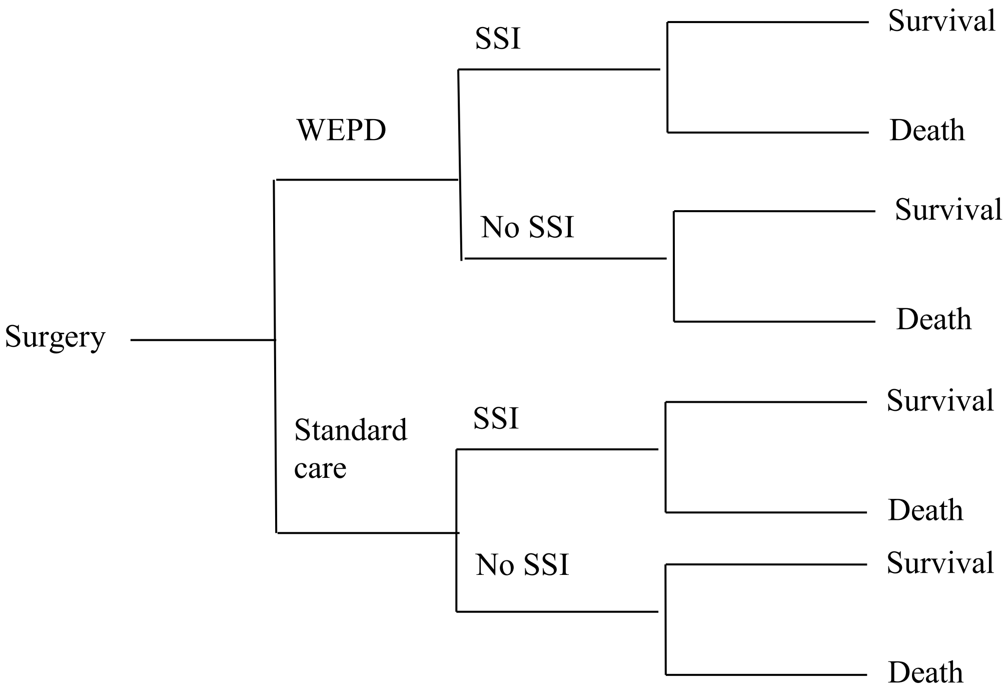


Figure 4.3 Structure of the alternative decision model (model 2)

4.2.2. Results

Study parameters

Probability values (point estimates and 95%CI where applicable) for the base-case analysis and for the alternative scenarios are presented in Table 4.3 together with the corresponding data sources. Resource use, health utility and unit cost data are presented in Table 4.4 together with the corresponding data sources. For the probabilistic analysis, transition probabilities and utility values have been assigned beta distributions, while costs have been assigned gamma distributions (343). No uncertainty was modelled around unit costs and resource utilisation parameters. Where no information was available on the variability around the point estimate, the standard error was assumed 0.1 of the mean for probabilities and utilities and 0.2 of the mean for costs in acknowledgement of usual right skewness of cost data (65).

Table 4.3 Probability values used in the decision model

Description	Point estimate (95% CI)	Source
Base-case		
Probability of developing a SSI after large bowel surgery, inpatient and readmission	0.095 (0.090 to 0.101)	Health Protection Agency, 2011 (192)
Relative risk of developing an SSI in the WEPD arm	0.600 (0.410 to 0.860)	Systematic review (Chapter 3)
Probability of having a SSI caused by MRSA	0.100*	Derived from Health Protection Agency, 2011 (192)
Probability of developing a superficial SSI after large bowel surgery	0.571*	Health Protection Agency, 2011 (192)
Probability of death after large bowel surgery, uninfected	0.061 (0.054 to 0.067)	Derived from Coello et al, 2005 (191)
Probability of death with superficial SSI after large bowel surgery	0.040 (0.020 to 0.059)	Derived from Coello et al, 2005 (191)
Probability of death with deep/organ-space SSI after large bowel surgery	0.105 (0.069 to 0.141)	Derived from Coello et al, 2005 (191)
Probability of detecting a SSI at readmission	0.093 (0.075 to 0.110)	Derived from Health Protection Agency, 2011 (192)
Probability of being discharged with a SSI	0.700*	Assumed, informed by consultation with clinicians
Probability of being referred to the hospital by the GP if SSI develops post-discharge	0.200*	Assumed, informed by consultation with GPs
Probability of recurrent SSI post-discharge	0.100*	Assumed, informed by consultation with clinicians
Scenario 1: small bowel surgery		
Probability of developing a SSI after small bowel surgery, inpatient and readmission	0.082 (0.071 to 0.095)	Derived from Health Protection Agency, 2011 (192)
Probability of developing a superficial SSI after small bowel surgery	0.518*	Derived from Health Protection Agency, 2011 (192)
Probability of death after small bowel surgery, uninfected	0.059 (0.052 to 0.065)	Derived from Coello et al, 2005 (191)
Probability of death with superficial SSI after small bowel surgery	0.069 (0.055 to 0.083)	Derived from Coello et al, 2005 (191)
Probability of death with deep/organ-space SSI after small bowel surgery	0.185 (0.149 to 0.221)	Derived from Coello et al, 2005 (191)

*: Where the 95% CI could not be calculated based on the information in the data source, the standard error was assumed to be 10% of the point estimate for the purpose of the probabilistic sensitivity analysis.

Table 4.4 Resource use, unit costs and utility data in the decision model

Description	Value	Source
Resource use - inpatient care		
Average length of stay after large bowel surgery, uninfected patients	11.3	Coello et al, 2005 (191)
Additional length of stay after large bowel surgery, patients with superficial SSI	7.8	Coello et al, 2005 (191)
Additional length of stay after large bowel surgery , patients with deep/organ-space SSI	12.6	Coello et al, 2005 (191)
Average length of stay after small bowel surgery, uninfected patients	11.5	Coello et al, 2005 (191)
Additional length of stay after small bowel surgery, patients with superficial SSI	12.9	Coello et al, 2005 (191)
Additional length of stay after small bowel surgery , patients with deep/organ-space SSI	13.4	Coello et al, 2005 (191)
Unit costs	**Value (£)**	
Cost of antibiotic and painkillers prescription from GP	19	NHS The Information Centre, 2011 (337)
Cost of critical care per spell, 3 organ support	1 400	NHS Reference Costs 2009-2010, 2011 (335)
Cost of inpatient day, uninfected	462	Derived from Coello et al, 2005 (191)
Cost of inpatient day, SSI	507	Derived from Coello et al, 2005 (191)
Cost of MRSA care per day	407	Derived from Elliott et al, 2010 (344)
Cost of GP visit	36	Curtis 2010 (336)
Cost of district nurse home visit	27	Curtis 2010 (336)
Cost of practice nurse procedure	10	Curtis 2010 (336)
Cost of WEPD, medium size	16.5	Manufacturer
Cost of medication – scenario 2: intensive primary care	41	Tanner et al, 2009 (195)
Utility values (EQ-5D)		
Baseline utility	0.800	Kind et al, 1999 (345)
Utility for uninfected patients	0.752	Janson et al, 2007 (332)
Utility for SSI patients	0.653	Derived from Janson et al, 2007 (332) and literature review (section 4.1)

Note: Resource use for primary care has been discussed above. Given the lack of published data, the number of GP visits, practice nurse visits and district nurse visits has been assumed and is described in the text of Chapter 4 (sub-section 4.2.1). Primary care resource use for scenario 2 was informed by Tanner et al, 2009.

Base-case analysis

In the base-case analysis (Table 4.5) the WEPD strategy was associated with an average cost of £5,196 and a benefit of 0.0606 QALYs, while standard care costs on an average £5,240 and yielded a benefit of 0.0605 QALYs. The WEPD appears to be less expensive and slightly more effective than standard care, which is thus dominated. Figure 4.4 presents the output of the Monte Carlo simulations on the cost-effectiveness plane; only 1,000 of the 10,000 incremental cost-incremental QALY pairs are presented. The WTP threshold was set at £20,000 per QALY gained. The distribution of the incremental cost-incremental QALY pairs covers all the four quadrants, but the majority of the pairs fall below and to the right of the WTP threshold, suggesting that the WEPD appears to be cost-effective compared to standard care. In the corresponding CEAC the WTP threshold has been varied in the range £0 to £100,000 per QALY gained (Figure 4.5). The WEPD has 86.6% probability of generating a positive net monetary benefit at a threshold of £20,000/QALY and 87.6% at £30,000/QALY.

Table 4.5 Results of the decision model cost-utility analysis

Scenario	Alternatives	Mean cost (£)	Mean effectiveness (QALYs)	ICER (£/QALY)
Base-case	WEPD	5,196	0.06061	Standard care is dominated
	standard care	5,240	0.06051	
Scenario 1: Small bowel surgery	WEPD	5,286	0.06062	Standard care is dominated
	standard care	5,330	0.06048	
Scenario 2: Intensive primary care	WEPD	5,221	0.06060	Standard care is dominated
	standard care	5,272	0.06051	
Alternative decision model (model 2)	WEPD	5,672	0.06051	Standard care is dominated
	standard care	6,056	0.06036	

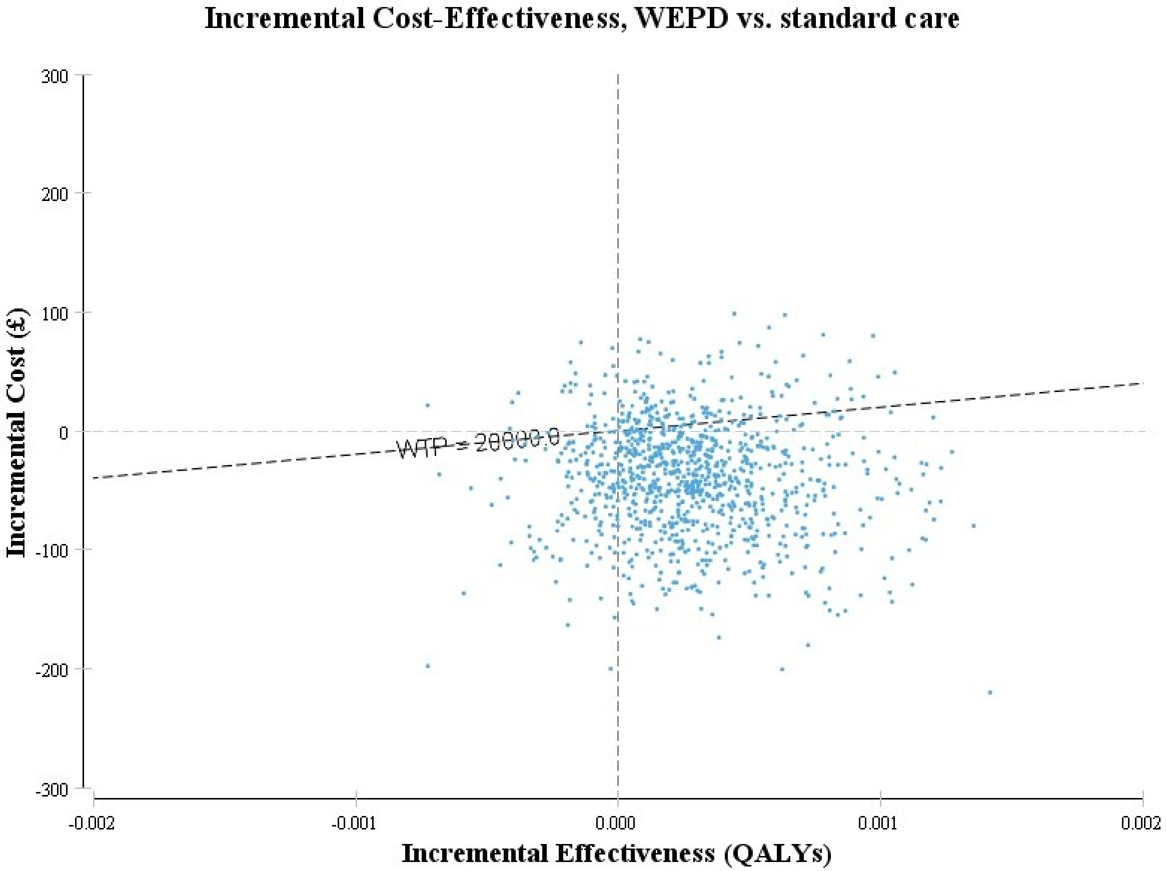


Figure 4.4 Decision model: probabilistic sensitivity analysis for the base-case – incremental cost-effectiveness scatter plot

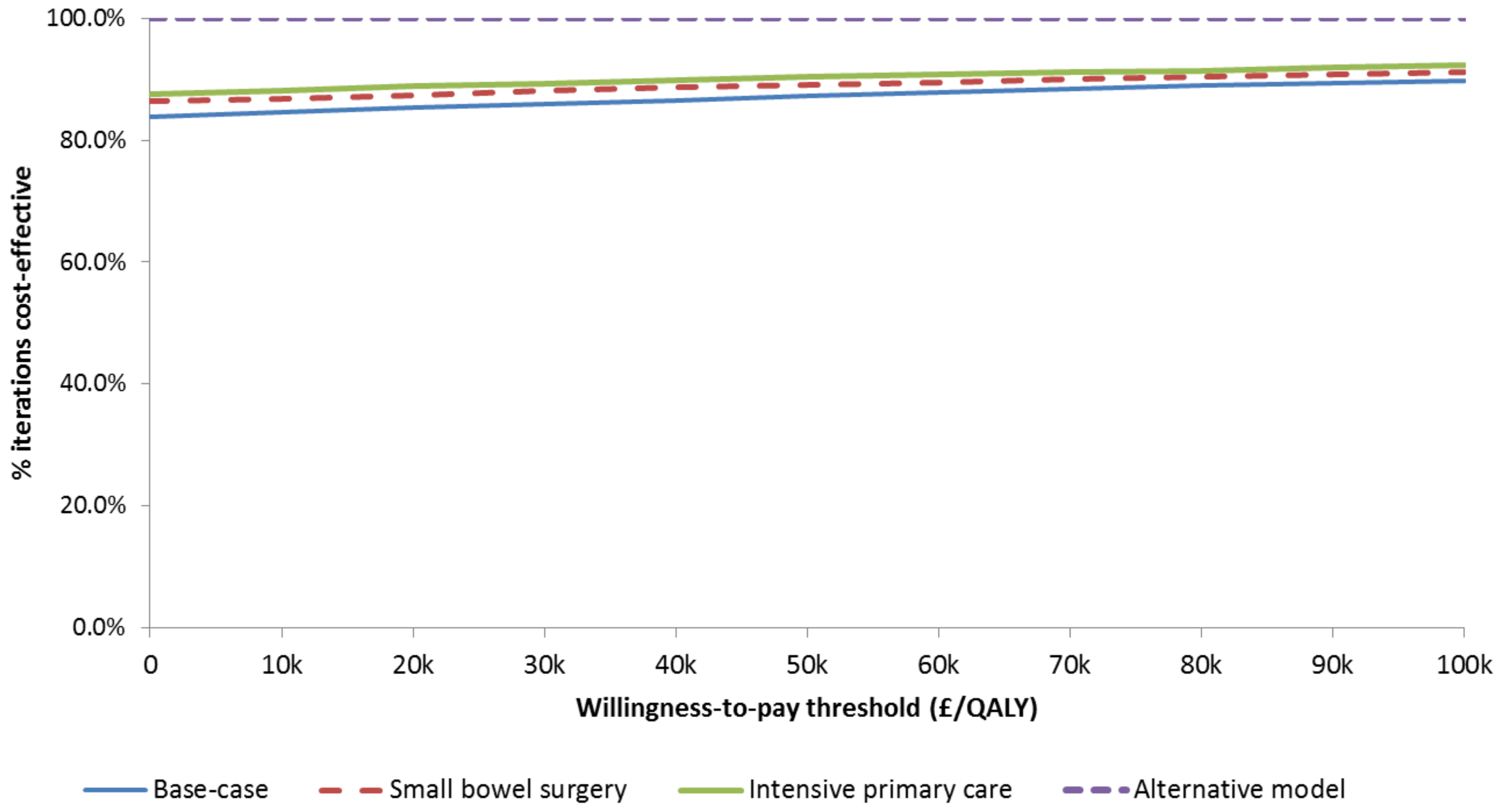


Figure 4.5 Decision model: cost-effectiveness acceptability curves for the base-case and alternative scenarios

Deterministic sensitivity analyses

Six deterministic sensitivity analyses were performed and their results are summarised in Table 4.6. The WEPD dominated standard care across the range of inspected discharge policies, but it must be noted that the incremental cost decreased as the probability of discharging patients with a SSI increased, from £62 when 50% of SSI patients are discharged with a SSI to £16 when all patients with a SSI are discharged before full recovery. The second analysis looked at the influence of GP behaviour and varied the probability of patients being referred to the hospital by the GP when the SSI develops after discharge. The WEPD dominated standard care across the range of inspected referral policies; the incremental cost increased from £34 when 10% of SSI patients are referred back to hospital to £122 when all SSI patients are referred back to hospital. In the third sensitivity analysis the utility decrement associated with having a SSI was varied from 0 (i.e. having a SSI does not affect quality of life) to 0.40 (0.10 in base-case). The WEPD dominated standard care for any value of the utility decrement larger than 0.02. There was very little variation in the incremental effectiveness, from 0.059 QALYs when the utility decrement is null to 0.056 QALYs when the utility decrement is 0.4. When inpatient length of stay for uninfected patients was varied from 6.1 to 20 days (11.3 days in base-case), standard care was also dominated across the range of investigated values: the incremental cost varied from £40 (6.1 days inpatient stay) to £50 (20 days inpatient stay).

Table 4.6 Results of the deterministic sensitivity analyses in the decision model

Parameter varied	Base-case value	Range tested	Effect on cost-effectiveness (WEPD vs. standard care)
Probability of discharging patients with SSI	0.7	0.5 to 1.0	Standard care is dominated across the range Incremental costs decrease across the range from £62 (0.50) to £16 (1.0)
Probability of GP referring SSI patients to hospital	0.2	0.1 to 1.0	Standard care is dominated across the range Incremental costs increase across the range from £34 (0.10) to £122 (1.0)
Utility decrement associated with SSI	0.1	0 to 0.4	Standard care is dominated for utility decrements larger than 0.02
Uninfected inpatient length of stay	11.3	6.1 to 20.0	Standard care is dominated across the range Incremental costs increase from £40 (6.1) to £50 (20)
Relative risk of SSI in the WEPD arm	0.6	0 to 1.0	Standard care is dominated for relative risk lower than 0.89 Standard care is cost-effective for relative risk between 0.89 and 0.90 Standard care optimal but cost-ineffective for relative risk higher than 0.90
Price of WEPD (£)	16.5	0 to 100	Standard care is dominated for prices of WEPD lower than £61 WEPD is cost-effective at £20,000/QALY for prices between £61 and £66 WEPD is optimal but cost-ineffective for prices higher than £66

Varying the relative risk of SSI in the intervention arm (thus modifying the clinical effectiveness of WEPD) across the entire range of possible values revealed that the WEPD strategy dominates standard care for all RR values lower than 0.89. Standard care becomes the optimal option when the RR is higher than 0.89. Ultimately, varying the price of the WEPD in the range 0 to £100 indicated that WEPD dominates standard care for prices lower than £61. Thus standard care becomes the optimal option for WEPD prices beyond £61. A willingness-to-pay threshold of £20,000 per QALY gained was considered in all interpretations of the sensitivity analyses results.

Scenario analyses

The first alternative scenario used UK-specific data for small bowel surgery (probability of SSI, postoperative mortality and inpatient length of stay) as opposed to large bowel surgery in the base-case. WEPD is associated with an average cost of £5,286 and an average benefit of 0.0606 QALYs, while standard care yielded an average cost of £5,330 and an average benefit of 0.0605 QALYs (Table 4.5). The CEAC indicated that the WEPD has 89.1% probability of generating a positive net monetary benefit at a threshold of £20,000/QALY and 90.1% at £30,000/QALY (Figure 4.5).

The second scenario investigated the effect of more intensive care received for the SSI in a primary setting, after discharge. The WEPD was associated with an average cost of £5,221 and an average benefit of 0.0606 QALYs, while standard care yielded an average cost of £5,272 and an average benefit of 0.0605 QALYs (Table 4.5). The WEPD has 89.0% probability of generating a positive net monetary benefit at a threshold of £20,000/QALY and 89.8% at £30,000/QALY (Figure 4.5).

Structural uncertainty

The alternative decision model (model 2) indicated that the WEPD strategy was associated with an average cost of £5,672 and a gain of 0.0605 QALYs, while standard care was associated with an average cost of £6,056 and a 0.0604 QALY gain (Table 4.5). WEPD thus dominates standard care as it is cost saving and more effective. CEACs suggest that WEPD is highly likely to be cost-effective across the range of reasonable WTP thresholds (Figure 4.5).

4.2.3. Discussion

Summary of findings

The results of the decision model suggested that, based on the best available data, the WEPD strategy appears to be cost-effective compared to standard care (i.e. not using the WEPD) when used in adult patients undergoing large bowel surgery. This finding was generally robust to a range of sensitivity analyses and scenario analyses as well as to an alternative model structure. WEPD was the dominant strategy across all the considered scenarios; in the base-case analysis the WEPD strategy was on average £43 less costly and brought an average additional benefit of 0.0001 QALYs compared to standard care. No other economic evaluations looking at the cost-effectiveness of WEPDs have been identified, so these results cannot be compared to any other study.

Varying the parameters reflecting the behaviour of health care providers i.e. probability of discharge with SSI, inpatient length of stay and GP referral attitude, did not affect the cost-effectiveness recommendation: the WEPD strategy dominated standard care across the range of plausible values. In the proposed decision model these parameters only bear an influence on costs: the more time spent as an inpatient (either by delaying the discharge or by encouraging hospital readmissions from GPs), the larger the incremental cost associated with the WEPD option. In other words, the longer a patient stays in hospital the more likely it is that preventing a SSI will be cost-saving. These findings are in line with intuition and previous research, which showed that inpatient care has the largest contribution to health care costs attributable to SSI (section 2.1). Indeed, the scenario analysis which considered intensive primary care resource utilisation returned similar results to the base-case in that WEPDs were highly likely to dominate standard care. The same finding was obtained in the scenario assuming that patients undergo small bowel surgery.

Impact of uncertainty

The structure of the base-case model attempted to incorporate the intricacies of SSI management, which rely on the interaction between secondary and primary health care services. Modelling SSI care is further complicated by a number of particularities which include the lack of reliable and rich data on SSI management as well as the difficulties of accounting for the various types of SSI and their implications on cost and patient outcomes. The model attempts to account simultaneously for SSI causative pathogens, SSI severity and the behaviour of health care professionals towards managing SSI patients. This is in line with the decision model for antibiotic prophylaxis after orthopaedic surgery in the UK, developed by Cranny *et al.* (314): the authors incorporated both the distinction between non-MRSA and MRSA SSI and that between superficial and deep/joint SSI, but did not include any primary care costs in their analysis. Three other decision models offered a US perspective: Slobogean *et al.* (318) investigated the cost-effectiveness of antibiotic prophylaxis for fractures and accounted only for the severity of SSI (superficial vs. deep). Neither the model of Thoma *et* *al.* (321) nor that of Lee *et al.* (315) differentiated between any type of SSI at all, using aggregate costs for SSI care.

The sensitivity analyses did not suggest that these assumptions affect the cost-effectiveness recommendation. An alternative, simpler decision model was constructed with the aim to ascertain whether the base-case model was unnecessarily complex and to investigate the extent to which complexity (or, equally, simplicity) in measuring and valuing resource utilisation affects the overall findings. Although the recommended method of accounting for structural uncertainty is constructing a general model and parameterising uncertainty directly in the model, a review found that most UK Health Technology Assessment (HTA) models accounted for structural uncertainty by providing parallel estimates for the alternative models (342). In this case, the conclusion of the alternative model was similar to that of the base-case analysis in suggesting that the WEPD option was highly likely to be cost-effective by dominating standard care and by offering an additional 0.0001 QALY gain. The difference between the two models was in terms of the incremental cost associated with WEPDs: the alternative model returned an average incremental cost of -£384 compared to -£43 in the base-case and higher than the cost differences indicated in any sensitivity analysis. This suggests that the WEPD strategy could be even more cost saving than originally thought, hence more cost-effective. However, if we reverse the perspective this finding may also suggest that accounting for subtle particularities of SSI care may actually prove any intervention to be less cost-effective than it may appear based on analyses informed by bulk costs. This suggests that incorporating SSI severity and causative pathogens in the model's cost inputs does make a difference and highlights the need for equally detailed HRQoL (specifically health utility) data.

Health utility evidence

The systematic review of the HRQoL data on SSI revealed there is very little reliable information available to describe in terms of utility the experience of SSI patients. The relevant model-based economic evaluations identified in the review relied for their SSI utility values either on clinicians' own judgements or on generic valuations which appear to have been informed by anyone but SSI patients. Only two studies used validated instruments, namely the SF-12 generic health survey, the GHQ and the IADL questionnaires. Mapping from non-preference based health measures to generic preference-based measures is possible (312), but the individual patient data was not available for the SF-12 study in order to generate utility scores, despite contacting the authors. Furthermore, no evidence was identified of a relevant mapping exercise for SSI utilities. No study used the EQ-5D instrument, which is currently recommended by NICE for the purpose of evaluating patient-level outcomes for economic evaluations in the UK (39). Nevertheless, one of the studies that were screened but excluded from the systematic review used EQ-5D to evaluate the HRQoL in surgical patients with MRSA complications, but reported data for a bundle of soft skin/tissue infections and not specifically for SSI (346). Nevertheless, the authors reported a utility decrement of 0.22, which was included in the sensitivity analyses accompanying the proposed model. In the light of all these considerations, the validity of the utility values identified in the systematic review and subsequently used in the decision model can be questioned. Nevertheless, the sensitivity analysis revealed that under the base-case assumptions using WEPDs is likely to be cost-effective for a utility decrement as little as 0.02. Considering that it has been suggested in the literature that the minimum clinically significant utility difference is 0.03 (347) and that SSI diagnosis is predominantly based on clinical signs, it is unlikely that this piece of information biased the overall cost-effectiveness recommendation in this case. However, this threshold relies on the substantial clinical benefit demonstrated by WEPDs: should this change, the need for more accurate utility data may become more stringent. An exploratory deterministic two-way sensitivity analysis explored the joint impact of variation in the relative risk of SSI in the WEPD arm and the utility decrement (Figure 4.6). For values of the RR below 0.89 and beyond 0.92 the optimal strategy is clear, namely the WEPD option or standard care, respectively. However, if the RR lies between 0.89 and 0.92, the exact utility decrement can be decisive in establishing the cost-effective alternative.

Several important questions regarding SSIs remain unexplored in the HRQoL literature. Do patients offer different valuations for SSI across various types of surgical interventions e.g. is the utility for SSI after orthopaedic surgery different to that of SSI after cardiac surgery, *ceteris paribus*? Do SSI utilities reflect the severity of infection e.g. is the utility for deep SSI lower than that of superficial SSI? And how do SSI utilities vary over time, especially for slow-healing infections? Reliable answers to these questions are pre-requisites for future economic evaluations of technologies and interventions aimed to reduce SSI.

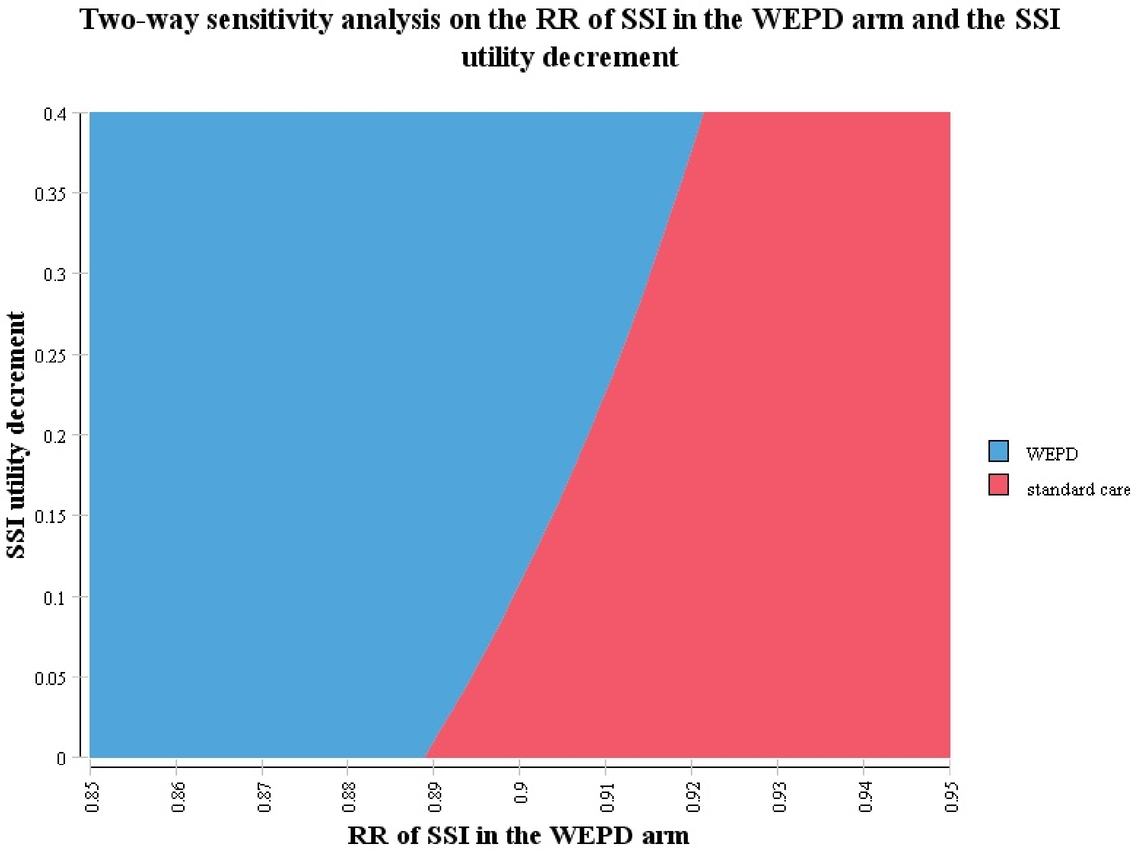


Figure 4.6 Decision model: two-way sensitivity analysis - joint impact of cost-effectiveness of the WEPD effectiveness and SSI utility decrement

Note: The figure indicates which option is cost-effective for the corresponding combinations of the two parameters, either WEPD (blue) or standard care (red).

Strengths and limitations

The decision model features several strengths: first, it accounts for evidence-based factors that influence the risk and burden of SSI (i.e. type of SSI, pathogenic agents, care received in secondary and primary settings). At the risk of challenging the parsimony principle (338, 339), the model incorporates all these considerations because they are supported by evidence in the literature and they have straightforward implications on the costs and outcomes associated with SSI management. Second, the model is informed by two systematic reviews, one on the clinical effectiveness of the intervention under evaluation (Chapter 3) and the other on the HRQoL associated with SSI (section 4.1). Third, the model is based on literature sources and official statistics that are as relevant as possible to the current clinical context of the UK. Fourth, a range of additional analyses have been conducted to test the robustness of the main findings.

It could be argued that a longer time horizon would have been more relevant, but the absence of reliable data would have led to making further assumptions; for instance, constant health utility across time had to be assumed for SSI patients. A short time horizon does not favour the WEPD because it can be expected that slow-healing SSIs are also the most costly and burdensome. No apparent consensus on the appropriate time horizon is available in the literature: for example, the decision model of Lee *et al.* (320), which looked at wound infection after caesarean delivery, also used a one month horizon, while other models took a lifetime perspective (314). The same line of reasoning applies for the assumption that MRSA and non-MRSA SSI cases spend the same number of inpatient days: the international literature suggests that MRSA-SSI patients are likely to have longer spells and to incur higher costs, thus favouring the cost-effectiveness of WEPDs. Although determined by limited available data, these two assumptions are undoubtedly conservative because they favour standard care and therefore have little potential to introduce bias in the model's conclusions.

The main limitation of the model is that several probability values have only been assumed and informed by discussions with relevant medical staff, given the lack of relevant published sources discussing the type of care delivered in a primary setting for SSI. Expert opinion (surgeons, GPs, district nurses and practice nurses) informed the number of GP, practice nurse and district nurse visits as well as the antibiotic regimen in primary care. However, the sensitivity analyses explored this limitation and found that the influence of these variables was little. Furthermore, the SSI surveillance programmes referred to in the model have only studied hospital-related care, while all the medical professionals approached in the development stage of this model conveyed the message that an important part of care is received after discharge, where little reliable data are currently available. Due to the lack of reliable data, the model also ignored the cost of consumables such as wound dressings and wound swabs. However, incorporating these costs would increase the cost of SSI care and thus favour the cost-effectiveness of WEPDs, making the current base-case estimate slightly conservative. Indeed, such resource items were incorporated in the total SSI care costs which informed the alternative decision model and suggested even larger cost savings than in the base-case analysis.

As pointed out before in section 4.1, few studies explicitly and reliably investigated SSI-related health utility and the utility decrement used in the base-case relied on the most reliable estimate. This casts some doubt over any utility decrement that can inform the model at this point. Moreover, the model assumed that the utility decrement does not differ with respect to the severity of the SSI (superficial vs. deep/organ), the causative agent (non-MRSA vs. MRSA) and the type of surgery (large bowel vs. small bowel) as there is yet no evidence in the literature to support this. The utility score for non-infected patients was informed by the study of Janson *et al.* (332): although the study sample comprised Swedish patients, the authors reportedly used the UK value set (348) to convert the EQ-5D scores into utilities.

A further limitation relates to the clinical effectiveness of the WEPD itself. The meta-analysis (Chapter 3) that informed the decision model identified 12 studies conducted over a span of more than 40 years, but they all had poor quality and the largest sample size was 360 patients. The threshold analysis suggested that WEPDs would still be cost-effective for RR lower than 0.89; in other words, as little as 11% relative decrease in SSI rate would be enough under the model's assumptions for the WEPD strategy to be cost-effective.

Finally, the findings of the model were found to change very little as a result of the uncertainty around most parameters. It has been outlined in the literature that robustness *per se* should not be regarded as a desirable property of decision models, as a model whose conclusions do not change when varying the input data may reflect a modelling error (339). A modelling error is unlikely to explain robustness in this case: the pivotal inputs to the cost-effectiveness of WEPDs, as demonstrated in the threshold analyses, are the low price of WEPDs relative to the cost of SSI care and the seemingly large clinical benefit associated with the use of WEPDs (reflected in the relative risk of SSI in the WEPD arm). While the price cannot be expected to change dramatically, the clinical effectiveness estimate is based on poor quality RCTs and more reliable evidence is still expected. When these estimates become available and are used to inform the model, a re-assessment of the cost-effectiveness drivers may offer further insights.

Relation to other studies

No other economic evaluations of WEPDs have been identified, so the decision model's findings cannot be directly compared to other results. The most recent study which offers data on SSI care in a primary setting in the UK was published by Tanner and colleagues in 2009 (195). Their study collected data on 29 SSI patients following colorectal surgery and found that primary care costs amount to about 15% of total SSI costs (on average £1,563 out of £10,523 per SSI patient), thus suggesting that the largest part of the SSI cost burden comes from inpatient care. Furthermore, district nurse visits only accounted for approximately 80% of primary care costs. These results support the model's finding that primary care costs are unlikely to influence the cost-effectiveness recommendation for WEPDs. However, the study published by Tanner *et al.* (195) reported total and average resource use and costs, respectively, without any mention of the variability around these quantities. For example, the authors reported a total of 623 inpatients days and 553 district nurse visits as part of SSI care, but gave no measure of variation around these estimates, so the reader is left without knowing how nurse visits were distributed in the study sample. Such variability is an important aspect, as illustrated by the older study of Davey *et al.* (349): out of seven patients with a SSI in primary care, one patient alone received 57 district nurse visits, another patient received two visits and the rest no visit at all. The absence of any measure of reported variability is the main justification why the results of Tanner and colleagues did not inform the base-case analysis; still, they informed one of the scenario analyses and the alternative decision model and their findings have been discussed above.

Further research

The findings of the decision model presented in this Chapter need to be interpreted within the larger perspective of health care decision making processes. It has been recognised that economic appraisal in health care needs to take an incremental and iterative approach where newly gathered evidence is interpreted and integrated with previous information to generate valid findings and new research questions (69, 350). The framework proposed by Sculpher *et al.* (69) suggested five stages in conducting economic evaluations of health technologies, namely: identifying decision problems; synthesis and modelling given available evidence; setting research priorities; primary research; and synthesis and modelling (Figure 4.7).

When applying the decision framework above onto the issue of using WEPDs to reduce SSI, it can be noted that the systematic review of clinical effectiveness of WEPDs (Chapter 2) and the decision model presented in this Chapter are part of stages 2 and 3, where existing evidence is synthesised and interpreted to form early judgements on the potential effectiveness and cost-effectiveness of WEPDs. All the available evidence suggests that WEPDs are likely to be both effective and cost-effective. Still, it must be acknowledged that existing data are of questionable quality (especially regarding clinical effectiveness), largely absent (especially HRQoL information) and based on a number of assumptions (e.g. pathways of care in a primary setting). These lay the premises for advancing the evidence generating process to stage 4, i.e. primary research, in order to offer reliable answers to the withstanding questions.

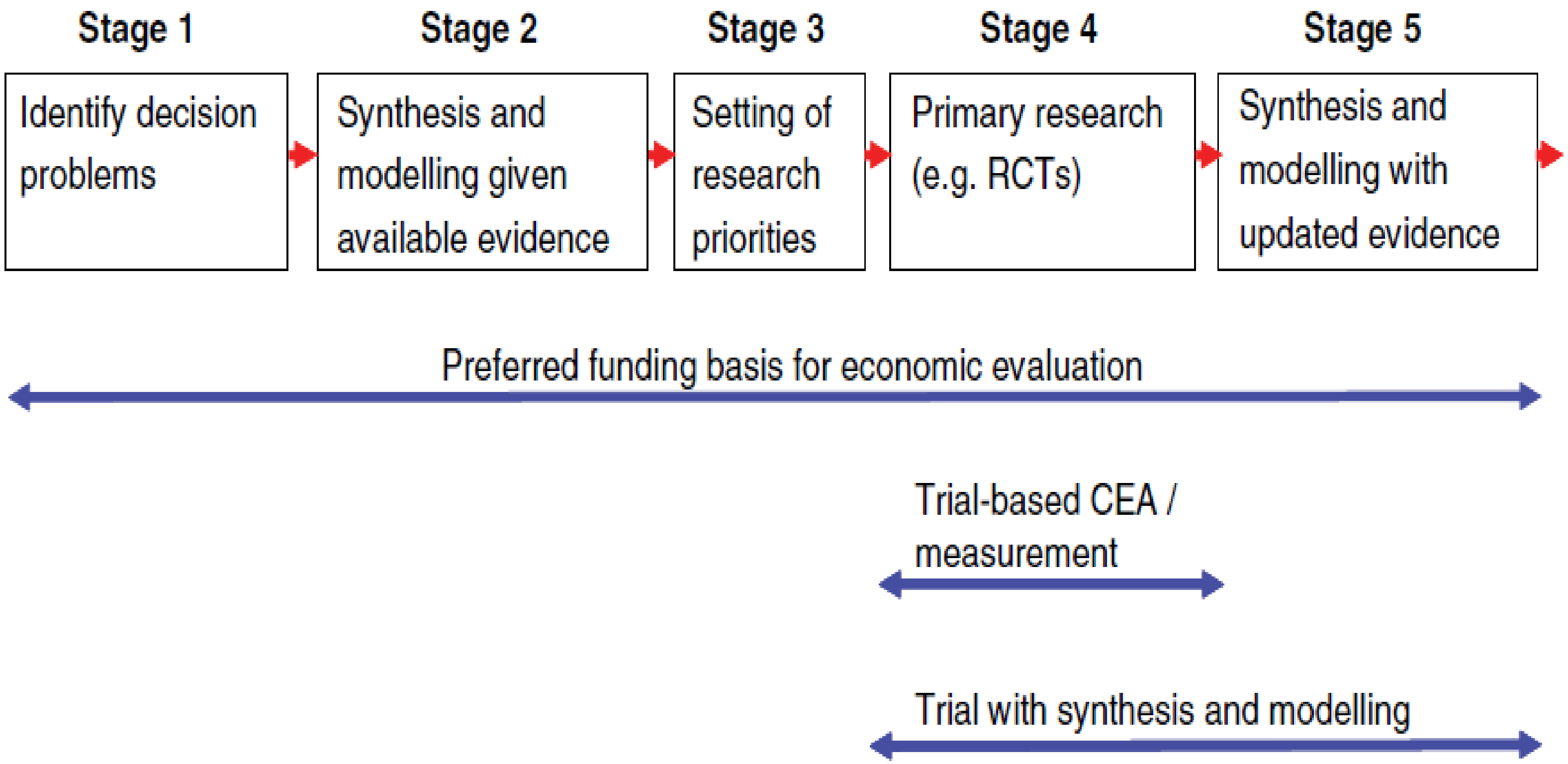

Stage 1
Stage 2
Stage 3
Stage 4
Stage 5
Identify decision problems
Synthesis and modelling given available evidence
Setting of research priorities
Primary research (e.g. RCTs)
Synthesis and modelling with updated evidence
Preferred funding basis for economic evaluation
Trial-based CEA / measurement
Trial with synthesis and modelling

Figure 4.7 The five stages of conducting economic evaluation of health care technologies

Source: Sculpher *et al.* (2006), reproduced with permission

In the light of the above considerations, a definitive RCT with a reasonably large sample size and embedded health-related quality of life data collection is required to offer reliable estimates of the clinical effectiveness of the WEPD and the patient burden of SSI. Beyond offering reliable estimates of effectiveness and cost-effectiveness, such a RCT could provide better information about the care received by SSI patients in a primary care setting and the pragmatic discharge policies of SSI patients (to be presented in Chapter 5).

The role of early modelling is not limited to warranting further research, as outlined in the iterative approach above, but also to focus future data collection on relevant processes. Therefore the importance of early model-based economic evaluations to inform the design of RCTs has long been recognised (351). In the case of evaluating the potential benefit of WEPDs in reducing SSI, the decision modelling exercise identified inpatient length of stay and the SSI utility decrement as potentially important cost-effectiveness drivers, which suggests that the RCT must ensure close patient follow-up within the relevant time horizon in order to capture the patient events occurring at the secondary care - primary care interface (especially readmissions) and the relevant HRQoL information.

4.3. Conclusion

Based on the best available evidence, including the estimated clinical effectiveness presented in Chapter 3, WEPDs are likely to be cost-effective when compared to standard care in reducing SSI rate after open abdominal surgery. This result was robust to the sensitivity and scenario analyses as well as to an alternative model structure. The clinical effectiveness of WEPDs emerged as the main driver of cost-effectiveness estimates. In line with the iterative approach to the economic evaluation of medical technologies, these findings warrant the conduct of a large, high quality RCT with an embedded economic evaluation that can offer reliable effectiveness and cost-effectiveness estimates on the benefit of WEPDs compared to standard care. The methods and findings of such a trial are presented in the following Chapter.

CHAPTER 5. COST-EFFECTIVENESS ANALYSIS OF WOUND-EDGE PROTECTION DEVICES VS. STANDARD CARE: THE *ROSSINI* TRIAL

ROSSINI trial results have been published as:

Pinkney TD, Calvert M, Bartlett DC, **Gheorghe A**, Redman V, Dowswell G et al. (2013). "Impact of wound edge protection devices on surgical site infection after laparotomy: multicentre randomised controlled trial (ROSSINI Trial)." BMJ **347**:f4305.

The previous Chapters demonstrated that, based on the best available evidence, WEPDs are likely to be both effective (Chapter 3) and cost-effective (Chapter 4) in reducing SSI rates in adults undergoing open abdominal surgery, when compared to standard care. As discussed at the end of Chapter 4, substantial uncertainty surrounded these findings due to the unsatisfactory methodological quality of existing studies. Further research was therefore warranted to provide high quality evidence of WEPD benefit. This Chapter presents the methods and results of the cost-effectiveness analysis in the ROSSINI trial, which compared the use of wound-edge protection devices (WEPDs) with standard care in adults undergoing laparotomy. The first section of the Chapter outlines briefly the methods and main results of the ROSSINI trial. The second section presents the methods and results of the within-trial economic evaluation of WEPDs informed by primary data collected alongside ROSSINI.

5.1. The ROSSINI trial

5.1.1. ROSSINI methods

The ROSSINI (Reduction of Surgical Site Infection using a Novel Intervention) trial aimed to assess the benefits to the patients and to the NHS of using WEPDs to reduce SSI in adult patients undergoing laparotomy. The main characteristics of the trial are briefly presented here, as ROSSINI methods have been described in detail elsewhere (352).

Objectives

The trial's primary objective was to determine the WEPD's effectiveness in reducing SSI rates 30 days after surgery. Secondary objectives were: to determine the effectiveness of the WEPD by degree of surgical wound contamination (clean, clean-contaminated, contaminated, dirty); to assess the impact of the use of WEPD on patient health-related quality of life; to assess the impact of the use of WEPD on length of stay in hospital; and to investigate the cost-effectiveness of the WEPD compared to standard care.

Trial design

ROSSINI was a multicentre, randomised, controlled, parallel group trial where adult patients undergoing laparotomy were randomised in a 1:1 ratio to either the control arm i.e. standard intra-operative care, or the intervention arm i.e. standard intra-operative care plus use of a WEPD during the intra-abdominal part of the operation. Patients undergoing laparotomy for any indication were included in order to maximise the generalisability of the findings. Patients less than 18 years of age, laparoscopic cases and patients who had had a laparotomy within the past three months were excluded (352).

Patients were randomised while in the anaesthetic room, immediately prior to surgery, using a secure online portal hosted by the Centre for Clinical Trials at the University of Birmingham. Stratification with embedded minimisation was employed according the following strata: urgency of surgery, likelihood of opening a viscus and likelihood of creating a stoma. The participating patients and all health care staff involved in post-operative care and wound assessments were blinded to the treatment allocation. ROSSINI recruited from general surgical units within NHS hospitals across England.

Outcomes

The primary outcome was occurrence of SSI within 30 days post-operatively, assessed according to the CDC criteria (Chapter 2) (200). The main hypothesis was that use of WEPD would reduce SSI rate by 50%, informed by the study of Horiuchi *et al.* (271). Assuming a conservative 12% SSI rate in the control arm and a 5% dropout rate, the target sample size was 750 patients.

Wound assessors undertook online training by completing an e-learning module and quiz to minimise the potential for inter-assessor variability in wound assessments. Secondary outcomes were: the degree of wound contamination; presence of major comorbidity; HRQoL; length of stay in hospital; health care utilisation and cost-effectiveness of WEPDs compared to standard care; and adverse events. HRQoL was measured using the validated EuroQol EQ-5D instrument (43) at three time points: at baseline (before surgery), at 5-7 days post-operatively and at 30-33 days post-operatively.

Analysis

All the analyses were based on the intention-to-treat (ITT) principle. The primary outcome was analysed using generalised linear models with logit link, binomial error and with surgeon as random effects (353). Continuous data were analysed with the use of mixed models, which include surgeons as random effects. The rates of adverse events were compared between groups by means of Fisher's exact test.

A prospective within trial cost-effectiveness analysis was undertaken from the perspective of the NHS. The chosen time horizon for the analysis was 30 days post-operatively. The incremental cost per additional QALY of the WEPD strategy compared to standard care was assessed to inform clinicians and policy makers of the cost-effectiveness of WEPDs. The detailed methods and results of the cost-effectiveness analysis are presented in the following section.

5.1.2. ROSSINI results

ROSSINI results are reported in detail elsewhere (354). Briefly, between February 2010 and January 2012 a total of 760 patients from 21 surgical centres across the UK were enrolled in the study and randomised to the WEPD (n=382) or control (n=378). 376 patients in the WEPD group and 373 patients in the control group received a laparotomy and were included in the study (Figure 5.1). The characteristics of ROSSINI patients are presented in Table 5.1.

ROSSINI results are presented in Table 5.2. In total, 184 patients experienced an SSI within 30 days of surgery, 91/369 (24.7%) of patients in the WEPD group and 93/366 (25.4%) in the control group (odds ratio (OR) 0.97 95% CI 0.69 to 1.36; p=0.85). The results were consistent across the assessments made at different time points within the study and by different observers, with both the formal clinician wound assessments and the patient self-reported data showing no difference (Figure 5.2).

A WEPD was used in four patients randomised to the control arm and was not used in 29 patients randomised to receive the device. A sensitivity analysis was undertaken to explore the effect of treatment cross-over on the estimate of WEPD effectiveness. In this 'best-case scenario' analysis (in which a maximal benefit from use of WEPD is assumed), all patients allocated to the control group but that received a device were assumed to have had an SSI within 30 days, conversely those patients randomised to WEPD who did not receive a device were assumed to have had no event. In this extreme case analysis the effect of WEPD was still statistically non-significant (OR 0.77; 95% CI 0.54 to 1.09, p=0.14).

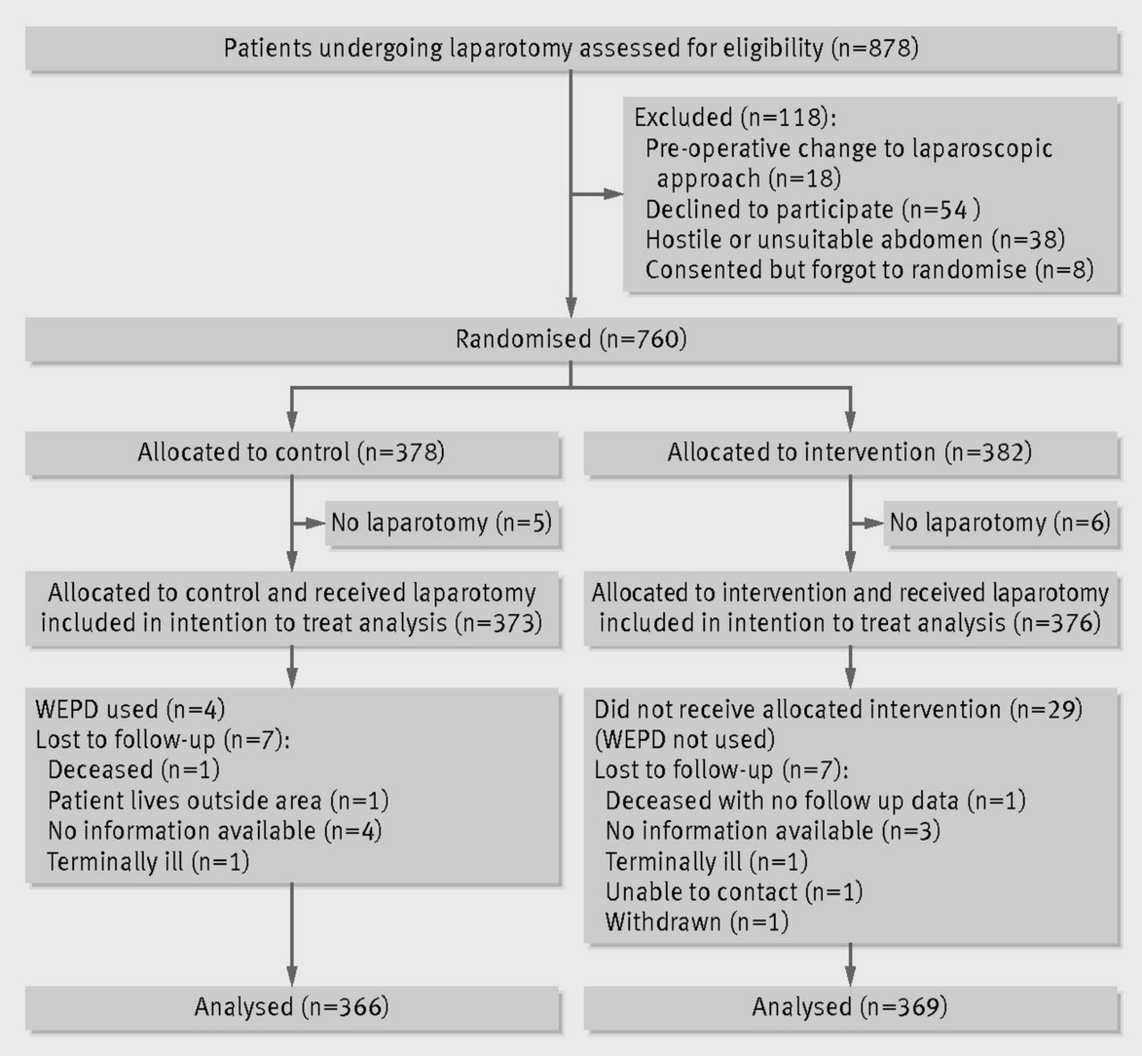

Patients undergoing laparotomy assessed for eligibility (n=878)
Excluded (n=118):
Pre-operative change to laparoscopic approach (n=18)
Declined to participate (n=54)
Hostile or unsuitable abdomen (n=38)
Consented but forgot to randomise (n=8)
Randomised (n=760)
Allocated to control (n=378)
Allocated to intervention (n=382)
No laparotomy (n=5)
No laparotomy (n=6)
Allocated to control and received laparotomy included in intention to treat analysis (n=373)
Allocated to intervention and received laparotomy included in intention to treat analysis (n=376)
WEPD used (n=4)
Lost to follow-up (n=7):
Deceased (n=1)
Patient lives outside area (n=1)
No information available (n=4)
Terminally ill (n=1)
Did not receive allocated intervention (n=29) (WEPD not used)
Lost to follow-up (n=7):
Deceased with no follow up data (n=1)
No information available (n=3)
Terminally ill (n=1)
Unable to contact (n=1)
Withdrawn (n=1)
Analysed (n=366)
Analysed (n=369)

Figure 5.1 CONSORT flow diagram for the ROSSINI trial

Source: Pinkney *et al.* (2013), reproduced with permission[7]

[7] Reprinted from BMJ, 347:f4305, Pinkney TD *et al.*, Impact of wound edge protection devices on surgical site infection after laparotomy: multicentre randomised controlled trial (ROSSINI Trial), Copyright (2013), with permission from BMJ Publishing Group Ltd.

Table 5.1 Patient characteristics in ROSSINI trial

Characteristic	WEPD (n=376)	Control (N=373)
Age (years)		
Median	66.37	64.23
Interquartile Range	54.79 to 74.69	55.51 to 72.83
Male gender (%)	200 (53.19%)	193 (51.74%)
Body Mass Index		
Median	26.50	26.00
Interquartile Range	23.10 to 30.00	23.05 to 29.07
Serum Albumin level		
Median	41.00	40.00
Interquartile Range	34.00 to 44.00	35.00 to 44.00
Diabetes (%)	62 (16.49%)	51 (13.67%)
Current smoker (%)	64(17.02%)	57 (15.28%)
On steroids or immunosuppressed (%)	35 (9.31%)	31 (8.31%)
Clinically jaundiced (%)	21 (5.59%)	20 (5.36%)
Documented MRSA colonisation (at any site) previously (%)	9 (2.39%)	10 (2.68%)
Operation urgency (%)		
Elective	181 (48.14%)	183 (49.06%)
Expedited	117(31.12%)	117(31.37%)
Urgent	75 (19.95%)	71 (19.03%)
Immediate	3 (0.80%)	2 (0.54%)
ASA Grade		
1	36 (9.57%)	49 (13.14%)
2	203 (53.99%)	186 (49.87%)
3	113 (30.05%)	95 (25.47%)
4	4 (1.06%)	7 (1.88%)
5	0 (0.00%)	1 (0.27%)
Unknown	20 (5.32%)	35 (9.38%)
Operation site (%)		
Large bowel	247 (65.69%)	237 (63.54%)
Small bowel	34 (9.04%)	48 (12.87%)
Hepatobiliary	77 (20.48%)	72 (19.30%)
Gastric	15 (4.02%)	8 (2.14%)
Cholecystectomy	1 (0.27%)	2 (0.54%)
Vascular	1 (0.27%)	0 (0.0%)
Abdominal hysterectomy	0 (0.0%)	2 (0.54%)
Unknown	1 (0.27%)	4 (1.07%)
Stoma created (%)	109(28.99%)	106 (28.42%)
Cancer resection (%)	223(59.31%)	219 (58.71%)
Skin Prep used (%)		
Chlorhexidine	136 (36.17%)	135 (36.19%)
Aqueous Betadine	215 (57.18%)	197 (52.82%)
Alcoholic Betadine	16 (4.26%)	29 (7.77%)
Towels/mops used on wound edges (%)	42 (11.17%)	78 (20.91%)
Type of surgery performed (%)		
Clean	24 (6.38%)	31 (8.31%)
Clean-contaminated	275 (73.14%)	268 (71.85%)
Contaminated	48 (12.77%)	48 (12.86%)
Dirty	29 (7.71%)	25 (6.70%)
Duration of surgery (hours)		
Median	3.0	2.73
Interquartile Range	2.0 to 4.0	2.0 to 4.0
NNIS index		
Median	1	1
Interquartile Range	0 to 1	0 to 1
Prophylactic antibiotic given (%)		
On Induction	321 (85.37%)	322 (86.33%)
During procedure	25 (6.65%)	18 (4.83%)
Catheters left in place (%)	6 (1.60%)	6 (1.61%)
Grade of operating surgeon (%)		
Consultant	302 (80.32%)	280 (75.07%)
Trainee	69 (18.35%)	82 (21.98%)
Grade of surgeon closing fascia (%)		
Consultant	186 (49.47%)	197 (52.82%)
Trainee	182 (48.40%)	157 (42.09%)

Table 5.2 Primary and secondary outcomes in ROSSINI trial

Outcome	WEPD	Control	Estimate (95% CI)	P value
Primary outcome				
Surgical site infection (SSI) within 30 days	91/369 (24.7)	93/366 (25.4)	0.97* (0.69 to 1.36)	0.85
Secondary outcomes				
Mean (SD) EQ-5D	0.69 (0.29)†	0.69 (0.30)‡	0.001§ (−0.04 to 0.05)	0.95
Median (IQR) length of hospital stay (days)	9 (6 to 15)	9 (6 to 14)	1.03¶ (0.88 to 1.19)	0.82
Degree of wound contamination:				
Clean	8/24 (33.3)	7/29 (24.1)	1.76* (0.40 to 7.70)	0.43
Clean-contaminated	61/269 (22.7)	63/263 (24.0)	0.94* (0.62 to 1.42)	0.76
Contaminated	10/48 (20.8)	15/48 (31.3)	0.601* (0.23 to 1.63)	0.31
Dirty	12/28 (42.9)	7/25 (28.0)	1.85* (0.50 to 6.87)	0.33

Legend:
IQR=interquartile range
*Odds ratio
†n=318
‡n=313
§Difference in means
¶Hazard ratio

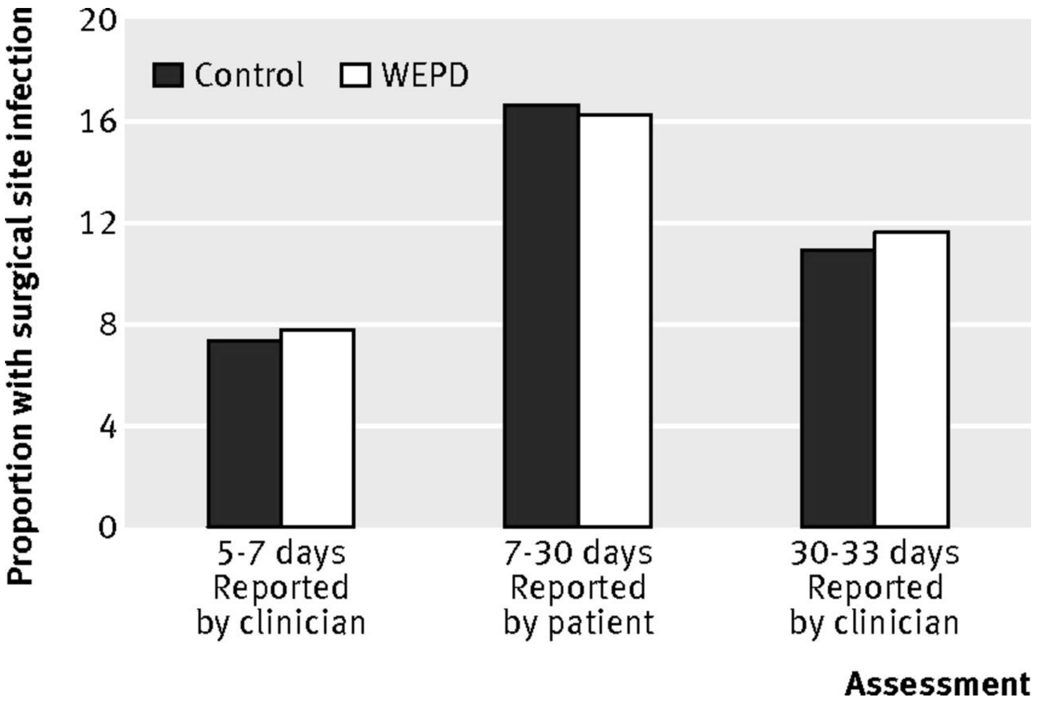


Figure 5.2 SSI rates by treatment group at various time points in the ROSSINI trial

Source: Pinkney *et al.* (2013), reproduced with permission[8]

[8] Reprinted from BMJ, 347:f4305, Pinkney TD *et al.*, Impact of wound edge protection devices on surgical site infection after laparotomy: multicentre randomised controlled trial (ROSSINI Trial), Copyright (2013), with permission from BMJ Publishing Group Ltd.

5.2. Economic evaluation of WEPDs vs. standard care alongside ROSSINI

The aim of the within-ROSSINI economic evaluation was to provide evidence on the incremental cost-effectiveness of WEPDs compared to standard care in reducing SSI when used in adults undergoing open abdominal surgery. The methods for conducting economic evaluations using clinical trials data have been described previously (62-64) and the principles have been outlined in Chapter 1 (section 1.2). The present economic evaluation is reported according to the CHEERS Statement 2013 (329) (Appendix 6).

5.2.1. Methods

Characteristics of ROSSINI patients were given in Table 5.1. ROSSINI was conducted in NHS hospitals. The trial-based economic evaluation took a health care provider perspective and thus considered only cost centres relevant for the NHS and Personal Social Services. The intervention under scrutiny was the use of a WEPD during surgery. The comparator was no WEPD use. In order to enhance the generalisability of the trial, the surgical teams were given the liberty to use retraction and SSI prophylactic procedures of their choice. The time horizon was 30 days post-operatively, in accordance with SSI monitoring in the English NHS(206). Given the short time horizon, no discounting was applied to costs and outcomes.

Health outcomes, preference-based outcomes and resource use data were collected from the participating sites using custom designed paper-based case report forms (CRFs), which were completed by patients or trial staff, as appropriate, at each site then managed centrally at the Centre for Clinical Trials at the University of Birmingham. Information on clinical outcome data was reported above (sub-section 5.1.2).

Health-related quality of life

HRQoL was assessed using the EuroQol EQ-5D 3L questionnaire (the English version and validated for use in the UK, Appendix 7), a standardised generic preference based instrument that describes a patient's health status using a single index value (43). EQ-5D has five dimensions (mobility, self-care, usual activity, pain/discomfort and anxiety/depression) and each dimension has three mutually exclusive levels the patient has to choose from (no problem, some problem or extreme problem). There are 243 different health states described by the EQ-5D, each health state being associated with a HRQoL weight derived from the preferences of a representative sample of the UK population using the time trade-off technique. The EQ-5D score is bounded to 0 (death) and 1 (perfect health), but negative scores are possible for states perceived to be worse than death (348).

There has been relatively little research in the HRQoL assessment of SSI. Very little is also known about the comparative validity of different quality of life instruments. The systematic literature review of SSI utility values (Chapter 4) identified a small number of studies, the majority of which used historical and unspecific (e.g. hip infection) utility values. EQ-5D was chosen as the HRQoL instrument in this study for the purpose of its relevance for the UK policy makers, particularly NICE (39).

The EQ-5D was administered to patients in ROSSINI at three time points: at baseline (prior to surgery), at 5 to 7 days post-operatively and at 30 to 33 days post-operatively. The first assessment was conducted in clinic, after the patient provided informed consent and before randomisation. The second assessment (5 to 7 days) was performed on the hospital ward if still inpatient or at discharge, as applicable. The third assessment (30 to 33 days) was performed on the hospital ward if still inpatient or, more often, in the outpatient clinic on the occasion of the scheduled follow-up visit.

Resource use

Data on resource utilisation of health care resources in both secondary and primary care settings was collected using the custom designed CRFs (Appendix 8). CRF6 recorded resource use items related to inpatient care, filled in by research nurses or dedicated trial staff at each site, using hospitals' databases and patient notes as appropriate. CRF4 recorded resource utilisation in primary care setting i.e. access to GP, practice nurse, district nurse and outpatient clinic, together with any medication received. This information was recorded by patients in clinic on the occasion of the scheduled follow-up visit at 30 to 33 days post-operatively.

For subjects who were diagnosed with an ongoing SSI or were still inpatients at this follow-up visit, an individualised follow-up procedure was set-up. This involved telephoning the respective patients and asking for their consent to contribute data to the follow-up procedure. Upon gaining consent, the trial office posted the primary care resource use CRF (CRF4a) and an EQ-5D questionnaire at the patient's home address. Patients were asked to complete the forms and post them back to the trial office using the freepost envelope provided in the pack. Due to the 30-day time horizon, information collected during the extended follow-up was used only in the cost analysis of SSI care and not in the present economic evaluation.

A wound-dressing diary was devised as a separate document shortly after recruitment started and ethical approval for its introduction in ROSSINI was issued on 7th September 2011, when more than 550 patients had been recruited. The aim of the wound dressing diary was to capture the effect of the WEPD on the total cost of dressings utilised in wound care, both in a secondary and primary setting. Due to regulatory delays, the wound-dressing diary was ultimately implemented as a pilot in four sites and the information collected as such was not included in the present economic evaluation.

Unit costs

Unit costs were valued in £ (2011 value). Inpatient care items were sourced from the NHS Reference Costs 2010-2011. Primary care items were sourced from the Personal Social Services Resource Unit (PSSRU) Unit Costs and Social Care 2010-2011 (333). Medication unit costs were taken from the British National Formulary 2011 (355).

All unit costs were average national costs (Table 5.3). Consistent with the NHS perspective, only resource use items affecting the NHS budget were considered. Total resource costs were obtained by summation of the individual resource costs for each category of resource item accessed by trial patients. Individual resource costs were obtained by multiplying the resource use by the corresponding unit costs.

Data analysis

The base-case analysis included all the patients with complete primary outcome data (information on SSI status). Any missing cost and HRQoL data as well as patient-level characteristics were imputed using the multiple imputations using chained equations method (MICE) (see paragraph *Missing data* below). The analysis included descriptive statistics for the resource use items, resource costs (both at aggregate and individual level) and HRQoL scores.

QALYs were calculated by multiplying the utility weight associated with each individual health state and the time spent in that health state. QALYs were calculated based on the baseline and 30-day EQ-5D assessments and were adjusted for baseline utility (356).

Table 5.3 ROSSINI trial: unit costs at 2011 value

Resource	Unit cost (£)	Source
WEPD (intervention)	15.1	Manufacturer
HOSPITAL CARE		
Day on general ward	311.0	NHS Reference Costs 2010/2011 (357)
Day in ITU	1,515.0	NHS Reference Costs 2007/2008* (358)
Day in HDU	856.0	NHS Reference Costs 2007/2008* (358)
PRIMARY CARE		
GP visit	36.3	Curtis 2011 (333)
Practice nurse visit	13.2	Curtis 2011 (333)
District nurse visit	73.0	Curtis 2011 (333)
Outpatient clinic visit	101.0	NHS Reference Costs 2010/2011 (357)
Medication (antibiotics, painkillers)	as appropriate	British National Formulary 2011 (355)

* The unit costs for a day in Intensive Therapy Unit (ITU) and a day in High Dependency Unit (HDU) were not available in NHS Reference Costs 2010/2011. The last available document where they were given explicitly was the 2007/2008 edition. For the purpose of this analysis, the 2007/2008 unit costs were updated to their 2011 value using the appropriate Hospital and community health services (HCHS) pay and price inflation (Curtis 2011).

The average differences in costs and outcomes, as well as the 95% confidence intervals around the point estimates and the ICER, were calculated using bias-corrected and accelerated (BCa) non-parametric bootstrap methods with 1,000 replications (54). The differences in costs and effects were plotted on the cost-effectiveness plane, a visual decision-aiding tool representing the incremental costs and effects of the intervention under evaluation relative to the next best option (47). One alternative is said to dominate another if both the average costs and average effects associated with it are relatively lower than another's. If the evaluation does not show a case of dominance, the ICER is calculated as the ratio between the difference in mean costs and the difference in mean QALYs between the intervention and the comparator (Chapter 1, equation 1.1). Cost-effectiveness acceptability curves (CEACs) were plotted, indicating the probability of the two alternatives to be cost-effective at varying thresholds of the decision makers' willingness to pay for an additional unit of outcome (58).

Sensitivity analyses were performed to check the robustness of cost-effectiveness findings, as follows:

1. A complete case analysis based on trial subjects with complete primary outcome, cost and HRQoL data.

2. Adjusted analyses for both base-case and complete case scenarios, where differences between the trial's arms were investigated using generalized linear models. Total costs and EQ-5D scores were modelled using the intervention (treatment arm) and other relevant baseline characteristics as covariates (identified with clinical input): treatment arm, baseline utility (only for adjusting incremental QALYs), plan to create a stoma, plan to create a viscus (defined as any internal organ), elective/emergency surgery, age, BMI, diabetes, current smoking status and SSI. The total cost and QALY values were regressed against the variables above using generalised linear models with an identity link (353). A gamma distribution was assumed for costs and a normal distribution was assumed for QALYs. All the analyses were performed using SAS 9.2® software (304) and R 2.15.3 software (359).

Missing data

Missing data on costs and health utilities were imputed using independent chained equations (MICE) methods (26). The imputations were performed using the *mice* package available in R statistical software (359). Resource costs and EQ-5D data were imputed using an algorithm which predicted the missing values based on a wide range of variables: patient and operative characteristics (age, BMI, smoking status, diabetes, plan to open viscus, plan to create a stoma, elective/emergency surgery, ASA grade, duration of surgery); hospital care cost items (cost of days on ward, cost of ITU days, cost of HDU days); primary care cost items (cost of GP visits, cost of GP-prescribed medication, cost of district nurse visits, cost of practice nurse visits and cost of outpatient clinic visits); and EQ-5D scores at baseline, at 5-7 days post-operatively and at 30-33 days post-operatively.

In addition, the MICE imputation model included age and SSI status, for which complete data were available. The predictive mean matching method was used to impute patient-level characteristics: following each cycle of the imputation model, the observed value which was the closest to the predicted value was chosen in order to ensure that only plausible values are imputed. Costs were bounded to be positive and EQ-5D scores were bounded between -0.594 and 1, in accordance with the UK scoring algorithm (348). Aggregate hospital costs, primary care costs and total costs were imputed based on the sum of individual cost items resulted from the imputation model to ensure their convergence. Twenty datasets (each obtained after 20 iterations/cycles of the imputation algorithm) were generated from the imputation process, and then entered the bootstrapping process.

5.2.2. Results

Resource use

The average utilisation of health care resources is presented in Table 5.4. There is no apparent difference between the two treatment groups for secondary care or primary care services, as confirmed by the corresponding p-values. The only notable exception is the number of practice nurse visits: patients in the standard care arm reported twice as many practice nurse contacts than WEPD patients. However, this may well be an artefact of the data collection process, as the information on primary care utilisation was reported by patients themselves, who may not have accurately discriminated between practice nurse and district nurse visits when reporting.

Table 5.4 ROSSINI trial: summary of resource use by treatment group, detailed

Resource use item		WEPD (n=369)	Standard care (n=366)	p-value
HOSPITAL CARE				
Inpatient days				
	N	359	358	
	Mean (SD)	12.55 (15.46)	11.56 (11.68)	0.3350
	SE	0.82	0.62	
	Median	9	9	
Days in ITU				
	N	369	366	
	Mean (SD)	0.93 (3.12)	1.06 (5.46)	0.6913
	SE	0.16	0.28	
	Median	0	0	
Days in HDU				
	N	369	366	
	Mean (SD)	0.60 (1.67)	0.55 (1.51)	0.6396
	SE	0.09	0.08	
	Median	0	0	
PRIMARY CARE				
GP visits				
	N	364	358	
	Mean (SD)	0.43 (0.81)	0.51 (1.03)	0.2474
	SE	0.04	0.05	
	Median	0	0	
District nurse visits				
	N	360	355	
	Mean (SD)	3.43 (7.24)	3.52 (6.94)	0.8644
	SE	0.38	0.37	
	Median	0	0	
Practice nurse visits				
	N	366	361	
	Mean (SD)	0.16 (0.70)	0.32 (1.21)	0.0355
	SE	0.04	0.06	
	Median	0	0	
Outpatient clinic visits				
	N	364	363	
	Mean (SD)	0.42 (1.09)	0.31 (0.71)	0.1205
	SE	0.06	0.04	
	Median	0	0	

This potential source of bias is further explored in Table 5.5, where the difference between the two arms is explored in terms of 'primary care contact points', a variable encompassing all types of care received in a primary care setting, and 'nurse visits', a variable which aggregates district nurse and practice nurse visits. Moreover, a large proportion of patients reported no GP visits or nurse visits within the 30 day time horizon. A secondary analysis explored the difference between treatment arms for patients who reported at least one primary care visit, in order to investigate whether differential proportions of zero values in the two groups mask any underlying difference (Table 5.5). The decision to conduct this secondary analysis was prompted by the large number of zero values for the number of practice and district nurse visits. Neither of the analyses revealed any difference between patients in the two arms in terms of the volume of care received.

Table 5.5 ROSSINI trial: summary of resource use by treatment group, overview

Resource use item		WEPD (n=369)	Standard care (n=366)	p-value
HOSPITAL CARE				
Inpatient days				
	N	359	358	
	Mean (SD)	12.55 (15.46)	11.56 (11.68)	0.3350
	SE	0.82	0.62	
	Median	9	9	
PRIMARY CARE				
Primary care points of contact (includes GP visits, all nurse visits and outpatient clinic visits)				
	N	350	347	
	Mean (SD)	4.38 (7.59)	4.47 (7.02)	0.8795
	SE	0.41	0.38	
	Median	1	2	
Nurse visits (includes district nurse visits and practice nurse visits)				
	N	357	352	
	Mean (SD)	3.54 (7.16)	3.74 (6.81)	0.6939
	SE	0.38	0.36	
	Median	0	0	

A large number of patients did not report any primary care visits (median is 0). The table section below only looks at patients who reported at least one primary care visit (GP, practice nurse, district nurse or outpatient clinic) and at least one nurse visit, respectively.

Resource use item		WEPD (n=369)	Standard care (n=366)	p-value
Primary care points of contact (includes GP visits, all nurse visits and outpatient clinic visits)				
	N	242	247	
	Mean (SD)	6.88 (8.57)	6.80 (7.70)	0.9163
	SE	0.57	0.51	
	Median	3	3	
Nurse visits (includes district nurse visits and practice nurse visits)				
	N	188	189	
	Mean (SD)	7.18 (8.83)	7.54 (8.04)	0.6937
	SE	0.66	0.61	
	Median	4	4	

Missing data

Despite very low levels of missing data for the primary outcome, the amount of missing data for resource utilisation and patient-reported outcomes was somewhat higher (Table 5.6). EQ-5D scores at 30 days post-operatively were not available for 14% of patients, while hospital and primary care data were unavailable cumulatively for less than 10% of patients (6.66%). Overall, 20.4% of patients had incomplete observations in terms of resource use or HRQoL data. However, there was no imbalance between the two arms with respect to the levels of missing data, which suggests that having complete observations was not influenced by receiving the intervention or not.

Data missingness was further explored by looking at differences in missing observations in subgroups defined by relevant patient, intraoperative and clinical characteristics. Table 5.7 compares the levels of missing EQ-5D data at 30-33 days by several patient-level variables and the results suggest there is no difference with respect to completeness of HRQoL information based on these characteristics. A similar analysis was carried out for resource use data (Table 5.8). It appears that there are differences in levels of missing cost data with respect to two patient-level variables: age and BMI. More specifically, there is more missing cost data in patients below 65 years compared to those above 65 years (9.9% vs. 5.7%, p=0.03). Patients with a higher BMI had more missing cost data than patients with BMI lower than 26.75 (12.5 vs. 6.4%, p<0.01).

It appears that SSI status i.e. having been diagnosed with a SSI or not, does not influence the level of missingness either for HRQoL or for resource use data. Thus there is no evidence to suggest that missing data was influenced by the primary outcome, while there is some evidence that several patient level-variables (age and BMI) may be associated with missingness. The assumption of data missing at random (MAR) appears thus to be plausible.

Table 5.6 ROSSINI trial: summary of missing data, by treatment group

Missing data item		Missing observations (% of trial arm)		
		WEPD (n=369)	Standard care (n=366)	Trial arm differences (p-value)
HOSPITAL CARE				
	Inpatient days	10 (2.7%)	8 (2.2%)	0.64
PRIMARY CARE				
	GP visits	5 (1.4%)	8 (2.2%)	0.39
	Practice nurse visits	3 (0.8%)	5 (1.4%)	0.47
	District nurse visits	9 (2.4%)	11 (3%)	0.63
	Outpatient clinic visits	5 (1.4%)	3 (0.8%)	0.48
PATIENT-REPORTED OUTCOMES				
	EQ-5D data, any time point	51 (13.8%)	53 (14.5%)	0.79

Table 5.7 ROSSINI trial: summary of missing EQ-5D data at 30-33 days by patient-level variables

Variable		Total observations	Missing observations (%)	p-value
Age	Age <= 65 years	383	61 (15.9%)	0.64
	Age > 65 years	352	43 (12.2%)	
Diabetic	Yes	111	17 (15.3%)	0.70
	No	624	87 (13.9%)	
Smoker	Ever smoker	370	49 (13.2%)	0.48
	Never smoker	365	55 (15.1%)	
BMI	BMI <= 26.75	559	82 (14.7%)	0.47
	BMI > 26.75	176	22 (12.5%)	
Duration of surgery	<= 170 minutes	375	56 (14.9%)	0.53
	> 170 minutes	360	48 (13.3%)	
ASA grade	ASA grade <= 2	482	63 (13.1%)	0.25
	ASA grade > 2	253	41 (16.2%)	
SSI status	SSI	551	73 (13.2%)	0.23
	No SSI	184	31 (16.8%)	

Note: Threshold values for variables Age, BMI and Duration of surgery are median values

Table 5.8 ROSSINI trial: summary of missing resource use data by patient-level variables

Variable		Total observations	Missing observations (%)	p-value
Age	Age <= 65 years	383	38 (9.9%)	0.03
	Age > 65 years	352	20 (5.7%)	
Diabetic	Yes	111	10 (9.0%)	0.63
	No	624	48 (7.7%)	
Smoker	Ever smoker	370	28 (7.6%)	0.74
	Never smoker	365	30 (8.2%)	
BMI	BMI <= 26.75	559	36 (6.4%)	<0.01
	BMI > 26.75	176	22 (12.5%)	
Duration of surgery	<= 170 minutes	375	33 (8.8%)	0.35
	> 170 minutes	360	25 (6.9%)	
ASA grade	ASA grade <= 2	482	40 (8.3%)	0.57
	ASA grade > 2	253	18 (7.1%)	
SSI status	SSI	551	40 (7.3%)	0.27
	No SSI	184	18 (9.8%)	

Note: Threshold values for variables Age, BMI and Duration of surgery are median values

Results of the base-case analysis

The base-case analysis used information from all patients with complete primary outcome data (n=735). Figure 5.3 depicts the results of the imputation process across iterations for variables total cost (c_final), baseline EQ-5D score (score0) and EQ-5D score at 30-33days (score EQ3); the imputation sets appear to converge for all the variables from the very beginning and no trend in the imputed values is apparent along the iterations, which suggests that the results of the imputation can be used with confidence (360).

Health-related quality of life

There were no significant differences between patients in the two groups with respect to EQ-5D scores at either time point. At 30 days postoperatively, intervention and control patients reported utility scores of 0.683 and 0.684, respectively (Table 5.9).

Costs

The use of the WEPD was associated with slightly higher inpatient costs than standard care (Table 5.9). Moreover, the difference in primary care costs was minimal. It appears that inpatient care accounts for the largest part of costs within the 30 day time horizon.

Table 5.9 ROSSINI trial economic evaluation base-case analysis: summary of costs and HRQoL data by treatment group

	Mean (SE)	
Variable	**WEPD (n=369)**	**Standard care (n=366)**
Cost of hospital care	5,089.32 (246.80)	4,812.39 (234.14)
Cost of primary care	315.89 (28.65)	317.88 (28.85)
EQ-5D score at baseline	0.751 (0.016)	0.752 (0.016)
EQ-5D score at 30 days	0.683 (0.016)	0.684 (0.016)

Note: SE values were calculated assuming a Gamma distribution for costs and a normal distribution for EQ-5D scores

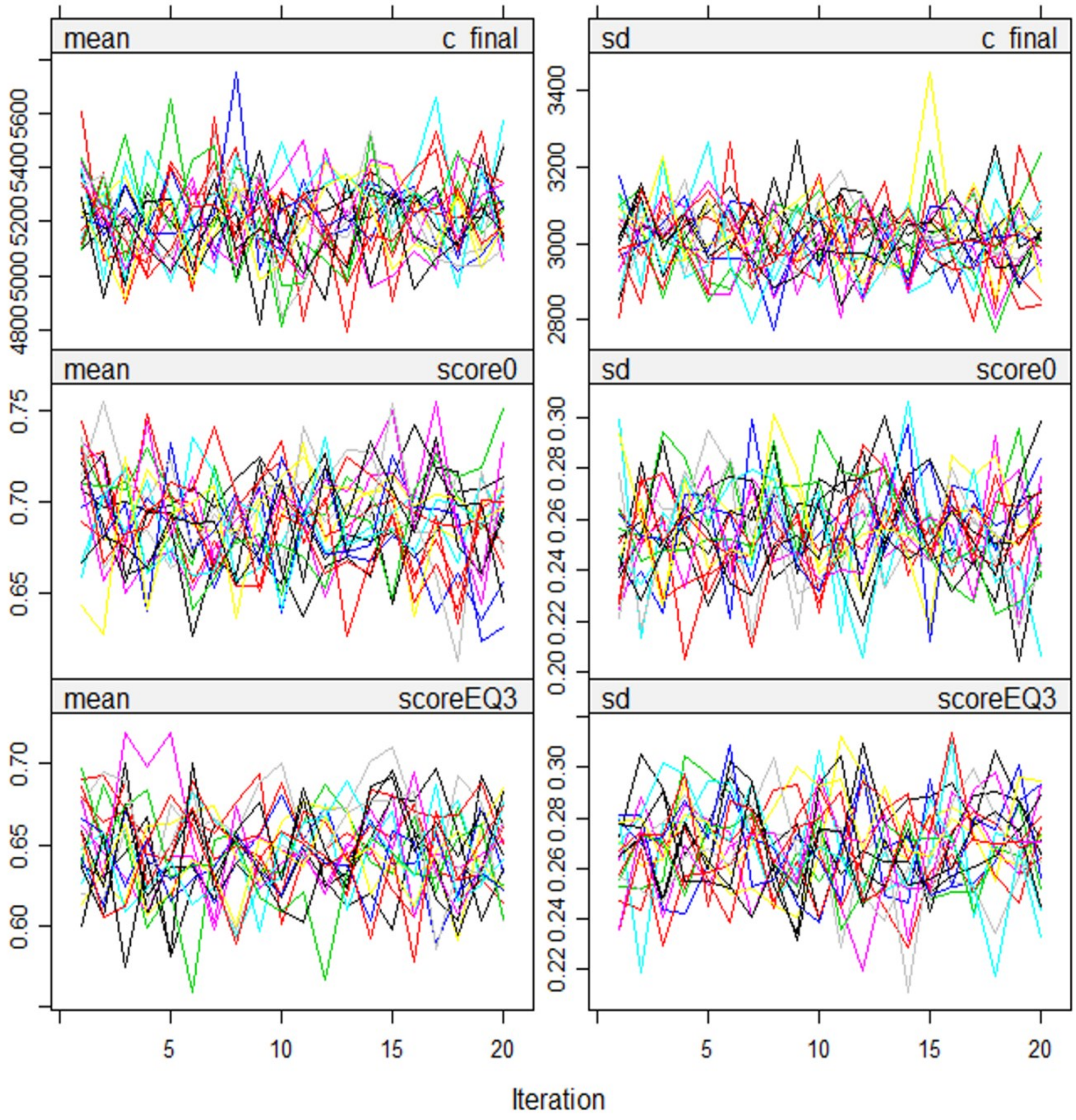


Figure 5.3 ROSSINI trial: multiple imputation diagnostics for three variables (total cost, baseline EQ-5D and final EQ-5D)

Note: Depicted variables (vertical axis) are total cost £ (c_final), baseline EQ-5D score (score0) and final EQ-5D score (scoreEQ3). For each variable the results of the multiple imputation exercise (20 sets) across 20 iterations (x axis) are depicted for the mean and standard deviation.

Cost-effectiveness

Patients in the WEPD arm accessed health care worth £5,420 on average, compared to £5,130 for patients in the standard care arm (Table 5.10). The use of the WEPD was associated with 0.02131 QALYs, compared to 0.02133 QALYs in the control group. Overall, the WEPD strategy was on average £290 more costly (95%CI -372 to 948) and 0.00002 QALYs (95%CI -0.0018 to 0.0017) less beneficial than standard care, thus suggesting that WEPD was dominated by standard care (Table 5.10). The distribution of the cost-effectiveness pairs on the cost-effectiveness plane are presented in Figure 5.4. Just as the confidence intervals and the cost-effectiveness plane suggest, there is a great amount of uncertainty around the point estimates of both incremental costs and incremental QALYs.

The cost-effectiveness acceptability curves (CEACs) indicate that the WEPD is approximately 20% likely to be cost-effective for a willingness-to-pay threshold within the £20,000-30,000/QALY range. Its probability of cost-effectiveness slowly increases with the WTP threshold but still remains below 40% at a threshold in excess of £1 million per QALY (Figure 5.5).

The results of the adjustment models for costs and QALYs are presented in Table 5.11 and Table 5.12, respectively; the coefficient estimates and their variances were combined using Rubin's rules (22). Adjusted estimates for total costs and QALYs are presented in Table 5.13.

Using the WEPD was associated on average £310 more costly (95% CI -273 to 1012) and more effective (0.00018 QALYs, 95% CI -0.0015 to 0.0019) compared with not using the device (Table 5.13). The associated CEAC suggests that the WEPD is approximately 16% likely to be cost-effective at a willingness-to-pay threshold between £20,000 and £30,000 (Figure 5.5).

Table 5.10 ROSSINI trial economic evaluation base-case analysis: mean difference in costs and outcomes by treatment group (unadjusted)

Variable	Mean (SE)		Mean difference (WEPD – standard care)	95% BCa CI	ICER
	WEPD (n=369)	Standard care (n=366)			
Total cost (£)	5,420.31 (246.16)	5,130.27 (233.74)	290.04	-371.70 to 948.49	WEPD is dominated
QALY	0.02131 (0.00141)	0.02133 (0.00139)	-0.00002	-0.0018 to 0.0017	

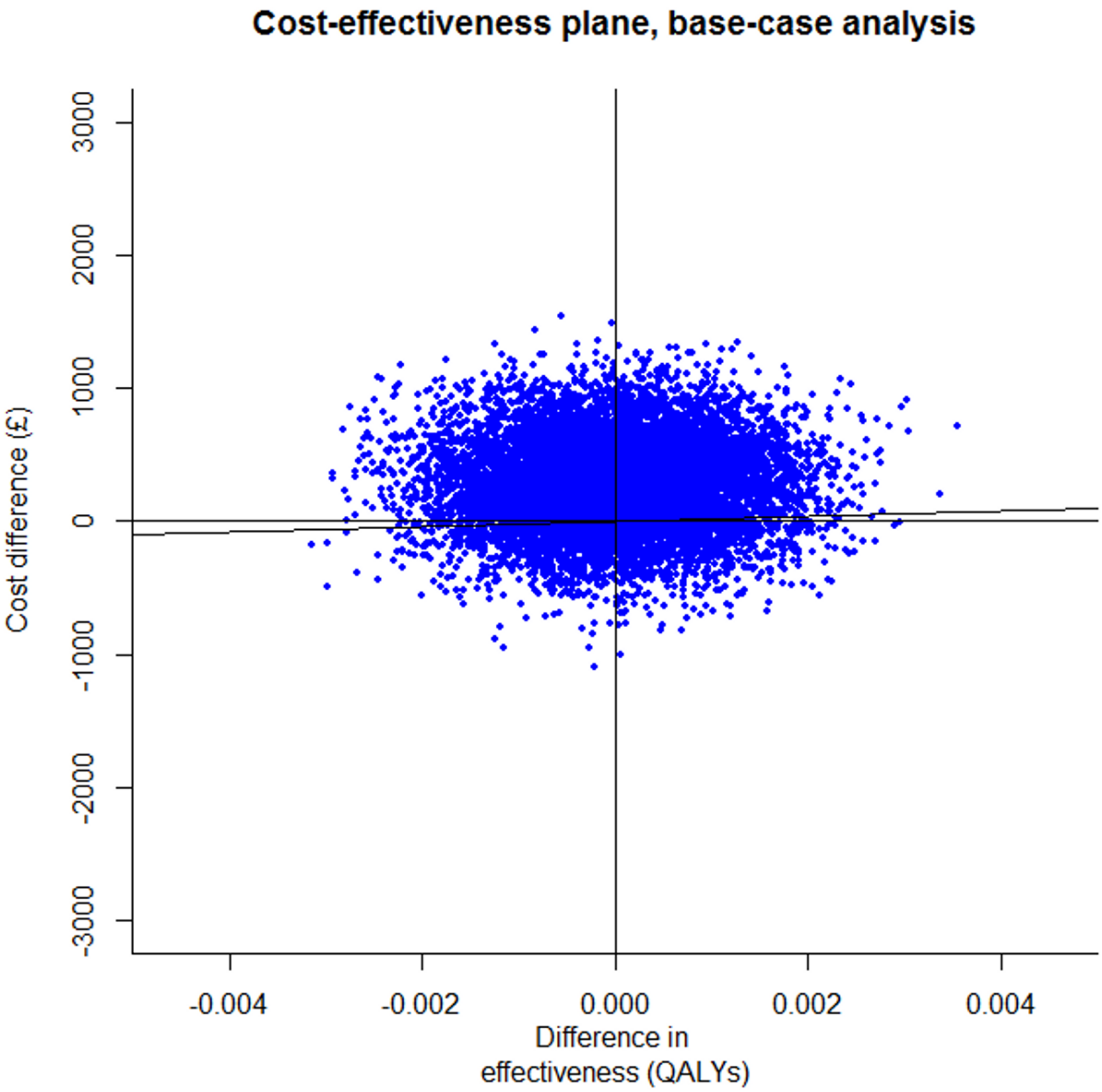


Figure 5.4 ROSSINI trial economic evaluation base-case analysis: cost-effectiveness plane

Note: Willingness-to-pay threshold set at £20,000/QALY

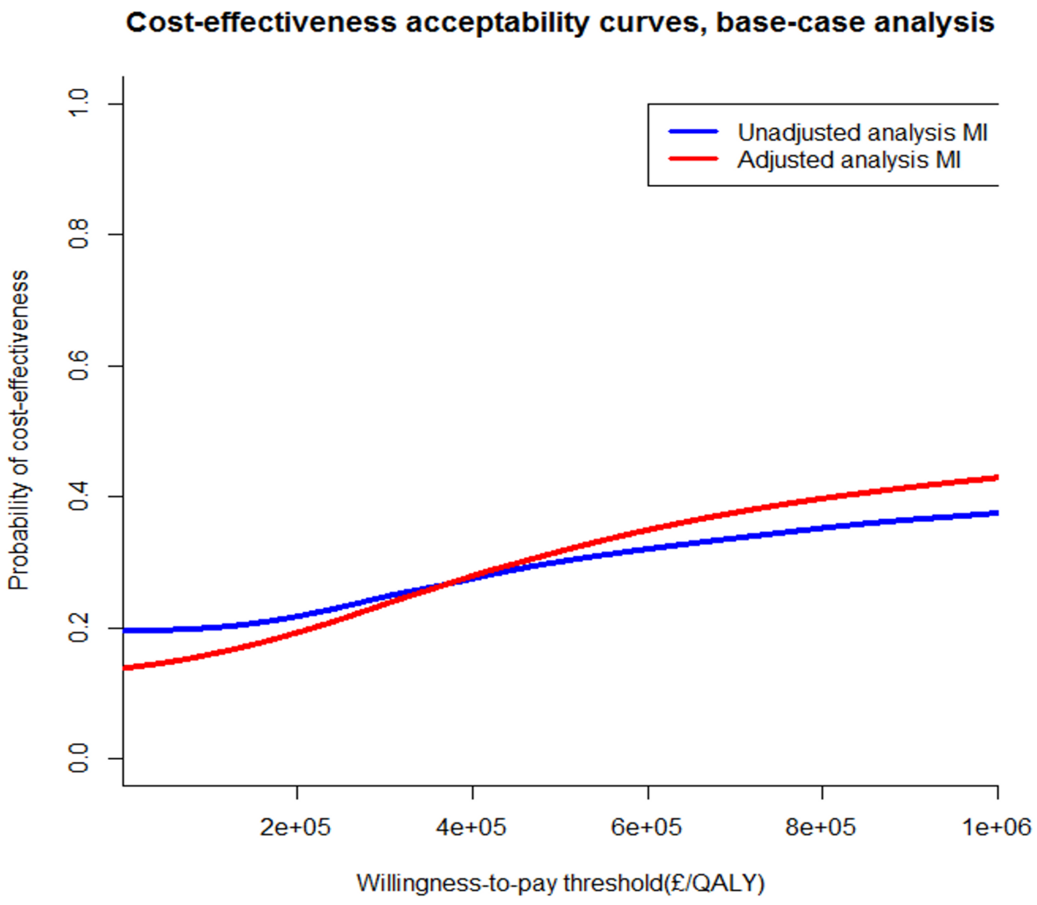


Figure 5.5 ROSSINI trial economic evaluation base-case analysis: cost-effectiveness acceptability curves

Table 5.11 ROSSINI trial economic evaluation base-case analysis: results of the adjustment model for total costs

Variable	Coefficient estimate	SE	p-value
Intercept	4,341.32	585.56	<0.001
Intervention arm	310.86	307.44	0.312
Plan to open viscus	-58.79	432.37	0.892
Plan to create stoma	481.86	415.58	0.246
Emergency surgery	-768.19	358.30	0.033
.Age (55-65)	1,403.09	442.56	0.001
.Age (65-75)	955.08	418.61	0.002
.Age(75-85)	886.01	484.04	0.007
.Age(85+)	-30.64	849.26	0.971
.BMI(23.2-26.7)	29.42	546.66	0.957
.BMI(26.7-30)	-332.19	526.31	0.529
.BMI(30+)	-22.65	599.42	0.969
Current smoker	665.98	461.45	0.149
Diabetic	451.90	472.99	0.339
SSI	952.35	395.91	0.016

Table 5.12 ROSSINI trial economic evaluation base-case analysis: results of the adjustment model for QALYs

Variable	Coefficient estimate	SE	p-value
Intercept	0.02458	0.00230	<0.001
Intervention arm	0.00018	0.00093	0.844
Baseline EQ-5D score	-0.03312	0.00184	<0.001
Plan to open viscus	-0.00031	0.00151	0.837
Plan to create stoma	-0.00012	0.00128	0.936
Emergency surgery	0.00050	0.00100	0.619
.Age (55-65)	0.00032	0.00134	0.809
.Age (65-75)	-0.00034	0.00138	0.804
.Age(75-85)	-0.00098	0.00166	0.553
.Age(85+)	-0.00630	0.00315	0.462
.BMI(23.2-26.7)	0.00018	0.00170	0.914
.BMI(26.7-30)	0.00082	0.00165	0.620
.BMI(30+)	-0.00330	0.00197	0.100
Diabetic	-0.00146	0.00136	0.286
Current smoker	-0.00027	0.00140	0.843
SSI	-0.00456	0.00114	<0.001

Table 5.13 ROSSINI trial economic evaluation base-case analysis: mean difference in costs and outcomes by treatment group (adjusted)

Variable	Mean difference (WEPD – standard care)	95% BCa CI	ICER
Incremental cost (£)	310.86	-272.88 to 1011.67	1,712k/QALY
Incremental QALY	0.00018	-0.0015 to 0.0019	

Results of the complete case analysis

A complete case analysis was also performed, using information from patients who had complete data on resource use, HRQoL (all time-points) and primary outcome (n=532). The average cost in the intervention arm was £5,049 while the average cost in the control arm was £4,812 (Table 5.14). The use of the WEPD was associated with an average HRQoL effect of 0.02038 QALYs, compared to 0.02070 QALYs in the control arm.

The bootstrapping exercise revealed that using the WEPD is on average £237 more costly (95% CI -407 to 892) and more effective (0.00032 QALYs, 95% CI -0.00235 to 0.00162) compared with not using the device (Table 5.14), yielding an ICER of approximately £740,000 per additional QALY. The distribution of the incremental cost-incremental effectiveness pairs on the cost-effectiveness plane still suggests a great amount of uncertainty around the point estimates (Figure 5.6). The associated CEAC suggests that the WEPD is approximately 23% likely to be cost-effective at a willingness-to-pay threshold between £20,000 and £30,000 (Figure 5.7).

The results of the complete case adjustment models for costs and QALYs are presented in Table 5.15 and Table 5.16, respectively. In the adjusted analysis using the WEPD was associated on average £369 more costly (95% CI -214 to 976) and less effective (-0.00016 QALYs, 95% CI -0.00218 to 0.00193) compared with not using the device (Table 5.17). The associated CEAC suggests that the WEPD is approximately 13% likely to be cost-effective at a willingness-to-pay threshold between £20,000 and £30,000 (Figure 5.7). The combined findings of the base-case and alternative analyses are presented in Table 5.18 and Figure 5.8.

Table 5.14 ROSSINI trial economic evaluation complete case analysis: mean difference in costs and outcomes by treatment group (unadjusted)

Variable	Mean (SE)		Mean difference (WEPD – standard care)	95% BCa CI	ICER
	WEPD (n=369)	Standard care (n=366)			
Total cost (£)	5,049.5 (232.9)	4,812.4 (229.4)	237.1	-406.90 to 891.87	£ 740k/QALY
QALY	0.02038 (0.00159)	0.02070 (0.00162)	0.00032	-0.00235 to 0.00162	

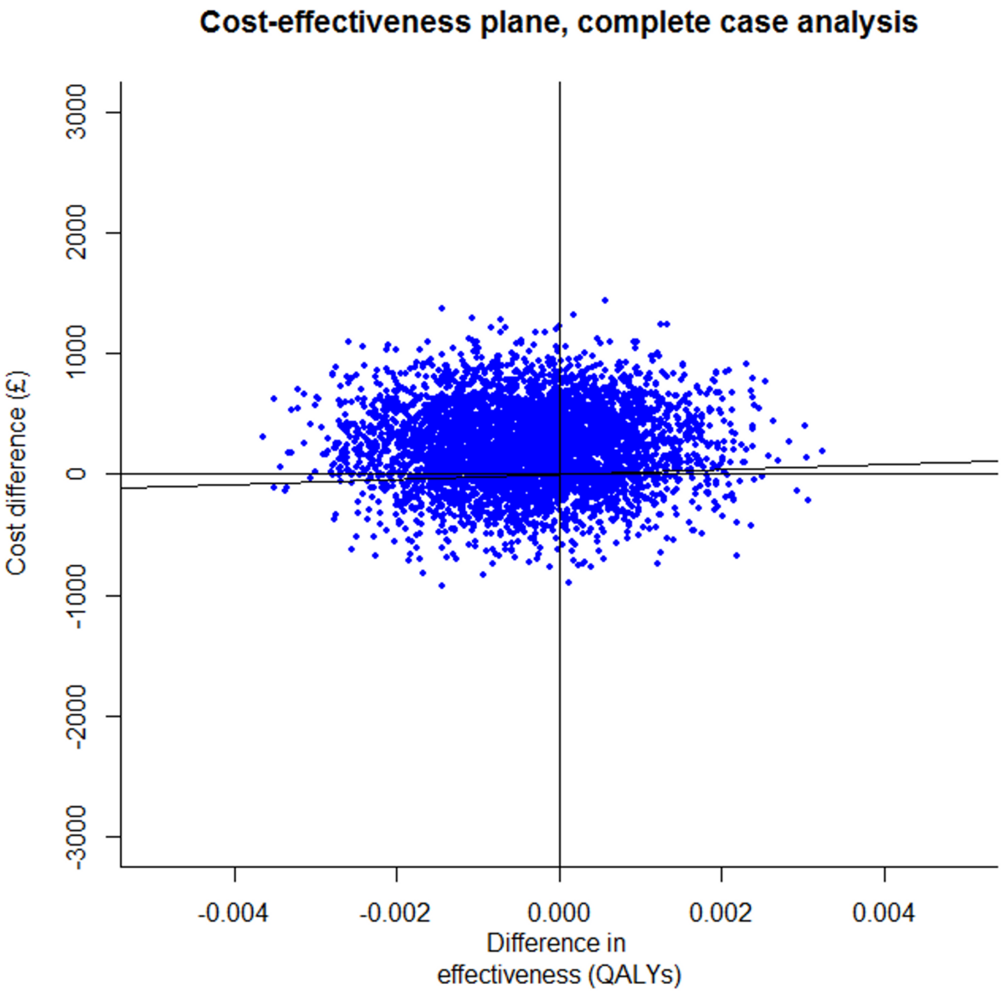


Figure 5.6 ROSSINI trial economic evaluation complete case analysis: cost-effectiveness plane

Note: Willingness-to-pay threshold set at £20,000/QALY

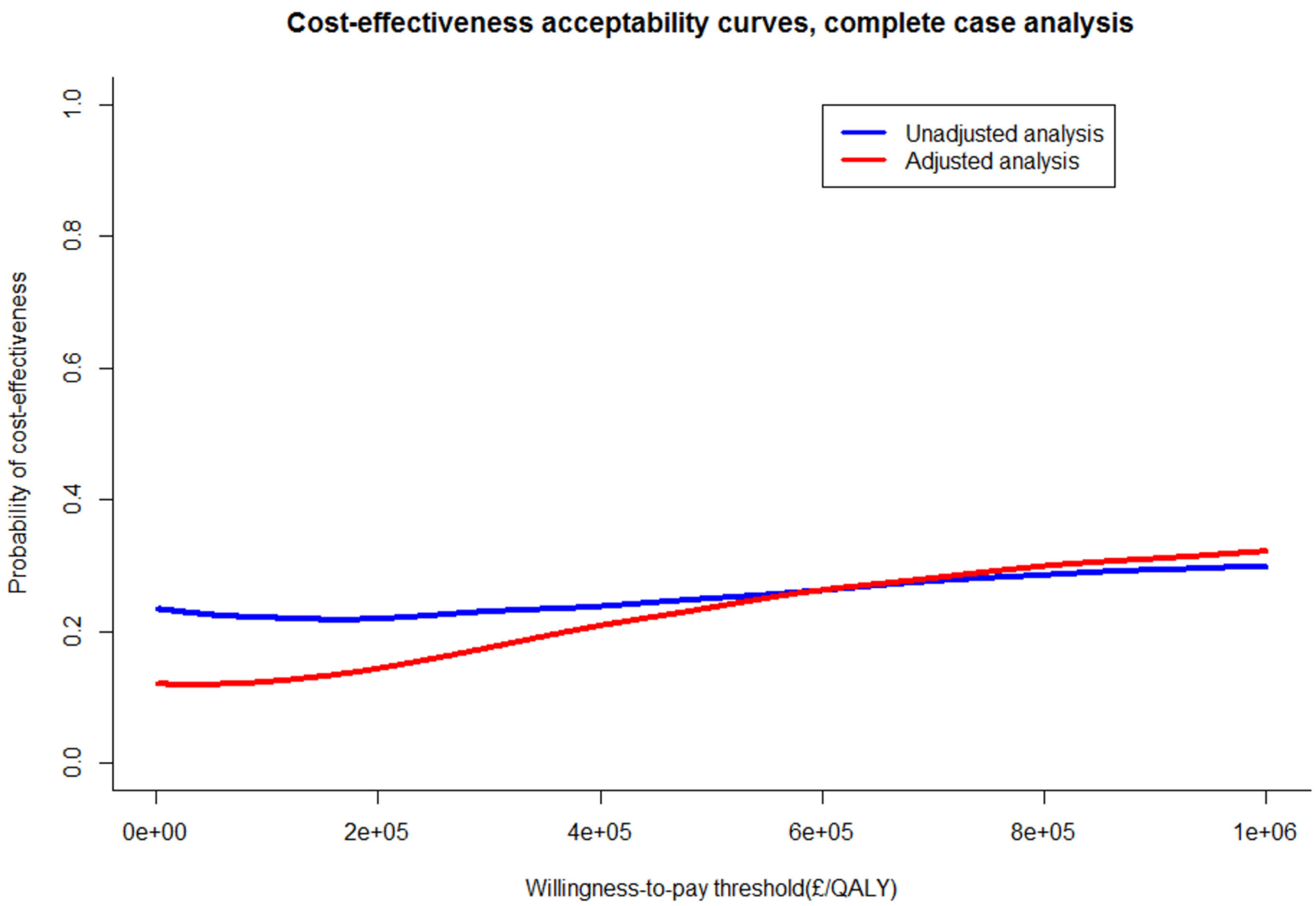


Figure 5.7 ROSSINI trial economic evaluation complete case analysis: cost-effectiveness acceptability curves

Table 5.15 ROSSINI trial economic evaluation complete case analysis: results of the adjustment model for total costs

Variable	Coefficient estimate	SE	p-value
Intercept	4,608.71	566.96	<0.001
Intervention arm	369.03	315.67	0.243
Plan to open viscus	3.51	452.47	0.994
Plan to create stoma	687.38	456.05	0.132
Emergency surgery	-789.97	322.84	0.015
.Age (55-65)	319.85	443.28	0.471
.Age (65-75)	550.01	432.50	0.204
.Age(75-85)	166.34	486.44	0.732
.Age(85+)	-655.91	801.68	0.414
SSI	591.51	400.64	0.140

Table 5.16 ROSSINI trial economic evaluation complete case analysis: results of the adjustment model for QALYs

Variable	Coefficient estimate	SE	p-value
Intercept	0.02151	0.00220	<0.001
Intervention arm	-0.00016	0.00102	0.877
Baseline EQ-5D score	-0.03168	0.00193	<0.001
Plan to open viscus	0.00035	0.00147	0.811
Plan to create stoma	0.00058	0.00140	0.677
Emergency surgery	0.00012	0.00107	0.909
.Age (55-65)	0.00082	0.00147	0.575
.Age (65-75)	0.00048	0.00139	0.731
.Age(75-85)	0.00121	0.00164	0.461
.Age(85+)	-0.00422	0.00322	0.189
SSI	-0.00452	0.00122	<0.001

Table 5.17 ROSSINI trial economic evaluation complete case analysis: mean difference in costs and outcomes by treatment group (adjusted)

Variable	Mean difference	95% BCa CI	ICER
Incremental cost (£)	369.03	-214.23 to 976.28	WEPD is dominated
Incremental QALY	-0.00016	-0.00218 to 0.00193	

Table 5.18 ROSSINI trial economic evaluation: summary of incremental costs and QALYs across the analysed scenarios

Scenario	Variable	Mean difference (WEPD – standard care)	95% BCa CI	ICER
Base-case unadjusted	Total cost (£)	290.04	-371.70 to 948.49	WEPD is dominated
	QALY	-0.00002	-0.0018 to 0.0017	
Base-case adjusted	Total cost (£)	310.86	-272.88 to 1011.67	£1,712k/QALY
	QALY	0.00018	-0.0015 to 0.0019	
Complete case unadjusted	Total cost (£)	237.1	-406.90 to 891.87	740k/QALY
	QALY	0.00032	-0.00235 to 0.00162	
Complete case adjusted	Total cost (£)	369.03	-214.23 to 976.28	WEPD is dominated
	QALY	-0.00016	-0.00218 to 0.00193	

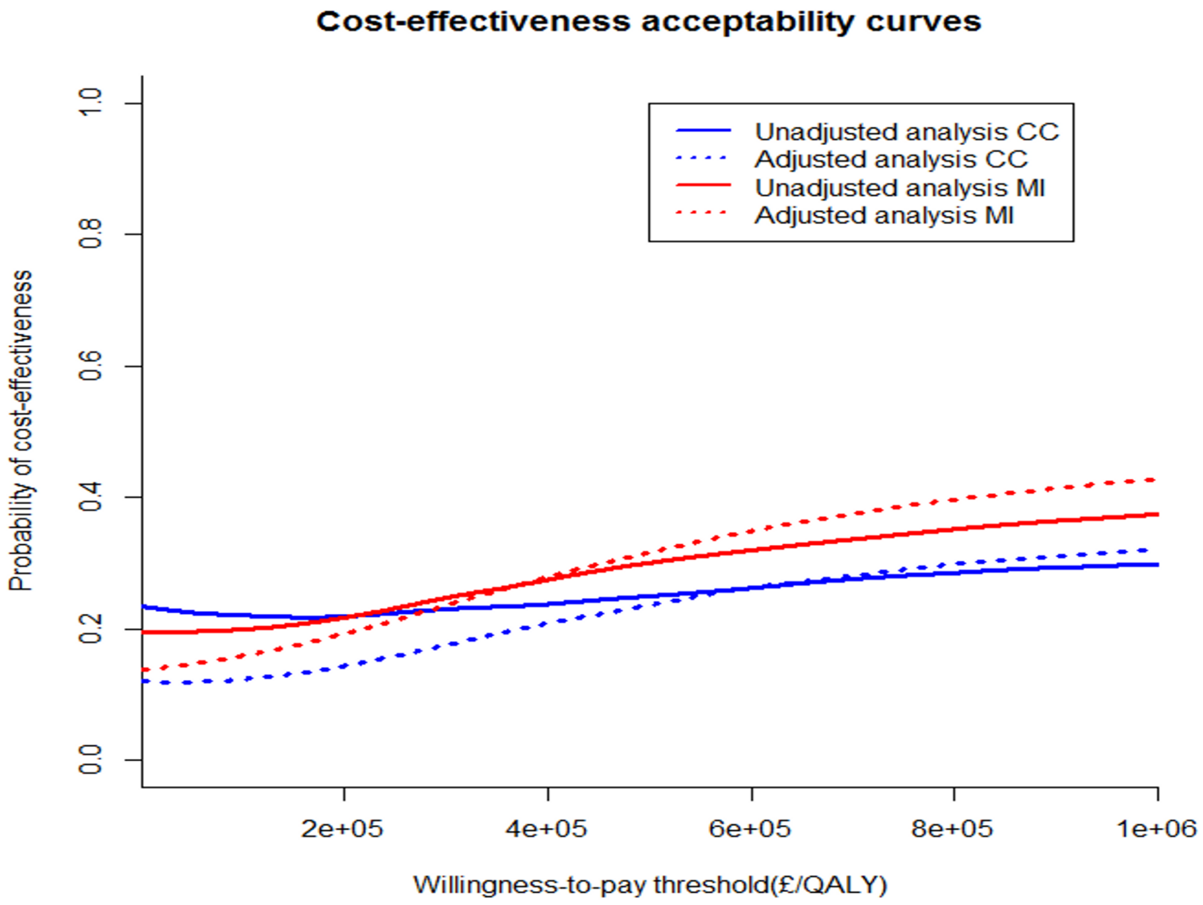


Figure 5.8 ROSSINI trial economic evaluation: comparison of cost-effectiveness acceptability curves

Legend: MI - Base-case analysis; CC - complete case analysis

The cost and HRQoL burden of SSI

The comparison of resource utilisation between SSI and uninfected patients revealed that hospital care was comparable across the two groups in terms of number of inpatient days (Table 5.19). However, SSI patients consumed more resource in primary care, after discharge, as they had significantly more GP visits (average 0.73 vs. 0.38, p=0.0003) and district nurse visits (average 6.54 vs. 2.45, p<0.0001) than uninfected controls. When primary care points of contact and nurse visits, respectively, were aggregated the differences remained statistically significance (Table 5.20).

On average, having a SSI was associated with an additional cost of £1,069 (95% CI £237 to £1,901) and a decreased utility of 0.12 (95% CI 0.07 to 0.17) at 30 days post-operatively (Table 5.21). SSI patients received more expensive care both in an inpatient (£739, 95% CI -76 to 1,555) and outpatient setting (£330, 95% CI £242 to £417).

Table 5.19 ROSSINI trial: summary of resource use by SSI status, detailed

Resource use item		SSI (n=184)	No SSI (n=551)	p-value
HOSPITAL CARE				
Inpatient days				
	N	181	536	
	Mean (SD)	13.02 (14.86)	11.73 (13.29)	0.3015
	SE	1.10	0.57	
	Median	9	9	
Days in ITU				
	N	184	551	
	Mean (SD)	1.21 (4.41)	0.93 (4.46)	0.4473
	SE	0.32	0.19	
	Median	0	0	
Days in HDU				
	N	184	551	
	Mean (SD)	0.59 (1.69)	0.57 (1.56)	0.8837
	SE	0.12	0.07	
	Median	0	0	
PRIMARY CARE				
GP visits				
	N	178	544	
	Mean (SD)	0.73 (0.09)	0.38 (0.03)	0.0003
	SE	0.09	0.03	
	Median	0	0	
District nurse visits				
	N	179	536	
	Mean (SD)	6.54 (8.64)	2.45 (6.16)	<0.0001
	SE	0.65	0.27	
	Median	2	0	
Practice nurse visits				
	N	178	549	
	Mean (SD)	0.33 (1.17)	0.21 (0.92)	0.2410
	SE	0.09	0.04	
	Median	0	0	
Outpatient clinic visits				
	N	179	548	
	Mean (SD)	0.52 (1.38)	0.31 (0.71)	0.0575
	SE	0.10	0.03	
	Median	0	0	

Table 5.20 ROSSINI trial: summary of resource use by SSI status, overview

Resource use item		SSI (n=184)	No SSI (n=551)	p-value
HOSPITAL CARE				
Inpatient days				
	N	181	536	
	Mean (SD)	13.02 (14.86)	11.73 (13.29)	0.3015
	SE	1.10	0.57	
	Median	9	9	
PRIMARY CARE				
Primary care points of contact (includes GP visits, all nurse visits and outpatient clinic visits)				
	N	170	527	
	Mean (SD)	8.03 (8.67)	3.26 (6.40)	<0.0001
	SE	0.66	0.28	
	Median	6	1	
Nurse visits (includes district nurse visits and practice nurse visits)				
	N	175	534	
	Mean (SD)	6.67 (8.29)	2.65 (6.19)	<0.0001
	SE	0.63	0.27	
	Median	3	0	
GP visits				
	N	178	544	
	Mean (SD)	0.73 (1.20)	0.38 (0.79)	0.0003
	SE	0.09	0.03	
	Median	0	0	

Table 5.21 ROSSINI trial: summary of costs and HRQoL data by SSI status

Variable	Mean (SE)		Difference (SSI - no SSI)	95% CI*
	SSI (n=369)	No SSI (n=366)		
Total cost	6,077.49 (381.86)	5,008.19 (182.09)	1,069.29	237.71 to 1900.87
Cost of inpatient care	5,506.01 (371.06)	4,766.21 (185.71)	739.79	-75.77 to 1555.35
Cost of outpatient care	564.01 (38.56)	234.35 (22.36)	329.65	241.97 to 417.34
EQ-5D score at baseline	0.718 (0.023)	0.762 (0.013)	-0.044	-0.096 to 0.008
EQ-5D score at 7 days	0.464 (0.028)	0.514 (0.015)	-0.049	-0.011 to 0.110
EQ-5D score at 30 days	0.594 (0.023)	0.714 (0.013)	0.119	0.067 to 0.172

Note: 95%CI computed assuming a Gamma distribution for costs and a normal distribution for EQ-5D scores

5.2.3. Discussion

Summary of findings

The results of the economic evaluation give a strong indication that using the WEPD in adults undergoing open abdominal surgery is unlikely to be cost-effective when compared to standard care i.e. no WEPD. In the base-case analysis the intervention was found to be more costly and less effective than standard care: the WEPD was associated with an incremental cost of £290 and a 0.00002 QALY loss, thus being dominated by standard care. Since the willingness to pay threshold for most medical technologies lies between £20,000-30,000 per QALY gained, the most likely recommendation is not to adopt the WEPD. Within this WTP interval, the WEPD was less than 30% likely to be cost-effective compared to standard care in all analyses (Figure 5.8). The recommendation was robust to a range of sensitivity analyses (Table 5.18).

There remains uncertainty around the point estimates of costs and HRQoL outcomes, reflected in the width of the confidence intervals. The estimation of uncertainty used non-parametric bootstrapping with n=1,000 replications and confidence intervals were calculated using the bias corrected and accelerated method, in line with methodological recommendations (54). It also is very unlikely that ROSSINI was underpowered: the pre-specified sample size in the statistical analysis plan (n=750), based on the best available evidence to date, assumed a 50% reduction in SSI and a 12% SSI rate in the study population. The trial recruited well and included n=735 patients in the final analysis, while the overall SSI rate was much higher than predicted, at 25.4%, therefore overall ROSSINI was rather overpowered for its original study question. This suggests there may be a large amount of variability in the cost and HRQoL gains associated with the use of the WEPD. In support of this hypothesis, the primary outcome also exhibited considerable uncertainty (OR 0.97, 95% CI 0.69 to 1.36).

Almost all differences between trial arms in terms of costs and EQ-5D scores were not statistically significant (Tables 5.4 and 5.9). The only exception is the cost of practice nurse visits, which appears to have been somewhat higher in the control arm than in the intervention arm as control patients received about twice as many practice nurse visits than WEPD patients (Table 5.4). However, this difference may well be an artefact because resource utilisation in primary care was informed by patient-completed forms and there may have been some confusion regarding the exact nature of the health care professional who led the visit, for example not differentiating between practice nurse and district nurse visits at the time of filling in the CRF. Indeed, when the total number of primary care contact points was summarised there was no difference between the two groups (Table 5.5). Furthermore, the cost of practice nurse visits only had a small contribution to the total cost and it is unlikely that it could have biased the overall results.

Sensitivity analyses

735 patients were included in the primary analysis of the trial based on availability of primary outcome data. Of these, n=532 patients had complete resource use, HRQoL and primary outcome data, leading to approximately 25% missing data for the purpose of the economic evaluation (Table 5.6). The largest proportion of the missing information referred to EQ-5D scores at baseline and at 30-33 days postoperatively (n=104 observations). No pattern of missingness was apparent upon inspection of the missing values. Patients below 65 years old and with a BMI larger than 26.75 had larger amounts of missing data than their counterparts, respectively, but this only applied to resource utilisation data and not to EQ-5D

scores. No other observed variables were associated with missingness. As such, the MCAR mechanism can be ruled out, while MAR may hold and thus informed the imputation exercise.

Results of the base-case were compared with the complete case analysis, which used information only from the 532 patients with complete data (Table 5.21). In the complete case analysis, incremental costs were lower than in the base-case (£237 vs. £290) and the QALY gain was positive (0.00032 vs.-0.00002). While the results are different between the two scenarios, they both lead to the same recommendation of not adopting the WEPD. The difference between the two results could be explained by the large amount of heterogeneity that appears to be inherent to the dataset. The 95% BCa confidence intervals for incremental costs and QALYs are comparable across the scenarios, thus suggesting that most of the variation in point estimates is due to natural variability, especially considering that the QALY gain is negligible.

The results of the adjusted analyses were similar to the unadjusted results: in the base-case analysis, the incremental cost increased from £290 to £311, while the QALY gain increased from -0.00002 to 0.00018. Although the WEPD is no longer dominated in the base-case adjusted analysis, it remains cost-ineffective when considering the NICE WTP threshold. In the complete case scenario, following adjustment the incremental cost increased from £237 to £369 and the incremental effectiveness decreased from 0.00032 to -0.00016 QALYs (Table 5.21). While the adjusted analyses did not change the final recommendation, the increase in incremental costs results deserves comment. The most plausible explanation is that the variables used for adjustment, which were recognised SSI risk factors, reflected a slight imbalance in the two trial arms as a result of the randomisation procedure (Table 5.1): patients in the WEPD arm were slightly older and a slightly higher proportion had diabetes. These may have impacted on the severity of the SSIs they acquired, which needed more intensive care. However the WEPD's probability of cost-effectiveness, as reflected by CEACs, was largely unaffected by the adjustment: Figure 5.8 depicts the CEACs for adjusted and unadjusted base-case and complete case analyses, respectively, and the variations are minimal. There is great uncertainty around the point estimates in all three scenarios but the WEPD appears to be cost-ineffective under all the considered scenarios.

Cost-effectiveness drivers

It appears that inpatient costs are the main drivers for the total cost (Table 5.9). In both trial arms inpatient costs accounted for approximately 94% of total costs incurred. Patients in the intervention arm consumed more resources in hospital care, which may be primarily explained by the fact that patients in the WEPD arm spent, on an average, an extra day in hospital compared to patients in the control arm, but this difference was not statistically significant (Table 5.5). There are no reasons to believe, however, that blinding was violated: the SSI rates at 5-7 days postoperatively were comparable between the two arms (Figure 5.1) and no significant difference between the two arms was highlighted in the time to first hospital discharge analysis (hazard ratio 1.03, 95%CI 0.88 to 1.10) (Table 5.2).

Primary care costs accounted for less than 10% in the total costs of postoperative management. This is in line with the findings of the analytic decision model (Chapter 4), which indicated that primary care costs would not be a major driver in the economic evaluation. The costs were comparable between the two groups and the difference was negligible. It appears that the largest part of primary care costs were due to district nurse visits, but again the differences between the two arms were negligible. The issue of practice nurse visits was discussed above and the possibility of an artificial difference cannot be ruled out.

Strengths and limitations

The economic analysis considered a range of scenarios to account for missing data and the effect of confounding variables on cost-effectiveness estimates. Multiple imputation using chained equations was used in the base-case analysis to account for missing data, in line with current recommendations (17, 26). Non-parametric bootstrapping and the bias corrected and accelerated method were employed to quantify the uncertainty around costs and outcomes in order to avoid any distributional assumptions (54).

Several limitations of the economic evaluation deserve consideration: the proportion of missing data, the time horizon, the complexity of SSI management and data collection. Just over 20% of patients had at least one cost of HRQoL missing data item, which may lead one to question the appropriateness of a complete case analysis. Nevertheless, the results of the base-case and complete case analyses are largely comparable and there are no reasons to believe that the proportion of missing data brings into question the results of the cost-effectiveness analysis.

A 30-day time horizon after surgery was chosen for the economic evaluation due to ROSSINI's design, where the primary outcome was the occurrence of SSI within 30 days post-operatively, in line with the international guidelines on SSI diagnosis (200, 206). A 30-day time horizon was also adopted in other decision models which evaluated interventions reducing SSI (314). In clinical practice two things must be considered: first, not all SSIs develop immediately after surgery. This was reflected in ROSSINI, which found that the majority of SSIs were diagnosed in the interval 7-30 days post-operatively (Figure 5.1).

Second, the time required for a SSI to heal is highly variable, ranging from several days to several months, depending on factors like the severity of the infection, the nature of the underlying pathogens and co-morbidities (see Chapter 4, section 4.2). 11% of patients still had an ongoing SSI at 30 days post-operatively. A further limitation refers to the complexity of SSI management, especially in primary care. NICE clinical guidelines on SSI care provide evidence that the weekly cost of wound dressings can be up to £100, depending on the type of wound and the type of dressing (243). Although an ethics amendment was put through to extend follow-up and to introduce a wound dressing diary for health care professionals and patients to complete to gather primary data on the type and frequency of dressings used, the procedures could not be implemented in due time because of regulatory delays; these aspects were eventually excluded from the analysis. Nevertheless, the WEPD did not show any sign of clinical benefit and there are little reasons to believe that it could reduce the burden of severe, long-term SSIs. The wound dressing diary was only piloted in four centres and the economic impact of their use remains unknown. However, if working under the assumption that district nurses are the health care professional most likely to apply the wound dressings in a primary care setting, the trial arms were more than comparable regarding the number of district nurse visits, which reduces the potential effect of not costing wound dressings (Tables 5.4 and 5.5).

5.3. Conclusion

Based on the findings of the ROSSINI trial, WEPDs are unlikely to be cost-effective in reducing SSI when compared to standard care. Total costs were higher in the intervention arm, mostly due to higher inpatient costs, but the differences were not statistically significant. Furthermore, HRQoL gains associated with the WEPD were negligible. There was a great deal of uncertainty around the point estimates of both incremental costs and incremental QALYs. WEPDs are approximately 20% likely to be cost-effective considering the NICE WTP threshold range. The results were robust to the adjustment for confounding factors and to a complete case analysis. The following Chapter discusses these findings in the context of previous evidence on WEPD effectiveness.

CHAPTER 6. REASSESSING THE EVIDENCE ON THE CLINICAL AND COST-EFFECTIVENESS OF WOUND-EDGE PROTECTION DEVICES: INCORPORATING THE *ROSSINI* TRIAL

The aim of this Chapter is to integrate the evidence of clinical and cost-effectiveness of wound-edge protection devices (WEPDs) compared to standard care, which has been presented previously in Chapter 3 to Chapter 5. The findings of the systematic review, decision model and ROSSINI analyses will be presented and contrasted. Ultimately, consolidated findings will be formulated.

6.1. Integrating the evidence on the clinical effectiveness of WEPDs

Two sources of evidence are available for the clinical effectiveness of WEPDs compared to standard care in reducing SSI after open abdominal surgery: the systematic review and exploratory meta-analysis presented in Chapter 3; and the ROSSINI trial, presented in Chapter 5. The exploratory meta-analysis included data from 12 trials (n=1,850 participants) and suggested that WEPDs are likely to be effective when compared to standard care (RR 0.60, 95%CI 0.41 to 0.86). On the other hand, the ROSSINI trial randomised 760 patients between two arms (WEPD vs. standard care), of which 735 were included in the final analysis. 91 patients in the intervention arm (n=369) and 93 patients in the control arm (n=366) developed a SSI after laparotomy, thus suggesting no benefit associated with the use of WEPDs (RR 0.97, 95%CI 0.76 to 1.25).

These two results are, thus, markedly different: while the systematic review suggested that WEPDs are largely effective, the ROSSINI trial demonstrated no benefit at all. The following paragraphs discuss the potential reasons and implications of this discrepancy.

First, the systematic review (Chapter 3) concluded that the methodological quality of the identified studies was generally poor: the sample sizes were generally small; there were concerns about the methods used for randomization and blinding; it was often unclear whether the reported outcomes had been pre-specified or not; and ten of the 12 included trials were single-centre. In addition, most studies reported insufficient information to allow appropriate judgements on the Cochrane 'Risk of bias' categories. Nevertheless, the included trials spanned a large time interval (over 40 years) along which the methodological standards of conducting evaluative research have improved considerably: indeed, the more recent trials had better quality than older ones. In the light of these arguments, the results of the meta-analysis were purely exploratory and this has been clearly specified at the outset.

The question then becomes how much can the results of such meta-analyses which included poor quality studies be trusted. A published commentary (361) to the systematic review outlined this very point and reinforced the need of more methodologically sound evidence to support the indicative results of the systematic review. No matter how methodologically sound its methods, a meta-analysis informed by poor quality data cannot offer a valid result.

Second, ROSSINI addressed most of the methodological limitations of earlier trials: it incorporated electronic randomisation with a minimisation procedure and stratification by three risk factors (elective/emergency surgery, intention to create a stoma, intention to open a viscus), group concealment, blinding of wound assessors and patients, and a robust follow-up protocol including training in wound assessment. In addition, the patient groups in ROSSINI were well matched with no significant over-representation of any patient or operative characteristic in either arm (Table 5.1). Moreover, ROSSINI found a baseline infection rate of 25.4%, significantly higher than the conservative 12% predicted baseline rate, which offered increased power to detect potential benefit. SSI was assessed according to internationally accepted CDC guidelines at three time points: 5-7 days post-operatively (clinician assessed), 7-30 days post-operatively (patient assessed) and 30-33 days post-operatively (clinician assessed). There was no significant difference between the intervention and control arms at either time point (Figure 5.1). Furthermore, no difference was discernible in time to discharge analysis, health-related quality of life and cost-effectiveness, which suggests that ROSSINI findings are robust.

Third, ROSSINI aimed to provide as generalisable results as possible by accurately reflecting standard clinical practice. As such, it recruited from a large number of centres in England (21 general hospitals). The inclusion criteria were deliberately broad and referred to any type of open surgery requiring laparotomy. In addition, the study protocol was not prescriptive in relation to the prophylaxis measures to be taken: these were left at the discretion of each surgical team to reflect current local practice, but were also recorded and no difference was apparent between the two arms. This differs from the studies included in the systematic review, which usually focused on a particular type of surgery (such as appendicectomy or colon surgery) and had strict protocols for perioperative care.

Finally, two WEPD designs are available: the single-ring one, which features a plastic drape expanding from the ring; and the double-ring one, with a plastic semi-rigid drape linking the two rings in the shape of a cylinder. Historically speaking, the single-ring design precedes and has been around for much longer than the double-ring one: as a result, it was used in nine of the 12 trials included in the systematic review. The ROSSINI trial tested the single-ring design, whose effectiveness had last been suggested by the RCT of Sookhai *et al.*(269). All three major manufacturers of WEPD devices of all designs were invited to take part in ROSSINI: 3M^{TM} (SteriDrape©); Applied Medical^{TM} (Alexis©); and Medical Concepts Development^{TM} (Vi-Drape©). Only 3M^{TM} responded to the invitation. On the other hand, the meta-analysis pooled data from studies irrespective of the design type they used. Sensitivity analyses suggested that there may be a difference in effectiveness favouring the double-ring design i.e. the exploratory subgroup analysis in the meta-analysis (section 3.2) and the published analysis of Edwards *et al.* (301). Still, this indication is inevitably affected by the poor quality of the trials informing it. Without a head-to-head comparison of the two devices and accurate knowledge of how pathogens infect the surgical wound, it is still difficult to ascertain what ROSSINI actually demonstrated: that the WEPD mechanism of action is invalid; or that the single-ring design is not effective.

In light of these arguments, ROSSINI results appear to be more robust and generalisable than those of the exploratory meta-analysis. If the quality of ROSSINI and that of previous studies had been comparable, the next step in the evidence generating process would have been to update the results of the meta-analysis by incorporating ROSSINI results. This is not the case: ROSSINI is the only good quality RCT investigating this research question and a *de novo* systematic review would most likely only include ROSSINI and discard the previous trials. For exploratory purposes, however, an investigation was undertaken into how bringing together the evidence from all available trials, including the two more recent trials published after the systematic review (both with methodological quality issues, as discussed at the end of Chapter 3) and ROSSINI, would impact the estimate of clinical effectiveness of WEPDs. As Figure 6.1 suggests, adding the recently available information did not change the point estimate of WEPD effectiveness compared to the original result (RR 0.60, 95% CI 0.43 to 0.85).

ROSSINI appears to be the best available evidence on the clinical effectiveness of WEPDs compared to standard care. As such, the use of WEPDs cannot be routinely recommended to reduce SSI after open abdominal surgery.

	Intervention		Control			Risk Ratio
Study or Subgroup	**Events**	**Total**	**Events**	**Total**	**Weight**	**M-H, Random, 95% CI**
1.5.1 Original meta-analysis						
Maxwell 1969	16	88	12	82	9.0%	1.24 [0.63, 2.47]
Williams 1972	10	84	10	83	7.7%	0.99 [0.43, 2.25]
Psaila 1977	8	46	10	47	7.6%	0.82 [0.35, 1.89]
Nystrom 1984	7	70	6	70	6.1%	1.17 [0.41, 3.30]
Gamble Hopton 1984	10	27	8	29	8.2%	1.34 [0.62, 2.89]
Batz 1987	1	25	7	25	2.4%	0.14 [0.02, 1.08]
Brunet 1994	6	73	18	76	7.4%	0.35 [0.15, 0.83]
Redmond 1994	11	102	27	111	9.3%	0.44 [0.23, 0.85]
Sookhai 1999	23	170	54	182	11.4%	0.46 [0.29, 0.71]
Horiuchi 2007	8	111	16	110	7.9%	0.50 [0.22, 1.11]
Lee 2009	1	61	7	48	2.3%	0.11 [0.01, 0.88]
Reid 2010	3	64	15	66	5.2%	0.21 [0.06, 0.68]
Subtotal (95% CI)		**921**		**929**	**84.4%**	**0.60 [0.41, 0.86]**
Total events	104		190			
Heterogeneity: Tau² = 0.20; Chi² = 24.08, df = 11 (P = 0.01); I² = 54%						
Test for overall effect: Z = 2.78 (P = 0.005)						
1.5.2 Recent studies						
Theodoridis 2011	0	115	3	116	1.2%	0.14 [0.01, 2.76]
Cheng 2012	0	34	6	30	1.3%	0.07 [0.00, 1.16]
ROSSINI 2012	91	369	93	366	13.0%	0.97 [0.76, 1.25]
Subtotal (95% CI)		**518**		**512**	**15.6%**	**0.33 [0.05, 2.16]**
Total events	91		102			
Heterogeneity: Tau² = 1.73; Chi² = 5.13, df = 2 (P = 0.08); I² = 61%						
Test for overall effect: Z = 1.16 (P = 0.25)						
Total (95% CI)		**1439**		**1441**	**100.0%**	**0.60 [0.43, 0.85]**
Total events	195		292			
Heterogeneity: Tau² = 0.22; Chi² = 36.33, df = 14 (P = 0.0009); I² = 61%						
Test for overall effect: Z = 2.89 (P = 0.004)						
Test for subgroup differences: Chi² = 0.37, df = 1 (P = 0.55), I² = 0%						

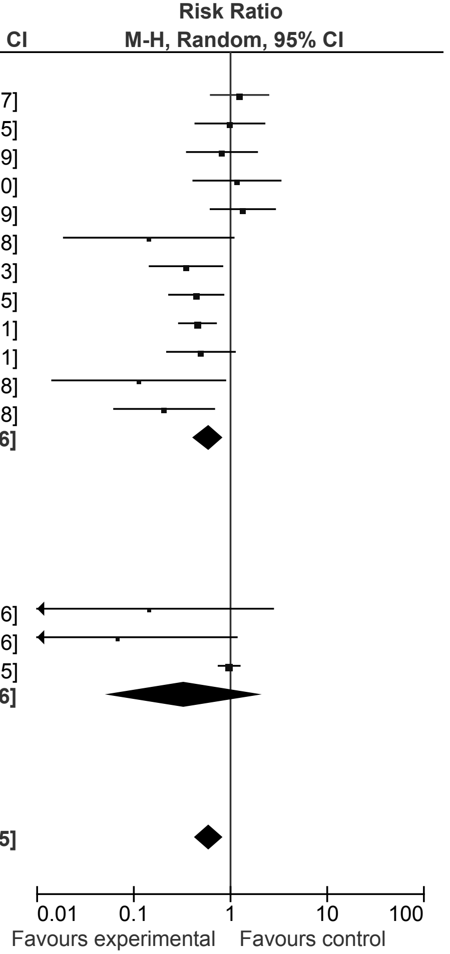


Figure 6.1 Exploratory meta-analysis with all the available evidence on the clinical effectiveness of WEPDs

6.2. Integrating the evidence on the cost-effectiveness of WEPDs

There are two available sources of evidence for the cost-effectiveness of WEPDs compared to standard care: the preliminary decision model, discussed in Chapter 4; and the economic evaluation alongside the ROSSINI trial, discussed in Chapter 5. The comparative results of the two evaluations are jointly presented in Table 6.1: while the decision model suggested that use of the WEPD dominates standard care across the entire range of explored scenarios, the trial-based cost-effectiveness analysis showed that the WEPD was either dominated or cost-ineffective, depending on the scenario considered. The unadjusted base-case analysis of ROSSINI data suggested that the WEPD option is on average £290 more expensive than standard care and generates 0.00002 less QALYs. When incorporating uncertainty in the results and assuming a decision maker's WTP threshold of £20,000/QALY, the decision model base-case analysis indicated that the WEPD is 86.6% likely to be cost-effective as opposed to 20% in the ROSSINI base-case analysis.

These are, again, contrasting results: the decision model depicts the WEPD as highly cost-effective, while the in-trial analysis indicates the opposite. This can also be observed in Figure 6.2, which depicts the cost-effectiveness acceptability curves from both the decision model scenarios and ROSSINI analyses. The decision model rests on the assumption that the WEPD is clinically effective, as suggested by the systematic review and meta-analysis in Chapter 3. However, as discussed in the previous section of this Chapter, there are reasons to believe that ROSSINI findings, however different, are more robust than those of the meta-analysis. In that respect, a direct comparison of the decision-model and ROSSINI results is difficult because the two are informed by utterly different clinical realities.

Table 6.1 Comparison of decision model and ROSSINI economic evaluation results

Source	Type of analysis	Alternative	Total cost (£)	Incremental Cost (£)	Total QALY	Incremental QALY	Decision
ROSSINI trial	Base-case unadjusted	WEPD	5,420	290	0.05975	-0.00010	WEPD is dominated
		Standard care	5,130	-	0.05985	-	
	Complete case unadjusted	WEPD	5,049	237	0.06073	-0.00096	WEPD is dominated
		Standard care	4,812	-	0.06169	-	
Decision model	Base-case	WEPD	5,196	-44	0.06061	0.00010	Standard care is dominated
		Standard care	5,240	-	0.06051	-	
	Scenario 1	WEPD	5,286	-44	0.06062	0.00014	Standard care is dominated
		Standard care	5,330	-	0.06048	-	
	Scenario 2	WEPD	5,221	-51	0.06060	0.00009	Standard care is dominated
		Standard care	5,272	-	0.06051	-	
	Alternative model	WEPD	5,672	-384	0.06051	0.00015	Standard care is dominated
		Standard care	6,056	-	0.06036	-	

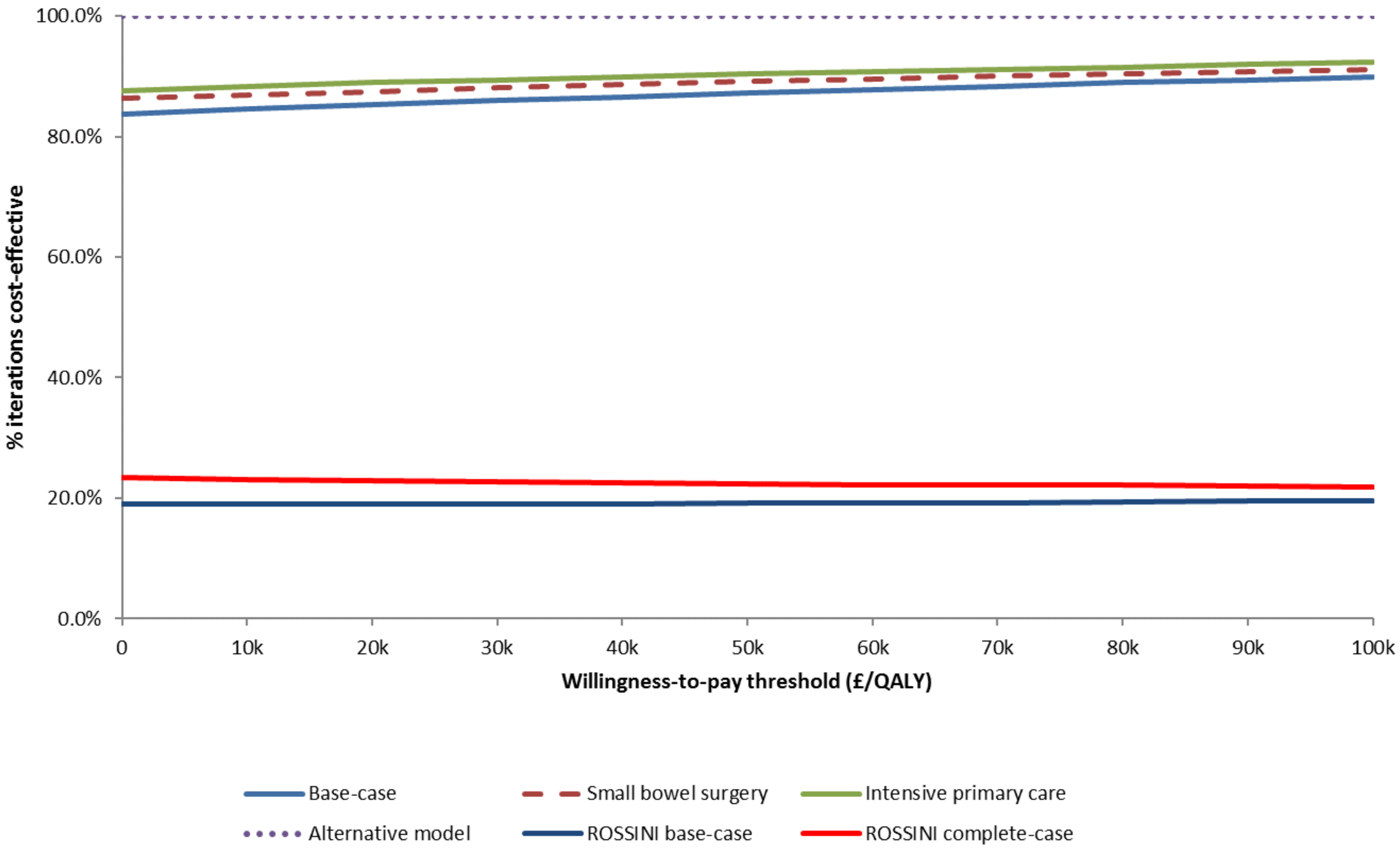


Figure 6.2 Comparison of cost-effectiveness acceptability curves: original decision model and ROSSINI results

In absolute terms, the decision model appears to have predicted reasonably well the magnitude of average costs in both arms, ranging in the interval £5,000 to £5,500 in both base-case analyses. However, the decision model mis-estimated the incremental costs as it predicted that the WEPD alternative would be cost-saving. In terms of predicting effectiveness, the decision model estimates of total QALYs are compatible with those resulted from ROSSINI; the discrepancy between positive incremental QALYs in the decision model and negative incremental QALYs in ROSSINI is most likely a result of the assumption that the WEPD was effective.

Two key assumption made in the decision model with respect to estimating QALYs appear to have been appropriate. First, the decision model assumed a utility decrement of approximately 0.05 as a result of uninfected surgery at 30-days postoperatively, calculated as the difference between baseline utility (0.800, informed by UK population norms for age group 55-64) and the EQ-5D score reported by Janson *et al.* (332) after open colon resection (0.752). On the other hand, the difference between baseline (0.762) and 30-day EQ-5D score (0.714) for uninfected patients was also 0.05, thus closely compatible with the model estimate. Second, the 0.1 utility decrement assumed in the model also appears to be defensible, as the comparison of utility scores between SSI and non-SSI patients in ROSSINI suggested a difference at 30 days postoperatively of 0.119 (95%CI 0.067 to 0.172).

In order to investigate the effect of the clinical effectiveness estimate on the results of the decision model, the original effectiveness parameter (RR 0.60, 95%CI 0.41 to 0.86) was replaced with ROSSINI's estimate of effectiveness (RR 0.97, 95%CI 0.76 to 1.25). The results are shown in Table 6.2: under the new assumption of no clinical benefit of the WEPD, the base-case and two scenarios of the decision model yield positive incremental costs and marginally positive QALY gains, thus suggesting that the WEPD option is more costly and more effective than standard care. However, the resulting ICERs are in excess of £1million/QALY, therefore the WEPD is clearly cost-ineffective at a WTP threshold of £20,000/QALY.

Figure 6.3 depicts the cost-effectiveness acceptability curves from the updated decision model scenarios and ROSSINI results. An alternative decision model was developed in Chapter 4 to explore structural uncertainty, and deliberately had a simpler structure than the base-case model. It appears that the alternative model offers the closest estimate of the WEPD's probability of cost-effectiveness to trial-based results. By contrast, the base-case decision model appears to yield results which are the least compatible with ROSSINI. This may prove to be an argument towards using a simpler model structure; however, it has to be fully acknowledged that the simpler model is much more dependent on the accuracy of data inputs. As shown in Figure 6.3, the alternative decision model clearly favoured the WEPD based on the evidence emerging from the systematic review. Moreover, the alternative decision model also overestimated the magnitude of total costs in both arms the most among all investigated scenarios (Table 6.2).

In light of all these findings, it appears that the base-case decision model produced a conservative estimate of the WEPD's cost-effectiveness. Of all the considered scenarios, it appears that the 'alternative', less sophisticated model, had a higher predictive value of cost-effectiveness than the other scenarios, although the latter better predicted absolute costs and effects, which may also be of interest to decision makers.

Table 6.2 Results of the decision model's reassessment using the ROSSINI clinical effectiveness estimate

Source	Type of analysis	Alternative	Total cost (£)	Incremental Cost (£)	Total QALY	Incremental QALY	Decision
ROSSINI trial	Base-case	WEPD	5,420	290	0.05975	-0.00010	WEPD is dominated
		Standard care	5,130	-	0.05985	-	
	Complete-case	WEPD	5,049	237	0.06073	-0.00096	WEPD is dominated
		Standard care	4,812	-	0.06169	-	
Decision model	Base-case	WEPD	5,274	13	0.06053	0.00001	WEPD not cost-effective at £20k/QALY
		Standard care	5,261	-	0.06052	-	
(updated with ROSSINI clinical effectiveness estimate)	Scenario 1	WEPD	5,344	11	0.06050	0.00001	WEPD not cost-effective at £20k/QALY
		Standard care	5,333	-	0.06049	-	
	Scenario 2	WEPD	5,293	11	0.06051	0.00001	WEPD not cost-effective at £20k/QALY
		Standard care	5,282	-	0.06050	-	
	Alternative model	WEPD	6,040	-14	0.06038	0.00001	Standard care is dominated
		Standard care	6,054	-	0.06037	-	

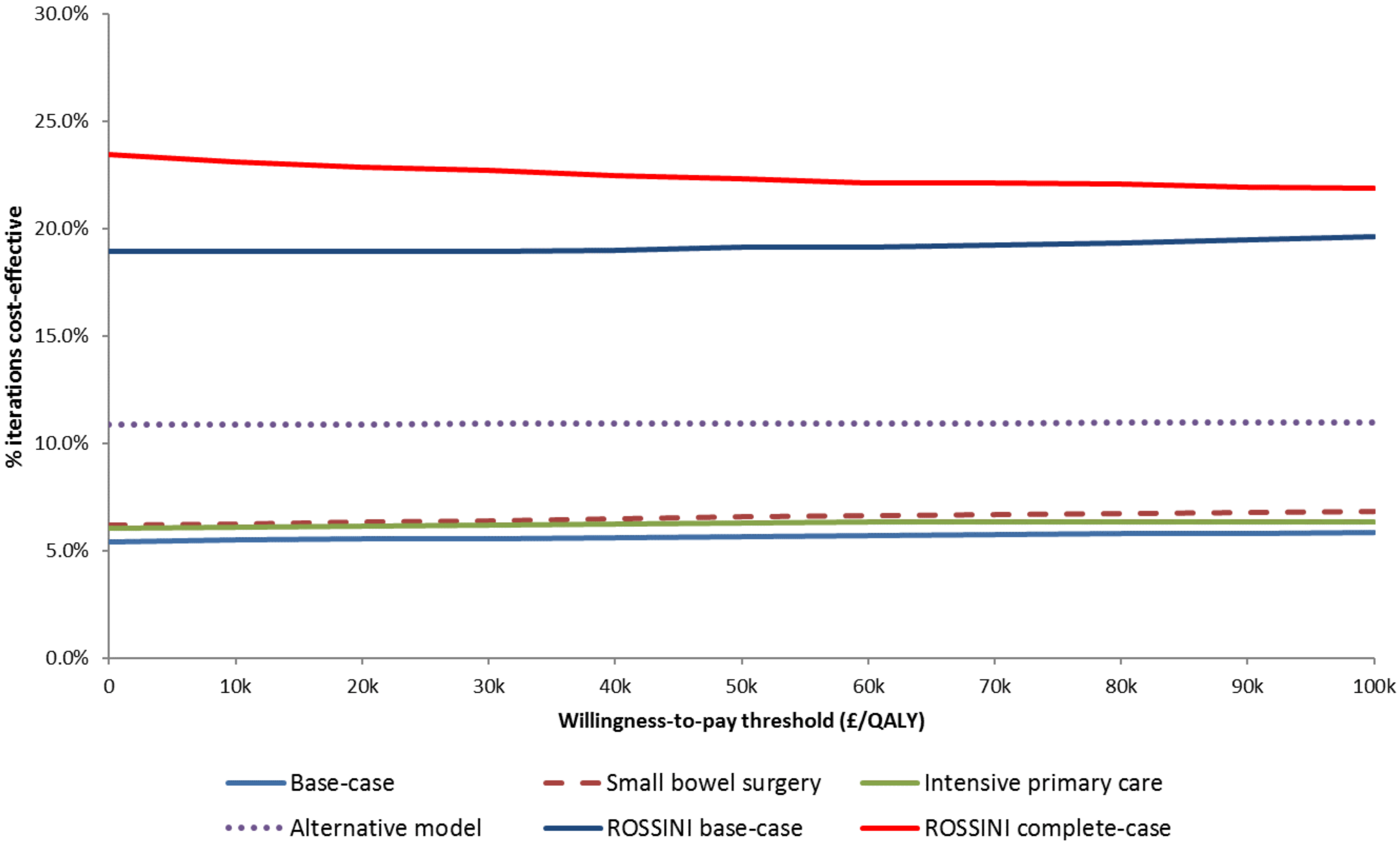


Figure 6.3 Comparison of cost-effectiveness acceptability curves: updated decision model and ROSSINI results

A limitation of both the decision model and ROSSINI was that the use of wound dressings was not accounted for. There was insufficient evidence in the literature to inform an appropriate model input and the ethical approval for a close monitoring of wound dressings in ROSSINI came in too late to be fully implemented. There are indications that wound dressings are an important component of wound care (243); this has also been reflected in the differential number of nurse visits between SSI and non-SSI patients (Table 5.19). Future studies investigating interventions aimed at reducing the burden of SSI should incorporate this element in the study design in order to obtain more accurate representations of the care received in a primary setting.

6.3. Conclusion

The evidence on the clinical effectiveness of WEPDs appears to be contradictory. However, the methodological quality of the ROSSINI trial is superior to that of previous smaller trials and thus ROSSINI results are more robust. For this reason, it is very likely that WEPDs are not associated with any significant benefit compared to standard care in reducing SSI in patients undergoing open abdominal surgery. Consequently, both the updated decision model and trial-based analyses suggested that WEPDs are also unlikely to be cost-effective. In light of this evidence, WEPDs cannot be currently recommended for routine use in the NHS.

Generalisability has been of the principal aims of ROSSINI design. The following Chapters will explore the potential impact of centre selection on trial results and will use ROSSINI as a case study to demonstrate a novel approach to quantify this impact.

CHAPTER 7. CENTRE SELECTION IN RANDOMISED CONTROLLED TRIALS IN THE UK: CURRENT AND OPTIMAL PRACTICE

The work presented in this Chapter has been published as:

Gheorghe A, Roberts TE, Ives JC, Fletcher BR, Calvert M (2013) Centre Selection for Clinical Trials and the Generalisability of Results: A Mixed Methods Study. PLoS ONE 8(2): e56560. doi:10.1371/journal.pone.00565

Section 1.3 of the Introduction argued for the necessity of understanding the current practice of centre selection for RCTs in order to ascertain whether there is a potential for bias in the clinical and economic results of contemporary trials. The remainder of the thesis presents an empirical exploration of the practice of centre selection (Chapter 7), followed by the development of a novel methodology (Chapter 8) which will be demonstrated using the ROSSINI trial (Chapter 5) as a case study. This Chapter presents the methods and results of an investigation into the current and ideal practice of centre selection for clinical trials in the UK.

7.1. Introduction

As pointed out in Chapter 1 (sub-section 1.3.1), the external validity (generalisability) of RCTs may be questioned (78, 83, 84). For example, evidence suggests that trial participants are often unrepresentative of the target population (87-92), which can introduce bias in the measures of effect (95). The choice of participating centres also has a role (78), especially in non-pharmacologic trials, as outcomes may be affected by factors like hospital volume (97) and practitioners' expertise (98). Ensuring the generalisability of RCT results may be particularly challenging for economic outcomes informing health policy changes. Whilst the relative clinical effect of an intervention has been historically assumed constant across settings, albeit not without challenges (78, 128, 129), this assumption may not hold for economic outcomes (130, 131). Modelling methods are one way to retrospectively address this limitation, but they rely on inferences made on a sample of centres whose representativeness to their jurisdiction is unknown (134, 155, 164, 167, 362).

Given that the sample of participating centres may impact on the generalisability of trial results, especially with respect to decision making based on cost-effectiveness evidence, the question arises as to whether the current practice of clinical trials design and conduct allows for such a bias to occur. This piece of research had two objectives: first, to establish which factors currently drive centre selection in trials; and second, to reveal what is perceived as good practice in terms of enrolling centres.

A mixed methods approach was employed: a systematic review of protocols of RCTs funded by the National Institute for Health Research - Health Technology Assessment (NIHR-HTA) programme was conducted; two focus groups with clinical trial professionals; and an online survey distributed to clinical trials professionals in the 48 UK Clinical Research Collaborative Clinical Trials Units (UKCRC CTUs) and 10 NIHR Research Design Services (RDS). Two steps were envisaged: first, to assemble a comprehensive list of considerations that trialists consider when including centres in RCTs (by means of the systematic review and focus groups); and second, to have these considerations inform a national survey of UK trialists on the topic of centre selection. As such, the survey design and content were informed by the systematic review and the two focus groups.

The approach targeted RCTs conducted with a clear view to influence policy and thus included studies with a built-in economic evaluation funded by the UK NIHR-HTA stream. It was judged that the systematic review alone would be insufficient to achieve the research objectives and it was decided that it should be complemented with focus groups and a survey of trialists for the following reasons: 1) there is a large amount of heterogeneity in the structure of HTA trial protocols and reporting criteria for selecting sites/clinicians is not a pre-requisite, so any such reporting is at the discretion of researchers; 2) there is evidence in the literature on poor adherence to trial protocols (363, 364); and 3) there is no guarantee that the trialists involved in writing the (sections relevant for centre selection of the) protocol are the ones who actually perform the selection in practice, so new considerations could be brought in the process. Considering all the above, the aim was to obtain a first-hand account of centre selection from trialists and compare it with the findings of the systematic review. Current and optimal practice were contrasted in order to explore trialists' views on the extent to which generalisability in centre selection should be explicitly considered in trial design. The following three sections present the results and methods of the systematic review, focus groups and online survey.

7.2. Systematic review of trial protocols

7.2.1. Methods

The objective of the systematic review was to investigate the process of centre selection in RCTs with a parallel economic evaluation in the UK. More specifically, the review aimed to answer two research questions:

1) How did RCT investigators report the rationale for selecting and including centres in the RCT? and

2) How did RCT investigators report the intention to use methods of addressing the generalisability by location of the trial-wide economic evaluation results?

The review included multi-centre RCTs with a parallel economic evaluation. Any type of economic evaluation was accepted as long as both costs and outcome data were collected alongside the RCT; it was not necessary for costs and outcomes to be formally combined in a cost-effectiveness or cost-benefit metric. Multinational RCTs were accepted if at least one participating centre in the RCT was in the UK. Only RCTs started after 1st January 2005 were included.

Studies with the following characteristics were excluded: RCTs without an explicit economic evaluation component described in the protocol; RCTs initiated before 1st January 2005; analytic decision modelling studies based on one or more RCTs; other types of studies apart from RCTs: cohort studies, case-control studies, follow-up studies, diagnostic accuracy studies; studies where UK centres did not participate; studies that involved animal subjects; studies for which the protocol was not available; and pilot RCTs or feasibility studies. No ethical approval was necessary for the systematic review.

Data sources

Trials were searched in the National Institute for Health Research - Health Technology Assessment (NIHR-HTA) Primary Research (trial) repository, available at the address http://www.hta.ac.uk/projectdata/PjtSearchResult.asp. This source lists details of publicly-funded studies commissioned by the NIHR. The search was performed in July 2011.

Study selection

The study selection process comprised three phases: in Phase 1, all the projects listed in the NIHR HTA Primary Research repository were scanned against the inclusion/exclusion criteria based on the information in the project abstract and, if necessary, the study protocol. In Phase 2, the protocols of all included RCTs and all accompanying publications, as listed on the NIHR website, were downloaded and scanned against the inclusion/exclusion criteria. In Phase 3, the RCTs included at the end of Phase 2 underwent data extraction.

The three-phase approach was applied to all the projects listed in the NIHR HTA Primary Research repository. For validation a second researcher (BF[9]) performed Phase 1 and Phase 2 on 20% of the projects which were randomly selected using the online random numbers generator www.random.org.

Data extraction

For all the included studies (Phase 3) the full study protocol and any accompanying publications were scanned for relevant information.The following information items were extracted: study authors; project start year; acronym of RCT; study design (parallel/cluster/cross-over/other); type of intervention (pharmacologic/non-pharmacologic); intervention; control; rationale for centre selection (free text); rationale for centre selection (yes/no); discussion on generalisability across locations of economic evaluation results (free text); discussion on generalisability across locations of economic evaluation results (yes/no/unclear).

The following definitions were used to ascertain whether a given trial protocol accounted for centre selection and generalisability of economic evaluation results, respectively:

Rationale for centre selection

'*Yes*': the protocol explicitly mentioned one or more reasons or considerations that justify or describe the choice of particular centres for the RCT to the detriment of others. The mere enumeration of recruiting centres did not fall into this category.

'*No*': the protocol did not identify any obvious consideration to justify or describe the choice of particular centres for the RCT to the detriment of others.

[9] Benjamin R. Fletcher, Research Associate, Primary Care Clinical Sciences, University of Birmingham

Discussion on generalisability of economic evaluation results

'Yes': the protocol explicitly mentioned that generalisability across locations would be addressed in subsequent analyses of the economic evaluation results and specified the methods that would be employed.

'Unclear': the protocol explicitly mentioned that generalisability across locations would be addressed in subsequent analyses of the economic evaluation results, but did not identify the methods that would be employed.

'No': the protocol did not mention explicitly that generalisability across locations would be addressed in subsequent analyses of the economic evaluation results.

The protocols were downloaded in pdf format from the NIHR website. The information on centre selection was identified using the 'Find' command embedded in Adobe Reader© with the following search terms entered separately: 'centre', 'site', 'clinic', 'hospital' and 'practice'. The information on generalisability across locations of economic evaluation results was identified using the 'Find' command embedded in Adobe Reader© with the following search terms entered separately: 'economic' and 'cost'. I performed data extraction for all the included studies. Another researcher (BF) checked data extraction for the included studies in the random sample of 20% considered for Phase 1 and 2.

Data analysis

Following the extraction of free text information relevant for the two review questions, the free text was analysed using the meta-summary method (365), such that the information was abstracted, reformulated and categorized into meaningful units i.e. themes, categories and sub-categories. A frequency effect size was calculated for each emerging category as the ratio between the number of studies containing that particular finding and the total number of included studies.

The qualitative data analysis was performed using NVivo 8 software (366): this involved developing the codes and manually coding the extracted free text information in all the included studies. A second researcher (BF) reviewed the code structure and the manual coding for all the included studies. Any disagreements were resolved by discussion.

Centre selection reporting can be confounded by the trial characteristics. In some types of trials, such as those testing non-pharmacologic interventions, one can expect variation in effectiveness across locations because the expertise and practice of local health care professionals influence the patient-level outcomes. Moreover, cluster RCTs may be more likely to justify the choice of participating centres (clusters) compared to parallel RCTs. In acknowledgement of these considerations, an exploratory analysis compared centre selection reporting across non-pharmacologic/pharmacologic RCTs and cluster/non-cluster RCTs to identify such disparities.

7.2.2. Results

365 projects in the UK NIHR HTA Primary Research portfolio were reviewed, of which 129 RCTs met the inclusion criteria; these had a target sample size total of more than 317,000 participants (Figure 7.1). The main reasons for study exclusion were initiation before January 2005 (n=233; 64%) and non-randomised design (n=46; 13%). The majority of included RCTs had a parallel design (n=112; 87%) and investigated non-pharmacologic interventions (n=96; 74%). The vast majority of the included studies compared the intervention against standard care or (an)other intervention(s) and only nine studies were

placebo controlled. Mental health was the best represented therapeutic area (n=25; 19%), while the proportions of studies in areas such as oncology, respiratory disorders, neurology and cardiology were comparable and ranged between 5 and 7% of total. Table 7.1 presents a descriptive summary of the included studies. Appendix 9 contains a full list of the included studies.

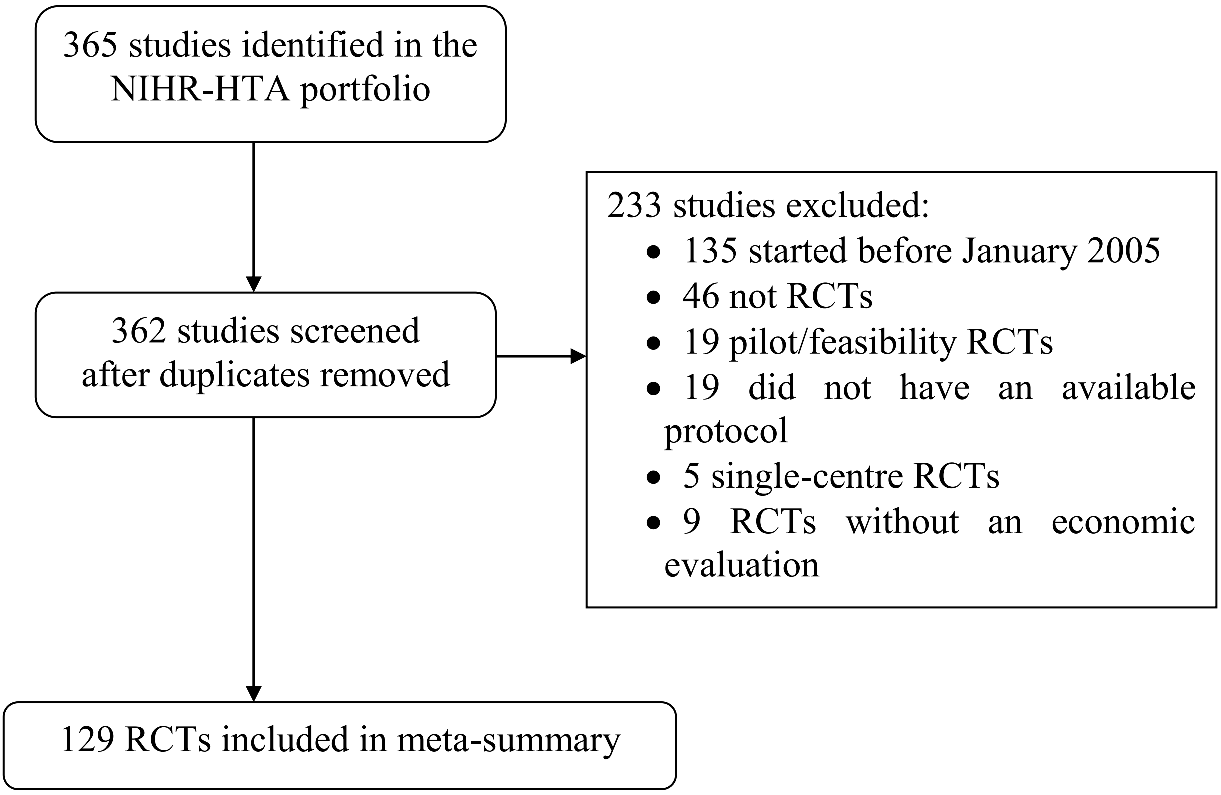

365 studies identified in the NIHR-HTA portfolio
362 studies screened after duplicates removed
233 studies excluded:
• 135 started before January 2005
• 46 not RCTs
• 19 pilot/feasibility RCTs
• 19 did not have an available protocol
• 5 single-centre RCTs
• 9 RCTs without an economic evaluation
129 RCTs included in meta-summary

Figure 7.1 Systematic review of trial protocols: Study inclusion flowchart

Table 7.1 Systematic review of trial protocols: characteristics of included RCTs

Characteristic	Number of studies (%, n=129)
International recruitment	
Yes	9 (7%)
No	120 (93%)
Design	
Parallel	112 (87%)
Cluster	14 (11%)
Factorial	3 (2%)
Intervention	
Pharmacologic intervention	33 (26%)
Non-pharmacologic intervention	96 (74%)
Comparator	
Placebo	9 (7%)
Standard care or other intervention(s)	120 (93%)
Therapeutic area	
Mental health	25 (19%)
Oncology	9 (7%)
Musculoskeletal disorders	9 (7%)
Respiratory disorders	8 (6%)
Obstetrics and gynaecology	8 (6%)
Behavioural medicine	8 (6%)
Neurology	7 (5%)
Infectious diseases	6 (5%)
Digestive tract disorders	6 (5%)
Cardiology	6 (5%)

Note: Other therapeutic areas with less than 5% of studies were (number of studies): obesity (5), diabetes (5), urology (5), haematology (5), circulatory disorders (5), dermatology (3), dentistry (3), emergency medicine (2), ageing (2) and five other miscellaneous areas.

Considerations for centre selection

Of 129 included trials, 78 (60%) reported one or more considerations related to centre selection. The meta-summary identified 53 unique centre selection considerations (Appendix 10) that were grouped into three themes comprising 13 categories (Table 7.2).

Theme 1: 'Diversity and Representativeness'

'Diversity and representativeness' refers to trialists' explicit concern for enrolling representative or diverse centres in the RCT. Although diversity and representativeness have different meanings, in this context both concepts strongly relate to ensuring that the trial is conducted in such conditions (for example in terms of population, health care setting, clinical practice) so that its results can be generalised at national level. As such they both denote an interest for generalisability and for this reason they were analysed together under one theme. In 31 studies (24%) the rationale for centre selection explicitly referred to the need for a diverse or representative sample. The considerations that trialists invoked with respect to ensuring diversity/representativeness pertained to three categories: population characteristics, health service delivery and centre setting.

'Population characteristics' refers to an interest for recruiting from centres which serve diverse populations. 14 study protocols (11%) included such considerations in their description of centre recruitment. Diversity was categorised according to terms which described three sub-categories, namely socio-economic status (n=10; 8%), ethnicity (n=9; 7%) and cultural background (n=1; 1%).

Table 7.2 Systematic review of trial protocols: centre selection considerations, results of the meta-summary

Themes	Frequency (effect size)
PROVIDED CONSIDERATIONS FOR CENTRE SELECTION	**78 (60%)**
Diversity and representativeness in terms of...	**31 (24%)**
Population characteristics	14 (11%)
Health service delivery	15 (12%)
Centre setting	15 (12%)
Centre characteristics	**57 (44%)**
Centre setting	4 (3%)
Health service delivery	16 (12%)
Trial intervention	31 (24%)
Research	19 (15%)
Centre size (catchment area/patient throughput)	22 (17%)
Trial participation	**37 (29%)**
Recruitment	17 (13%)
Trial constraints (time, budget)	5 (4%)
Ensuring trial processes and requirements	24 (19%)
Support for running the trial	7 (5%)
Willingness	9 (7%)
DISCUSSED THE GENERALISABILITY OF ECONOMIC EVALUATION RESULTS	**18 (14%)**
Methods of addressing the generalisability of economic evaluation results	
Sensitivity analyses	13 (10%)
Multilevel modelling	2 (2%)
Collecting costs from representative centres	2 (2%)
Regression modelling	2 (2%)

Model interpretation: Of 31 RCTs (24% of total) which mentioned at least one consideration for centre selection pertaining to diversity and representativeness, 14 RCTs (11% of total) were concerned with diversity in terms of population characteristics, 15 RCTs (12% of total) mentioned diversity in terms of health service delivery and 15 RCTs (12% of total) referred to diversity in terms of centre setting.

Examples of relevant excerpts are given below:

"*The four centres [...] serve a population that includes people from a variety of different ethnic communities.*" [ID3]

"*We will use Census and deprivation data to select a higher number of practices located in low socio-economic areas to ensure full representation of smokers from areas of high deprivation*" [ID14]

"*To further increase generalisability recruitment will be from urban and rural settings, including a large urban setting with a culturally more diverse population (Bristol) and across localities with the full range of deprivation indices expected in the UK*" [ID74]

'Health service delivery' denotes an interest towards ensuring recruitment from centres with a wide range of health care provision characteristics. 15 study protocols (12%) explicitly documented this intention, most often in relation to the types of organisations or practitioners (n=9; 7%), but also regarding patient case-mix (n=2; 2%), intervention throughput (n=1; 1%) and the range of services offered (n=4; 3%). The category is illustrated by several relevant excerpts below:

"*In the UK, study centres will be UK Dermatology Clinical Trials Network (UK DCTN) dermatologists in a mixture of district general and teaching hospitals.*" [ID59]

"*We will recruit obese adults from GP practices, exercise on prescription schemes, commercial weight loss programmes, gyms and the community. The multiple sources of recruitment should increase the generalisability of the study results.*" [ID98]

"*Coverage by the screening programme in [...] is similar to that for England as a whole.*" [ID118]

15 protocols (12%) included considerations pertaining to 'centre setting', thereby referring to recruitment from locations with a wide range non-health care delivery characteristics, such as urban-rural mix (n=8; 6%), geographical region (n=3; 2%), size (n=1; 1%) and type of community (n=1; 1%). Examples include:

"*Recruitment will take place in eight secondary care referral centres in the UK serving a variety of ethnic and social groups and including both urban and peri-urban dwellings.*" [ID117]

"*We aim to involve all regions of the UK.*" [ID102]

"*To aid generalisability participants will be from a range of community settings in the four study sites* including [...]" [ID83]

Theme 2: 'Centre characteristics'

The 'centre characteristics' theme refers to clarity when the trial aim is to enrol centres with particular features (such as location, size and research activity) that reflect a centre's day-to-day setting and activity. The content of some of the categories and sub-categories under this theme overlaps with that of the units in the previous theme 'Diversity and Representativeness'. The difference between the two resides in *why* the individual considerations were specified: in 'Diversity and Representativeness', certain characteristics were invoked to enhance generalisability; on the other hand, considerations counted as 'Centre characteristics' were not explicitly invoked with the aim of ensuring generalisability. Most often they relate to elements of study design and to the nature of the clinical question the trial addresses. In some cases no explanation is apparent as to why trialists preferred certain characteristics over others, such as hospitals of a given dimension.

57 studies (44%) provided such a rationale for centre selection. Five categories emerged, namely: 'centre setting', 'health service delivery (research ready)', 'trial intervention', 'research' and 'centre size'.

Protocols which specified a particular 'centre setting' (n=4; 3%) referred to aspects such as geographical location (n=2; 2%) and deprivation status (n=1; 1%). One RCT recruited in a particular centre because it was the only available facility in the region:

"*The University Hospitals of [...] NHS Trust is the only facility within the county of [...] providing inpatient emergency medical care to the inhabitants of [...]*" [ID131]

The 'Health service delivery (research-ready)' category groups considerations which relate to centres' health care provision characteristics (n=16; 12%). In addition, some of these characteristics put the centre in a good position to undertake research activities (these are to be distinguished from particular requirements for trial participation, which are discussed separately in a dedicated theme). Most protocols specified they would recruit from NHS centres (n=7; 5%) and are interested in organisations/practitioners with a clear interest in the clinical question under investigation (n=6; 5%).

"All centres will be NHS Trusts" [ID128]

"Each participating centre (and investigator) has been identified on the basis of: [...]; having at least one lead clinician with a specific interest in, and responsibility for, supervising and managing children who present with acute exacerbations of asthma" [ID15]

"Each participating centre (and investigator) has been identified on the basis of: lead clinicians in radiology, respiratory medicine, pathology and surgery with a specific interest in the management of early lung cancer" [ID37]

Other protocols mentioned more specific considerations, such as being a centre of excellence (n=1; 1%) and the centre having received satisfactory peer review (n=1; 1%).

"The [...] hub in made up of [...] hospital, a Centre of Excellence in the Treatment of Musculoskeletal disease and a Rheumatology centre for the region." [ID24]

"The criteria of participation for a centre are as follows: [...] 2. Centre received a satisfactory peer review within last 2 years" [ID43]

The 'Trial intervention' category (n=31; 24%) relates to centre-specific considerations which are relevant for the intervention being investigated in the RCT. This was the best represented category both in this theme and the entire analysis: approximately a quarter of the included studies and approximately 40% of the ones that reported any explicit centre selection consideration invoked such intervention-related characteristics. Most protocols referred to the generic suitability of implementing the intervention in the enrolled centres (n=16; 12%) and the amount of experience centres/practitioners have in delivering the intervention (n=13; 10%). Experience was referred to equivocally: while most protocols asked for a given level of previous experience, several of them specifically looked for centres where the intervention had never been delivered before (n=6; 5%).

"The entry criteria for a site to participate in the [...] trial are that participating surgeons must have inserted at least 3 fistula plugs." [ID104]

"All centres will have carried out a minimum of over 250 BYPASS procedures before entering patients into the trial." [ID128]

"The following criteria must be met for a site to participate in [...] - a site must: [...]; not be providing early, goal-directed, protocolised resuscitation as part of standard resuscitation practice" [ID73]

Several protocols went further and required a given level of demonstrated performance in service delivery (n=5; 4%):

"The choice of centres has been informed by a national audit of ureteric stone management undertaken by the British Association of Urological Surgeons (BAUS) Section of Endourology in 2007 (co-led by our group)" [ID65]

"They must be able to provide CT scans of sufficient quality to the study centre in Newcastle." [ID70]

'Research' comprises considerations which speak about a given centre's specific research capabilities, as judged by the trialists. 19 protocols (15%) included such criteria, which most often referred to the centre being part of a research network/group (n=10; 8%) and the centre having previous research experience (n=10; 8%).

"In the UK, study centres will be UK Dermatology Clinical Trials Network (UK DCTN) dermatologists" [ID59]

"all but one [centres] have close association with the Mental Health Research Network if the National Institute for Mental Health (England)" [ID3]

"All three centres have a strong record of research in primary care and experience of, and commitment to, mental health trials." [ID20]

"The Southern hub consists of eight NHS trusts that have previously participated in RA hand research." [ID24]

Finally, the 'Centre size' category (n=22; 17%) refers to trials which explicitly targeted centres of a given size. Most protocols defined size based on patient throughput (n=11; 9%), while others invoked catchment area (n=7; 5%) or the actual size of the centre (n=5; 4%).

"Participating stroke units provide organised stroke care to a population of over 1.5 million and admit over 3,000 stroke patients per year." [ID119]

"All centres will be NHS Trusts, with surgical units carrying out at least 50 bariatric surgery operations per year." [ID128]

"The Midlands hub consists of three large acute trusts in the region" [ID24]

Theme 3: 'Trial Participation'

The 'Trial participation ' theme groups criteria that are meant to ensure a centre's successful integration in the RCT processes. They reflect trialists' desire to recruit from centres which are likely to successfully deliver the research within the specified time frame and budgetary constraints. This translates to clarity in the protocol, on the one hand, about the recruitment targets, trial processes and particular requirements that centres are expected to meet upon participation; and, on the other hand, about the centres' own commitment to participate in the RCT, reflected in their willingness and the support they would receive from other stakeholders. Furthermore, the centre selection process is influenced by the trial's own constraints, independent of centre characteristics, such as budget, calendar and regulatory requirements. The theme thus includes five sub-categories, as follows: recruitment; particular trial constraints; ensuring trial processes and requirements; support; and willingness. In total 37 studies (29%) reported such considerations.

17 trial protocols (13%) explicitly targeted centres that can recruit patients in a timely manner (n=10; 8%) and that have access to the relevant study population (n=8; 6%). The level

of detail to which recruitment was specified varied: some protocols were fairly generic (e.g. protocol ID9), while others were much more prescriptive (e.g. protocols ID59 and ID80). Moreover, there were instances when trialists inspected the centres' previous recruitment rates or patient throughput data to inform their inclusion decision (e.g. protocols ID124 and ID31).

"Hubs will be selected upon the basis of: [...] identifying that they will be able to recruit the required number of patients;" [ID9]

"Each centre will need to recruit approximately 7 participants over a 3 year recruitment period to meet the recruitment target." [ID59]

"Criteria for selection of trial sites & clinicians: [...] the site has the potential to recruit at least 10 patients within the 12-18 month recruitment period;" [ID80]

"Referral rates for all the clinical sites for people with OCD range from between 60-100 patients per year. To ensure recruitment we have checked waiting lists in both primary and secondary care in our clinical sites and waiting lists range from 4 to 18 months." [ID124]

"Hospitals have been selected on the basis of recruitment rates in previous trials." [ID31]

The 'Trial constraints' category refers to protocols (n=5; 4%) which explicitly selected centres based on constraints that the RCT as a whole faced, such as time frame (n=1; 1%), the cost of including a centre (n=2; 2%) and proximity to study site (n=2; 2%).

"We will focus our efforts on recruiting from these practices in the first instance to reduce the number of practices required and to reduce costs." [ID95]

" In order to recruit a sufficient number of centres within the time frame of the project, it will be implemented in three different regions, one in Wales, one in England and one in Scotland, with a combined population of 5-6 million" [ID19]

"We will approach around 100 AOTs and CMHTs in total that are based within reasonable distance of the study sites so that regular travelling to the teams in realistic." [ID60]

Most protocols mentioned considerations related to the centre 'Ensuring trial processes and requirements' (n=24; 19%). The majority of these fell into two sub-categories, namely: ensuring compliance with trial procedures and regulatory requirements (n=17; 13%) and having the required time, staff and facilities to undertake the trial (n=16; 12%). Several examples are presented below.

"Centre/clinician inclusion criteria: [...] 3.Local principal investigator who acknowledges and agrees to conform to the administrative and ethical requirements and responsibilities, in compliance with Good Clinical Practice and regulatory requirements." [ID35]

"Centre/clinician inclusion criteria: [...] 4. the centre has an adequate number of experienced staff to conduct the trial properly and safely according to GCP i.e. to be able to be trained to follow the treatment protocol required and record all of the assessments at the appropriate times as described in Sections 7 and 8" [ID55]

"Each participating centre (and investigator) has been identified on the basis of: [...] ensuring that enough time, staff and facilities are available for the study; " [ID37]

A smaller number of protocols included more specific considerations, such as ensuring communication with the trial office (n=6; 5%), arranging patient follow-up (n=1; 1%) and identifying local champions to advance trial delivery (n=1; 1%).

"Each participating centre (and Investigator) has been identified on the basis of: [...] providing information to all supporting staff members involved with the trial or with other elements of patient management;" [ID89]

"*Each site must identify emergency medicine, critical care medicine and acute medicine "champions"."*[ID73]

The 'Support' category includes protocol statements which referred to prospective centres being supported by relevant stakeholders to participate in the trial (n=7; 5%). Most accounts envisaged support from the centre's management body (n=4; 3%), but local commissioners (n=1) and the relevant research network (n=1) were also mentioned.

"Each service will have: [...] 3. written agreement to participate from the service manager." [ID19]

"Each participating centre (and investigator) has been identified on the basis of: [...] support from the Trust's CEO;" [ID37]

"We will prioritise invitations to centres that have support from their local commissioners" [ID38]

Finally, 'Willingness' refers to trialists explicitly mentioning a centre's willingness as a relevant consideration towards its participation. Several formulations are apparent, such as willingness to randomise, willingness to perform the intervention and willingness to participate.

"The only exclusion criteria are lack of willingness to participate and [...]" [ID60]

"Trial sites will be selected on the basis of the following criteria: willingness to participate in the study; [...]" [ID95]

Two exploratory analyses were performed to compare centre selection reporting across non-pharmacologic/pharmacologic and cluster/non-cluster RCTs (Appendix 10). Pharmacologic RCTs did better in reporting centre selection considerations compared to non-pharmacologic RCTs (67%, n=22/33 vs. 58%, n=56/96). Non-drug trials performed much better in including diversity and representativeness considerations (27%, n=26/96 vs. 15%, n=5/33), especially those pertaining to population characteristics and health service delivery. Nevertheless, a larger proportion of drug trials identified specific centre characteristics (55%, n=18/33 vs. 41%, n=39/96) and trial participation considerations (42%, n=14/33 vs. 24%, n=23/96).

There was only a small proportion of cluster RCTs in the sample (14 trials vs. 115 non-cluster trials), but almost all of them (93%, n=13/14) included centre selection considerations in the protocol, as opposed to non-cluster designs (57%, n=65/115). Cluster RCTs were more prescriptive about centre selection than non-cluster RCTs across all the three themes, particularly 'Diversity and representativeness' (43%, n=6/15 vs. 22%, n=25/115) and 'Trial participation' (57%, n=8/15 vs. 25%, n=29/115). One of the largest discrepancies related to ensuring diversity/representativeness in terms of health service delivery, where very few non-cluster trials included considerations (8%, n=9/115 vs. 43%, n=6/15).

Methods of addressing the generalisability of economic evaluation results

18 RCT protocols (14%) mentioned explicitly the intention to address the generalisability of economic evaluation results (Table 7.2). The majority of these (n=13 studies) mentioned they would perform "sensitivity analyses" to explore the extent to which their findings are applicable to other settings. Two studies each referred to multilevel modelling methods, collecting costs from representative centres and regression modelling, the latter without giving any further details.

7.2.3. Discussion

Summary of findings

365 studies were screened and 129 met the inclusion criteria. The meta-summary identified 53 centre selection considerations and 4 strategies to explore the generalisability of economic evaluation results. The centre selection considerations were grouped in three themes i.e. diversity and representativeness, centre characteristics and trial participation, and 13 categories.

Of 129 trial protocols reviewed, 78 (60%) provided at least one explicit centre selection consideration. Approximately a quarter of them (n=31; 24%) referred to diversity and representativeness, while more studies invoked particular centre characteristics (n=57; 44%) and trial participation considerations (n=37; 29%). In terms of ensuring generalisability, the emphasis was comparable across population characteristics (11%), health service delivery (12%) and centre setting (12%). 18 protocols (14%) mentioned the intention to explore the generalisability of economic evaluation results.

Interpretation of findings

The considerations for centre selection appear to be currently under-reported in RCT protocols, thus making it difficult to ascertain the characteristics of the sample of participating centres at the design stage. Most explicit considerations concern particular centre characteristics, often in relation to the trial intervention and the centre size, and specific trial requirements, particularly recruitment and meeting regulatory requirements. These findings indicate that the trialists' main concern is to include centres which can support the attainment of the trial's successful completion. It appears that some centres may be perceived by trialists as being more aligned than others with the values and practical requirements of conducting clinical research, either in terms of capability or intent. An assessment of centre-level research capability was beyond the scope of this research, but such an investigation would help answer whether trialists' perceptions are objective i.e. some centres are indeed more 'research-ready' than others, or not. In terms of intent, of particular interest is the willingness sub-theme, which was nuanced as 'willingness to randomize', 'willingness to participate in a particular trial' or 'willingness to perform the intervention'. There is considerable overlap between these formulations; furthermore, it is possible that willingness may also include broader considerations which were not explicitly named here, such as 'willingness to take part in a particular trial' or 'willingness to take part in research'. The scope of the meta-summary is limited in this respect, but willingness will be discussed again in the following sections.

The choice of trial protocols as data sources for the review was informed by existing evidence on selective or biased reporting in trial publications (367, 368). For example, previous reviews assessing the compatibility between protocols and reports found that the latter either omit or distort information pertaining to outcomes (369), statistical methods (370) and eligibility criteria (371). In that respect, protocols can be judged to better reflect trialists' intentions. Still, it has to be acknowledged that there is no guarantee that protocol specifications will be fully enacted. Although the CONSORT statement requires all protocol deviations to be reported (27), a systematic review found that reporting of protocol violations is poor, especially with respect to enrolment and randomisation (372). I did not look for such protocol deviations in this study.

Meta-summary has been used in health care research, both in specific clinical areas (373) and to address methodological questions. For example, Limkakeng *et al.* explored the attitudes towards medical research in emergency settings (374) and Fletcher *et al.* explored the barriers to clinician recruitment in RCTs (375). To my knowledge, however, this is the first time when meta-summary was used to analyse information in clinical trial protocols. It has been noted that the distinction between meta-summary and meta-synthesis is a fine one because synthesis is inherent to any summarizing effort (365); the 'interpretive' component of this particular analysis was intentionally kept to a minimum because the aim of this research was to identify the features of centre selection rather than to generate a working theory of centre selection, therefore the findings should be viewed in this light.

The effect size of individual considerations must be interpreted with caution as it may not reflect the true importance of a given consideration, as perceived by trialists, or the extent to which it actually informs the centre selection process in practice. Instead, it can indicate the extent to which each consideration is perceived as important enough to be mentioned explicitly in the trial protocols.

It emerged from the analysis that RCT protocols in the NIHR portfolio have a heterogeneous structure: there was no common format that they followed; not all of them had a table of contents; and information on centre selection considerations was not confined to a particular section. For example, several protocols had a dedicated section entitled

'Inclusion/exclusion criteria for clinicians [or sites]', where such considerations were presented in detail (for example ID35, ID 55, ID 95, ID 112 and ID122). However, these instances were few in this sample of trial protocols. It is likely that had there been a standard protocol format that all RCTs had to comply with, the findings would have been different. This heterogeneity could be the result of the lack of unified guidance on protocol design. A recent systematic review of guidelines for RCT protocols found substantial variations in the recommendations and enforced the need for an evidence-based, systematically developed document (376). The release of the SPIRIT 2013 statement (15) and the availability of protocol writing tools (377) may contribute to closing this gap.

The uptake of analytical methods to explore the generalisability across locations of economic evaluation results appears to be low. Since the extensive systematic review of Sculpher *et al.* was published in 2004 (127), more sophisticated methods accounting for hierarchical structure of the data have been proposed (164, 167). ISPOR currently endorses bivariate hierarchical modelling as the preferred approach to address such concerns at the analysis stage (134). Although these methods have been demonstrated and are especially useful in the context of multinational RCTs, this is unlikely to be the reason why they haven't been adopted more in the UK setting since their proponents have constantly argued that the methods are equally applicable to within-country settings, as well.

Strengths and limitations

The NIHR repository was chosen as the data source because it allowed access to the original, full-text study protocols, thereby allowing a close investigation of trialists' explicit intentions towards centre selection. This resource has been used before for trial methodology research: for example, Jones *et al.* examined the documentation of 48 NIHR-funded trials to identify whether their design was informed by systematic reviews (378). Several advantages are associated with this data source. First, it offers access to information about the version history and protocol development. Second, it includes publicly funded RCTs with a distinct view to influence policy and practice, therefore utterly relevant for the generalisability issue; it is reasonable to assume that funders are interested in the trials obtaining representative results that are directly relevant to the NHS context. Conversely, this also acts as a limitation because the sample may not be representative of the clinical trial practice in the UK. However, it can be argued that RCTs funded through alternative streams (e.g. industry or charity-led studies) may have less of an explicit concern in ensuring generalisability. Therefore these findings may actually overestimate the current interest towards generalisability in centre selection.

The review only included trials with an explicit economic evaluation component. Before considering this as a limitation, two further points must also be considered. First, the UK decision making body (NICE) requires evidence of cost-effectiveness before advising on the nationwide adoption of a medical technology. The economic evaluation component is therefore mandatory for such policy changes, which makes it extremely relevant in the context of generalisability and we attempted to incorporate it accordingly. Furthermore, and lending strength to the previous consideration, only 9 trials out of the 365 trials considered were excluded from the systematic review because they did not have an explicit economic evaluation component (Figure 7.1). This suggests that their exclusion is unlikely to have biased the sample and confirms that most UK trials do indeed evaluate economic outcomes. The 1st January 2005 inclusion threshold was chosen as such because the seminal HTA publication concerning the factors affecting the generalisability of economic evaluations was published in 2004 (127).

It was not always possible to make a clear distinction between the emerging sub-themes therefore there is some overlap between and within several codes. For example, the recruitment requirements and the time frame of trial have been coded individually, although in practice they are clearly interdependent. Furthermore, as discussed above, 'willingness' is a broad sub-theme which may include willingness to do research, willingness to participate in a particular RCT at a given time, willingness to randomise against a particular intervention and so on. These considerations may often be intertwined and a textual analysis can only draw artificial distinctions between such concepts. The present analysis was guided by the explicit information provided in the protocols. In that respect, the results reveal only what the trialists thought appropriate to include in the protocol.

It can be argued that the potential bias associated with centre selection is more relevant to some trials (e.g. primary care and surgery trials) than to others (e.g. drug trials), therefore the meta-summary may have overestimated the extent of centre selection misreporting by pooling together various types of RCTs. However, the sample was dominated by non-pharmacologic trials and an exploratory subgroup analysis (Appendix 10) revealed that pharmacologic trials actually did better than non-pharmacologic trials in reporting centre selection considerations (67% vs. 58%), but, as expected, were less concerned with generalisability (15% vs. 27%). The study sample included a high proportion of non-pharmacologic trials, which may limit the applicability of the findings.

An exploratory sub-group analysis was performed to investigate the differential reporting of centre selection considerations in cluster RCTs and non-cluster RCTs, respectively (Appendix 10). The effect sizes suggest that cluster RCTs perform better than non-cluster RCTs in reporting centre selection considerations (93% vs. 57%), especially in relation to representativeness (43% vs. 22%) and trial participation (57% vs. 25%). Such a finding is in line with the interest towards accounting for setting-dependent effects in cluster trials, but the small number of such RCTs in the sample i.e. 14 out of 129, preclude any strong inferences to be made.

7.3. Focus groups

The objective of the focus groups was to complement the centre selection considerations which had emerged from the systematic review of trial protocols so as to ensure that no relevant considerations are missed. Due to the potential discrepancies between study protocols and study conduct and the lack of structure in the NIHR-HTA protocols, it was considered that eliciting trialists' views on the centre selection process could identify and fill any gaps in the findings of the systematic review, thereby leading to a consolidated list of considerations to inform a national survey (section 7.4 below).

Focus group methodology was useful here because it allowed the capture of data that resulted from discussion and negotiation (379), and thus helped distinguish between factors that affected participants as a group and those that were specific to individuals. The focus groups did not aim to reach consensus on the practice of centre selection, but to identify as many relevant issues and considerations as possible in order to inform the design and content of the online survey.

7.3.1. Methods

The Science, Technology, Engineering and Mathematics Ethical Review Committee at the University of Birmingham favourably reviewed this phase of the study (Ref. no. ERN_11-0792).

Email invitations to participate in focus groups were circulated in August 2011 to all staff affiliated with the Birmingham Centre for Clinical Trials, comprising three distinct trials units: Cancer Research UK Trials Unit, Birmingham Clinical Trials Unit and Primary Care Clinical Research Trials Unit (Appendix 11). Participants who expressed an interest to participate were distributed a link to an online poll where they could mark their availability for all working days in September 2011.

Participants were asked for written informed consent (Appendix 12) prior to their participation in the study, which was provided on forms approved by the Ethical Review Committee. Participants were not reimbursed for their participation, but lunch was provided.

Two focus groups were conducted (n=6 and n=4 participants, respectively – please see details about group composition below) exploring trialists' thoughts and experiences of centre selection with the aim to identify potential reasons for centre selection not already identified in the systematic review. The first focus group was attended by a clinical investigator, four trial managers and one health economist; and the second focus group was attended by a trial manager, two trial methodologists and a biostatistician. There were at least two trialists from each of the three trials units which comprise the Birmingham Centre for Clinical Trials. Each focus group was moderated by one experienced qualitative researcher (JI[10] and NG[11], respectively) and co-moderated by myself. I took detailed notes throughout and both sessions were audio recorded for later transcription.

Discussions were structured using a topic guide (Appendix 11) that ensured key issues were explored (380). While mainly informed by the systematic review of trial protocols, the topic guide also inquired about several aspects outside the scope of the review, such as the relative importance of these considerations, the time frame of making such decisions and the relevant professionals involved in the decision-making process. The topic guide was developed and agreed upon by myself, JI and both my supervisors. Participants were also able to direct the content of the discussion, allowing unanticipated themes to arise. At the

[10] Jonathan C. Ives, Senior Lecturer in Biomedical Ethics, Primary Care Clinical Sciences, University of Birmingham

[11] Nicola K. Gale, Lecturer in Medical Sociology, Health Management Services Centre, University of Birmingham

beginning of each focus group, the participants were presented by the moderator with a scenario describing a generic parallel RCT with a concurrent economic evaluation (RCT-EE). The topic guide then inquired about the considerations the participants would find relevant when selecting centres to participate in such a study, and who is more likely to be the major driver behind this decision.

Discussions were transcribed verbatim and analysed using simple conventional content analysis (381), in which the data were coded and arranged into meaningful organizational units, from which themes were derived that described the participants' views. The analysis was performed by myself and reviewed entirely by another (JI).

7.3.2. Results

There were no disagreements between the two researchers (myself and JI) on the coding of focus group data. Four overarching themes and nine sub-themes emerged during the analysis. The four themes were: considerations that influence the decision of including a centre in a RCT-EE; professionals involved in the centre selection process; characteristics of the centre selection process; and the role of health economics in RCT-EEs. The following paragraphs present the content of these themes in more details, supported by selected quotations. Participants' identities have been coded to preserve anonymity and the provenance of each quotation is marked as either FG1 (focus group 1) or FG2 (focus group 2).

Considerations that influence the decision of including a centre in a RCT-EE

The participants touched upon a large number of issues influencing the decision to include a centre in a RCT-EE. The umbrella term 'considerations' was used here to encompass both centre-level characteristics and external influences, as explained below. Three sub-themes were identified: minimum requirements, preference-based considerations and non-preference based considerations. The sub-themes are defined below.

Minimum requirements

There was general consensus across the focus groups about minimum requirements that a centre must fulfil in order to qualify for inclusion in the RCT-EE.

"Participant 1 (FG1): There are minimum requirements for every trial [general agreement], you have to say 'Yes' to these questions...and we always ask these questions at the beginning because we don't want to spend three months with a site to find out that they haven't got a radiotherapy person [general agreement] or the person is on maternity leave and might be back for a year and a half or whatever."

Three such fundamental issues became apparent: the existence of available resources, an interest towards the trial and having access to the relevant patient population.

"Participant 2 (FG1): We take anyone who's got the space, the staff and is enthusiastic"

"Participant 1 (FG1): Who's willing, who can do it and have they got the patient group..."

In terms of available resources, a range of requirements was mentioned. Having relevant specialist staff was the most often mentioned topic, but more pragmatic issues such as having an Internet connection or available physical space were also discussed.

"Participant 1 (FG2): And if you need a specialist member of staff, like somebody to be able to deliver that treatment, then if you don't have anybody at a particular centre with that specialist training, then that centre will already be eliminated out of your..."

"Participant 3 (FG1): Sometimes it's physical space...space for storage of drugs or equipment, space for parking."

Having the necessary licenses for delivering the intervention (e.g. environmental license for radiotherapy) was also noted.

"Participant 1 (FG1): So the first selection was: 'Have you got this environmental license?'. If you have, then we'll go and ask 'Are you interested?'. If you don't have an environmental license and you're not going apply for an environmental license or you haven't got a radiotherapy therapeutic team, then you said 'No' to the first questions and there's no point going further."

The second major category of minimum requirements was related to serving a patient population relevant to the study question. Socio-demographic indicators such as deprivation and ethnicity were cited to lead the centre selection process towards particular areas, where applicable.

"Participant 1 (FG2): I definitely want to know that the patients that are coming through from those centres were what we were expecting and that there wasn't some bizarre ... why they were at a certain end of the scale, say better or more poorly than you'd expect."

"Participant 2 (FG2): [...] if you're working in a particular ethnic group or you want to specifically target people in areas of high deprivation, for example, then you're obviously not going to go somewhere fancy and posh to do that you're going to go to the place where those patients generally are or get referred to or are treated."

Displaying an interest towards the study question was widely seen as a key factor.

"Participant 4 (FG1): I think it has got a lot to do with who's actively interested in taking part, it's the main thing. They're not going to finish unless they're interested, no matter how good their research staff is. If there's no one there who's interested, they're not going to put patients in. So there's got to be their interest there."

"Participant 2 (FG1): [...] so I think that comes back to their original buy in and enthusiasm. So, for me that's quite a big part of...if you can work out which GPs or hospital doctors or whatever are the ones that are enthusiastic and actually buy into your study, then that would be a good way of selecting people that would recruit successfully, I think, and ethically, hopefully."

Preference-based considerations

A different set of considerations are those for which participants expressed unambiguous preferences towards or against and which may clearly influence their decision to do research in one centre or another. As a result, preference-based considerations refer either to centre characteristics that are sought by researchers or to characteristics that make a centre undesirable for inclusion in a RCT-EE.

a. Desirable centre characteristics

A topic that was recurrent throughout the focus groups was the ability to recruit patients in the study. Investigators clearly wanted to conduct research in centres that can deliver in terms of recruitment targets and recruitment time. Although this particular analysis did not have a quantitative remit, 'being a good recruiter' was by far the most often mentioned topic across both focus groups.

"Participant 3 (FG2): We usually start by sending brief questionnaires to all our existing collaborators, saying 'We're thinking about this, are you interested? Do you have people, you know, who does speech language therapy or whatever it is we're studying? And how many people will you be able to recruit?'"

"*Participant 2 (FG1): Your main focus is to get patients into your study and that's you main...your study either succeeds or fails on whether you get your patients or not.*"

Another topic that was touched upon was the local clinicians' understanding of clinical trials, and it was suggested that this helps the communication between trial centre and local centres.

"Participant 1 (FG2): I think it can have an impact on recruitment because if you've got a clinician who understands the whole clinical trial background and the whole reasoning for randomisation etc. they can sell that to a patient"

Building on the previous point, having a good communication relationship with centre staff was referred to as desirable.

"Participant 2 (FG2): Isn't it also more about communication with the people who are at the potential centres...because you're never going to recruit anybody unless you've got a working relationship that you can use and they can build on as well."

"Participant 4 (FG2): Well the people on the ground... picking the poke, aren't they? Because you don't know that until you start the study...the actual people who are going to be doing work."

A convenient location of the involved centres was deemed preferable, especially in cases where site visits have to be carried out within short time intervals (e.g. for collecting biological samples).

"Participant 5 (FG1): I think location is one. We have a small one...in [study name] we don't need that many practices and...because it's blood samples coming back to University and, you know, we don't need huge numbers of practices so we're only going for the really local ones so that we can be going out managing them on a fairly regular basis."

It also emerged that the engagement of the staff involved in research is highly desirable. This refers, on the one hand, to the trial's question being meaningful to them and, on the other hand, to being able to 'sell the trial to patients'.

"Participant 3 (FG2): I think it's that and I think also it's about the PIs at the sites have to buy into your trial. However how you try and minimise it, it's always extra work for them [general agreement] so it's got to be meaningful to them and you've got to make it as attractive to them as possible, make it as simple but also it's got to be a question they recognise needs answering and they want answering."

The computer systems compatibility between the trial centre and local practices appears to play a role as well in centre selection.

"Participant 2 (FG1): Plus I think, for us, sometimes we've selected on kind of what computer system they've got, for example [...]"

A final point relates to generalisability: the group discussions revealed the aspiration that included centres retained generalisability in terms of the target population.

"Participant 1 (FG2): [...] by selecting certain centres you need to make sure that you're not, that you're still retaining that generalisability to that population"

"Participant 4 (FG2): [...] none of us would argue against generalisability because it's obviously something desirable"

b. Undesirable centre characteristics

A number of characteristics would keep investigators from approaching a particular centre or from actually enrolling them in the study. The majority of these considerations mirror the desirable characteristics detailed in the previous section. One of these undesirable features was the lack of interest for the clinical question of the study.

"Moderator: This is again blue sky, naive thinking, but are there any practical situations that couldn't be overcome by increased resources?
Participant 3 (FG1): Yes. People not being engaged [agreement]. If someone's not interested, it doesn't matter what amount of money there is. You got to be engaged at a really high level early on..."

Having a difficult communication relationship with local staff was also mentioned.

"Participant 4 (FG2): So I just think: 'I don't want to do any more research in [centre name] if I can avoid it because I'd rather do the research than have these stupid discussions'"

The clinicians not agreeing with the intervention has been highlighted as a potential barrier in undergoing research in a particular centre.

"Participant 1 (FG1): [...] the two things that will not start a trial if it's got unlimited...endless money is: patients not want to go to and basically if the doctors don't agree with the intervention."

Processing paperwork slowly was cited as an undesirable feature: investigators strongly stressed that they would explicitly avoid locations with a known history of taking a long time to obtain research and development (R&D) approvals or to send completed trial documents.

"Participant 2 (FG1): But when it takes, you know, six to nine months to get approvals for a straightforward study and you know you're not going to choose to go to one where it's going to take you 18 months to do the same thing...so it definitely factors into your choice."

"Participant 3 (FG1): I don't want to count the time waiting in the R&D department to turn around a piece of paper or find out... I want the real research time to be counted, the time taken from the admin process to get happening, the whole things."

For all the considerations detailed above, both desirable and undesirable, participants were generally not in disagreement about their influence on the centre selection decision. This was not the case for previous experience in conducting research, where participants' views differed with respect to its desirability.

"Participant 3 (FG2): The other thing that is really ideal is knowing you've got a site with a particular PI who has a track record in clinical trials, who really understands what it is that we're doing and why it is important."

"Participant 2 (FG1): [...] in my experience some of the best recruiters we've had were practices that have never taken part in research before. They might be really small, single-handed practices of...just interested in the study, have never done before, and they've actually been far easier and far more successful than some of the big, established, well-known practices. So I think, for me, I'm very less...ok, I'm motivated by sort of what population they've got, what area they're in and how easy it is to get going there, but I don't necessarily look at whether they've done research ever before."

Non-preference based considerations

The third sub-theme within factors that influence the decision of including a particular centre in the study has been labelled 'non-preference based considerations': this describes a collection of categories that are not necessarily amenable to personal preferences i.e. they are not intrinsically desirable or undesirable by trialists; they are neutral considerations that are factored in the process of enrolling centres.

A rich category includes a range of centre characteristics. One of them is merely the type of centre that the study requires (e.g. GP practices or hospitals).

"Participant 2 (FG2): [...] because those two types of centres [GP surgeries and hospitals] are so different, the criteria that you'd have to work to in order to select or to establish centres are really different."

The degree to which local staff feels incentivised to participate, the position of the centre on the rural-urban continuum and local staff fluctuations (e.g. due to maternity leaves or changing jobs) have all been mentioned as being important.

"Participant 3 (FG2): Or that can be difficult if people move jobs or people go on maternity leave or they get ill or all sorts of things"

Another important category included requirements of the funding and regulatory bodies and their impact on the trial conduct.

"Participant 1 (FG2): [...] we'll have to have the PI send an email to say that he will support this study if it was endorsed by CRUK"

"Participant 2 (FG2): [...] but high up level the Department of Health and the NIHR [National Institute for Health Research] and all these...all they do is look at the spreadsheets and they look at the figures and they say 'Oh! You're not recruiting, you're not meeting your target. Do something about it!'"

The local research environment i.e. the number and nature of other studies conducted concurrently at a centre were named to influence participation. 'Competition' and 'trial fatigue' have been used to describe this phenomenon.

"Participant 2 (FG2): [...] then maybe a less prestigious centre somewhere else would be better for your overall recruitment because they wouldn't have this competition and this pressure on them to see all these patients and to process them through a load of different studies"

"Participant 2 (FG2): I think the trial fatigue thing is important. Because if you are, I don't know, for a lot of our studies we tend to use the same centres because we do know the people there and we've worked with them before, but if they are the kind of specialist centres in some cases that attract a lot of the patients with a given condition, but they attract a lot of the research studies as well."

A large number of considerations relate to the trial itself. Issues like cost and time constraints, the rarity of the disease and logistics were touched upon.

"Participant 3 (FG2): So depending on what you need in terms of how many patients and how long a time period you got to collect them and how rare a condition it is will feed into what we're looking for in sites"

"Moderator: I was going to ask if that was primarily visiting cost considerations or out of convenience or practical considerations, as well?
Participant 2 (FG1): It's all of that, all of that...all of the above, really. I guess you have limited amount of time to spend on a study. If you spend a whole day driving somewhere and driving back, you know..."

A fertile sub-category here consists of study design considerations: participants repeatedly mentioned that factors such as intervention design, total sample size, total number of required centres and the pool of eligible centres (e.g. for a highly specialised therapy) are all factored in the decision.

"Participant 3 (FG2): [...] And so if you only have one therapist, once you've recruited someone in that arm of the trial, they probably can't manage another person on that arm of the trial...there are two other arms, but they can't go for another randomisation until that person's finished their treatment, so it can make a huge difference if you've got two speech and language therapists...just kind of how the local setups are...so depending on how intense the treatments are going to be...they can really affect...
Participant 1 (FG2): Select a centre for you in itself, almost... [general agreement]"

Patient convenience is also thought to be important when selecting a centre and related to issues such as travel distance and incurred costs (e.g. on-site parking).

"Participant 1 (FG2): I think also there's a cost to the patient...The one I know, it was mentioned in one of our TMG [Trial Management Group] meetings about recruitment of a certain centre where the car parking was astronomical and because patients had to come five consecutive days, the car parking was adding up. There wasn't generally...literally down to how much car parking was having an impact on whether the patient chose to go into the trial or not, you just wouldn't believe. And there's obviously ways around that, you can do...you can sort things out, but unless you're aware of them...."

"Participant 3 (FG1): If it's healthy patients then the things that seem to matter to them are about: geography - if it's coming to your own GP surgery, that's fine. If they have got to go to a community hospital three miles away, well ok. But ask them to go to a hospital 10-15 miles away and they think...they just say 'No'."

Professionals involved in the centre selection process

The participants were asked to identify the types of decision makers involved in identifying and selecting centres for a trial. The lead clinical investigator, the trial coordinator and research networks all appear to have a prominent role in the selection process.

"Participant 1 (FG1): It's not here because basically what will happen with R&D departments is...if the local doctor is interested he will make things happen [smiles, general agreement]."

"Participant 3 (FG2): [...] research networks...ours is usually [research network name] for dementia and neurodegenerative diseases, they're also actively looking for sites for us"

"Participant 2 (FG1): [...] and then, as a trial manager, you say 'Well, we did a trial on this last year and centres X,Y and Z were very good, as well. "

Participants also mentioned the participation of health economists, statisticians, Trial Management Group (TMG) members as a group and members of the Data Monitoring Committee (DMC) as being involved in the decision.

"Participant 1 (FG2): So it's that key sort of group, isn't it, who make those decisions [agreement], who are responsible for writing the protocol, responsible for getting the trial design right, responsible for putting the grant application in that cohort"

"Participant 3 (FG2): We tend to have these TMGs and the TMG will often, being brought together from a previous study, so you bring with it the people who are experienced from a previous study. I think through that TMG you can identify who are going to be good centres to open up first and then which other centres in the second phase of signing are opened."

"Participant 1 (FG2): I mean we've just had a DMC for a study yesterday and a centre that hadn't been...it's quite special, it's treatment so there's only two centres so far in the country that are involved...But the actual chair of the DMC made a proposition for a particular centre and a particular person to be involved and one of the major action points that came out of the DMC was to target this centre that nobody had never thought of before."

Characteristics of the centre selection process

Although not specifically asked about the mechanics of centre selection, participants provided throughout their discussions a wide range of insights regarding how the process of identifying and enrolling centres unfolds. Four sub-themes emerged: (i) identifying centres; (ii) information resources; (iii) the nature of the selection process; and (iv) time considerations.

Identifying centres

Two main types of approaches to identifying eligible centres are apparent: a top-down approach, where investigators purposely scan the clinical and research communities; and a bottom-up approach, where the trial centre is being approached by interested centres themselves. The principal activities related to the top-down approach involve assessing and eliciting interest for the trial.

"Participant 1 (FG2): And if in the mean time you've been in a conference and you've managed to do a bit of publicity and somebody approaches you, then that centre will then be discussed."

"Participant 3 (FG2): We usually start by sending brief questionnaires to all our existing collaborators, saying "We're thinking about this, are you interested?"

Obtaining information about particular centres and offering incentives in some cases were also mentioned.

"Participant 3 (FG2): And that now is not... because of the UK CRN everybody has to be...for our trials we have to be uploading them onto their website monthly about all our recruitment and all the rest of it. It's not just known from your own experience who is and who isn't good recruiters, but presumably your portfolios can potentially look at that...across all the trials there [agreement] supporting across the country. I think that sort of information is becoming a lot more transparent and readily available [general agreement]."

"'Participant 2 (FG2): We're doing this study, you're going to get x amount of service support costs if you help us' because you're not getting anything out in practice unless you pay them for it."

As far as the bottom-up approach is concerned, participants have mentioned a number of times that there are cases when centres want to be part of the trial and contact the trial centre.

"Participant 1 (FG2): [...] we've had these situations before where somebody's approached us to set them up"

"Participant 3 (FG2): We've also had R&D's approach us, research and development units from the hospital trusts, who are looking on the UK CRN portfolio websites..."

A particular situation was described as 'natural selection' and refers to centres being eligible in relation with the study requirements.

"Participant 1 (FG1): You want to start off with 30 or 40 centres but you very soon know that actually I'm constrained on the first five because I get an email back within seven days, rather than seven months or seven weeks. So you actually have...you may select 30 or 40 centres but then, by natural selection, the centres that have a decent system will be open first..."

A remark that was made in relation with bottom up selection refers to centres themselves playing the decisive role in trial participation.

"Participant 3 (FG1): It's not so much about us choosing them, it's them choosing us [agreement]. Whether they say 'Yes' to us..."

"Participant 4 (FG2): What we're saying is that we make choices about which sites we'd like, but they make choices whether they wish to participate in reality [agreement]."

<u>Information resources</u>

A variety of resources are used by investigators to identify and gather information about potential centres. Databases of practices and clinicians, past and existing collaborators, history of trial participation, informal networks and personal contacts were all mentioned as means to inform the selection process.

"Participant 3 (FG1): In the trials unit we keep databases of all the practices we have ever done any research with and how difficult were the trials that they've done."

"Participant 3 (FG2): We first, and again this is mostly hospital based, I don't know if it's relative to this scenario. We start out with people we already work with."

"Participant 1 (FG2): [...] the trial coordinator may have a mail shot from a previous trial, that kind of mailing list."

"Participant 4 (FG2): And then there are lots of things that are about history...you bring in your history, your knowledge and informal networks that tell you that things have changed somewhere or got worse somewhere or got better somewhere"

<u>Nature of the selection process</u>

Participants in both focus groups agreed that centre selection is best characterised by the challenge to merge ideal and pragmatic considerations. Moreover, it emerged that a purely rational selection is implausible.

"Participant 4 (FG2): I don't think that any of us could claim that we use purely scientific criteria or we picked sites randomly because we know that wouldn't necessarily work. But at the same time none of us would say it's an art, none of us would say that we just make it up as we go along, it's so creative [some laughter]. It's somewhere between the two, isn't it, a craft really that ascribing that you have certain desirable characteristics [agreement] and you maximise them..."

Time

In terms of the timing of selecting centres for a trial, two ideas were expressed: first, the sooner the selection is planned, the better.

"Participant 1 (FG2): I just can't emphasise enough that the earlier you start considering these things [general agreement]. You just can't think about these things too early [laughter]. Because if you don't get your centres right, you are not going to have a successful trial, so it has to be considered upfront."

And second: enrolling centres is an ongoing process throughout the trial, as unexpected changes may appear such as some centres recruiting slowly or even dropping out from the trial altogether.

"Participant 3 (FG2): [...] and then adjust it as you go through [agreement]...the weird and wonderful things that you never thought possible in your trial [laughter] with remarkable regularity."

The role of health economics in RCT-EEs

The fourth theme that emerged from the discussions was the position of health economics within such a RCT. Two sub-themes were identified in the analysis: the role of health economics in the trial; and generalisability of economic evaluation results.

The role of health economics

It was agreed that health economics is a secondary consideration in running the trial in general and in selecting centres in particular. There are little reasons to believe that health economics issues may influence the choice of a particular centre.

"Moderator: [...] but in terms of determining if you might select a new centre in order to...
Participant 3 (FG2): Not solely based on health economics, no... [agreement]"

"Participant 2 (FG2): It's not to say health economics is the poor relation or anything [laughter] but they are generally secondary considerations, I think it is fair to say [agreement]."

Generalisability

Building on the previous point, there is a concern about having generalisable economic evaluation results from the RCT (particularly in relation with costs), but as previously mentioned this is less likely to play an active role in centre selection compared with the preference and non-preference based considerations detailed above.

"Participant 4 (FG1): That's the same at PCT level, you know...because if you know a particular region is particularly awkward when you're trying to get approvals from them then I tend to avoid going back there. [agreement]
Participant 3 (FG1): Not worth going.
Participant 4 (FG1): No, it's not, it's not worth it. So...
Participant 6 (FG1): That's a worry, because that's affecting the generalisability of the results."

"Participant 2 (FG1): And all the members of the trial team - the statisticians, the health economists - basically...we just have to make the best of what we get [laughter, agreement] and then prompt in our discussion how the generalisability was...or lack of generalisability, we may have to deal with it."

7.3.3. Discussion

Summary of findings

The focus groups identified a wide range of themes and sub-themes pertaining to centre selection in RCTs. An important distinction that trialists made is that between key considerations, which have been denoted ‘minimum requirements’, and other considerations which factor in the centre selection decision. Results suggest that trialists tend to seek certain centre-level characteristics and avoid others. These preferences appear to be largely driven by pragmatic imperatives, such as the proximity to trial office, administrative ease and the expectation of successfully recruiting patients. The importance of ensuring generalisability by means of centre selection was acknowledged, mostly in relation to the trial population, but there was general agreement that it is a secondary consideration in the centre selection decision.

Interpretation of findings

A large number of the centre selection considerations which emerged from the focus groups were identified in the systematic review of trial protocols. These include: the level of motivation of centre staff, meeting recruitment targets, research experience, the regulators', funding bodies' and sponsors' requirements. This enforces the findings of the systematic review of trial protocols and supports the relevance of these considerations for current practice. Furthermore, it appears that, overall, the sample of trial protocols in the review contained most of the relevant considerations, thus suggesting that protocol texts serve their purpose as a reflection of how trialists will actually proceed. However, this must still be interpreted cautiously and on a case by case basis in light of the great amount of heterogeneity in trial protocol structure and content (sub-section 7.2.3).

There were several novel considerations that the focus groups revealed. These include patient convenience and the state of the local research environment. More importantly, they allowed insights in the process of centre selection, specifically on the professionals involved and the time frames.

No other centre-level variables were mentioned in relation to generalisability apart from ensuring a representative patient population. However, other centre-level variables are also known to influence the generalisability of trial findings, as well, such as the experience and training of health care professionals, local economic environment and the managerial performance of the centre’s leadership (section 1.3). It may be the case that research findings on these latter topics have not penetrated enough in the trialists’ community so as to become a prominent concern.

Although not originally designed towards this end, a major contribution of the focus groups was to reveal a classification of centre selection considerations, which wouldn’t have been possible on the basis of the systematic review alone. Thus, there is a set of minimum considerations (resources, willingness and access to the relevant population) which must be met before any reasoning takes place. Further, there are desirable and undesirable centre-level characteristics. Preference appears to be established in relation to the expectation of meeting pragmatic requirements such as processing documents, ensuring communication and meeting recruitment targets. Furthermore, there are considerations towards which no explicit preference was revealed, but which must be accounted for in the centre selection decision. Besides the research environment, this latter category includes study design elements, patient convenience and further centre characteristics such as the type of centre, thus suggesting that the approach to selecting centres has a strong trial-specific component.

The focus groups also revealed what appears to be a tension between pragmatism and ideal practice. This applies both to study design in general and to health economics considerations in particular. On the one hand, trialists’ accounts often emphasised the pressures and requirements that centre selection and the trial in general must meet. On the other hand, there were indications that generalisability is obviously seen as desirable, albeit not often acted upon. The focus groups were not designed to explore this topic further, but this finding contributed to the design of the survey (section 7.4) and is explored further below.

The focus groups did not aim to inform a standalone theory of centre selection and trial conduct, but merely to complement the list of centre selection considerations which emerged from the systematic review of trial protocols. Nevertheless, elements of the process of conducting RCTs surfaced throughout the discussions. For example, trialists highlighted the types of pressures they are faced with when planning and conducting the studies. The most prominent of these related to meeting the recruitment targets, time and budget constraints, the requirements of the funding bodies and sponsors. This suggests that one potential reason why generalisability is currently a secondary consideration is that regulatory bodies do not explicitly require it. With respect to the time frame of centre selection, it was agreed that early planning is essential in targeting the appropriate locations, but also that centre inclusion is a continuous activity throughout the RCT. This reality is especially relevant in relation to two issues: first, participants agreed that unpredictable developments are very likely throughout the lifetime of a RCT. Second, there has equally been agreement around the fact that centre selection cannot be an entirely rational or entirely subjective effort. In light of these observations, it becomes apparent that any developments in centre selection practices must incorporate trialists' need for flexibility and permanent adjustment to changing conditions throughout the lifetime of the study.

Although the focus groups did not aim to generate consensus on the current or ideal practice, there was spontaneous agreement between the participants in relation to a large number of centre selection considerations. This lends strength to these findings and suggests that they are largely applicable to the wider clinical trials community. The only consideration where explicitly opposing views were expressed referred to approaching centres with or without research experience.

Strengths and limitations

Due to time constraints, only trialists from trials units based at the University of Birmingham were invited to participate. It is, therefore, inevitable that participants' contributions are influenced by the institutional culture to which they belong. The values and practices in other UK clinical trials units may be different. Nevertheless, the breadth of the participants' professional roles, the diverse therapeutic focus of the trials unit that they represented as well as the excellence status of Birmingham Centre for Clinical Trials (382), are arguments towards the relevance of the results.

The number of focus group participants was limited due to practical considerations. Despite invitations to participate having been sent to more than 50 staff at the Birmingham Centre for Clinical Trials, common availability was identified only for ten of them, respectively. Given that the findings of the focus groups would inform the development of the online survey and that the survey was due to be sent out early January 2012, the aim was to conduct the focus groups not later than September 2011 in order to allow sufficient time for data analysis and survey development. It is possible that self-selection occurred and the sample of focus group participants predominantly included professionals with an interest in trials methodology and centre selection in particular. Trial managers were overrepresented in the sample, with four out of ten participants. Still, there was general agreement that trial managers appear to have an important and continuous role in the centre selection process.

7.4. Survey

The principal aim of the survey was to elicit UK trialists' views on the considerations which inform centre selection for RCTs. In particular, it was of interest to explore the role of generalisability concerns for centre selection. The secondary aim was to identify trialists' perspectives on optimal practice.

7.4.1. Methods

The considerations emerging from the systematic review and the focus groups informed an online survey circulated to trialists at the UK Clinical Research Collaborative (UKCRC) registered Clinical Trials Units (CTUs) and NIHR Research Design Services (RDS). The survey had two sections: the first section asked the respondents about the *current practice* of centre selection for RCTs in terms of influential considerations and key professionals involved in the process; and the second section used the same questions to elicit respondents' views about what should constitute *optimal practice* (Appendix 14). The structure of the survey was informed by the main themes of the focus group analysis, as follows: respondents were asked to assume that the minimum centre requirements for participation in the trial were met i.e. access to the study population and required time, staff and facilities for running the RCT; the first two questions asked about preferable and neutral considerations relevant for centre selection, respectively; and the third question asked about the professionals involved in the centre selection decision. These three questions were used both in the current and optimal practice sections of the survey. In addition, the 'current practice' section also inquired about the participants' views on the current role of health economics considerations in centre selection.

For each question concerning centre selection considerations, the participants had to choose a minimum of three and a maximum of five items they considered to be most important for centre selection from a comprehensive list. No explicit ranking was required. All questions had a free text field where participants could input additional information. Prior to distribution the survey was piloted with the focus group participants, who commented on its structure and content.

A secure web-link to the survey was distributed by email to the direct email addresses (not via automated distribution list) of directors and deputy directors of all 48 UKCRC CTUs and ten NIHR RDS, who were invited to complete the questionnaire and forward it to relevant staff within their units i.e. through a snowballing approach (Appendix 15). The CTUs and their directors/deputy directors were identified by accessing the UKCRC CTU website (www.ukcrc-ctu.org.uk). When the (deputy) directors did not have a CTU domain specific email address, their academic email address was used. Relevant staff' explicitly referred to: clinical investigators, trial coordinators/trial managers, statisticians, health economists and any other academic position (e.g. research associate, research fellow). One reminder email was circulated two weeks after the initial distribution. The online survey was distributed on 24th January 2012 and data collection ended on 27th February 2012.

Only the complete responses were included in the analysis, which was performed using STATA 10 software (Stata Corp, College Station TX, US). In addition to descriptive statistics for the response items, a response consistency analysis was conducted to identify which response items were selected for optimal practice but not for ideal practice and vice versa. This made it possible to identify which considerations which were subject to differences between current and optimal practice in terms of perceived importance.

The Science, Technology, Engineering and Mathematics Ethical Review Committee at the University of Birmingham have favourably reviewed this study (Ref. no. ERN_11-1347). Respondents were asked for informed consent on the first page of the survey and before contributing any information. The survey was anonymous: the only personal information items referred to the participants' professional role and their experience (years) in the design and/or conduct of RCTs.

7.4.2. Results

77 responses were received, of which 70 were complete and entered the analysis. One further response was received in April 2012, after the database had been locked, and it was discarded. Trial managers were the best represented professionals (n=21; 30%). Most respondents (n=49; 70%) had been involved in the design and/or conduct of RCTs for more than five years (Table 8.3).

Table 7.3 Survey: profile of survey participants

Characteristic	Respondents (%, n=70)
Professional role	
Clinical investigator	9 (13%)
Statistician	13 (19%)
Trial coordinator	21 (30%)
Health economist	5 (7%)
Clinical trials methodologist	7 (10%)
Epidemiologist	1 (1%)
Other academic position	7 (10%)
Other professionals	7 (10%)
Experience in design/conduct of RCTs	
Less than 2 years	3 (4%)
Between 2 and 5 years	18 (26%)
Between 5 and 10 years	19 (27%)
More than 10 years	30 (43%)

Overview of results

In current practice, the most desirable centre characteristics were: the ability to recruit patients, centre staff displaying interest in the RCT and good communications with the trial office (Table 7.4). Most respondents reported that including a centre in a RCT is influenced by the centre staff's motivation to participate in the RCT (n=52; 74%) and the local research environment i.e. trial fatigue and competing trials (n=48; 69%). Ensuring generalisability in terms of population characteristics and clinical practice were mentioned by 33% (n=23) and 29% (n=20) of respondents, respectively, while 7% (n=5) of them referred to the generalisability of economic evaluation results. The trial coordinator and the chief investigator appear to be the key drivers in the process of centre selection. 26% of respondents reported that health economics considerations have a limited influence in centre selection, while 74% reported no such influence.

In optimal practice, the majority of survey participants indicated the ability to recruit (n=52; 74%) as desirable, followed by ensuring generalisability in terms of clinical practice (n=42; 60%), population characteristics (n=40; 57%) and economic evaluation results (n=32; 46%), respectively. Most respondents indicated that trial-design characteristics e.g. sample size and number of centres required, and centre staff motivation for the RCT should influence centre selection. Trial management group members as a team should ideally drive centre enrolment.

Table 7.4 Survey: current and optimal centre selection for RCTs (n=70)

Survey questions	**Current practice**		**Optimal practice**	
	N	%	N	%
1. Desirable centre characteristics				
Ability to recruit patients	61	87%	52	74%
Understanding RCTs	10	14%	16	23%
Good communication with trial office	37	53%	26	37%
Convenient geographical location	17	24%	3	4%
Having support from local commissioners	16	23%	10	14%
Part of a relevant research network	11	16%	9	13%
Ability to obtain necessary approvals timely	33	47%	25	36%
Showing interest in the RCT	44	63%	28	40%
Computer systems are compatible with the trial centre	4	6%	1	1%
Retains/contributes to generalisability (population characteristics)	23	33%	40	57%
Retains/contributes to generalisability (clinical practice)	20	29%	42	60%
Retains/contributes to generalisability (economic evaluation)	5	7%	32	46%
Centre staff have experience with conducting RCTs	28	40%	23	33%
2. Considerations influencing the process of centre selection				
Centre staff are motivated to participate	52	74%	41	59%
Centre staff know the Chief Investigator	29	41%	4	6%
Geographical setting (rural vs. urban)	8	11%	18	26%
Requirements of funding/regulatory bodies	13	19%	14	20%
State of local research environment	48	69%	24	34%
Recruiting time frame of the RCT	27	39%	31	44%
Budget of the RCT	21	30%	14	20%
Efficiency of local R&D department	26	37%	17	24%
Disease rarity	9	13%	17	24%
Trial-design characteristics	40	57%	52	74%
Patient convenience	6	9%	22	31%
3. Professionals driving the process of centre selection				
Chief Investigator	38	54%	19	27%
Trial coordinator/Trial manager	45	64%	33	47%
Research networks	16	23%	24	34%
Trial statistician	0	0%	1	1%
Trial health economist	1	1%	6	9%
Trial Management Group members as a team	25	36%	41	59%
Data Monitoring Committee members	0	0%	2	3%

Table 7.5 Survey: number of chosen items by question

Survey questions	Number of available items*	Number of items allowed	Current practice	Optimal practice
			Average number of items (SD)	Average number of items (SD)
Desirable centre characteristics	14	Min 3, max 5	4.44 (0.73)	4.42 (0.80)
Considerations influencing the centre selection process	12	Min 3, max 5	4.07 (0.82)	3.72 (0.87)
Professionals driving the centre selection decision	7	Min 1, max 2	1.84 (0.37)	1.87 (0.34)

- Excluding the ‘Other’ free text option, which was available for all questions

Table 7.5 presents the number of items chosen for each question. The respondents appear to have selected comparable number of items for each question across current and optimal practice. There were slightly more items included in current (mean 4.07, SD 0.82) compared to optimal practice (mean 3.72, SD 0.87) for considerations which influence the centre selection process.

Detailed results

The following paragraphs present detailed results for each survey question (Table 7.4). In addition, the results of a response consistency analysis are presented, which describe the extent to which current and optimal practice choices agree at respondent level (Table 7.6).

a) Desirable centre characteristics

Most respondents reported the ability to recruit patients (87%) and displaying interest in the RCT (63%) as characteristics they want to see in centres. compatibility of computer systems with the trial office (6%) and contributing to the generalisability of economic evaluation results (7%) were least reported. Approximately a third of respondents reported an explicit interest in the centre contributing to generalisability in terms of population characteristics (33%) and clinical practice (29%). Suggested characteristics outside the provided list included a track record in recruitment, expertise in the given disease area and staff engagement (Appendix 16).

In ideal practice, the ability to recruit patients was still a leading consideration for most respondents (74%), followed by the generalisability in terms of clinical practice (60%), population characteristics (57%) and economic evaluation results (46%). Few respondents indicated computer systems compatibility (1%) and convenient geographical location (4%).

Table 7.6 Survey: response consistency

Survey questions	**Response consistency (%)**		
	Ideal, NOT current	Consistent	Current, NOT ideal
1. Desirable centre characteristics			
Ability to recruit patients	4%	79%	17%
Understanding RCTs	14%	80%	6%
Good communication with trial office	13%	59%	29%
Convenient geographical location	3%	74%	23%
Having support from local commissioners	4%	83%	13%
Part of a relevant research network	9%	80%	11%
Ability to obtain necessary approvals timely	10%	69%	21%
Showing interest in the RCT	9%	60%	31%
Computer systems are compatible with the trial centre	1%	93%	6%
Retains/contributes to generalisability (population characteristics)	29%	67%	4%
Retains/contributes to generalisability (clinical practice)	34%	63%	3%
Retains/contributes to generalisability (econ. evaluation)	41%	56%	3%
Centre staff have experience with conducting RCTs	11%	70%	19%
2. Considerations influencing the process of centre selection			
Centre staff are motivated to participate	9%	67%	24%
Centre staff know the Chief Investigator	0%	64%	36%
Geographical setting (rural vs. urban)	20%	74%	6%
Requirements of funding/regulatory bodies	14%	73%	13%
State of local research environment	7%	51%	41%
Recruiting time frame of the RCT	23%	60%	17%
Budget of the RCT	13%	61%	26%
Efficiency of local R&D department	16%	80%	4%
Disease rarity	24%	69%	7%
Trial-design characteristics	27%	69%	4%
Patient convenience	6%	90%	4%
3. Professionals driving the process of centre selection			
Chief Investigator	4%	64%	31%
Trial coordinator/Trial manager	3%	77%	20%
Research networks	21%	69%	10%
Trial statistician	1%	99%	0%
Trial health economist	7%	93%	0%
Trial Management Group members as a team	3%	97%	0%
Data Monitoring Committee members	4%	93%	3%

Legend:
'Current, NOT ideal' – item was selected in 'Current practice', but not in 'Ideal practice'
'Ideal, NOT current' – item was selected in 'Ideal practice', but not in 'Current practice'
'Consistent' – item responses are the same in 'Current practice' and 'Ideal practice'

Additional ideal characteristics included the ability to collect resource use data and good working relationships between research staff and local service providers (Appendix 16).

41%, 34% and 29% of respondents who did not indicate generalisability in terms of economic evaluation results, clinical practice and patient population, respectively, as preferred characteristics in current practice did so in optimal practice (Table 7.6). Conversely, 31% of participants indicated ‘showing interest in the RCT’ as relevant in current practice, but not so in ideal practice; 23% did the same for ‘convenient geographical location’. The largest degree of consistent responses was for the compatibility of computer systems with the trial office.

b) Considerations influencing the centre selection process

The majority of trialists indicated staff’s motivation to participate in the RCT (74%) and the state of local research environment (69%) were influential considerations for centre selection in current practice. Few respondents suggested that patient convenience (9%) and the disease rarity (13%) as relevant. Free text responses also referred to the centre relationship with the study CI, the support that the centre receives and whether centre staff perceive the research question as being important to them (Appendix 16).

In ideal practice, trial design characteristics (74%) and staff motivation (59%) were most often seen as important considerations, while the centre’s staffs knowing the Chief Investigator was rarely included in the respondents’ choices (6%). The importance of a meaningful clinical question was mentioned again in the free text comments (Appendix 14). Accounting for trial design characteristics and disease rarity were the considerations that most trialists did not include in current practice, but did so when referring to ideal practice (27% and 24%, respectively). On the other hand, 41% of respondents indicated that the state of the local research environment is currently important, but didn’t include it in ideal practice; 36%

of respondents similarly indicated the centre's staff knowing the Chief Investigator (Table 8.6).

c) Professionals driving the centre selection process

Most respondents indicated that the trial manager (64%) and the Chief Investigator (54%) currently drive centre selection. It was suggested in the free text comments (Appendix 14) that sponsors (especially commercial sponsors) and the trial statistician may also be involved in the process (the latter in relation to cluster RCTs). In ideal practice, the Trial Management Group as a team was seen as the key personnel that should be responsible for centre selection (59%). Several free text responses emphasised the role of local investigators (Appendix 16).

Response consistency analysis suggested an increased role for research networks in ideal practice as opposed to current practice (21%); furthermore, 31% of respondents who indicated the Chief Investigator as a major driver in current practice did not maintain their choice in ideal practice (Table 7.6).

d) Health economics considerations and centre selection

18 respondents reported that health economics considerations influence centre selection decision to a limited extent (26%), while 52 reported no such influence (74%). Free text comments indicated that health economics concerns are usually minor and rarely given separate consideration (Appendix 16).

e) Other comments

The final page of the survey invited the respondents to share any comments or feedback about the survey in a free text field (Appendix 16). Several comments reported that the questions appeared difficult to understand and/or difficult to answer, mainly because they could be approached from a multitude of angles e.g. methodological or pragmatic. A further comment suggested that not defining what the questions meant by 'ideal practice' made answering difficult.

7.4.3. Discussion

Summary of findings

The survey results suggest that considerations such as meeting recruitment targets and having good working relationships with front line investigators appear to drive centre selection for RCTs in current practice. The importance of ensuring generalisability in terms of the population and, more broadly, centre characteristics is acknowledged by trialists and ideally should be more explicitly incorporated in practice than it currently is. The Chief Investigator and Trial Manager are key professionals in the decision-making process, but ideally the process should involve more the TMG as a team. Health economics considerations appear to play a minor role in centre selection and they are incorporated as 'socio-economic characteristics' of the centres.

Interpretation of findings

The survey was divided in two identical and distinct sections i.e. current practice and ideal practice, in order to investigate further the tension which became apparent during the focus groups. The survey results are consistent with the focus groups findings in highlighting this discrepancy between what trialists currently do and what they think they should do in ideal practice. In addition, the results specify better where this tension lies. As such, there are aspects of centre selection which trialists perceive should be different in ideal practice: generalisability considerations and patient convenience should be incorporated more; the role of the TMG and the trial team in the decision should be more prominent; trial fatigue and previous knowledge of centre staff should play a lesser role (Figure 7.2). Conversely, there also seem to be considerations which currently receive the attention they deserve; these include the ability to recruit successfully, identifying highly motivated centres and ensuring good communication.

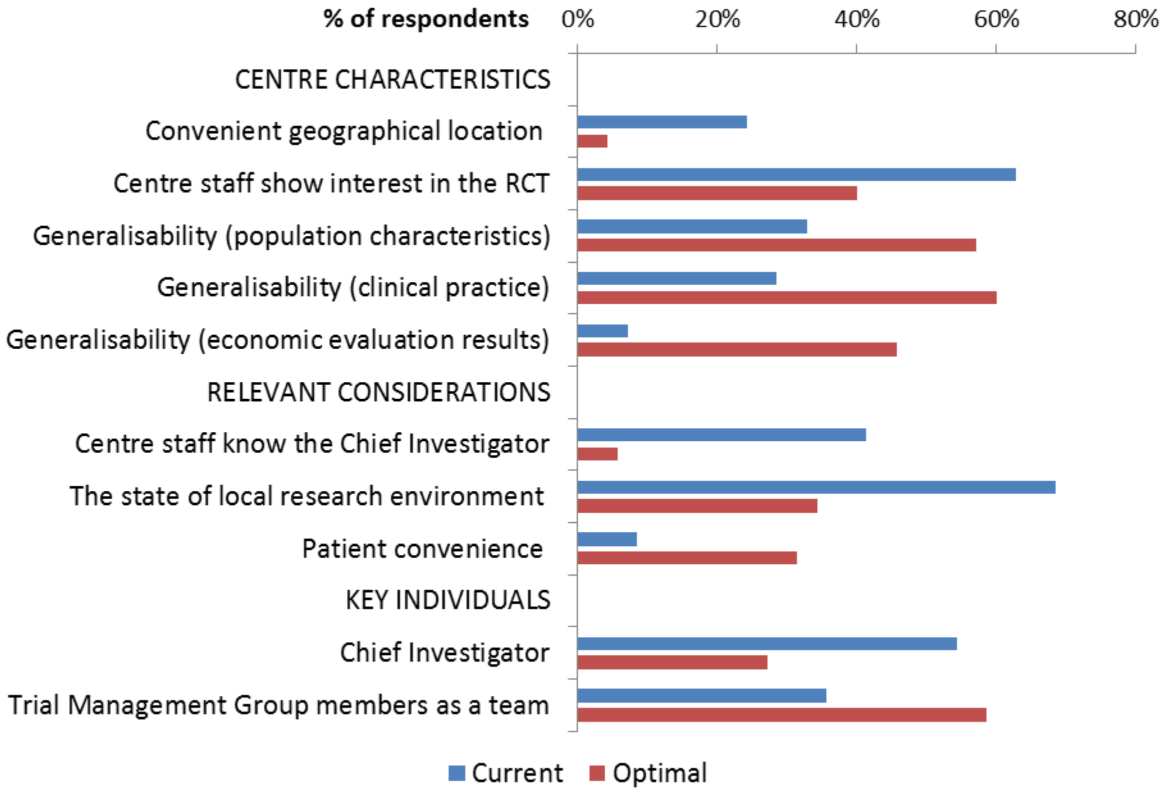


Figure 7.2 Survey: discrepancies between the current and optimal practice of centre selection for RCTs.

Note: Only survey items with a difference larger than 20% of responses between current and optimal practice are displayed.

The survey findings on the importance of generalisability are compatible with those of the meta-summary presented in section 7.2: approximately 30% of survey respondents reported that generalisability considerations are currently taken into account; similarly, 24% of RCT protocols included at least one consideration aimed at ensuring a diverse or representative sample of participating centres. Although the two metrics i.e. the survey response frequency and the met-summary effect size, were not designed to be directly comparable and the aim of this research was not quantitative *per se*, it is reassuring that they appear to illustrate the same reality: the majority of RCTs do not explicitly account for generalisability in centre selection. It must be noted, of course, that given the UK focus of this research, it is very likely that a large proportion of the survey respondents may also have been responsible for designing and writing the protocols included in the systematic review. This argument is particularly notable when considering the high proportion of experienced trialists (more than 5 years) in the survey sample (Table 7.3). However, it does not limit the validity of the findings in the absence of any indication that either the included protocols or the sample of trialists were unrepresentative.

Strengths and limitations

The survey asked the respondents to choose between a minimum and a maximum number of considerations. The entire list of considerations was not left open for choice because all the options were relevant for trial design, at least for methodological purposes, and it is likely that very few options would have been left out. A full ranking exercise was also ruled out due to the cognitive burden, as two questions had more than 12 considerations each. In the absence of clear guidance on such a matter in the survey literature, the choice was to allow approximately a third of the available considerations (8, 12 and 14, respectively), thus obtaining three to five options open for choice, without explicit ranking.

The sample size is a limitation and ideally more respondents would have answered the survey. However, this could not be controlled because of the heterogeneous websites of the CTUs and RDS, such that individual contacted details were not always available. This context led to relying on unit directors and deputy directors to distribute the survey link to the indicated professionals. This snowballing approach was the main reason why a survey response rate could not be calculated. This limitation can be partly justified by the lack of prior knowledge about the process of centre selection for RCTs, so I was interested in the views of a wide range of trialists. With a more specific question the survey sample could have focused on fewer professional roles, as did, for example, McPherson *et al.*, who recently published the results of a survey where they inquired statisticians in UK CTUs on their approach to randomization (383). The limited sample size was the main reason why subgroup analyses by age and professional role were not performed.

In psychology, priming refers to previous experience of a stimulus influencing later responses that stimulus (384). Although a priming effect is possible when comparing current and ideal practice in this survey, the results are not consistent with such an effect: on the one hand, the centre's ability to recruit patients and staff's motivation to participate in the RCT were the most prominent both in current and optimal practice, which testifies their importance for trialists. On the other hand, the largest relative increase in importance from current to optimal practice was for the three generalisability items. The invitation email (Appendix 15) and the survey (Appendix 14) did not mention generalisability as a research interest.

7.5. Discussion of the mixed methods study

A mixed methods approach was used to explore centre selection considerations in current and ideal practice of conducting RCTs in the UK. Mixed methods have been used before in the context of trial methodology, more often with a focus on informing the design or evaluation of particular studies (385-388). For instance, Brady *et al.* used a combination of medical records review, semi-structured interviews with staff and a validated questionnaire alongside a feasibility study to inform the design of a definitive trial of a complex oral health care intervention (386). However, mixed methods have also been used to address broad methodological questions. Kaur *et al.* recently developed an online survey on barriers and facilitators to RCT recruitment using an approach similar to the one described in this Chapter: first they conducted a literature review and identified a list of relevant factors which led to the initial version of the survey; and afterwards the survey underwent a succession of piloting stages until the final version was agreed upon (389). Hamm *et al.* used an online survey of trialists whose results informed the topic guide of semi-structured interviews to identify the barriers in conducting unbiased trials in paediatric care in Canada (390).

The findings of the three methods used in this study are generally in agreement. As such, there was significant overlap between the centre selection considerations identified in the systematic review of protocols and focus group discussions with trialists. The survey results confirmed the reported tension between current and ideal practice that became apparent during the focus groups. Furthermore, the meta-summary effect size and survey response frequency for the generalisability items in current practice were compatible in suggesting that the large majority of RCTs do not currently recruit centres with generalisability in mind.

Pragmatic considerations such as recruitment and communication seem to drive the centre selection process in current practice. In ideal practice, however, trialists acknowledged concerns such as generalisability of results and patient convenience. There appears, thus, to be a tension between what trialists report as currently being done and what they think ought to be done. Ensuring generalisability is one of the objects of this tension in the sense that its importance is acknowledged, but other considerations currently take precedence.

Two sets of explanations are possible. First, it may be that generalisability is not currently a prime consideration because it rightfully isn't an overarching concern. As one of the free text commentaries suggested (Appendix 16), there may be no substantial difference between recruiting and non-recruiting centres, which would make the issue of representativeness in centre selection rather trivial. However, there is little literature available to substantiate this claim and the little available evidence suggests that most evaluative research takes place in university centres (100). The two focus groups made apparent trialists' concerns that some centres are more suited to recruiting than others and that it is important to approach the 'right' ones. This suggests that centres are indeed different; therefore the selection process can make a difference both to the RCT's completion and its findings. Furthermore, approximately 75% of the RCTs included in the systematic review did not explicitly account for generalisability when including centres and only two studies used a random process. More often than not, the RCT protocols included in the review included statements such as "We recruited from a representative sample of centres [...]" without any other details on how the investigators assessed representativeness and what were the characteristics of their reference sample (for example ID16, ID24, ID76 and ID87). In the light of these issues, it can be concluded that there is still insufficient evidence to claim that generalisability should not be a prime concern on the basis of no significant differences between recruiting and non-recruiting centres.

Second, the RCT funders' interest in generalisability may yet not be compelling enough for trialists to modify the practice in this direction. As pointed out in sub-section 7.2.3, the structure of the RCT protocols included in the systematic review was highly heterogeneous and there were no set headings on either centre characteristics or centre selection processes, leaving the reporting of such considerations at investigators' discretion. The funders' lack of explicit interest towards these issues could be explained by the absence of evidence on why generalisability across locations for within-country studies is important.

The importance of centres' willingness to participate is a particularly interesting finding of this research. While it emerged as a relevant consideration in the meta-summary, focus group discussions went further and suggested that willingness to participate is essential for centre selection. The survey results confirmed the importance of local staff showing interest in the trial and of their motivation in current and ideal practice. When corroborated, these findings have two implications: first, trialists perceive motivation to participate as key for successful trial completion; second, and most importantly, centres have different levels of engagement, which under specific conditions makes some more desirable than others.

Variation in willingness to participate can have multiple causes: for example, not being ready to randomise against or perform a particular intervention suggests that some clinicians are not in equipoise, a key ethical requirement of RCTs (73). If anything, this can be interpreted as a healthy concern; if the large majority clinicians and patients agreed on the relative merits of one intervention against the other prior to obtaining evidence, there would be no RCTs at all because nobody would agree to (be) randomise(d) at 50:50 probability against an inferior intervention (391). Furthermore, the pragmatic barriers to research participation as perceived by clinicians have been documented and include: time constraints, lack of training, concerns about the research impact on the doctor-patient relationship and answering an interesting clinical question (392). Nevertheless, generic solutions for health care organisations to address such shortcomings have been suggested (393) (e.g. selecting research questions that are of interest to clinicians, setting a transparent reimbursement schedule for research tasks and provide technological support to practices) and innovative business models to guide the design and conduct of RCT processes have been proposed (394). Another source of controversy may be that some centres are highly sought after in the research community and have limited capacity to take on new projects; this explanation is supported by trialists' perception that the characteristics of the research environment should be less relevant for centre selection in ideal practice than it currently is (Figure 7.2). It remains unclear whether current research activities surpass research capacity or research is simply concentrated around selected centres while others are idle. If the latter is the case, it may constitute the foundation of a self-enforcing limitation, as the centre selection process is influenced, to some extent, by previous research experience. If such centres are currently left out of research, they are likely to be left out in the future, too. Finally, an extreme explanation of variation in research uptake would be the sheer refusal to take part in research in the absence of any capacity constraints. In that respect, the obligation to participate in research has been firmly established in the bioethical literature (395, 396) and, thus, such an attitude cannot be justified. Future research on the validity and relative extent of these considerations may guide research commissioners' efforts to mitigate them.

The finding that pragmatism takes precedence before generalisability may at first seem obvious, but this rather enforces its importance because this is, to my knowledge, the first time that the tension between pragmatism and generalisability is explored based on evidence from RCT protocols and trialists. The focus of this research was on generalisability and how its role is perceived among the other relevant trial considerations. However, this is not to say that generalisability and pragmatism are in direct competition. First, the results do not suggest that pragmatism should be downplayed, but rather report that trialists' perceptions appear to indicate that there is clearly room for addressing generalisability more conscientiously. Second, several potential reasons why generalisability is currently regarded as less important than pragmatism were discussed; none of them implies that trialists face an informed choice between the two, mostly because there is currently little guidance towards incorporating generalisability.

There is evidence in the literature on the positive impact of guidelines on the quality improvement of clinical trial design and reporting (115). However, few guidelines explicitly refer to representativeness and centre selection. On the design side, the SPIRIT 2013 statement contains a 33-item checklist which acts as a guideline for the minimal content of a RCT protocol (15). The characteristics of participating centres are required under the 'Study setting' and 'Eligibility criteria' items, but there is no explicit requirement to address generalisability in selection or in data analysis (397). On the reporting side, only one CONSORT extension requires explicit reporting of the extent to which participating centres and practitioners are representative to wider settings (116). Thus, it appears that generalisability currently receives insufficient attention in trial design and analysis, which may explain why trialists currently regard it as a secondary consideration. It must be acknowledged that a stronger focus on generalisability in widely recognised guidelines is no guarantee of improved practice: Zarin *et al.*'s analysis of the ClinicalTrials.gov records revealed that much less controversial methodological decisions, such as the selection of a single primary outcome, are sub-optimally implemented (398). However, explicitly incorporating generalisability in trial guidelines could contribute, in time, to the conduct of increasingly valid and relevant trials.

This mixed methods study focused on UK publicly-funded RCTs. The generalisability of the findings to privately-funded trials and to other countries, respectively, is unknown and can be viewed as a limitation. Nevertheless, it is likely that the interventions evaluated in industry-led RCTs are often drug therapies as opposed to complex interventions and, as such, patient-level characteristics are more important than centre-level ones. In terms of the international scope of the results, there is little evidence of centre selection practices in other settings to enable informed comparisons. Only the replication of various components of this research in other research settings can add an international perspective to these findings.

7.6. Conclusion

The rationale for centre selection appears to be underreported in RCT protocols in the UK. Enrolling a representative sample of recruiting centres, which can ensure or contribute towards the generalisability of trial findings, is currently a secondary consideration in centre selection. Pragmatic considerations such as meeting recruitment targets and ensuring good communication take precedence. Trialists acknowledge the importance of generalisability and would ideally incorporate it more in the centre selection process.

Generalisability across settings is currently insufficiently present in major guidelines on conducting and reporting research, which may explain the current state of affairs. More importantly, there is a need for evidence as to whether the sample of centres participating in a RCT can influence its clinical and economic results. In the next Chapter a method that can address, to some extent, this need is proposed.

CHAPTER 8. ENHANCING THE GENERALISABILITY OF TRIAL-BASED ECONOMIC EVALUATIONS USING A GENERALISABILITY INDEX (GIX)

The conclusion of the previous Chapter was that the inclusion of centres in RCTs with a view to ensuring generalisability is currently acknowledged as being important but is rarely implemented. The impact of this suboptimal practice on the generalisability to the jurisdiction level of the results of trial-based economic evaluations is unknown. In this Chapter a real-world example illustrating this impact is presented and a novel methodology is proposed that can assess and potentially enhance the generalisability of trial results. The cornerstone of this methodology is the Generalisability index (Gix), which is a measure of representativeness and can be computed at centre- and trial-level. The application of the Gix will be demonstrated using a case study drawing on the ROSSINI trial, which was presented at length in Chapter 5.

8.1. A real-world example

The ROSSINI trial, which was presented at length in Chapter 5, will be used to illustrate the potential impact of the sample of recruiting centres on trial results. ROSSINI recruited patients from 21 UK hospitals and randomised 760 patients. The embedded economic evaluation took an NHS perspective and evaluated the cost-utility of the device compared to standard care over a 30-day post-surgery time horizon.

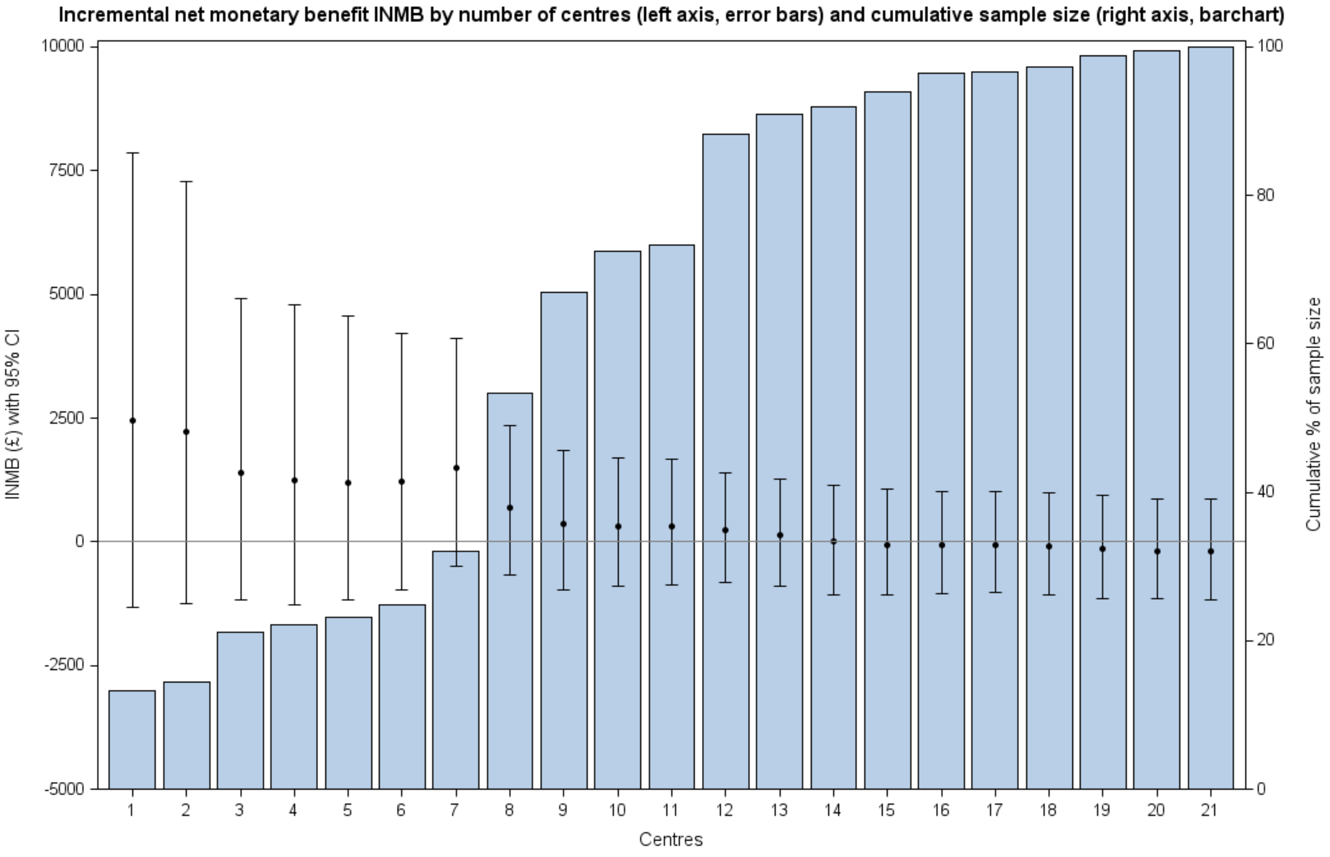
Method

The working hypothesis was that different samples of participating centres yield different overall cost-effectiveness estimates. In order to test the hypothesis, the 21 recruiting centres in ROSSINI were treated as the complete population of centres and standard cost-effectiveness methods were applied on incremental sub-samples of 1, 2, 3 ... 21 centres. Centres were considered in the chronological order in which they started contributing patients; data from all patients in a particular centre were analysed. For example, the third sub-sample included all patients recruited from the first three recruiting centres; the seventh sub-sample included all patients recruited from the first seven recruiting centres and so on. The 21st sub-sample included all patients and is equivalent to the trial-wide analysis.

Results

The error bars in Figure 8.1 depict the point estimate of the incremental net monetary benefit (55) for each incremental sub-sample of participating centres together with the 95% BCa confidence intervals. The incremental net monetary benefit (INMB) was calculated using a willingness to pay threshold of £20,000, in accordance with NICE guidance (39). Based on information from patients recruited in the first 13 centres (~ 90% sample size), it is apparent that the point estimate of the INMB is positive, suggesting that the intervention may be cost-effective, only to eventually become slightly negative, suggesting that the intervention is not cost-effective when complete trial data were analysed. The width of the confidence interval gradually decreases with sample size, as expected.

Figure 8.1 Illustration of the changing incremental net monetary benefit estimate in ROSSINI as recruitment progressed



Note 1: Centres have been considered in the temporal sequence that they started contributing patients in.For example, the INMB for 3 centres refers to the economic evaluation informed by data from all patients recruited at the first 3 recruiting centres.
INMB at 21 centres represents trial-wide results.
Note 2: INMB was calculated for £ 20,000 per QALY. A positive INMB suggests that the intervention is cost-effective, while a negative INMB suggests the contrary.

Discussion

This example suggests not only that cost-effectiveness estimates differ among centres within the same country, but, more importantly, that the sample of participating centres can influence the cost-effectiveness decision. Had a couple of other major recruiters been included, trial-wide results could have been quite different. While there is constantly considerable uncertainty around the INMB, reflected in the width of the confidence intervals, the changing point estimates and upper/lower confidence bounds impact the shape of cost-effectiveness acceptability curves (CEACs) and potentially the cost-effectiveness acceptability frontier (CEAF). For interventions which are borderline cost-(in) effective, the sample of centres may change the decision makers' belief in the cost-effectiveness likelihood of an intervention. It must be acknowledged that in the particular case of ROSSINI recruiting from a different sample of centres is unlikely to have had a major impact on the final results because of the lack of clinical effectiveness and great uncertainty in clinical and economic results. However, it appears that the sample of participating centres introduces variation in trial results and the magnitude of this variation deserves further exploration.

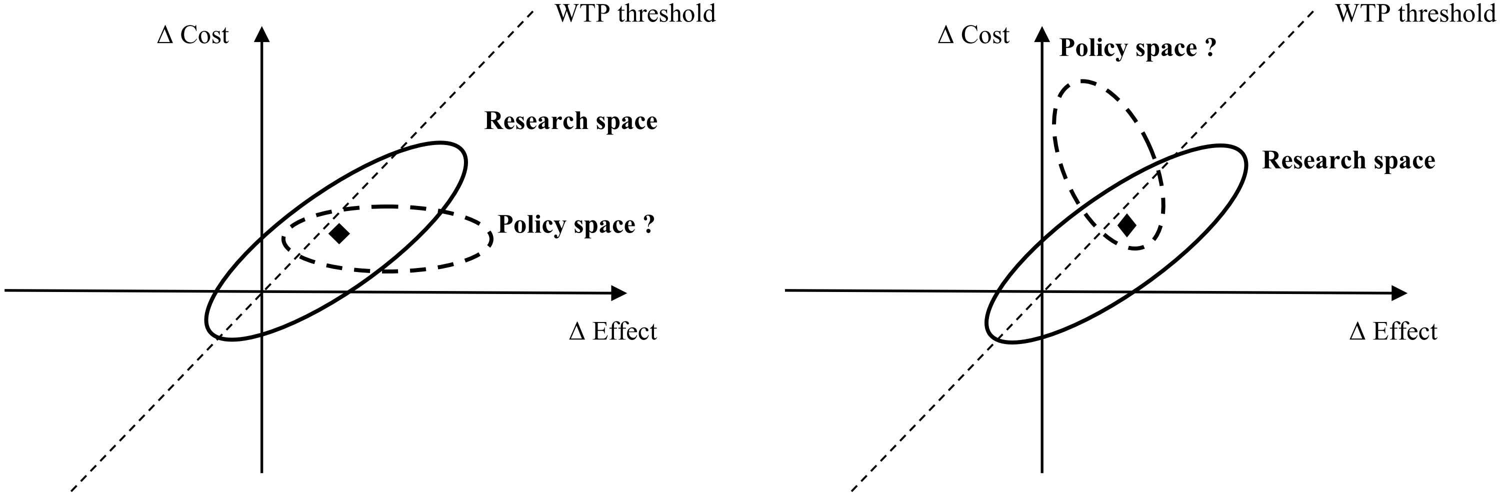
8.2. Suggested way forward and proposed plan

The previous section argued that the sample of participating centres may influence the trial-based cost-effectiveness findings. This section introduces the research plan that aims to address the current lack of knowledge in relation to the impact of recruiting centres and generalisability.

The first point to be made stems from the fact that decision makers often have to make nationwide decisions informed by research findings from a sample of centres. Thus two conceptual decision spaces can be described: the policy space, defined as all centres in the jurisdiction that have the potential to use the intervention being evaluated; and the research space, defined as all the centres participating in the given RCT. The overarching problem is the difficulty to quantify the overlap between these two types of decision spaces.

Figure 8.2 depicts two hypothetical scenarios where the research space is described by the cost-effectiveness point estimate from the RCT and the associated 95% confidence ellipse derived from a bootstrapping exercise. It must be acknowledged that current methods of expressing uncertainty around the cost-effectiveness estimator, such as the cost-effectiveness plane, CEACs and CEAFs, compare *types* of policy scenarios and not *real-world distributions* of policy scenarios. The same applies for decision models informed by RCT findings, which are nowadays instrumental to more and more health technology assessments. Base-case and sensitivity analyses influence the point estimate of the cost-effectiveness metric and the uncertainty around it, but these quantities can only refer to one policy scenario at a time. In reality, policy makers are interested in evaluations of real-world distributions of scenarios (e.g. distributions of centres and patient populations) which reflect the policy contexts they face. The available methods cannot address the relationship between the two decision spaces and the only reasonable assumption is that the policy space is likely to contain the trial-based point estimate. However, a decision informed by the research space may or may not apply to the policy space (Figures 8.2a and 8.2b). The distinction between the two types of decision spaces is important because it can guide researchers towards providing decision makers with the estimate they really want, i.e. 'the cost-effectiveness of an intervention when implemented across a specifiable (real) population of scenarios (centres)' as opposed to 'the cost-effectiveness of an intervention when implemented in a perfectly homogeneous population of scenarios (centres)'.

Figure 8.2 The research space and the policy space



Panel A: Hypothetical scenario where the centres participating in the RCT are representative of centres within the jurisdiction which have the potential to use the intervention of interest. The extent and direction of the overlap in relation to the acceptable WTP threshold suggest that decisions based on the cost-effectiveness estimate from the research space can be extrapolated to the policy space. In other words, it is likely that the intervention will still be cost-effective when implemented in other centres in the jurisdiction that were not part of the RCT.

Panel B: Hypothetical scenario where the centres participating in the RCT are not representative of centres within the jurisdiction which have the potential to use the intervention. The extent and direction of the overlap in relation to the acceptable WTP threshold may lead to different policy decisions. In other words, the results of the research space cannot be so obviously applied to the policy space because in a significant number of non-participating centres the intervention may not be cost-effective.

Building on this conceptual distinction and on the limitations of current generalisability methods (sub-section 1.3.3), a legitimate research aim appears to be investigating how the sample of centres included in a given RCT influence the trial-wide cost-effectiveness results. First, there is a need for reliable evidence on how centres are currently included in RCTs. This has been addressed and discussed at large in Chapter 7: the results of the mixed-methods study suggested that the majority of UK publicly-funded RCTs do not explicitly aim to recruit from a representative sample of centres (399). Furthermore, it emerged from focus groups (section 7.3) and a survey of UK trialists (section 7.4) that ensuring generalisability should be considered when approaching trial centres, but pragmatic considerations, such as the proximity to trial office and a history of successful recruitment, currently take precedence.

Second, there is a need to operationalise 'generalisability'; one way to achieve this is to propose a centre-level generalisability index which measures the extent to which a given centre is representative to a larger population of centres. Such an index would allow trialists to evaluate and ensure representativeness at trial design stage.

This research may be beneficial from a multitude of angles. First, empirical evidence will test the assumption that centres are representative for the jurisdictions they represent and thus potentially warrant the validity of current adjustment methods. Second, a methodology based on the generalisability index can be envisaged to assist retrospective modelling techniques in assessing the external validity of the trial *as it was designed* and in pursuing more and more precise cost-effectiveness estimates, to a reasonable level. Third, an advance in clinical trials recruitment would be made possible by providing a method to identify at the trial design stage the centres which are of more interest than others in terms of extrapolating economic evaluation results. Such a method may also inspire the centre selection process in multi-centre studies that do not necessarily have an economic evaluation component.

8.3. Methods

The working hypothesis is that current methods of centre selection for RCTs result in unrepresentative samples of centres, which may lead to biased estimates of both effectiveness and cost-effectiveness. This Chapter proposes a novel quantitative measure of representativeness called the generalisability index (Gix). The Gix measures the extent to which a given centre and a given trial are representative of the jurisdiction to which they belong. The aim is to establish whether a measure of generalisability, such as the Gix, is associated with the extent to which a trial's results are generalisable to the jurisdiction where it recruited from.

The research has two objectives: first, to define a conceptual framework for the Gix and to consider how it can be applied at the centre, RCT and jurisdiction level. The proposed conceptual framework is illustrated using a real-world multi-centre RCT, namely the ROSSINI trial, which was presented at length in Chapter 5. The second objective is to investigate, by way of a simulation study, how biases in the treatment effect and cost-effectiveness estimates vary depending on RCT-level measures of representativeness.

This section will first provide an overview of the methodological approach and then will discuss the steps in detail. The proposed Gix is a measure of representativeness which can be defined at two levels:

- at the centre-level, the centre-Gix (Gix$_c$) measures the extent to which a given centre is representative of its jurisdiction (e.g. NHS England and Wales) according to several relevant characteristics.
- at the RCT-level, the trial-Gix (Gix$_t$) measures the extent to which the sample of centres and corresponding patients enrolled in a given RCT are representative of the jurisdiction-wide distribution of centres and patient throughput.

The RCT-level and centre-level Gix indices are compared to the jurisdiction-wide distribution of the Gix$_c$, summarised by the Gix$_j$. The purpose of introducing these metrics is to assess the extent to which trial recruitment, both at centre and patient level, is representative of the jurisdiction it recruits from. There is a conceptual distinction between the three types of metrics: the centre- and trial-Gix are measures of representativeness at centre and study level, while Gix$_j$ is a metric that the study-level representativeness is judged against. It must be noted that the Gix is currently designed to be meaningful in the context of a specific research question or therapeutic area.

The conceptual outline of applying these concepts to evaluate the generalisability of trial results is the following:

i. First, the ‘jurisdiction’ is defined by identifying the relevant centres i.e. all centres where the intervention under investigation can be applied and is expected to be implemented if found to be clinically and cost-effective.

ii. Second, the dimensions of generalisability are decided upon and inform the calculation of the Gix_c for all centres in the jurisdiction. A jurisdiction-wide distribution of Gix_c values is generated, which is summarised (Gix_j) using a metric such as the median, geometric mean or weighted (patient throughput) mean.

iii. Third, considering a RCT identified by participating centres and their respective patient recruitment, the Gix_t is calculated as the weighted (patient recruitment) mean of the Gix_c values of participating centres. The assumption is that trial recruitment is equivalent to patient throughput as a measure of patient volume for reasons that will become obvious in sub-section 8.3.5 below.

iv. Finally, the Gix_t is compared to the jurisdiction-wide distribution of Gix_c, more specifically to its summary measure Gix_j, by calculating the standardised mean difference.

The following sub-sections present the steps above in detail with the exception of the first step i.e. defining the jurisdiction, which is assumed to be straightforward.

8.3.1. The dimensions of the Gix

As the Gix is a measure of representativeness, appropriate measures or indicators of representativeness must be determined. Two large systematic reviews identified a large number of factors which may influence the generalisability of economic evaluation results (127, 136). In addition to these reviews, a pragmatic literature search was conducted to identify further centre-level characteristics which were investigated in relation to between-centre variation in health care costs and outcomes. Given the ROSSINI trial was to be used as a case study a pragmatic decision was made to focus on those factors which may affect hospital care.

The following characteristics were included in the Gix, based on their potential to influence the effectiveness and cost-effectiveness results of the assumed RCT: centre size/capacity; teaching function; economic environment; cost performance; and degree of specialisation (Table 9.1). These are termed the *dimensions* of the Gix (Box 9.1) and the evidence around them is discussed in more detail below:

Centre size/capacity. Provider capacity has often been investigated in relation to health care costs and outcomes. There is evidence from a large number of studies that larger providers, both physicians and hospitals, are associated with better health outcomes (97). In terms of cost, studies from China (400) and US (401-404) as well as several multi-national investigations (405, 406) suggested that hospital size may be (usually positively) associated with health care cost. However, the issue remains controversial and is unlikely to extend to all clinical specialties as studies in Italy (407), France (408) and US (409, 410) have not found any significant effect of hospital capacity.

Teaching status. Teaching hospitals appear to deliver superior health outcomes than non-teaching hospitals (138, 143). There is also evidence of an association between teaching status and health care costs (400, 411-414). Similarly to capacity, controversy remains as there are also studies which did not identify such a relationship (173, 415).

Specialisation. Both comprehensive systematic reviews (127, 136) identified the provider's experience, skills, training and learning curve characteristics as potential factors that affect generalisability. Daidone and D'Amico found that specialisation is negatively associated with inefficiency in Italian hospitals and proposed a hospital-level specialisation index bounded at 0 and 1 which quantifies the proportion of patient episodes of a particularly type seen in a given hospital (416). In their recently published study of 153 English hospitals, Gutacker *et al.* found that specialisation was positively associated with superior HRQoL outcomes following hip replacement, but not so for knee replacement, groin hernia repairs or varicose veins surgery (417).

Market environment. It has been argued that hospital reimbursement arrangements incorporate additional payments for providers facing higher costs for reasons outside their control (418). For this purpose, the Department of Health uses the Market Forces Factor (MFF), a metric which accounts for three main categories of capital costs: labour (non-medical staff and medical staff), land and buildings (419). The MFF is set to average at 1.0 so that organisations with an index higher than 1.0 face input costs higher than the average, while organisations with an index lower than 1.0 face input costs lower than the average provider. Kristensen *et al.* incorporated the MFF in their analysis of cost variation in diabetes care across English hospitals and found a significant positive association between the index and inpatient costs, which explained the largest amount of cost variation (420). Laudicella *et al.* found similar results when looking at costs across English obstetrics departments (173).

Cost performance. The MFF, which is ultimately a measure of provider exposure to environmental factors, is used to calculate a metric of provider performance, namely the Reference Cost Index (RCI). RCI is a measure of relative efficiency across NHS organisations and shows the relative cost of a given NHS trust’s casemix compared to the cost of delivering that casemix at national average cost (421). Providers with costs equal to the national average score 100; higher cost providers score above 100 and lower cost providers score below 100.

Other factors that were considered for inclusion in the Gix were: staff mix; staff specialisation; and urban/rural setting. They were not pursued due to lack of readily available and interpretable data. Such a development may be the object of future research. Nevertheless, the current choice of dimensions appears to be reasonable in terms of relevance to the English NHS context.

Table 8.1 Dimensions of the centre-Gix

Dimension	Operationalised as	Data source
Size/capacity	Number of beds	NHS The Information Centre – Hospital Estates and Facilities Statistics 2011/2012 (422)
Teaching function	Teaching status	NHS The Information Centre – Hospital Estates and Facilities Statistics 2011/2012 (422)
Economic environment	Market Forces Factor (MFF)	Department of Health - Reference Cost Index 2011/2012 (423)
Cost efficiency	Reference Cost Index (RCI)	Department of Health - Reference Cost Index 2011/2012 (423)
Specialisation	% of relevant finished consultant episodes (FCEs) from total FCEs in one calendar year	Health & Social Care Information Centre - Hospital Episode Statistics 2011/2012 (424)

Box 8.1 The dimensions included in the centre-level Gix

1. *Centre size/capacity* – a measure of volume, thereby reflecting potential economies of scale;
2. *Teaching status* – a measure of technical expertise;
3. *Specialisation* – a measure of concentration, also reflecting potential economies of scale and learning curve effects;
4. *Market environment* – a measure of the organisation’s external environment;
5. *Cost performance* – a measure of the organisation’s efficiency.

8.3.2. The centre-level Gix

As outlined above, the aim of the centre-Gix is to quantify the extent to which a given centre is representative of its jurisdiction. More specifically, the centre-Gix measures how representative a centre is compared to all the other centres in the jurisdiction of interest where the given intervention could be implemented.

Once the relevant dimensions are identified, constructing the centre-Gix entails two steps: 1) quantifying how representative a given centre is of the jurisdiction, according to each of the Gix dimensions; and 2) aggregating these measures of representativeness across dimensions to obtain a centre-Gix. The steps are detailed below.

To quantify representativeness, data for the included dimensions are collected for all the centres in the jurisdiction. For each dimension, centres are dichotomised into those which fall into the middle 80 percentile range (10th to 90th percentile) and those falling outside of it. This range was chosen arbitrarily as it was judged that 80% of observations would reasonably describe ‘commonness’. The influence of this assumption on the results is investigated further in sensitivity analyses (sub-section 8.4.4).

For each dimension reflected by a continuous variable, a score of 1 is assigned to centres lying in this range and 0 otherwise. For dimensions reflected by dichotomous variables (e.g. teaching status – teaching hospital or non-teaching hospital), 1 is assigned to centres in the predominant category (e.g. non-teaching) and 0 to the other. For each dimension, a score of 1 thus denotes a centre which for that dimension is fairly typical, while 0 denotes atypical centres in that dimension. The dichotomised score s_i *(a)* for a continuous dimension *a* can, thus, be defined as:

$$\mathrm{s_i(a)} = \begin{array}{c} \mathbf{1, if\ P_{10}(a) \leq a_i \leq\ P_{90}(a)} \\ \mathbf{0, if\ P_{10}(a) > a_i\ or\ a_i <\ P_{90}(a)} \end{array} \text{(8.1),}$$

where:

$s_i(a)$ – centre-level dichotomised score on dimension *a* for centre *i*;

a_i – centre-level raw value on dimension *a* for centre *i*; and

$P_{10}(a)$, $P_{90}(a)$ – 10th and 90th percentiles for dimension *a*.

The dichotomised score *s_i (a)* for a dichotomous dimension *a* can be defined as:

$$s_i(a) = \begin{cases} \mathbf{0, if\ p(a_i) \leq 0.5} \\ \mathbf{1, if\ p(a_i) > 0.5} \end{cases} \quad (8.2),$$

where:

$p(a_i)$ – proportion of centre-level raw value a_i in the total number of observations for dichotomous dimension *a*.

The measures of representativeness for the dimensions are aggregated by summation to obtain a centre-level measure of representativeness. The resulting centre-Gix takes discrete values between 0 (outlier, most uncommon for all five dimensions) and D, where D is the number of dimensions considered (the centre is common across all D dimensions).

Thus, the formula for the centre-Gix is:

$$\mathbf{Gix_{c_i}} = \sum_{j=1}^{D} \mathbf{s_i(a_j)} \ (8.3),$$

where:

Gix_{c_i} – the centre-Gix for centre *i*;

D – the number of centre-level dimensions in the Gix (in the base case above, D=5); and

$s_j(\boldsymbol{a_j})$ – centre-level dichotomised score (1 or 0) for dimension $\boldsymbol{a_j}$, derived from equations (8.1) or (8.2), as appropriate.

Equation (8.3) can be extended to incorporate differential weightings across the included dimensions (equation 8.4). Such weightings may reflect the relative influence of the dimensions on clinical and cost-effectiveness results as indicated by the available evidence, expert opinion or belief. For example, if capacity is thought to correlate stronger with costs than teaching status, their contribution to the index can reflect this by assigning a larger weight w_i to capacity.

$$\mathbf{Gix}_{c_i} = \sum_{j=1}^{D} s_i(a_j) w_i(a_j) \text{ (8.4),}$$

where:

$w_i(\boldsymbol{a_j})$ – the weight of dimension $\boldsymbol{a_j}$, subject to $\sum_{i=1}^{D} \mathbf{w_i}(\mathbf{a_j}) = 1$; and

D – the number of centre-level dimensions in the Gix.

The unweighted centre-Gix takes discrete values from 0 to D, where D is the number of dimensions incorporated in the Gix. In this case five dimensions have been considered, so the centre-Gix ranges from 0 to 5. Centres with high Gix_c values (close to 5) can be considered ‘common’ across most of the dimensions when compared with the rest of the centres in the jurisdiction; conversely, centres with low Gix_c values (positive and close to 0) are outliers across most dimensions and can, thus, be considered ‘less common’.

8.3.3. The trial-level Gix

The centre-level Gix outlined above can be used to compute a trial-level Gix (Gix_t), which measures the extent to which a given RCT recruits in a representative manner from all the available centres. A representative recruitment might be measured straightforwardly in terms of patient volumes i.e. recruiting trial participants across centres so as to reflect the patient throughput across jurisdiction centres in current clinical practice. Alternatively, representative recruitment might be measured in terms of patient case-mix, which also accounts for the complexity of each case.

For the base case of this analysis, patient numbers were considered a measure of recruitment representativeness. As such, the trial-Gix is calculated as the weighted mean of the participating centres' $\mathrm{Gix_c}$ values, where the weights q_i are the proportions of patients recruited from each centre (equation 8.5).

$$\mathbf{Gix_t} = \sum_{i=1}^{n} \mathbf{Gix_{c_i}}\, \mathbf{q_i} \text{ (8.5),}$$

where:

$\mathrm{Gix_t}$ – the trial-Gix;

n – the number of centres participating in the trial;

$\mathbf{Gix_{c_i}}$ – the centre-Gix for centre *i*; and

$\mathrm{q_i}$ – the proportion of patients recruited from centre *i* relative to the trial sample size.

The trial-Gix is bounded by the minimum and maximum values of the centre-Gix i.e. theoretically by 0 and D. As such, in the base case it ranges from 0 to 5, where values close to 0 denote trials which recruit predominantly from centres with a low centre-Gix, and values close to 5 denote trials which recruit most patients from centres with a high centre-Gix. $\mathrm{Gix_t}$ is a descriptive measure of the characteristics of a trial's recruitment, but it cannot inform on the extent to which trial recruitment is a good reflection of the clinical practice landscape at jurisdiction level. To enable such an assessment, a jurisdiction-wide measure of representativeness is necessary.

8.3.4. The jurisdiction-level Gix

The jurisdiction-Gix ($\mathrm{Gix_j}$) summarises the distribution of the $\mathrm{Gix_c}$ values across the jurisdiction. In the base-case, the weighted mean has been chosen as a summary statistic for this distribution, where the weights are given by centre-level patient throughput. Alternative summary measures could have been the median or the geometric mean. The jurisdiction-Gix measures the extent to which patients in a jurisdiction come from more or less representative centres, as reflected by the centre-Gix.

The $\mathrm{Gix_j}$ is calculated in exactly the same manner as the $\mathrm{Gix_t}$, with two differences (equation 8.6): the sample of centres is the entire pool of relevant centres in the jurisdiction, not just the ones participating in the trial i.e. all centres where the intervention is expected to be implemented; and the patient weights are not given by local patient recruitment tallies, but by the local patient throughput in a specified time frame e.g. number of finished consultant episodes (FCEs) in a given year. It is, thus, assumed that trial recruitment and patient recruitment in usual practice are equivalent measures of patient volume.

$$\mathbf{Gix_j} = \textstyle\sum_{i=1}^{N} \mathbf{Gix_{c_i}}\, \mathbf{q_i} \quad (8.6),$$

where:

$\mathrm{Gix_j}$ – summary measure of the jurisdiction-wide $\mathrm{Gix_c}$ values;

N – the total number of eligible centres in the jurisdiction;

$\mathrm{Gix_c}$ – the centre-Gix; and

$\mathrm{q_i}$ – the proportion of usual practice patient throughput at centre *i* relative to the jurisdiction-wide patient throughput.

Just as the trial-Gix, the $\mathrm{Gix_j}$ is bounded by the minimum and maximum values of the centre-Gix. In the base case the jurisdiction-Gix ranges from 0 to 5. If closer to 5, this means that more patients in the jurisdiction receive care in centres lying within the middle 80 percentile range for all five dimensions i.e. in ‘more common’ centres. Conversely, if closer to 0 this means that more patients receive care in centres lying outside the middle 80 percentile range for all five dimensions i.e. in ‘less common’ centres.

In order to assess how representative a given trial is to the jurisdiction where it recruited, a comparison between the trial-Gix and the jurisdiction-wide distribution of Gix_c values must be made by means of its summary measure Gix_j. As presented above, the two metrics are calculated using analogous formulae, take values on the same scale and have similar interpretations. Box 8.2 describes the interpretation of a hypothetical numerical example.

Box 8.2 A hypothetical example and interpretation of the trial- and jurisdiction-Gix

Suppose the centre-Gix (Gix_c) incorporates five dimensions and thus ranges from 0 ('less common' centres) to 5 ('more common' centres). Also suppose a given jurisdiction where the weighted (patient throughput) mean Gix_j is 4.2. This suggests that most patients (or patient-episodes, depending on the calculation details) in the jurisdiction are seen in centres with high Gix_c values. A trial with Gix_t of 1.9, for example, recruited the majority of patients from 'less common' centres and cannot be considered a close reflection of patient throughput in the jurisdiction.

In summary, the steps for constructing the three types of Gix for a given research question are as follows:

- Define the jurisdiction, the population of centres, and each centre size, as all potential centres in the jurisdiction where the intervention could be implemented.
- Identify the relevant centre-level dimensions e.g. capacity, teaching function, staff training etc.
- Extract centre-level data for each of the dimensions of the Gix.

- Calculate the dimension level scores (1 or 0) by categorising individual centres as within/outside the middle 80th percentile range i.e. 10th to 90th percentile.
- Calculate each centre-Gix by summing the dimension level scores.
- Determine the distribution of the jurisdiction-wide Gix_c and its summary measures e.g. weighted mean (Gix_j) and SD.
- Calculate the trial-Gix as the weighted (patient recruitment) mean of centre-Gix.
- Calculate the standardised mean difference between Gix_t and Gix_j.

The three sub-sections above introduced the centre-, trial- and jurisdiction-Gix. As explained at the beginning of the Methods section, these concepts were developed to assess the extent to which a given trial is representative of its jurisdiction and to further allow investigating whether this extent affects the generalisability of trial results. The following sub-section outlines how the three types of Gix can be used to investigate the generalisability of trial results.

8.3.5. Using the trial-Gix to evaluate the generalisability of trial results

The objective of this investigation is to establish whether the representativeness of the sample of centres participating in a trial, as reflected by the trial-Gix, influences the generalisability of trial results. In order to estimate the association between the trial-Gix and trial results, simulation methods were used to artificially construct a large number of samples of centres from a real-world trial. Multiple simulated RCTs were thus obtained and the relationship between their Gix_t values and their estimates of effectiveness and cost-effectiveness was investigated.

In this context, 'generalisability of trial results' refers to the accuracy of trial estimates compared to the jurisdiction 'true values'. As pointed out at the beginning of the Chapter (sections 8.1 and 8.2) , the motivating concern of this investigation is that trials which do not recruit representatively may systematically produce biased estimates in relation to jurisdiction-wide decision-making requirements. The prime difficulty when attempting to quantify this type of bias is that no reference point can be identified *a priori*. In other words, no jurisdiction-wide effectiveness and cost-effectiveness estimates are available and thus the 'true values' against which trial results should be compared are unknown.

The simulation method eliminates the need for knowledge of the 'true' clinical and cost-effectiveness parameters. The key assumption here is that the real-world trial is the jurisdiction where we assume that there are no other centres apart from those participating in the trial. The analogy is that with a country where every single centre participated in the trial and thus the nation-wide results are known.

The chosen case study was the ROSSINI trial, which was presented at length in Chapter 5. A brief outline of its design and results is given in Box 8.3. Five thousand trials were simulated by sampling centres with replacement from the ROSSINI trial and all recruited patients were included from each sampled centre (52). For example, if a centre was sampled three times, all its patients would appear three times in the simulated trials. Standard analytical methods were then applied to derive estimates of clinical (odds ratio) and cost-effectiveness estimates (incremental cost; incremental QALYs; and the probability of cost-effectiveness at £20,000/QALY) for each simulated trial. The probability of cost-effectiveness for each simulated trial was calculated using non-parametric bootstrapping with 2,000 iterations. Bootstrapped bias corrected and accelerated (BCa) 95% confidence intervals were also calculated for odds ratios, incremental costs and incremental QALYs in order to investigate the relationship between the trial-Gix and the precision of the estimates (51, 53). Bootstrap methods were described in more detail in Chapter 1. The result was a dataset of

5,000 simulated trials, each with its clinical and cost-effectiveness estimates as well as the trial-Gix.

Box 8.3 Overview of the ROSSINI trial

The ROSSINI trial (Reduction of Surgical Site Infection using a Novel Intervention) is a parallel double-randomised trial comparing a wound-edge protection device (WEPD) with standard care in reducing the rate of surgical site infection in adult patients after open abdominal surgery (Chapter 5). The study recruited 769 patients from 21 hospitals across the UK. For the case study presented in this Chapter, the complete-case dataset was used i.e. patients with complete data on the clinical outcome, cost and health-related quality of life at baseline and 30 days post-operatively (585 patients from 21 hospitals). The trial-wide analyses indicated evidence of neither clinical effectiveness (OR 1.13, 95%CI 0.77 to 1.66) nor cost-effectiveness (14.1% probability of being cost-effective at £20,000/QALY) for the intervention. These estimates are, thus assumed to be the ‘true values’ of clinical and cost-effectiveness in the simulation study. Of note, this estimate is slightly different from the one reported earlier in section 5.1 because it is based on a complete case dataset.

The centre-Gix was calculated as per formula (8.3). The distributions for the five dimensions of the centre-Gix (Table 8.1) were constructed using only the information available from the 21 hospitals in ROSSINI because, for the purpose of the simulation, the 21 centres constitute the pool of centres in the jurisdiction. Specialisation was calculated for each hospital as the proportion of FCEs in lower digestive tract interventions (most interventions in ROSSINI fell in this category) in the total number of FCE in 2011/2012. The data sources for all the five dimensions are indicated in Table 8.1.

Patient recruitment in ROSSINI was assumed to be equivalent to usual patient throughput in the jurisdiction as a measure of patient volume. The Gix_j was calculated as the weighted (patient recruitment) mean of the centre-Gix (formula 9.5). For each simulated trial, the trial-Gix was calculated according to formula (9.6). The standardised mean difference was calculated by subtracting the Gix_j from each Gix_t individual value and dividing the difference by the standard deviation of the Gix_j. This led to the standardised trial-Gix, which is a measure of the trial-Gix departure from the Gix_j. Small positive and negative values of the standardised trial-Gix signify that the trial-Gix is close to the jurisdiction-wide Gix and therefore it can be assumed that the trial is a fairly accurate representation of it. Conversely, extreme positive and negative values of the standardised trial-Gix reflect more extreme trial-Gix values and mean that the given trial is less representative of the jurisdiction as a whole. The standardised trial-Gix was categorised in terms of multiples of standard deviations (SD), as follows: -1 SD (-1.25 SD to -0.75 SD); -0.5 SD (-0.75 SD to -0.25 SD); 0 (-0.25 SD to 0.25 SD); 0.5 SD (0.25 SD to 0.75 SD); and 1 SD (0.75 SD to 1.25 SD).

The relationship between the standardised trial-Gix and generalisability was investigated by exploring the clinical and cost-effectiveness estimates (both the point estimates and the width of the confidence intervals) of the simulated trials across categories of the standardised trial-Gix. The simulations and analyses were performed in R 2.15.3 statistical software (359). Uncertainty in the point estimates was reflected by calculating bootstrapped 95% confidence intervals (percentile method, based on 2,000 iterations) (54) for the clinical and cost-effectiveness outcomes across the five categories of the standardised trial-Gix.

Sensitivity analysis

The base-case analysis referred to the centre-Gix informed by five dimensions (capacity, teaching status, market forces, cost performance and specialisation) and dichotomised centre-level information using the middle 80th percentile range. Additional scenarios were analysed to test the influence of these methodological choices, as presented below.

The influence of Gix content

While keeping the dichotomisation approach constant, three alternative Gix specifications included different combinations of centre-level dimensions:

- Gix3 includes three dimensions: capacity, teaching status and specialisation;
- Gix4a includes four dimensions: capacity, teaching status, market context and specialisation; and
- Gix4b includes four dimensions: teaching status, market context, cost performance and specialisation.

The influence of Gix construct

Three alternative Gix formulations varied the approach to constructing the index:

- Gix4z includes only the four continuous centre-level dimensions (excluding teaching status). The raw centre-level values for each dimension were standardised by subtracting the average from the individual value and dividing the difference by the standard deviation. Each centre retained the *absolute* z-score for each dimension to quantify the departure from the mean; the four absolute z-scores were summed to obtain the centre-level Gix. High values of centre-Gix4z denote centres which are more extreme, while low values (closer to 0) denote proximity to the mean i.e. more ‘common’ centres. Similarly, simulated trials with high trial-Gix4z values recruited the majority of patients from more extreme centres, while simulated trials with low trial-Gix4z (closer to 0) denote trials which recruited predominantly for more ‘common’ centres.

- Gix90 included the five dimensions in the base-case Gix, but dichotomised centre-level information by considering the middle 90th percentile range (5th percentile to 95th percentile); and
- Gix50 also included the five dimensions in the base-case Gix, but dichotomised centre-level information by considering the middle 50th percentile range or inter-quartile range (25th percentile to 75th percentile).

8.4. Results

8.4.1. The centre-level Gix

The centre-level characteristics across the five representativeness dimensions and the centre-Gix for ROSSINI hospitals are presented in Table 8.2. For the sample of ROSSINI centres, the minimum centre-Gix is 2 and the maximum is 5.

8.4.2. The jurisdiction-level Gix

The mean Gix_c in the ROSSINI trial i.e. Gix_j, is 3.90 (median Gix_c 4.00), which suggests that most patients in the 'jurisdiction' come from representative centres, as reflected by the centre-Gix. The weighted mean and the median are comparable, thereby suggesting that the weighted mean is a defensible choice to summarise the jurisdiction distribution of Gix_c.

Table 8.2 Centre-level Gix for ROSSINI centres

Centre	Raw dimension values						Categorised dimension values					Centre
ID	Patients	Beds	Teaching	MFF	RCI	Spec	Beds_ix	Teaching_ix	MFF_ix	RCI_ix	Spec_ix	Gix
1	47	1,019	1	0.96	117	0.05	0	0	1	0	1	2
2	79	423	0	0.95	102	0.04	1	1	1	1	1	5
3	35	696	1	0.94	101	0.04	1	0	1	1	1	4
4	5	765	0	0.95	99	0.03	1	1	1	1	0	4
5	5	472	0	0.95	105	0.05	1	1	1	0	1	4
6	98	692	0	0.95	95	0.03	1	1	1	1	1	5
7	10	458	0	0.97	94	0.05	1	1	1	0	1	4
8	1	708	0	0.96	97	0.03	1	1	1	1	1	5
9	4	532	0	0.96	97	0.03	1	1	1	1	1	5
10	42	992	0	0.95	95	0.04	0	1	1	1	1	4
11	145	886	1	0.93	96	0.05	1	0	0	1	1	3
12	44	419	0	0.95	103	0.05	0	1	1	1	1	4
13	5	544	0	1.01	100	0.05	1	1	1	1	1	5
14	18	666	1	0.97	101	0.04	1	0	1	1	1	4
15	6	756	0	0.94	94	0.05	1	1	1	0	1	4
16	13	508	0	0.93	96	0.05	1	1	0	1	1	4
17	13	458	0	1.10	99	0.09	1	1	0	1	0	3
18	1	358	0	1.02	103	0.06	0	1	0	1	1	3
19	10	436	0	0.96	99	0.07	1	1	1	1	1	5
20	3	568	0	0.96	95	0.07	1	1	1	1	0	4
21	1	916	1	0.95	102	0.03	1	0	1	1	0	3
Median		568	0	0.96	99	0.05						
10th percentile		423	0	0.94	95	0.04						
90th percentile		916	1	1.01	103	0.07						

Legend: Teaching (1 - teaching hospital; 0 - non-teaching hospital); MFF - Market Forces Factor; RCI - Reference Cost Index; Spec - specialisation as % of lower digestive tract finished consultant episodes (FCEs) in the total number of FCEs per hospital in 2011/2012. Categorised dimension values (_ix) are 1 if raw value within 10th-90th percentile and 0 otherwise; for binary variables (Teaching_ix), values are 1 for most common category and 0 otherwise.

8.4.3. The simulated trials

5,000 trials were simulated from the ROSSINI trial. 49 simulated trials were discarded because the probability of cost-effectiveness could not be calculated, resulting in 4,951 simulations analysed. The distributions of the odds ratio, the probability of cost-effectiveness at £20,000/QALY, the incremental costs and the incremental QALYs across the simulated trials are presented in Figure 8.3. Table 8.3 compares the descriptive characteristics and results of the simulated RCTs with ROSSINI estimates. The median values of the parameters from the simulated trials approximate the ROSSINI values, thus suggesting that the simulations are a reasonable representation of the original data.

8.4.4. The standardised trial-Gix and generalisability

Table 8.4 presents the distribution of design characteristics i.e. sample size, the number of recruiting centres, the event rate (incidence of SSI in the RCT population) and the randomization ratio (intervention arm: control arm), across categories of the standardised trial-Gix. It is apparent the trial-Gix groups are similar in terms of the number of recruiting centres, SSI incidence and the relative number of patients in each arm. Trials with higher trial-Gix values seem to have slightly smaller sample sizes than the other categories.

Figure 8.3 Distribution of effectiveness and cost-effectiveness estimates across the simulated RCTs

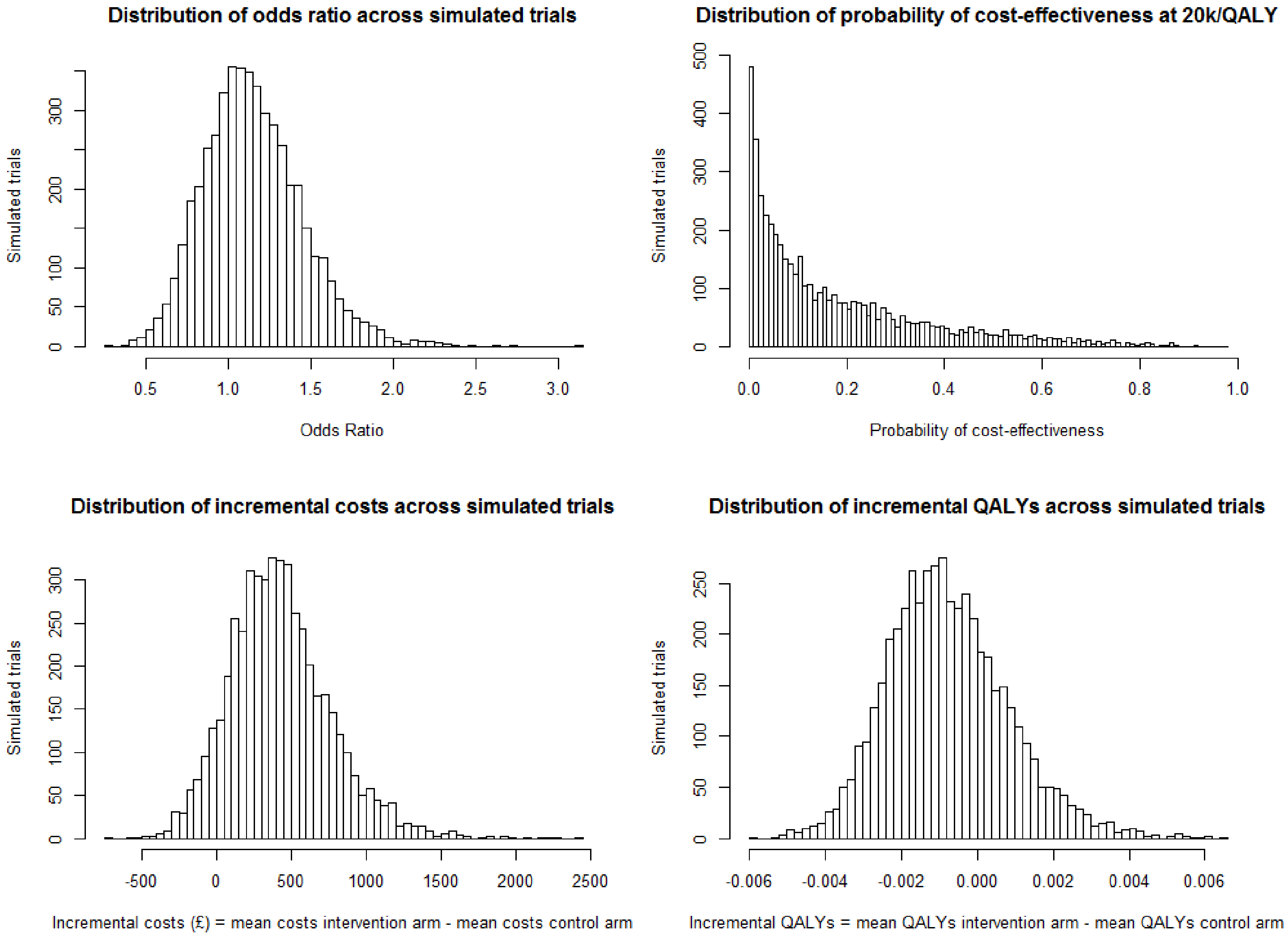


Table 8.3 Comparative characteristics and results of ROSSINI and the simulated RCTs

Parameter	Simulated trials Mean [Median, IQR]	ROSSINI Mean
Descriptive characteristics		
Sample size	584 [572, 463 to 697]	585
Number of unique centres	13.5 [13, 13 to 14]	21
Event rate (SSI incidence)	0.244 [0.241, 0.229 to 0.256]	0.241
Intervention: control ratio	0.99 [1.00, 0.96 to 1.04]	1.00
Trial-Gix	3.91 [3.91, 3.67 to 4.17]	3.90
Results		
Odds ratio	1.15 [1.12, 0.95 to 1.33]	1.13 (95%CI 0.77 to 1.66)
Probability of cost-effectiveness at £20k/QALY	0.18 [0.11, 0.04 to 0.27]	0.14
Incremental costs (£)	425.30 [397.50, 200.10 to 627.40]	376.37
Incremental QALYs	-0.00076 [-0.00090, -0.00186 to 0.00021]	-0.00089

Table 8.4 Characteristics of the simulated trials across categories of standardised trial-Gix

Parameter		Categories of standardised trial-Gix				
		-1 SD (-1.25 to -0.75 SD)	**-0.5 SD (-0.75 to -0.25 SD)**	**0 (-0.25 to 0.25 SD)**	**0.5 SD (0.25 to 0.75 SD)**	**1 SD (0.75 to 1.25 SD)**
Number of simulated RCTs		160	1210	2131	1208	241
Sample size						
	Mean	568.2	607.6	594.0	562.2	538.7
	Median	558.5	603.5	584.0	546.0	531.0
	IQR	452.5 to 672.2	483.2 to 722.0	472.0 to 709.0	446.0 to 674.0	447.0 to 630.0
Number of centres						
	Mean	12.6	13.4	13.7	13.4	12.9
	Median	13	13	14	13	13
	IQR	12 to 13	12 to 14	13 to 15	12 to 14	12 to 14
Event rate (SSI incidence)						
	Mean	0.247	0.241	0.244	0.247	0.243
	Median	0.243	0.238	0.241	0.246	0.243
	IQR	0.227 to 0.264	0.225 to 0.252	0.229 to 0.255	0.234 to 0.260	0.233 to 0.252
Intervention: control ratio						
	Mean	0.98	0.99	1.00	1.01	1.01
	Median	0.98	1.00	1.00	1.01	1.01
	IQR	0.95 to 1.02	0.96 to 1.03	0.97 to 1.04	0.97 to 1.05	0.97 to 1.05

Table 8.5 summarises the clinical effectiveness and cost-effectiveness estimates in the simulated RCTs across the five categories of the standardised trial-Gix. Two observations can be made in relation to the odds ratio and the incremental QALYs: first, both quantities exhibit a monotonic pattern (decrease and increase, respectively) across the standardised trial-Gix subgroups, from the -1 SD to 1 SD subgroup; this suggests that simulated trials with higher trial-Gix values produce results more favourable to the intervention than trials with lower trial-Gix. Second, the average odds ratio and incremental QALYs are the closest to the 'true values' in ROSSINI (OR 1.13 and -0.00089 incremental QALYs) for the 0 subgroup (-0.25 SD to 0.25 SD); this suggests that simulated RCTs with a trial-Gix close to the jurisdiction-Gix give the closest results to the 'true' values. No trend is discernible for incremental costs and the probability of cost-effectiveness at £20,000/QALY.

The distributions of the effectiveness and cost-effectiveness parameters across standardised trial-Gix subgroups are depicted as box plots in Figures 8.4 to 8.7. The box plots confirm the identified trends for odds ratio and incremental QALYs, as well as the lack thereof for incremental costs and the probability of cost-effectiveness. Figure 8.8 depicts the bootstrapped BCa 95% confidence intervals calculated around the subgroup point estimates across the four clinical and cost-effectiveness estimates. The non-overlapping confidence intervals for odds ratio and incremental QALYs confirm the result of the box plots, where there is still no apparent standardized Gix group-dependent effect for incremental costs and the probability of cost-effectiveness.

Table 8.5 Results of the simulated RCTs across categories of standardised trial-Gix

	Categories of standardised trial-Gix				
Parameter	**-1 SD**	**-0.5 SD**	**0**	**0.5 SD**	**1 SD**
	(-1.25 to -0.75 SD)	**(-0.75 to -0.25 SD)**	**(-0.25 to 0.25 SD)**	**(0.25 to 0.75 SD)**	**(0.75 to 1.25 SD)**
Number of simulated RCTs	160	1210	2131	1208	241
Odds ratio					
Mean	1.58	1.30	1.14	1.01	0.91
Median	1.53	1.28	1.12	0.99	0.88
IQR	1.35 to 1.75	1.11 to 1.46	0.97 to 1.30	0.83 to 1.18	0.77 to 1.03
Probability cost-effective					
Mean	0.21	0.17	0.17	0.20	0.17
Median	0.17	0.10	0.11	0.13	0.10
IQR	0.06 to 0.29	0.03 to 0.24	0.03 to 0.26	0.04 to 0.31	0.03 to 0.24
Incremental costs (£)					
Mean	362.5	447.7	430.1	401.3	434.1
Median	345.0	430.4	394.4	364.5	400.7
IQR	205.7 to 547.8	215.6 to 642.1	205.0 to 630.7	156.5 to 583.9	215.6 to 627.7
Incremental QALYs					
Mean	-0.00203	-0.00150	-0.00081	0.00003	0.00033
Median	-0.00213	-0.00157	-0.00093	-0.00005	0.00016
IQR	-0.00302 to -0.00110	-0.00233 to -0.00150	-0.00180 to 0.00006	-0.00108 to 0.00105	-0.00077 to 0.00139

Figure 8.4 Clinical effectiveness estimates in simulated RCTs across categories of standardised trial-Gix

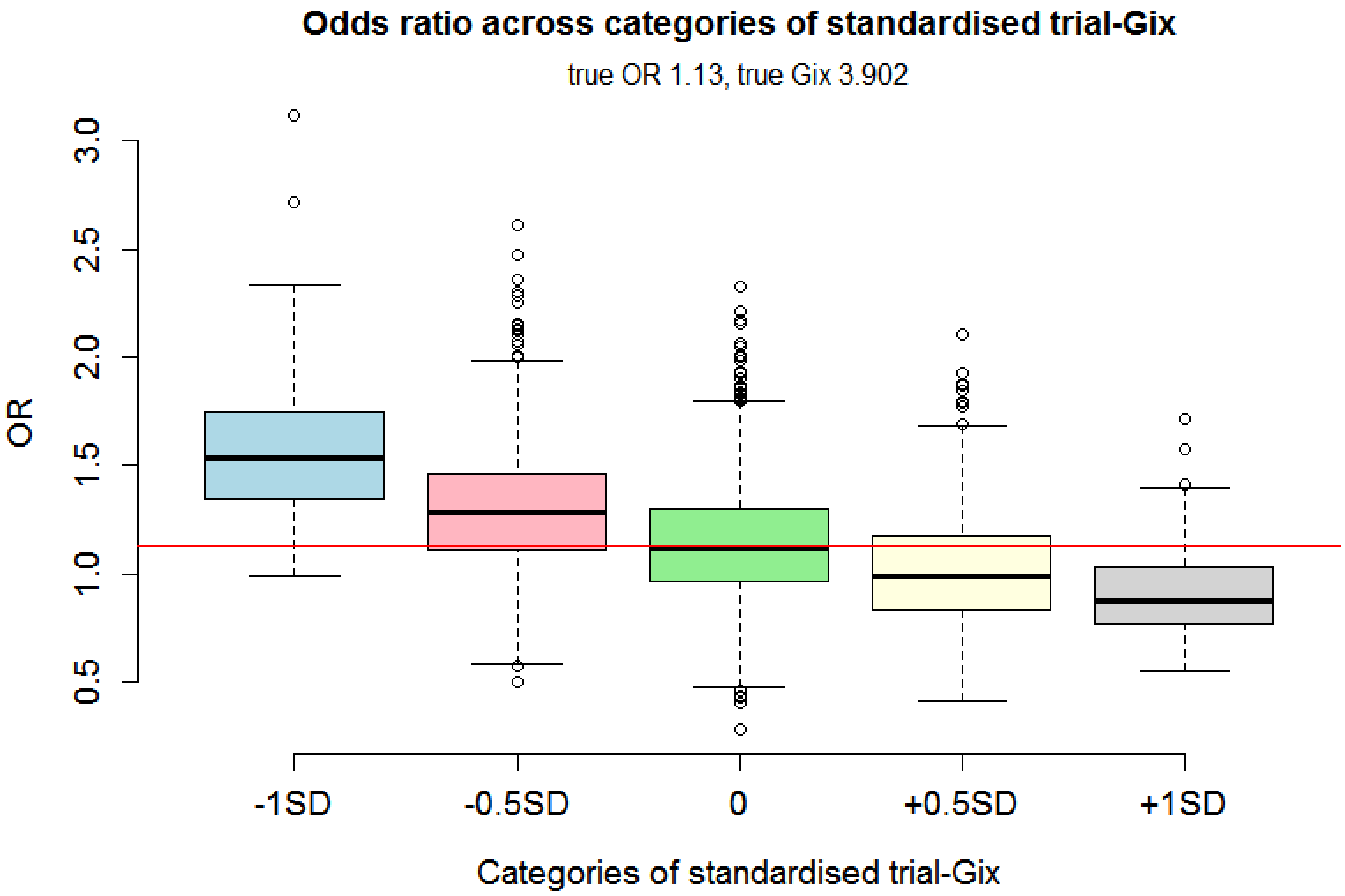


Figure 8.5 Incremental costs in simulated RCTs across categories of standardised trial-Gix

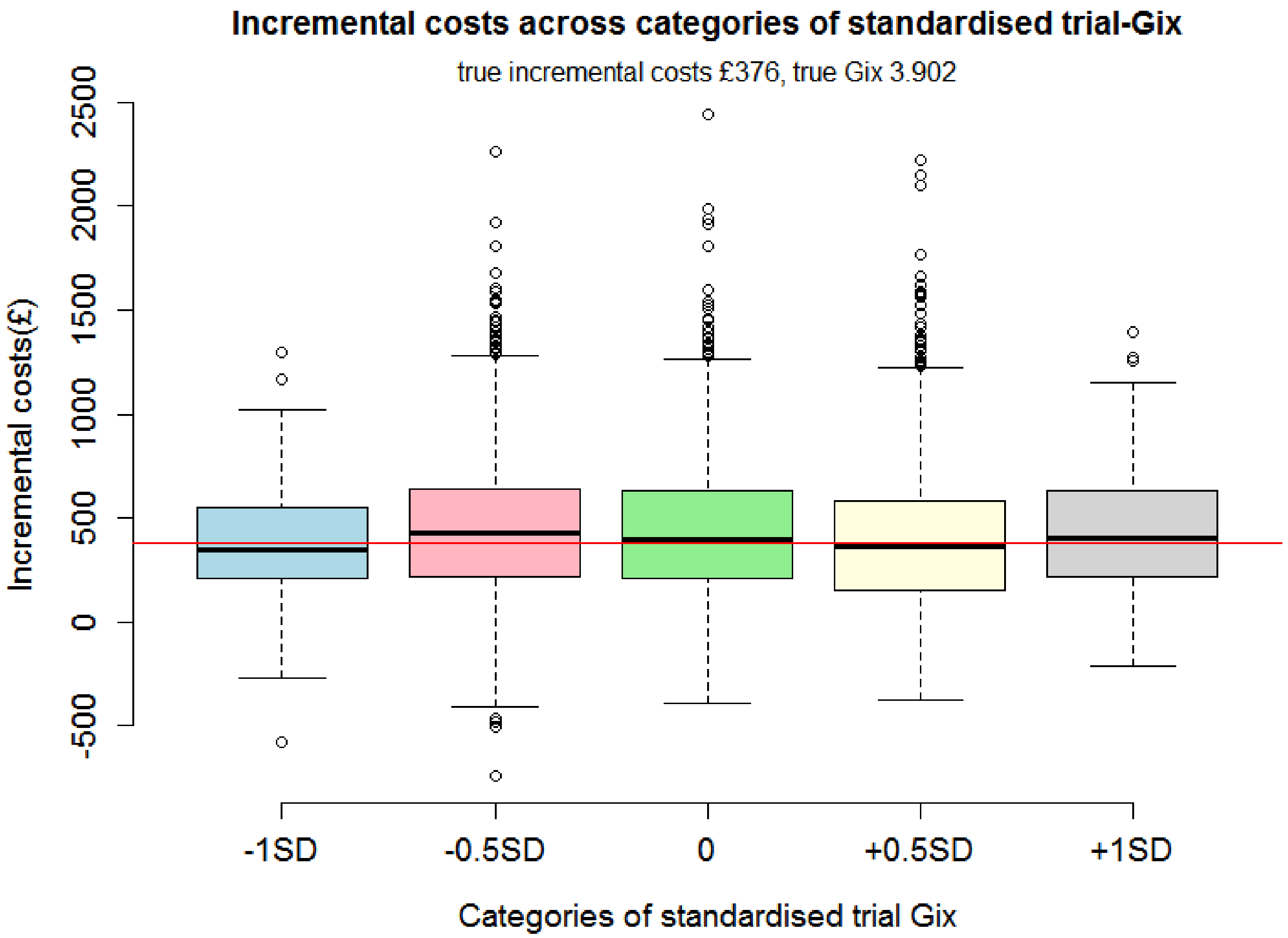


Figure 8.6 Incremental QALYs in simulated RCTs across categories of standardised trial-Gix

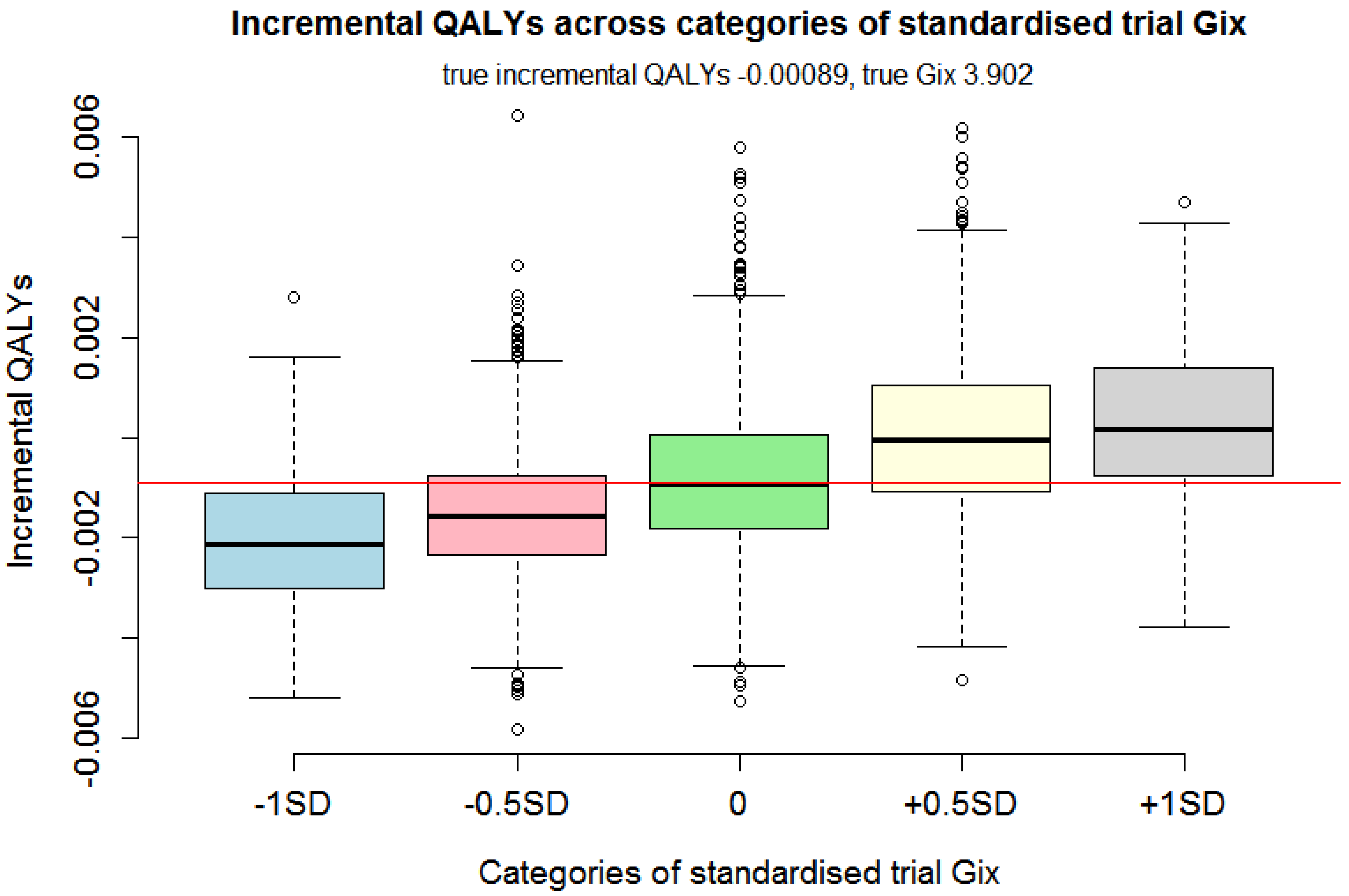


Figure 8.7 Probability of cost-effectiveness in simulated RCTs across categories of standardised trial-Gix

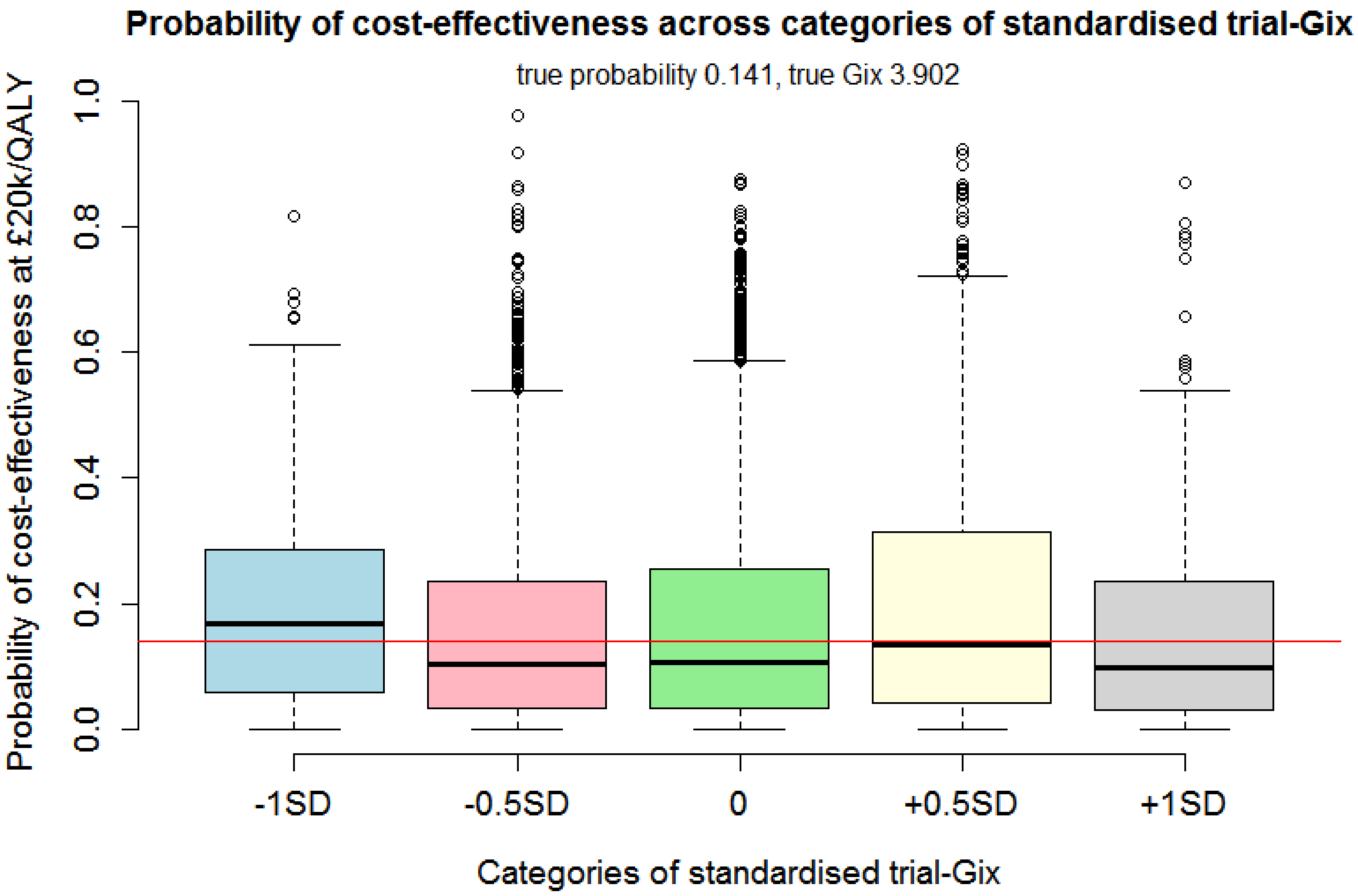


Figure 8.8 Bootstrapped 95% confidence intervals for point estimates across categories of standardised trial-Gix

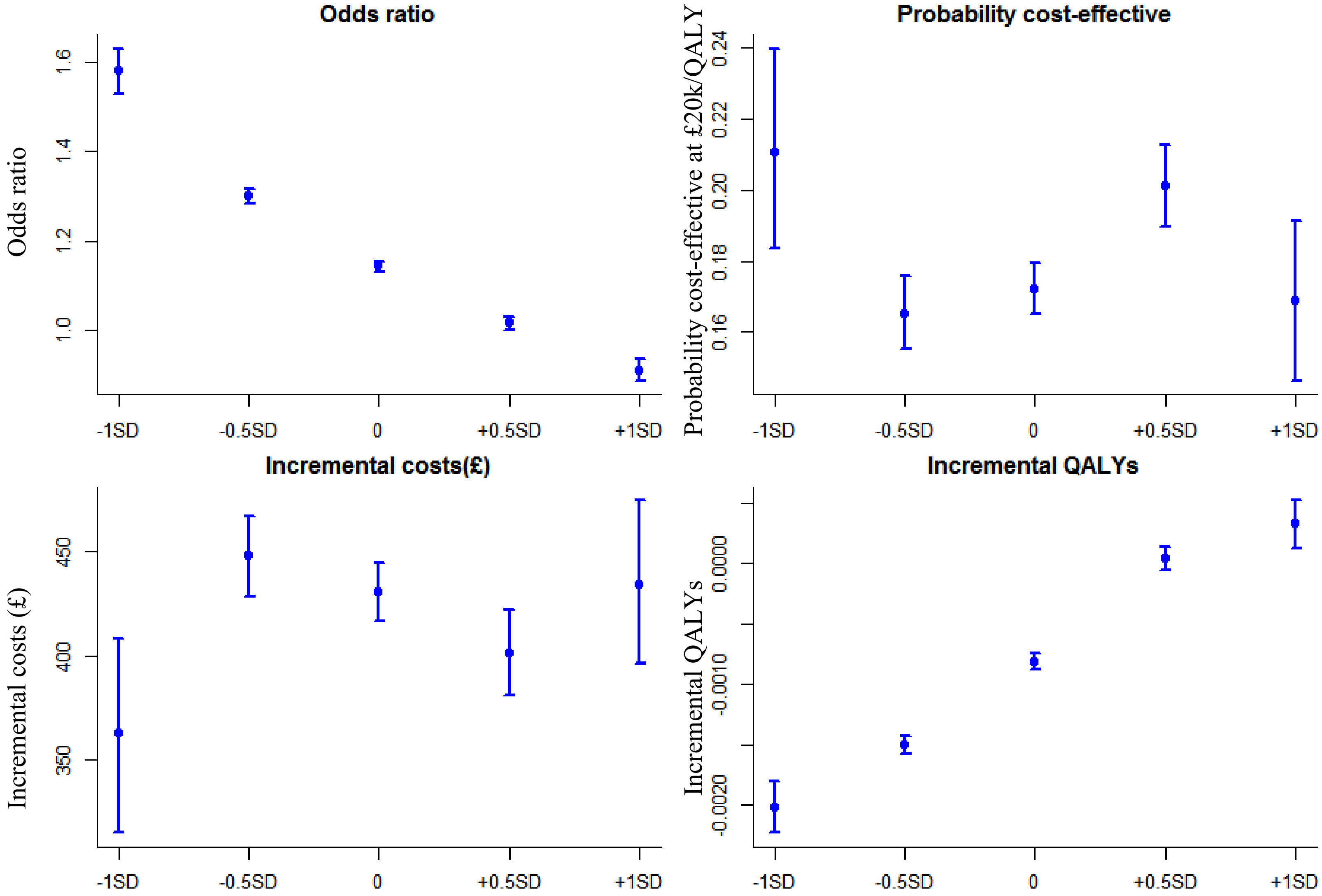
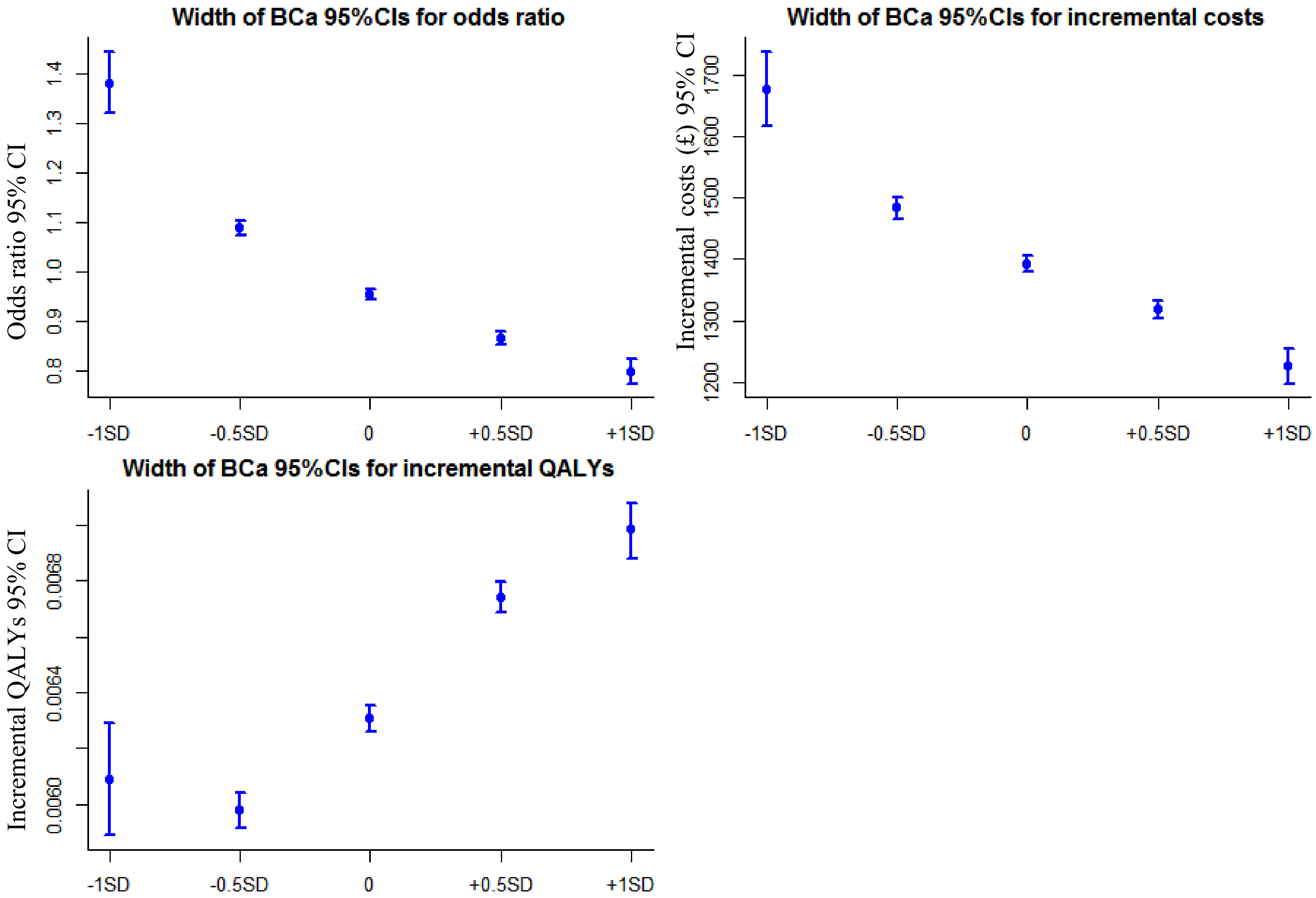


Figure 8.9 Bootstrapped 95% confidence intervals for the precision of point estimates across categories of standardised trial-Gix



The relationship between the standardized trial-Gix and the width of the bootstrapped BCa 95% confidence intervals for the odds ratio, incremental costs and incremental QALYs was also analysed (Figure 8.9). It appears that simulated trials with higher trial-Gix values produce more precise estimates of odds ratio and incremental costs. The trend is reversed for incremental QALYs, where trials with the lowest trial-Gix values had more precise estimates of incremental QALYs.

Sensitivity analysis

The simulation results for the alternative indices are presented in Table 8.6 and the box plots are displayed in Appendix 17. Figures A17.1 to A17.3 present box plots of point estimates for clinical and cost-effectiveness results across categories of standardised Gix3, Gix4a and Gix4b, respectively, which were based on various combinations of centre-level dimensions. As with the base-case standardised trial-Gix, simulated trials with higher trial-Gix values favour the clinical effectiveness and the incremental QALY benefit of the intervention for all the alternative indices. There is no discernible effect on incremental costs and probability of cost-effectiveness.

Table 8.6 Results of the simulated RCTs across various standardised trial-Gix formulations

	Categories of standardised trial-Gix				
Parameter (mean, median)	**-1 SD (-1.25 to -0.75 SD)**	**-0.5 SD (-0.75 to -0.25 SD)**	**0 (-0.25 to 0.25 SD)**	**0.5 SD (0.25 to 0.75 SD)**	**1 SD (0.75 to 1.25 SD)**
Gix3					
Simulated RCTs	43	1040	2818	1011	39
Odds ratio	1.78 (1.67)	1.33 (1.32)	1.13 (1.12)	1.00 (0.98)	0.94 (0.89)
Probability cost-effective	0.38 (0.37)	0.19 (0.13)	0.17 (0.10)	0.19 (0.11)	0.17 (0.09)
Incremental costs (£)	134.1 (157.6)	423.2 (407.4)	438.3 (404.1)	402.7 (372.3)	445.3 (441.6)
Incremental QALYs	-0.00201 (-0.00207)	-0.00156 (-0.00164)	-0.00076 (-0.00090)	0.00006 (-0.00003)	0.00082 (0.00006)
Gix4a					
Simulated RCTs	109	1188	2167	1304	183
Odds ratio	1.46 (1.45)	1.31 (1.29)	1.15 (1.12)	1.02 (0.99)	0.92 (0.89)
Probability cost-effective	0.12 (0.06)	0.15 (0.09)	0.18 (0.11)	0.21 (0.14)	0.19 (0.14)
Incremental costs (£)	467.5 (486.1)	459.3 (434.4)	418.0 (384.3)	406.6 (374.5)	398.3 (320.8)
Incremental QALYs	-0.00214 (-0.00226)	-0.00153 (-0.00160)	-0.00079 (-0.00088)	-0.00005 (-0.00015)	0.00037 (0.00006)
Gix4b					
Simulated RCTs	231	1149	1904	1313	349
Odds ratio	1.40 (1.37)	1.31 (1.27)	1.16 (1.13)	1.03 (1.00)	0.89 (0.87)
Probability cost-effective	0.11 (0.06)	0.16 (0.10)	0.18 (0.12)	0.21 (0.15)	0.17 (0.10)
Incremental costs (£)	522.1 (496.5)	447.8 (418.8)	410.5 (381.6)	409.8 (359.8)	425.7 (411.2)
Incremental QALYs	-0.00189 (-0.00201)	-0.00133 (-0.00144)	-0.00081 (-0.00092)	-0.00015 (-0.00026)	-0.00011 (-0.00021)
Gix4z					
Simulated RCTs	715	1065	1039	732	481
Odds ratio	0.97 (0.96)	1.06 (1.05)	1.16 (1.15)	1.25 (1.25)	1.33 (1.34)
Probability cost-effective	0.11 (0.06)	0.13 (0.08)	0.17 (0.11)	0.20 (0.15)	0.27 (0.22)
Incremental costs (£)	515.0 (491.4)	468.0 (427.6)	416.2 (375.1)	383.0 (351.1)	299.1 (258.9)
Incremental QALYs	-0.00049 (-0.00055)	-0.00073 (-0.00094)	-0.00079 (-0.00093)	-0.00092 (-0.00110)	-0.00095 (-0.00098)

			Categories of standardised trial-Gix		
Parameter (mean, median)	**-1 SD**	**-0.5 SD**	**0**	**0.5 SD**	**1 SD**
	(-1.25 to -0.75 SD)	**(-0.75 to -0.25 SD)**	**(-0.25 to 0.25 SD)**	**(0.25 to 0.75 SD)**	**(0.75 to 1.25 SD)**
Gix90					
Simulated RCTs	76	963	2832	1074	0
Odds ratio	1.68 (1.65)	1.40 (1.38)	1.14 (1.12)	0.92 (0.90)	n/a
Probability cost-effective	0.37 (0.33)	0.24 (0.18)	0.16 (0.10)	0.15 (0.08)	n/a
Incremental costs (£)	207.7 (182.4)	365.7 (333.3)	431.2 (407.3)	479.0 (438.1)	n/a
Incremental QALYs	-0.00155 (-0.00179)	-0.00137 (-0.00149)	-0.00077 (-0.00090)	-0.00012 (-0.00021)	n/a
Gix50					
Simulated RCTs	0	1063	2570	1129	184
Odds ratio	n/a	1.18 (1.15)	1.15 (1.12)	1.14 (1.12)	1.16 (1.16)
Probability cost-effective	n/a	0.14 (0.08)	0.17 (0.11)	0.23 (0.16)	0.23 (0.16)
Incremental costs (£)	n/a	408.0 (404.0)	425.8 (399.9)	430.2 (368.6)	481.6 (437.9)
Incremental QALYs	n/a	-0.00202 (-0.00205)	-0.00096 (-0.00096)	0.00042 (0.00045)	0.00194 (0.00187)

Figures A17.4 to A17.6 present box plots of point estimates for clinical and cost-effectiveness results across categories of standardised Gix4z, Gix90 and Gix50, respectively, where the 'commonness' of a centre was constructed differently compared to the base-case trial-Gix. The distributions of Gix90 and Gix50 across the simulated trials are skewed to the left and right, respectively, which explains the fact that extreme subgroups (+1SD and -1SD, respectively) are not populated with simulated values. The standardized Gix90 appears to exhibit the pattern identified in the base-case: trials with higher standardized trial-Gix produce clinical effectiveness and incremental QALY estimates which are more favourable to the intervention. Furthermore, the opposite pattern appears to apply to incremental costs and probability of cost-effectiveness, as well: the box plots in Figure A17.5 suggest that trials with higher trial-Gix values have slightly higher incremental costs and, overall, the intervention is less likely to be cost-effective despite being more likely to be clinically effective. As for Gix50, no differences between subgroups are apparent with the exception with incremental QALYs, in line with the base-case pattern (Figure A17.6).

The standardised Gix4z, which uses the sum of absolute z-scores to quantify centre-level 'commonness', displays a similar pattern: trials with low Gix4z values (recruiting predominantly from centres with low centre-Gix4z scores i.e. 'common' centres) favour the clinical effectiveness and QALY gains of the intervention. However, a notable difference from all the other indices is that patterns are discernible for incremental costs and the probability of cost-effectiveness at £20,000/QALY, as well. Simulated trials with low trial-Gix4z also have higher incremental costs and a lower probability of the intervention to be cost-effective. Conversely, trials with higher values of the trial-Gix4z (recruiting predominantly from centres with high centre-Gix4z scores i.e. more extreme centres) have lower incremental costs and, due to a higher proportion of negative incremental costs i.e. cost savings, also a higher probability of cost-effectiveness. Furthermore, the incremental QALY differences across the subgroups are much smaller than in the base-case.

8.5. Discussion

Summary of findings

The representativeness of the sample of participating centres, as reflected by the standardised trial-Gix, appears to influence both the accuracy and precision of RCT results, thereby representing a potential source of bias. In this case study, trials which recruited from more representative centres (high values of standardised trial-Gix) tended to overestimate the benefits of the intervention, both in terms of clinical effectiveness and QALY gains. The simulated trials with lower values of the standardised trial-Gix i.e. trials which enrolled patients from more 'extreme' centres, underestimated the benefits of the intervention. There was generally no discernible association between the trial-Gix and incremental costs or overall cost-effectiveness, with the exception of the Gix4z, which used only continuous variables and absolute z-scores instead of dichotomised scores, and the Gix90, which used a wider percentile range to define centre 'commonness' (5th to 95th percentile).

Furthermore, centre selection also appears to influence the precision of trial estimates. Simulated RCTs with a high trial-Gix produced narrower confidence intervals for the odds ratio and incremental costs, but wider confidence intervals for incremental QALYs. This may be because such 'common' centres deliver more uniform care, but receive a wide variety of patients which translates in the way they perceive their health improvements. The findings were generally robust to various specifications of the centre-Gix. The content of the Gix i.e. the included dimensions, and its construct i.e. the approach to aggregating centre-level information, did not seem to affect the direction of the results, although there were differences in magnitude.

Analytical considerations

The approach presented in this Chapter should be viewed as a proof-of-concept exercise from several viewpoints. First, the choice of dimensions for the centre-Gix was based on their likely impact on effectiveness and cost-effectiveness results, as suggested in the literature. Other variables may well find their place in the Gix. In order to test how dependent the findings were on the choice of these five dimensions, parallel indices were constructed in sensitivity analyses based on various combinations of these variables. While the magnitude of the results differed, their interpretation was entirely consistent with the base-case, thereby suggesting that the results are robust to the Gix content.

Moreover, the dimensions of the Gix are specific to a particular research context. The content of the Gix proposed in this case study refers primordially to hospital care and, as such, is mainly applicable to research questions where inpatient care is the most important determinant of health and economic outcomes. For other RCT settings, such as primary care or palliative care studies, the Gix may include entirely different dimensions, such as average length of consultation or Quality and Outcomes Framework scores (425).

A legitimate concern when aggregating multiple dimensions in an index is autocorrelation. In this case, the potential is even higher because the MFF, a marker of the market environment, is an input in the formula for the RCI, a marker of the provider’s cost performance. The correlation matrices of the five centre-level dimensions based on the 21 ROSSINI hospitals are presented in Table 8.7. The highest correlation coefficients are for capacity - teaching status (0.59) and market context – specialisation (0.68); however, these values do not suggest a high degree of correlation among the dimensions for the purpose of constructing the centre-Gix. As it appears, the amount of correlation after dichotomisation is even lower (Table 8.7).

Second, the proposed construct of the Gix is still simple, though intuitive. More complex approaches can be envisaged. For example, the dimensions' scores could be weighted in accordance with existing evidence of their relative influence on effectiveness and cost-effectiveness measures. The current construct cannot offer an accurate description of a 'representative' centre: there are centres which are 'common' under a number of dimensions, but much less so in others. The choice of the middle 80th percentile range was arbitrary, but in sensitivity analyses alternative ranges were tested (middle 90th percentile and inter-quartile range). The interpretation of the results was generally consistent with the one presented above, thus suggesting that the findings are robust to reasonable choices of the 'commonness' range. The differences between the choices of the 'commonness' interval appear to influence the discriminatory power of the trial-Gix across the relevant outcome measures, and thus the magnitude of the differences between subgroups of the standardised trial-Gix. Using the middle 90th percentile range discriminates strongly for clinical effectiveness i.e. the differences in odds ratio estimates between the extreme subgroups are larger than in the base-case scenario, but also for incremental QALYs, incremental costs and even the probability of cost-effectiveness. The opposite can be observed when using the middle 50th percentile range to dichotomise centre-level information: the differences in odds ratio estimates, incremental costs and probability of cost-effectiveness across subgroups are negligible, while the variation in incremental QALYs is notable. This suggests that a wider 'commonness' range may improve the discriminatory power of the trial-Gix, albeit not for all outcomes. The choice of an appropriate threshold appears, thus, to deserve close consideration in future research.

Table 8.7 Correlation among the five centre-level dimensions in ROSSINI hospitals

	Capacity (beds)	Teaching status	Market context (MFF)	Cost performance (RCI)	Specialisation (%FCE)
Raw standardised values					
Capacity (beds)	1.0000000	0.5866778	-0.38512524	0.13883975	-0.45896312
Teaching status	0.5866778	1.0000000	-0.23304984	0.42840681	-0.27203558
Market context (MFF)	-0.3851252	-0.2330498	1.0000000	0.09585704	0.67675033
Cost performance (RCI)	0.1388397	0.4284068	0.09585704	1.0000000	0.04967891
Specialisation (%FCE)	-0.4589631	-0.2720356	0.67675033	0.04967891	1.0000000
Dichotomised scores					
Capacity (beds)	1.0000000	0.01355815	0.07352941	0.07352941	-0.23529412
Teaching status	0.01355815	1.0000000	0.01355815	0.01355815	0.01355815
Market context (MFF)	0.07352941	0.01355815	1.0000000	-0.23529412	0.07352941
Cost performance (RCI)	0.07352941	0.01355815	-0.23529412	1.0000000	-0.23529412
Specialisation (%FCE)	-0.23529412	0.01355815	0.07352941	-0.23529412	1.0000000

Legend:
MFF – Market Forces Factor
RCI – Reference Cost Index
%FCE – proportion of lower GI tract finished consultant episodes in 2011/2012

Third, the choice of a simulation approach particularly emphasizes the proof-of-concept nature of this research. It has been argued that the usefulness of simulation studies is in investigating bias in the estimates of interest by exploiting the fact that the true values are known (426). The role of the simulations was, in this case, to create a hypothetical research environment which could 1) examine the bias in clinical and cost-effectiveness outcomes across various samples of centres; and 2) inform an assessment of the validity and usefulness of the Gix. The primary focus of this investigation was accuracy rather than precision, therefore 'true values' and an empirical distribution of estimates had to be generated. Two approaches are available to meet these requirements i.e. to inform the 'true values': the first is to use an existing case study. A similar strategy was used by Deeks *et al.*, who investigated the bias in two non-randomised studies by simulating 8,000 and 14,000 RCTs, respectively, based on the participating patients in the original studies and then comparing the distributions of the odds ratios between the randomised and non-randomised experiments (427).

An alternative approach involves generating the jurisdiction and the trial data *de novo* using simulation methods. At the expense of a more complex specification and computationally intensive algorithm, the latter allows a finer control of the model parameters. This has been applied, for example, by Gomes *et al.* in the context of refining economic evaluation methods for cluster-RCTs (428). McCarron *et al.* adopted a similar approach when testing the performance of different Bayesian models to combine the results of randomised and non-randomised studies (429). The case study method was preferred for the purpose of this thesis when considering the data available from the ROSSINI trial and its suitability for the research question given the number and variety of recruiting locations.

Multi-level modelling has been used to investigate centre-level variations in cost-effectiveness results (164, 430) and bivariate hierarchical modelling, a Bayesian extension of multi-level modelling, is currently the recommended approach with this aim (134). It must be acknowledged that such techniques are best suited to produce centre-level incremental cost and effect estimates based on trial-wide results. The main research question addressed in this investigation was different: its focus was on how the sample of centres influences the overall trial-wide results as opposed to how trial-wide results can refine centre-level estimates. The question was formulated as such in light of addressing the practicality of decision-making processes: policy makers would often have to make jurisdiction-wide decision based on information collected from a sample of locations within the jurisdiction or even from an entirely different jurisdiction. The decision at hand is whether the research recommendations as suggested by the existing evidence are relevant, to varying extents, across the entire jurisdiction. Under this context, precise centre-level cost and benefit estimates are of limited value in the absence of a methodological framework that can assess similarity between locations and, in a broader context, the amount of overlap between the research space and the policy space (section 8.2). Only such information can inform the transferability of clinical and cost-effectiveness results and it is this knowledge gap that the Gix attempted to address. For this reason, the use of multi-level modelling was not considered a pre-requisite for this investigation.

Aggregating the centre-Gix dimensions into a single index was necessary for constructing the trial-Gix. It can be argued that the dimensions could be left disaggregated. While this may be an option, the focus of this research was to investigate the properties of the Gix approach through comparing *RCTs* (rather than *centres*) by means of obtaining an empirical sampling distribution of RCT characteristics and result estimates. It would have been challenging to characterise the distributions of simulated RCTs characteristics and outcomes without an aggregated metric. As a future perspective, once the concept of bias due to sample un-representativeness is demonstrated in a range of settings, accepted and addressed at design stage, it may not be always necessary to aggregate the dimensions in order to compare centres and identify the suitable ones. For example, centre-level 'profiles' can be constructed by simply juxtaposing the centre-level scores and an informed decision can be made based on the analysis of individual dimensions.

Interpretation of findings

The results of the case study have two important implications: first, there appears to be a discernible relationship between the sample of participating centres and the accuracy of trial estimates in relation to the jurisdiction point estimate. Consequently, there seems to be support towards the primary hypothesis of this research: there is a potential for bias in trial results (compared to jurisdiction-wide values) resulting from an inappropriate selection of centres. This statement is backed by the identifiable trend in trial results' accuracy as a function of the measure of trial representativeness (standardised Gix_t) – this trend was consistently observed for clinical effectiveness and incremental QALYs. In this context, the counterfactual is that no evidence towards such a relationship can be produced, as it appears to be the general case for incremental costs and probability of cost-effectiveness (although some alternative indices, such as Gix90 and Gix4z, actually did show such an association).

Second, the results indicate that recruiting RCT centres such that the characteristics of the sample of centres resemble those of the jurisdiction as a whole leads to point estimates of trial results which are closest to 'true values'. The main implication is that trial recruitment must closely mirror the jurisdiction clinical practice in order to obtain accurate estimates. Knowledge about both patient throughput and the distribution of centre characteristics at jurisdiction level are pre-requisites for making an informed choice on the appropriate sample of centres. Several important implications become apparent: first, ensuring merely 'a mix' of centres may be insufficient if the mix doesn't reflect the actual joint distribution of centre characteristics; second, the local recruitment contribution is an equally important consideration and must also be factored in the assessment of representativeness; third, recruiting predominantly from 'average' centres i.e. centres with high Gix_c values leading to high Gix_t values, does not seem to be an acceptable compromise – in the case study, such trials overestimated the 'true values' of clinical effectiveness and QALY gains.

There appears to be a relationship between the sample of centres and the precision of trial estimates. However, accuracy and precision do not seem to be correlated strongly in terms of centre selection. For example, simulated trials which closely reflect recruitment at jurisdiction level produce the most accurate odds ratios, but the width of their bootstrapped 95% confidence intervals is higher than that of trials with high standardized Gix_t i.e. trials which recruited predominantly from 'common' centres. The reverse can be said about incremental QALYs, where trials with high standardized Gix_t produced the widest confidence intervals across all subgroups. Furthermore, while no effect has been detected on the accuracy of incremental costs, the precision of cost estimates also seems to improve with recruiting predominantly from 'common' centres. The reasons for these non-uniform variations across outcomes are yet unknown and deserve further exploration in subsequent research. The emerging picture is, however, that a trade-off must be reached between accuracy and precision: there is no category of trials which provides optimal results under both these dimensions, but trials with Gix_t close to the Gix_j (the '0' subgroup) appear to represent an acceptable solution as they produce the most accurate estimates with moderate precision.

As pointed out above, bias due to recruiting from unrepresentative samples of centres appears to exist. Nevertheless, the magnitude of this bias in point estimates across the categories of standardized Gix_t varied across the different index specifications and across outcomes. It is of great interest to ascertain whether the identified bias actually matters for policy purposes i.e. can it alter a reimbursement decision or not? In this case study, the spectrum of odds ratios and incremental QALYs across types of trials was wide enough to produce conflicting recommendations as to the relative merits of the intervention. However, only replicating such studies in other trials, clinical settings and jurisdictions will reliably assess the impact of this bias on policy decisions.

Gix90 and Gix4z were the only evaluated indices for which a pattern was apparent across all the trial results (Figures A17.4 and A17.5). This suggests that the definition of 'commonness' is crucial to the application of the Gix. However, interpreting the findings of these indices is not straightforward. For example, simulated trials with high trial-Gix4z values i.e. recruiting from more extreme centres (+0.5SD and +1SD subgroups), underestimated the clinical benefits of the intervention as well as the incremental costs and QALYs. Incremental costs are underestimated to a larger extent than incremental QALYs, the overall effect being that the intervention appears to be much more cost-effective at £20,000/QALY than it actually is. Conversely, simulated trials recruiting from 'common' centres (-0.5SD and -1SD subgroups) overestimated the intervention's clinical benefit as well as the incremental costs and incremental QALYs, thereby suggesting that the intervention is much less cost-effective than it actually is.

The simulation results for the Gix4z are compatible with a health care setting where the clinical benefit of the intervention is proportional with the costs of care: higher incremental costs are correlated with better clinical and HRQoL outcomes, and vice versa. It is yet unknown how much of this relationship is due to the construct of the Gix4z and how much due to the intrinsic nature of the dataset. The interpretation of the Gix90 results is analogous.

Nevertheless, it must be acknowledged that Gix4z and Gix90 are the only indices which behave as one may hypothesise in the sense of showing consistent associations across both clinical and economic outcomes. The construct of the Gix is, thus, a fertile topic for further exploration in subsequent research. It may be that dichotomization discards so much cost information that any association between incremental costs and the trial-Gix is lost in the base-case Gix.

Strengths and limitations

The approach outlined in this Chapter is, to my knowledge, the first attempt to quantify representativeness at centre and trial level. Furthermore, results appear to be consistent across various content combinations of centre-Gix dimensions and constructs, as discussed above. As such, the proposed approach is novel and appears to be robust.

There are several potential limitations. First, the proposed Gix may be regarded as difficult to interpret. The centre-Gix depends on the jurisdiction-wide distributions of its dimensions. It is unknown how the centre-Gix behaves in jurisdictions with two well-balanced types of centres which take extreme values on most relevant dimensions. For example, in a hypothetical country with equally represented rural, low-staffed, small size hospitals and urban, high-staffed, teaching hospitals. Under the current method, in such a situation comparable proportions of low extremes and high extremes would be artificially coerced in the middle 80th percentile and thus be acknowledged as 'common centres', although they share few characteristics.

Second, categorising the five representativeness dimensions results in losing a large amount of information. Furthermore, potentially similar centres may receive different categorisations because of the choice of the 'commonness' range. However, this problem will always occur when synthesising continuous and categorical variables, as it is the case here. When only continuous variables are considered, the use of z-scores to quantify departure from a measure of location would be more efficient; this approach was explored in sensitivity analysis and results were generally in line with the base-case and all the other indices. The base-case for the method deliberately included both continuous and categorical variables in order to anticipate the practical challenges of constructing the index. Dichotomising information appears to be defensible in the presence of a reasonable number of dimensions, but future research on the merit of alternative approaches is needed. A further challenge relates to incorporating multi-categorical variables in the Gix: under the current approach, further categorisation would be required to incorporate such variables, resulting in further loss of information.

Third, results are generally inconsistent across the studied outcomes i.e. no apparent centre effect on incremental costs or on overall cost-effectiveness with the exception of Gix90 and Gix4z. This may be in part caused by the costing methodology in the ROSSINI trial, where nationally averaged unit costs for inpatient, outpatient and primary care were employed, or by the intrinsic nature of the trial data. Furthermore, it has been challenging to identify a suitable cost-effectiveness metric for the purpose of this investigation because the classical metrics for decision-making i.e. the ICER and the INB, were not appropriate. The difficulties around interpreting ICERs have been highlighted in the literature (431). The INB has the advantage of aggregating incremental costs and incremental effects in one easily interpretable metric (55), but the INB point estimate does not communicate information on uncertainty (e.g. for a given comparator, it cannot be said that an intervention with an INB of £8,000 is twice as likely to be cost-effective relative to an intervention with an INB of £4,000) and therefore is not ideal for direct comparisons. The probability of cost-effectiveness at a meaningful WTP threshold, as reflected by the CEAC, has been chosen for the purpose of this investigation (58). However, there are no other sources that I am aware of where such a metric has been used to explore confounding factors. Future applications of this approach to other RCTs should also investigate the relative merit of alternative cost-effectiveness metrics.

Although there is evidence to support the relevance of the five centre-level variables which were included in the generalisability index (Box 8.1), these were identified from a pragmatic literature and their inclusion in the Gix was, to an extent, arbitrary. This must be acknowledged as a limitation. Ideally the dimensions would have to be identified and developed by applying a robust methodology. Such an approach would include conducting a systematic literature review, assessing the strength of evidence and expert opinion elicitation. However, two issues must be further considered: first, sensitivity analyses around the content of the Gix were performed by testing various combinations of the five dimensions and the findings were consistent across the scenarios. Second, the available systematic reviews on centre-level factors have identified a very large number of such considerations - up to 77 in the publication by Goeree *et al.* (136). In the absence of a classification of these factors' relative importance, it can be argued that any choice of dimensions for the Gix would presently be informed, to a large extent, by the investigators' experience and intuition. Nevertheless, future work should include an update of the existing reviews and the development of a framework to allow the rational identification of candidate variables for the Gix.

The current version of the Gix does not include population characteristics as a dimension. As such, the proposed version assumed a homogenous patient population across the jurisdiction and, furthermore, that patients enrolled at each centre were an unbiased sample from this population. This is a strong assumption. The main reason why such information wasn't incorporated in the present version was the difficulty to establish the association between area-specific demographic characteristics and health care outputs. For example, knowing that the prevalence of obesity, a risk factor for surgical infection, varies across locations cannot automatically inform the adjustment of hospital outputs without investigating the impact of such differences on health care utilization patterns. Detailed information on each centre's case-mix, such as that provided by the Hospital Episode Statistics data (432), would address this knowledge gap. The time constraints did not allow for such data to be available for this research, but future research may well incorporate it. The proposed framework is sufficiently flexible to accommodate such developments.

The simulation method sampled centres with replacement from ROSSINI centres and included all patients recruited at a given centre. This is equivalent to assuming that all relevant patients in a centre participated in the RCT. Such an extreme assumption is a limitation. It would be interesting to incorporate random patient selection at centre level in future developments of the method.

The ROSSINI trial was used as a case study to demonstrate the application of the Gix. However, as pointed out in Chapter 5, evidence suggested that it was very unlikely that the intervention under scrutiny (WEPD strategy) was effective or cost-effective; moreover, the uncertainty around the trial findings was substantial. As such, the simulation method propagated this uncertainty and it is possible that simulation results reflect more the propagated noise than true signal. For this reason, replicating the Gix methodology for other trials and in other clinical areas is a necessary future step.

Implication of findings

These results demonstrated that the sample of RCT centres can affect the accuracy and precision of trial results relative to jurisdiction-wide ‘true’ values. The main implication is that just because a trial was conducted in a given jurisdiction, this does not necessarily make its results representative of that jurisdiction. This result can be of interest to policy makers, research commissioners and researchers.

Two types of practical applications of the Gix concept can be thought of. First, as the trial-Gix is calculated using the centre-Gix, these results suggest that there may be a potential for the centre-Gix to inform the design of trials. This would involve selecting centres rationally and adjusting local recruitment rates appropriately such that the resulting trial-Gix has as close a value as possible to the jurisdiction-wide Gix, thereby producing more generalisable results.

Furthermore, a retrospective application of the Gix can also be envisaged. One of the current knowledge gaps is that extrapolations of cost-effectiveness results are difficult in locations which did not participate in the original trial because it is difficult to specify what 'similar centres/locations' actually mean. The centre-Gix could fill this gap by providing a rational and quantitative measure of similarity between centres.

Further research

This research produced a promising result, but also has several limitations which should be the object of future research. First, a systematic approach is needed to identify centre-level dimensions and to propose a framework, together with a set of transparent criteria, for quantifying the strength of the evidence for each dimension included in the index. This would also allow a rational, evidence-based weighting system for the Gix dimensions.

Second, further approaches to the Gix construct should be explored. In particular, it would be of great interest to expand the approach beyond binomial variables in order to avoid dichotomisation and preserve more information. Categorising raw values was the methodological choice to combine categorical and continuous variables. The conceptual development of the Gix should investigate and propose alternative ways of summarising centre-level information. Furthermore, the definition of 'commonness' appears to be equally as important.

Third, the scope of this demonstration was limited to a single case study RCT. The behaviour of the Gix in other settings must be further explored before making a judgement on its usefulness and potential for incorporation in study design. Two related directions are apparent: first, it would be of great interest to replicate this analysis using the same Gix across a range of surgical RCTs, just as ROSSINI, and establish whether findings are robust in suggesting between-centre variation in cost-effectiveness. Second, extending the approach to other health care settings and types of interventions with adapted generalisability indices would provide information on the extent of generalisability bias across clinical research areas.

Finally, generalisability has often been discussed in the economic evaluation literature in the context of international trials due to the purported significant differences across countries in terms of health system characteristics, macroeconomics environment, demography and other factors. The approach presented here looked at a within-country study and the results suggested that there is scope for bias due to centre selection for this type of studies, relative to jurisdiction-wide estimates. It will be worth extending the proposed framework to multinational studies, potentially by introducing a further level of the Gix: centre-Gix, country-Gix and multi-country-Gix (e.g. Europe).

8.6. Conclusion

The generalisability index (Gix), a measure of centre- and trial-representativeness, has been proposed. Using a real-world RCT (the ROSSINI trial) as a case study, the simulation results indicated a relationship between the trial-Gix and the clinical and incremental HRQoL benefits of the intervention, thereby suggesting that recruiting from an unrepresentative sample of centres can bias trial results compared to the jurisdiction-wide point estimates. Furthermore, trials whose recruitment closely mirrored that of the jurisdiction produced the most accurate estimates.

The findings were generally robust to alternative scenarios concerning different approaches to the content and construct of the Gix. From a methodological standpoint, further research should focus on devising a systematic way to including centre-level variables in the Gix and exploring potential construct approaches to the Gix, especially by avoiding dichotomisation and defining centre ‘commonness’. Furthermore, the results must be replicated in other trials, clinical areas and health care settings before making a judgement on the usefulness of the Gix.

These preliminary findings suggest that trial results can be biased compared to jurisdiction ‘true values’ due to unrepresentative centre selection. The Gix appears to be a useful tool for identifying and quantifying this bias. The Gix has the potential to develop as an instrument to assist trialists in improving their sampling for generalisability purposes and to inform decision makers on the generalisability of a given trial’s findings.

CHAPTER 9. DISCUSSION AND CONCLUSION

This Chapter summarises and integrates the findings presented throughout the thesis. The main objective of this work, as detailed previously in section 1.4, was to evaluate the implications of the current practice of centre selection to RCTs in the UK for the generalisability of trial results. A real-world example (the use of WEPDs vs. standard care to reduce SSI after open abdominal surgery) was used as a case study to support the methodological investigation. First, the existing evidence on WEPDs was synthesised (Chapter 3) and new clinical and cost-effectiveness evidence was generated (Chapter 4 and Chapter 5). Second, an empirical investigation of the impact of centre selection on trial results was conducted (Chapter 7), followed by a demonstration of the proposed methodology using the ROSSINI trial of WEPDs vs. standard care in the NHS as an example (Chapter 8).

9.1. Summary of principal findings

9.1.1. The benefits of WEPDs compared to standard care in reducing SSI

The clinical question at the heart of this thesis concerned the benefit of WEPDs in reducing SSI compared to standard care. WEPDs have been used informally to reduce SSI for more than 40 years, but the evidence around them had never been summarised. A systematic approach to produce evidence for decision-making was employed. First, a systematic review of existing studies (Chapter 3) and a preliminary cost-effectiveness decision model (Chapter 4) were conducted, followed by a primary data collection exercise to provide definitive evidence (Chapter 5).

The systematic review and meta-analysis presented in Chapter 3 suggested that WEPDs were likely to be effective compared to standard care, but it was noted that the quality of the 12 included RCTs was generally poor. The results of the cost-effectiveness decision model presented in Chapter 4, which was informed by the clinical effectiveness systematic review, indicated that WEPDs were also likely to be cost-effective compared to standard care in the UK setting. As such, conducting a large, high quality RCT was warranted to address the methodological shortcomings of previous studies and produce robust evidence on the effectiveness and cost-effectiveness of WEPDs in the UK.

The ROSSINI trial (Chapter 5) was a multi-centre RCT which demonstrated no benefit associated with the use of WEPDs in reducing SSI after open abdominal surgery (OR 0.97, 95%CI 0.69 to 1.36). The economic evaluation alongside ROSSINI was conducted from the perspective of the NHS and employed a 30-day post-operatively time horizon. The base-case analysis used multiple imputation to account for missing cost and HRQoL data. In addition, complete case and adjusted analyses were performed. The results of the base-case analysis suggested that the WEPD option had only an approximately 20% probability of being cost-effective at a WTP threshold of £20,000 to £30,000/QALY (39), thus standard care was the cost-effective alternative. The result was robust to the sensitivity analyses. As such, the ROSSINI trial showed no evidence of clinical or economic benefit associated with the use of WEPDs compared to standard care.

9.1.2. Empirical evidence on centre selection for RCTs in the UK

External validity was one of the main aims of ROSSINI: the inclusion criteria were deliberately broad; the standard care protocol was left at the discretion of the local operating teams; and the trial recruited from 21 centres across England, both university and general hospitals. However, an argument was made (section 1.3) that centre selection may influence the generalisability of trial findings and thus warranted further research. A significant gap in the current literature was that the characteristics of the sample of participating centres had not been formally incorporated in current analytical techniques and the assumption that centres were randomly selected from a jurisdiction was not supported by any evidence. The mixed methods study (Chapter 7) revealed that centre selection for RCTs in the UK is primarily determined by pragmatic considerations such as proximity of the centre to the trial office, having a positive recruitment history, complying with regulatory requirements and the ability to maintain good communication. The systematic review of NIHR trial protocols (section 7.2) demonstrated that reporting the reasons which guided the choice of participating centres is currently suboptimal and could be largely improved.

Having a representative sample of centres at jurisdiction-level with a view to ensure the generalisability of trial results appears to be among the considerations which are factored in the centre selection decision. However, only 30% of the protocols included in the review mentioned it and considerations relating to particular trial participation criteria (e.g. regulatory requirements, staff training and research experience) were more often specified. Centre selection does not appear to be a random process, either: of the 129 RCT protocols included in the systematic review, only two used random sampling to select participating centres.

The focus groups (section 7.3) and the online survey of trialists (section 7.4) identified a tension between the current and ideal practice of centre selection (discussed in section 7.5). While current practice seems to be driven by pragmatism in meeting recruitment targets, complying with funders' and sponsors' requirements and running the trial within the designated budget, in ideal practice considerations such as generalisability, patient convenience and a collaborative approach to decision making within trials (e.g. a broader role for the TMG as opposed to the Chief Investigator) should be emphasised. Trialists appear to acknowledge the importance of generalisability and of enrolling from such a sample of centres so that the generalisability of trial findings is ensured, but other concerns currently take precedence.

In light of the above, there are reasons to believe that the majority of RCTs in the UK do not explicitly aim to enrol a representative sample of centres. This may introduce bias in trial results relative to the jurisdiction-wide 'true values'; in other words, it is possible that trials recruit from samples of centres which are not representative of the jurisdiction and produce results which cannot be generalised to the entire jurisdiction.

9.1.3. The influence of centre selection on trial results

The Generalisability index (Gix) was proposed as a measure of representativeness of a given centre and of a given trial relative to the population of centres in a given jurisdiction. The Gix incorporates a number of centre-level characteristics such as size/capacity, teaching status and market environment and quantifies the extent to which a centre or a sample of centre is 'common' or extreme relative to the distribution of these characteristics at jurisdiction-level. Using simulation methods and the ROSSINI trial as a hypothetical jurisdiction where the 'true values' of clinical and cost-effectiveness were known, 5,000 RCTs were simulated. The relationship between the standardised trial-Gix and trial results was investigated in order to ascertain whether the sample of participating centres affects clinical and economic trial outcomes (section 8.4).

The principal finding was that the characteristics of the sample of participating centres influence trial results (section 8.5). Simulated trials which produced results closest to the 'true values' were those whose trial-Gix was the closest to the summary measure of the jurisdiction-wide distribution of centre characteristics i.e. trials whose participating centres were a close reflection of the population of centres in the jurisdiction. Conversely, simulated trials with a more extreme trial-Gix compared to the jurisdiction-wide summary measure either underestimated or overestimated the results. Clinical effectiveness and HRQoL improvements, measured in incremental QALYs, were most sensitive to variation in relation to the Gix. The effect was much less clear for incremental costs and the probability of cost-effectiveness at £20,000/QALY. The explanation is yet unclear, but it may be related to the costing methodology in ROSSINI, which used nationally averaged unit costs for the NHS.

The findings were robust to a number of sensitivity analyses which varied both the content (centre-level characteristics) and construct (approach to summarising and aggregating centre-level information) of the Gix. The alternative formulations of the Gix displayed different discriminatory powers across the clinical and economic outcomes and, thus, highlighted the impact of the definition of 'commonness' on the simulation results.

9.2. Interpretation and implications of findings

Chapter 6 integrated the findings of the ROSSINI trial, the previously reported RCTs and the preliminary cost-effectiveness decision model. The systematic review and the decision model had suggested that WEPDs may be effective and cost-effective, respectively, while ROSSINI results indicated the opposite. However, ROSSINI had superior methodological quality, was more generalisable than previous RCTs and is, thus, likely to have generated more robust results. In light of these considerations, it is likely that WEPDs are neither effective nor cost-effective compared to standard care in reducing SSI and cannot be recommended for routine use in the NHS.

The findings of the generalisability research (Chapters 7 and 8) suggested that it is inappropriate to assume that a given RCT's results can be generalised to the jurisdiction where it recruited from without a careful assessment of the characteristics of the participating sample of centres. As such, the importance of recruiting patients from RCTs from a representative sample of centres jurisdiction-wise becomes obvious. This finding must be interpreted with caution in that the Gix methodology does not aim to assess the internal validity of RCTs and is not designed to assess the accuracy of trial estimates. If the risk of bias is minimised (sub-section 1.3.1), it is likely that trial estimates are correct for the sample of centres which contributed data, but may not be generalised to the entire jurisdiction. As such, just because a RCT recruited in country A and suggested that a given intervention may be clinically or cost-effective, it cannot be straightforwardly assumed that the given intervention is effective in country A. In order to make such an assessment, information on how the sample of participating centres reflect the jurisdiction as a whole is needed. In perspective, the assumption of exchangeability, crucial to the application of current cost-effectiveness refinement methods (155), does not seem to hold as different samples of centres systematically appear to produce markedly different results. Consequently, there may be some merit in revisiting the currently recommended hierarchical trial analysis methods for cost-effectiveness (134) in light of this finding.

The focus of the research has initially been on trial-based economic evaluations and the Gix was designed toward this end (section 1.4). However, the most striking results were observed for clinical effectiveness (section 8.5). It must be, thus, acknowledged that any proposed index will capture the influence of centre selection on both types of trial outcomes.

The proposed Gix is not only a tool that can demonstrate the influence of the sample of centres on trial results, but also has the potential to address it. As suggested in section 8.2, two types of practical applications can be envisaged. The first is retrospective and refers to the analysis of trial data: the Gix could be used, on the one hand, to assess the relevance of a given RCT to the jurisdiction where it recruited from, and, on the other hand, to predict the clinical and cost-effectiveness of an intervention in locations which did not contribute patients in the RCT. The ultimate result would be to obtain a jurisdiction-wide estimate of clinical and cost-effectiveness. Conceptually, this process implies measuring the overlap between the research space and the policy space and then characterising the policy space based on information contained in the research space (section 8.2). The Gix could act, thus, as a link between the two domains. Such an application would be of interest from a policy perspective, where decision makers are often faced with making jurisdiction-wide judgements based on evidence collected from a limited number of settings.

The second practical application of the Gix is prospective and refers to trial design. As the Gix can characterise both a centre and a sample of centres in relation to the population of centres in the jurisdiction, it can be used as a tool in RCT design so as to purposively select the centres and corresponding recruitment rates which are likely to maximise the generalisability of trial findings. Furthermore, there is scope for dynamically monitoring throughout the trial how differential recruitment across participating centres affects generalisability. However, it must be noted that a pre-requisite for such an application is the validation of a given Gix for the therapeutic area where it is used (e.g. complications after general surgery, cardiovascular disease prevention etc.) to ensure that the Gix has sufficient predictive value of the trial's deviation from jurisdiction-wide 'true values'. Such a development would of interest to trialists and policy makers as it would allow them to plan the need for future RCTs so that their results will be as relevant as possible to the policy context. The approach would compel decision makers to think in advance of the locations where they are interested in applying the economic results of a clinical trial and also of how health services could be re-shaped in particular regions so that an intervention becomes locally cost-effective.

It is important to note that the proposed Gix has not been imagined as a universal metric. As such, it cannot be applied in its current form across all therapeutic areas and all settings. It was rather designed to be flexible enough so as to be adapted to particular research requirements in terms of its content (i.e. the dimensions it includes). Although demonstrated here in the context of SSI after abdominal surgery, its applications to other settings such as primary care may be very different (section 8.5), but the principles of the proposed method still apply.

Finally, demonstrating that the sample of centres may indeed impact the generalisability of trial findings could also improve the reporting of considerations for centre selection both in trial protocols and trial publications. This would allow the readership to make an informed judgement on the extent to which trial findings are indeed representative of their jurisdiction.

9.3. Strengths and limitations

The thesis took a systematic approach to synthesise and produce high-quality evidence towards the clinical and economic benefit of WEPDs in reducing SSI after open abdominal surgery. There were no previous systematic reviews of WEPD clinical effectiveness and no economic evaluations of WEPDs against any comparator. As such, the systematic review of clinical effectiveness (Chapter 3), the SSI decision model (Chapter 4) and ROSSINI economic evaluation (Chapter 5) were all novel. Furthermore, the SSI utility systematic review (section 4.1) synthesised for the first time health utility data in relation to SSIs, while ROSSINI generated the first EQ-5D estimates for SSI patients (Chapter 5).

The methodological result of the thesis i.e. the characteristics of the sample of recruiting centres influences trial results, has strong intuitive appeal. Furthermore, this research moves forward the current understanding of the topic under a number of aspects. First, it highlights the importance of having an operational definition of representativeness. As pointed out in the literature review on generalisability of economic evaluations (sub-section 1.3.2), the absence of such a definition has made transferability efforts across locations difficult as the degree of similarity between locations could not be established (155). The Gix was devised as such a measure and, thus, addresses the gap by proposing a set of measurable centre-level characteristics which can be aggregated into a single metric. Second, it underlines the importance of having a quantitative measure of representativeness. The systematic review of HTA trial protocols (section 7.2) revealed that there is some interest in generalisability and diversity as far as participating centres are concerned, but this interest was not substantiated by demonstrating the appropriateness of the participating sample (for example, by comparing the proportion of rural hospitals enrolled in the study compared to the jurisdiction-wide distribution of rural hospitals). The construct of the Gix allows such a comparison to be made and can inform an assessment of representativeness relative to the jurisdiction of interest. Third, the Gix methodology broadens the scope of 'similarity' beyond centre-to-centre comparisons and highlights the importance of comparing the characteristics of the sample of participating centres in the RCT to the jurisdiction of interest. As such, the focus of the generic research question in generalisability shifts from obtaining locally adjusted cost-effectiveness estimates to assessing the extent to which trial-wide results can be extrapolated to the jurisdiction level. Fourth, the proposed Gix demonstrated that trial results can vary in relation to centre-level characteristics. This is an advance in the current literature, which has been focusing on patient-level characteristics. Fifth, the approach can address both clinical and economic trial outcomes and can, thus, be extended to the generalisability of clinical effectiveness estimates. Finally, the case study demonstrated that within-country variations of trial clinical and cost-effectiveness results due to the sample of recruiting locations. This is significant as the majority of the contributions in the literature to date have focused on cross-country comparisons (section 7.2).

The research had several limitations. First, the limitations of the ROSSINI economic evaluation were discussed in detail in sub-section 5.2.3 and concerned the appropriateness of the 30-day time horizon, the amount of missing data and not capturing several resource use items such as wound dressings. Nevertheless, it was argued that it is unlikely that these limitations affected the final recommendation. Due consideration to collecting such information should be made in future SSI management studies.

Second, the mixed method study had a number of shortcomings, which were discussed in detail in sub-sections 7.2.4, 7.3.4 and 7.4.4. Of note, the systematic review only focused on publicly funded trials; the focus groups were conducted with trialists at a single institution; and the relatively small sample size of the online survey did not allow any inferential statistics to be computed. Nevertheless, the findings of the three methods were highly compatible both with one another and with previous literature findings (in relation to reporting of generalisability items), therefore it is unlikely that the emerging data lacked validity or relevance.

Third, the limitations in the development of the Gix and the simulation study were acknowledged in section 8.4. In particular, the choice of the centre-level variables (dimensions) included in the Gix and the methods of aggregating the dimensions in a single index can be a subject of debate. Nevertheless, the sensitivity analyses tested a wide range of scenarios concerning these choices and the results were generally robust to variations in the formulation of the Gix.

Of note, this research adopted a case study approach to demonstrate the usefulness of the Gix. As discussed in section 8.4, an alternative approach would have been to specify the characteristics of a hypothetical jurisdiction and the design of a hypothetical trial, then conduct the simulations (428). This would have allowed a fine control of study parameters at the expense of allowing more programming time and making supplementary assumptions. However, the two approaches are complementary and identifying bias due to centre selection in a completely hypothetical design would require validation using a real-world case study. In this case, an effect was identified using a real-world trial and the robustness of the finding deserves further exploration using a more complex simulation design.

Finally, in this research the methodological choice was to aggregate the relevant centre-level characteristics in a single generalisability metric (Gix). Such an approach has merits in that it allows head-to-head comparisons between centres and opens the way towards more informative study designs and trial analyses, as discussed above. However, it can be equally argued that the approach is reductionist because it attempts to capture centre-level variation, an obviously broad concept, into a single number. This sort of trade-off is common when attempting to measure complex processes. A relevant example at a larger scale is the controversy stirred by the publication of the World Health Report 2000 (433), which ranked national health systems according to their efficiency. The ethical, methodological and statistical underpinnings of the analysis came under close scrutiny and criticism (434). Among the lessons learned from the experience, of particular importance was the difficulty of capturing and representing all contextual factors relevant for health system performance in a composite measure that was, in addition, difficult to explain to the relevant public (435). As a matter of course, there is no substitute for judgement when reviewing and making policy decision based on contextual data. This challenge applies to the issue of assessing generalisability, as well. An alternative approach to the Gix would be presenting centre-level data in disaggregated form and allowing researchers and policy makers to make their own decision regarding the relative importance of centre-level domains to describe ‘common’ centres and outliers. Such a course of action deserves future scrutiny. Nevertheless, proposing the Generalisability index, however imperfect a measure, may prove to be beneficial if only by bringing the issue of generalisability at the forefront of academic research just as the World Health Report 2000 did for health systems performance evaluation (436) and, thus, catalyse advancements in this area.

9.4. Research in context

SSI remains a major concern for surgical patients. The most recent evidence concerning the interventions which may reduce SSI was summarised by NICE in an evidence update report published in June 2013 (437), which does not replace or supersede the current UK clinical guidelines (243). The use of WEPDs was thereby cited as a promising avenue where further research is needed and mentions the ROSSINI trial, but the document was published just before ROSSINI results were published and, thus, does not incorporate them. Of the 19 types of SSI reducing interventions reviewed, the majority either showed no effect or required further research, and only one (the use of antimicrobial-coated sutures vs. uncoated sutures for wound closure) has the potential to impact the current SSI guidance.

However, even if the effect of the potential interventions was convincingly demonstrated, it remains yet unclear whether strict adherence to current clinical protocols is likely to reduce SSI. For example, the Surgical Care Improvement Program (SCIP) was introduced in 2005 in US hospitals as a three-phase pay-for-performance mechanism for surgeons with the aim of reducing morbidity and mortality after surgery (438). However, Hawn *et al.* (439) analysed the relationship between the adherence to SCIP in 112 Veterans’ Affairs hospitals in the US (2005 to 2009) and SSI occurrence; they found that while adherence to SCIP improved gradually over time, risk-adjusted SSI rates remained constant and there was no association between SCIP adherence and either patient-level or hospital-level SSI rates, which suggests the need for a reassessment of current SSI-reduction policies.

Several directions for development have been proposed. First, it has been suggested that instead of a 'one size fits all' type of SSI prevention protocol (which both NICE and CDC currently adopt), an approach tailored to the characteristics of the relevant patient population may be more suitable, for example by recognising the role of body mass index as a risk factor (440). In addition, a strict adherence to the protocol may not be the only alternative: it has been demonstrated that adherence to individual SCIP measures is not associated with better SSI outcomes, but adherence to all SCIP measures (potentially a surrogate for superior coordination and team work) was (441). Such an interpretation is supported by the finding that a more empowering, collaborative interaction between the relevant healthcare professionals was effective in reducing SSI rates in colorectal patients (442). Second, the understanding of SSI pathophysiology may also require substantial updating: Lawson *et al.* (443) analysed 27,000 US patients and found that the risk factors for superficial and deep/organ-space SSIs differed in magnitude and significance. As such, the authors suggested that the two SSI categories could be viewed as two distinct disease processes, which may be instrumental for future research initiatives and practice guidelines. This is in line with the considerations discussed previously in sub-section 2.1.1, which highlighted the need for refining SSI definitions in the future.

In summary, the progress in SSI prevention has been slow to date and WEPDs do not appear to bring a favourable contribution. More importantly, however, the approach to SSI prevention is facing important challenges, as outlined above, and needs to incorporate recent research findings to produce cross-sectoral, personalised guidance to reduce SSI.

Such an approach echoes the wider interest towards acknowledging context-specific factors in order to ensure the generalisability of RCT findings. As it has been presented previously (sub-section 1.3.1), patient characteristics have captured most of the attention in this research area. The illustration of the Generalisability index (Gix) approach (Chapter 8) using a surgical RCT highlighted the fact that centre characteristics may play a major role in determining trial results and, thus, affecting their generalisability. I have no knowledge of other initiatives to quantify the representativeness of trial centres for generalisability purposes. There have been attempts, however, to isolate systematically contextual factors which may influence the findings of surgical trials. For example, Pibouleau *et al.* (444) surveyed 87 surgeons to prioritise hospital-specific factors which could influence the applicability of trial results of four orthopaedic procedures; the selected determinants were: the number of participating centres, the centres' surgical volume, the number of participating surgeons and the experience/training of surgeons. This is in accordance with the Gix formulation presented in Chapter 8, which incorporated measures of capacity and specialisation. However, the index did not consider practitioner-level information (e.g. surgeon-level characteristics). Such a development would add a further layer of complexity and may be the subject of further investigation.

As discussed in section 1.3, RCTs have several strong limitations with respect to generalisability. As far as economic evaluation results are concerned, updating pre-trial modelling results with trial data (69), potentially in a Bayesian framework (445), has been advocated as the appropriate approach. Under this paradigm, RCT findings can be viewed as merely an input in model-based economic evaluations. As such, other sources of information can also be integrated in the updated model to support the generalisability of findings. For instance, Freemantle and Hessel (446) argued that data from observational studies (e.g.

observational databases) may provide valuable information on current practice, especially on the characteristics of target population and local clinical processes, and can thus complement RCT results as inputs for economic models.

Incorporating contextual characteristics in the design and analysis of RCTs must become a priority for clinical research. However, such characteristics should not be merely acknowledged by way of reporting in trial protocols (15) and publications (27). As Bonell *et al.* (79) pointed out, researchers must also integrate process evaluations alongside RCTs and generate evidence-based theories on how contextual effects influence the intervention processes and how results may differ in other locations in order to enhance the generalisability of RCTs. Research funders, decision makers and journal editors can catalyse these advances by requiring and supporting increasingly complex and rigorous incorporation of generalisability issues in clinical research.

9.5. Future research

The research findings of this work and their limitations open several research directions. First, in terms of SSI reduction, Chapter 3 and Chapter 5 suggested that a head-to-head comparison between the existing WEPD designs would clarify whether a mechanical barrier between the wound margins and the exterior is indeed a valid SSI prevention strategy. However, in the light of current results, the commercial interests of the manufacturers would probably not be aligned with such a comparison Chapter 4 highlighted the paucity of available SSI health utility data and ROSSINI was only the first study to generate such evidence; future studies conducted in alternative settings must address this information gap. Furthermore, the discussion in Chapter 5 raised the need for more in-depth SSI management research to account for the complexities in patient care pathways beyond the 30-day time horizon.

Despite their apparent ineffectiveness, several important questions remain remained unanswered with regard to WEPDs, as well. For instance, WEPDs may still have a protective effect, but the pathogens responsible for SSI could be introduced to the wound when the device is not in place – either when opening the wound before the device is placed, or especially when closing the wound after the device has been removed, when gloves and instruments will clearly be dirtier. Evaluating WEPDs in conjunction with a changing gloves protocol before wound closure may thus be worth considering. It will also be interesting to compare ROSSINI findings with those of the upcoming BaFO trial (305), which has a similar pragmatic design.

Second, there is a need for better understanding the centre selection process. The mixed methods study (Chapter 7) could only provide a descriptive account for the purpose of the thesis, but it appears that more needs be known about how centre selection takes place in practice, what are the implications of overcrowding research in 'preferred' locations at the expense of neglecting others and how to ensure that all centres where an intervention is likely to be implemented have equal opportunities and capabilities to participate in research. Furthermore, the development of a theoretical framework for centre selection would guide researchers, practitioners and policy makers.

Third, the Gix requires further validation and refinement according to the priorities specified in section 8.4. It must be established whether the findings of this study can be replicated in other research areas and in other settings before considering a formal inclusion of the Gix approach alongside the available methodological tools addressing generalisability of trial results and, in particular, of trial-based economic evaluations.

9.6. Conclusion

The findings of this research suggest that WEPDs cannot be considered effective or cost-effective in reducing SSI after open abdominal surgery, therefore their use in the NHS cannot be recommended. More research on the impact of SSI on patient HRQoL and the cost of long-term care for SSI patients is warranted. This clinical context was used a case study to support a methodological investigation of the importance of centre selection for the generalisability of trial results. It was demonstrated that the characteristics of the sample of participating centres can influence trial results and can affect the generalisability of clinical and economic findings to the jurisdiction where the trial recruited from. Such a result may be of interest to researchers, health policy makers, funders and research commissioners. The Generalisability index could be a valuable tool in quantifying this type of bias and in designing RCTs with superior external validity. The robustness of the findings across therapeutic areas, clinical settings, geographic locations and formulations of the Gix is a fertile subject for further research.

LIST OF REFERENCES

1. Piantadosi S. Clinical Trials: A Methodologic Perspective. 2 ed. Hoboken, N.J: Wiley-Interscience; 2005.

2. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. ICH E9 - Statistical Principles for Clinical Trials. 1998 Accessed 2012/06/28. Available from: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf.

3. Calvert M, Wood J, Freemantle N. Designing "Real-World" trials to meet the needs of health policy makers at marketing authorization. Journal of Clinical Epidemiology. 2011;64(7):711-7.

4. Williamson P, Altman D, Blazeby J, Clarke M, Gargon E. The COMET (Core Outcome Measures in Effectiveness Trials) Initiative. Trials. 2011;12(Suppl 1):A70.

5. Bland M. An Introduction to Medical Statistics. 3rd ed. Oxford: Oxford University Press; 2000.

6. Campbell MJ, Julious SA, Altman DG. Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. BMJ. 1995;311(7013):1145-8.

7. Higgins JPT, Green S. Cochrane Handbook of Systematic Reviews of Interventions. Version 5.0.1 [updated September 2008]. Available online at www.cochrane-handbook.org: The Cochrane Collaboration; 2008.

8. Mills E, Chan A-W, Wu P, Vail A, Guyatt G, Altman D. Design, analysis, and presentation of crossover trials. Trials. 2009;10(1):27.

9. Senn S. Cross-over Trials in Clinical Research. 2nd edition ed. Chichester: John Wiley & Sons; 2002.

10. Montgomery A, Peters T, Little P. Design, analysis and presentation of factorial randomised controlled trials. BMC Medical Research Methodology. 2003;3(1):26.

11. Murray DM, Varnell SP, Blitstein JL. Design and Analysis of Group-Randomized Trials: A Review of Recent Methodological Developments. American Journal of Public Health. 2004;94(3):423-32.

12. Kairalla J, Coffey C, Thomann M, Muller K. Adaptive trial designs: a review of barriers and opportunities. Trials. 2012;13(1):145.

13. Gallo P, Chuang-Stein C, Dragalin V, Gaydos B, Krams M, Pinheiro J. Adaptive Designs in Clinical Drug Development—An Executive Summary of the PhRMA Working Group. Journal of Biopharmaceutical Statistics. 2006;16(3):275-83.

14. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. ICH E6 - Guideline for Good Clinical Practice. 1996 Accessed 2012/06/29. Available from: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E6_R1/Step4/E6_R1__Guideline.pdf.

15. Chan A-W, Tetzlaff JM, Altman DG, Laupacis A, Gøtzsche PC, Krleža-Jerić K, et al. SPIRIT 2013 Statement: Defining Standard Protocol Items for Clinical Trials. Annals of Internal Medicine. 2013;158(3):200-7.

16. Douglas GA, Bland JM. Missing data. BMJ. 2007;334.

17. Carpenter JR, Kenward M. Missing data in randomised controlled trials — a practical guide. 2008 Accessed 23/09/2013. Available from: http://www.hta.nhs.uk/nihrmethodology/reports/1589.pdf.

18. Molenberghs G, Kenward M. Missing Data in Clinical Studies. West Sussex: Wiley; 2007.

19. Fielding S, Maclennan G, Cook J, Ramsay C. A review of RCTs in four medical journals to assess the use of imputation to overcome missing data in quality of life outcomes. Trials. 2008;9(1):51.

20. Little RJA, Rubin DB. Statistical Analysis With Missing Data. New York: Wiley; 1987.
21. Schafer JL, Graham JW. Missing data: Our view of the state of the art. Psychologial Methods. 2002;7(2):147-77.
22. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: Wiley; 1987.
23. Graham JW, Schafer JL. On the performance of multiple imputation for multivariate data with small sample size. In: Hoyle R, editor. Statistical strategies for small sample research. Thousand Oaks, CA: Sage; 1999.
24. Bodner TE. What Improves with Increased Missing Data Imputations? Structural Equation Modelling. 2008;15(4):651-75.
25. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338.
26. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine. 2011;30(4):377-99.
27. Schulz KF, Altman DG, Moher D. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c322.
28. Campbell MK, Piaggio G, Elbourne DR, Altman DG. Consort 2010 statement: extension to cluster randomised trials. BMJ. 2012;345.
29. Zwarenstein M, Treweek S, Gagnier J, Altman DG, Tunis S, Haynes B, et al. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. BMJ. 2008;337.
30. Calvert M, Blazeby J, Altman D, et al. Reporting of patient-reported outcomes in randomized trials: The consort pro extension. JAMA. 2013;309(8):814-22.
31. Drummond MF, Sculpher MJ, Torrance GW, O'Brien BJ, Stoddart GL. Methods for the Economic Evaluation of Health Care Programmes. 2 ed. Oxford: Oxford University Press; 2005.
32. Brouwer WBF, Culyer AJ, van Exel NJA, Rutten FFH. Welfarism vs. extra-welfarism. Journal of Health Economics. 2008;27(2):325-38.
33. Cohen GA. Equality of what? On welfare, goods and capabilities. In: Nussbaum MC, Sen AK, editors. The Quality of Life. Oxford: Clarendon Press; 1993.
34. Culyer AJ. Commodities, characteristics of commodities, characteristics of people, utilities, and the quality of life. In: Baldwin S, editor. Quality of life Perspectives and policies. London: Routledge; 1990. p. 9-27.
35. Culyer AJ. The normative economics of health care finance and provision. Oxford Review of Economic Policy. 1989;5(1):34-56.
36. Birch S, Donaldson C. Valuing the benefits and costs of health care programmes: Where's the 'extra' in extra-welfarism? Social Science and Medicine. 2003;56(5):1121-33.
37. Williams A. The cost-benefit approach. British Medical Bulletin. 1974;30:252-6.
38. Mooney G. Economics, medicine and health care. Hemel Hempstead: Wheatsheaf; 1992.
39. National Institute for Health and Clinical Excellence. Guide to the Methods of Technology Appraisal 2013. London: NICE; 2013.
40. Gafni A, Birch S. Incremental cost-effectiveness ratios (ICERs): The silence of the lambda. Social Science & Medicine. 2006;62(9):2091-100.
41. Torrance GW, Thomas WH, Sackett DL. A utility maximization model for evaluation of health care programs. Health Services Research. 1972;7:118-33.
42. Fanshel S, Bush JW. A health status index and its applications to health services outcomes. Operations Research. 1970;18(6):1021-66.
43. The EuroQol Group. EuroQol - a new facility for the measurement of health-related quality of life. Health Policy. 1990;16(3):199-208.
44. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. Journal of Health Economics. 2002;21(2):271-92.
45. Torrance GWFDHF, W.J.; Barr, R.D.; Zhang, Y.; Wang, Q.;. Multiattribute Utility Function for a Comprehensive Health Status Classification System: Health Utilities Index Mark 2. Medical Care. 1996;34(7):702-22.

46. Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: A Quality-of-Life Instrument for Use in International Clinical Trials in Oncology. Journal of the National Cancer Institute. 1993;85(5):365-76.

47. Black WC. The CE plane: a graphic representation of cost-effectiveness. Medical Decision Making. 1990;10(3):212-4.

48. McCabe C, Claxton K, Culyer AJ. The NICE Cost-Effectiveness Threshold: What it is and What that Means. Pharmacoeconomics. 2008;26(9):733-44.

49. Rawlins M, Barnett D, Stevens A. Pharmacoeconomics: NICE's approach to decision-making. British Journal of Clinical Pharmacology. 2009;70(3):346-9.

50. Briggs A, Fenn P. Confidence Intervals or Surfaces? Uncertainty on the Cost-effectiveness Plane. Health Economics. 1998;7(8):723-40.

51. Briggs A, Wonderling DE, Mooney CZ. Pulling cost-effectiveness analysis up by its bootstraps: a non-parametric approach to confidence interval estimation. Health Economics. 1997;6(4):327-40.

52. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. London: Chapman and Hall; 1993.

53. Efron B. Better Bootstrap Confidence Intervals. Journal of the American Statistical Association. 1987;82(397):171-85.

54. Briggs AH, Mooney CZ, Wonderling DE. Constructing confidence intervals for cost-effectiveness ratios: an evaluation of parametric and non-parametric techniques using Monte Carlo simulation. Statistics in Medicine. 1999;18(23):3245-62.

55. Stinnett AA, Mullahy J. Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis. Medical decision making : an international journal of the Society for Medical. 1998;18(2 Suppl):S68-S80.

56. Van Hout BA, Al MJ, Gordon GS, Rutten FFH. Costs, effects and C/E-ratios alongside a clinical trial. Health Economics. 1994;3(5):309-19.

57. Lothgren M, Zethraeus N. Definition, interpretation and calculation of cost-effectiveness acceptability curves. Health Economics. 2000;9(7):623-30.

58. Fenwick E, O'Brien BJ, Briggs A. Cost-effectiveness acceptability curves - facts, fallacies and frequently asked questions. Health Economics. 2004;13(5):405-15.

59. Groot Koerkamp B, Hunink MGM, Stijnen T, Hammitt JK, Kuntz KM, Weinstein MC. Limitations of Acceptability Curves for Presenting Uncertainty in Cost-Effectiveness Analysis. Medical Decision Making. 2007;27(2):101-11.

60. Barton GR, Briggs A, Fenwick E. Optimal cost-effectiveness decisions: the role of the cost-effectiveness acceptability curve (CEAC), the cost-effectiveness acceptability frontier (CEAF), and the expected value of perfection information (EVPI). Value in Health. 2008;11(5):886-97.

61. Jakubczyk M, Kamiński B. Cost-effectiveness acceptability curves – caveats quantified. Health Economics. 2010;19(8):955-63.

62. Glick HA, Doshi JA, Sonnad SS, Polsky D. Economic Evaluation in Clinical Trials. Oxford: Oxford University Press; 2007.

63. Petrou S, Gray A. Economic evaluation alongside randomised controlled trials: design, conduct, analysis, and reporting. BMJ. 2011;342.

64. Gray AM, Clarke P, Wolstenholme J, Wordsworth S. Applied Methods of Cost-effectiveness Analysis in Health Care. New York: Oxford University Press; 2011.

65. Polsky D, Glick H. Costing and Cost Analysis in Randomised Trials: Caveat Emptor. Pharmacoeconomics. 2009;27(3):179-88.

66. Briggs AH, Claxton K, Sculpher M. Decision Modelling for Health Economic Evaluation. Gray A, Briggs AH, editors. Oxford: Oxford University Press; 2006.

67. Brennan A, Chick SE, Davies R. A taxonomy of model structures for economic evaluation of health technologies. Health Economics. 2006;15(12):1295-310.

68. Barton P, Bryan S, Robinson S. Modelling in the economic evaluation of health care: selecting the appropriate approach. Journal of Health Services Research & Policy. 2004;9(2):110-8.

69. Sculpher MJ, Claxton K, Drummond M, McCabe C. Whither trial-based economic evaluation for health care decision making? Health Economics. 2006;15(7):677-87.
70. Miettinen OS. Theoretical epidemiology. New York: J. Wiley; 1985.
71. Juni P, Altman AD, Egger M. Assessing the quality of controlled trials. BMJ. 2001;323(7303):42-6.
72. Verhagen AP, de Vet HCW, de Bie RA, Kessels AGH, Boers M, Bouter LM, et al. The Delphi List: A Criteria List for Quality Assessment of Randomized Clinical Trials for Conducting Systematic Reviews Developed by Delphi Consensus. Journal of Clinical Epidemiology. 1998;51(12):1235-41.
73. Freedman B. Scientific Value and Validity as Ethical Requirements for Research: A Proposed Explication. IRB: Ethics and Human Research. 1987;9(6):7-10.
74. Campbell DT. Factors relevant to the validity of experiments in social settings. 1957. 1957;54(4):297-312.
75. Murphy EA. The logic of medicine. Baltimore: Johns Hopkins University Press; 1976.
76. Sackett DL. Bias in analytical research. J Chron Dis. 1979;32:51-63.
77. Dekkers OM, Elm Ev, Algra A, Romijn JA, Vandenbroucke JP. How to assess the external validity of therapeutic trials: a conceptual approach. International Journal of Epidemiology. 2009.
78. Rothwell PM. Treating Individuals 1 - External validity of randomised controlled trials: "To whom do the results of this trial apply?"'. Lancet. 2005;365(9453):82-93.
79. Bonell C, Oakley A, Hargreaves J, Strange V, Rees R. Assessment of generalisability in trials of health interventions: suggested framework and systematic review. BMJ. 2006;333(7563):346-9.
80. Bornhoft G, Maxion-Bergemann S, Wolf U, Kienle G, Michalsen A, Vollmar H, et al. Checklist for the qualitative evaluation of clinical studies with particular focus on external validity and model validity. BMC Medical Research Methodology. 2006;6(1):56.
81. Flather M, Delahunty N, Collinson J. Generalizing results of randomized trials to clinical practice: reliability and cautions. Clinical Trials. 2006;3(6):508-12.
82. Treweek S, Zwarenstein M. Making trials matter: pragmatic and explanatory trials and the problem of applicability. Trials. 2009;10:37.
83. Black N. Why we need observational studies to evaluate the effectiveness of health care. BMJ. 1996;312.
84. Silverman SL. From randomized controlled trials to observational studies. American Journal of Medicine. 2009;122(2):114-20.
85. Croft P, Malmivaara A, Van Tulder MW. The pros and cons of evidence-based medicine. Spine. 2011;36(17):E1121-5.
86. Nallamothu BK, Hayward RA, Bates ER. Beyond the Randomized Clinical Trial: The Role of Effectiveness Studies in Evaluating Cardiovascular Therapies. Circulation. 2008;118(12):1294-303.
87. Moore DA, Goodall RL, Ives NJ, Hooker M, Gazzard BG, Easterbrook PJ. How generalizable are the results of large randomized controlled trials of antiretroviral therapy? HIV Medicine. 2000;1(3):149-54.
88. Hoel AW, Kayssi A, Brahmanandam S, Belkin M, Conte MS, Nguyen LL. Under-representation of women and ethnic minorities in vascular surgery randomized controlled trials. Journal of Vascular Surgery. 2009;50(2):349-54.
89. Kalata P, Martus P, Zettl H, Rodel C, Hohenberger W, Raab R, et al. Differences between clinical trial participants and patients in a population-based registry: the German Rectal Cancer Study vs. the Rostock Cancer Registry. Diseases of the Colon & Rectum. 2009;52(3):425-37.
90. Maasland L, van Oostenbrugge RJ, Franke CF, Scholte Op Reimer WJ, Koudstaal PJ, Dippel DW, et al. Patients enrolled in large randomized clinical trials of antiplatelet treatment for prevention after transient ischemic attack or ischemic stroke are not representative of patients in clinical practice: the Netherlands Stroke Survey. Stroke. 2009;40(8):2662-8.
91. Costa DJ, Amouyal M, Lambert P, Ryan D, Schunemann HJ, Daures JP, et al. How representative are clinical study patients with allergic rhinitis in primary care? Journal of Allergy & Clinical Immunology. 2011;127(4):920-6.

92. Falagas ME, Vouloumanou EK, Sgouros K, Athanasiou S, Peppas G, Siempos II. Patients included in randomised controlled trials do not represent those seen in clinical practice: focus on antimicrobial agents. International Journal of Antimicrobial Agents. 2010;36(1):1-13.
93. Masoudi FA, Havranek EP, Wolfe P, Gross CP, Rathore SS, Steiner JF, et al. Most hospitalized older persons do not meet the enrollment criteria for clinical trials in heart failure. American Heart Journal. 2003;146(2):250-7.
94. Hestbech MS, Siersma V, Dirksen A, Pedersen JH, Brodersen J. Participation bias in a randomised trial of screening for lung cancer. Lung Cancer. 2011;73(3):325-31.
95. Bartlett C, Doyal L, Ebrahim S, Davey P, Bachmann M, Egger M, et al. The causes and effects of socio-demographic exclusions from clinical trials. Health Technology Assessment. 2005;9(38):1-152.
96. Steg P, López-Sendón J, Lopez de Sa E, et al. External validity of clinical trials in acute myocardial infarction. Archives of Internal Medicine. 2007;167(1):68-73.
97. Halm EA, Lee C, Chassin MR. Is Volume Related to Outcome in Health Care? A Systematic Review and Methodologic Critique of the Literature. Annals of Internal Medicine. 2002;137(6):511-20.
98. Devereaux PJ, Mohit B, Mike C, Victor MM, Deborah JC, Salim Y, et al. Need for expertise based randomised controlled trials. BMJ. 2005;330.
99. National Institute for Health and Clinical Excellence. IPG322 Negative pressure wound therapy for the open abdomen: guidance 2009 Accessed 2012/11/23. Available from: http://guidance.nice.org.uk/IPG322/Guidance/pdf/English.
100. Britton A, McKee M, Black N, McPherson K, Sanderson C, Bain C. Threats to applicability of randomised trials: exclusions and selective participation. Journal of Health Services & Research Policy. 1999;4(2):112-21.
101. DeLong ER, Coombs LP, Ferguson TB, Peterson ED. The Evaluation of Treatment When Center-Specific Selection Criteria Vary with Respect to Patient Risk. Biometrics. 2005;61(4):942-9.
102. Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials. J Chron Dis. 1967;20:637-48.
103. Tunis SR, Stryer DB, Clancy CM. Practical Clinical Trials. JAMA: The Journal of the American Medical Association. 2003;290(12):1624-32.
104. Glasgow RE, Magid DJ, Beck A, Ritzwoller D, Estabrooks PA. Practical clinical trials for translating research to practice: design and measurement recommendations. Medical Care. 2005;43(6):551-7.
105. Roland M, Torgerson DJ. Understanding controlled trials: What are pragmatic trials? BMJ. 1998;316.
106. MacPherson H. Pragmatic clinical trials. Complementary Therapies in Medicine. 2004;12:136-40.
107. Williamson P, Altman D, Blazeby J, Clarke M, Devane D, Gargon E, et al. Developing core outcome sets for clinical trials: issues to consider. Trials [Electronic Resource]. 2012;13(1):132.
108. Karanicolas PJ, Montori VM, Schünemann HJ, Guyatt GH. "Pragmatic" clinical trials: from whose perspective? Annals of Internal Medicine. 2009;150(12):JC6-2.
109. Thorpe KE, Zwarenstein M, Oxman AD, Treweek S, Furberg CD, Altman DG, et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. Canadian Medical Association Journal. 2009;180(10):E47-E57.
110. Kent D, Kitsios G. Against pragmatism: on efficacy, effectiveness and the real world. Trials. 2009;10(1):48.
111. Andrew E, Anis A, Chalmers T, et al. A proposal for structured reporting of randomized controlled trials. JAMA. 1994;272(24):1926-31.
112. Kane RL, Wang J, Garrard J. Reporting in randomized clinical trials improved after adoption of the CONSORT statement. Journal of Clinical Epidemiology. 2007;60(3):241-9.
113. Plint AC, Moher D, Morrison A, Schulz KF, Altman DG, Hill C, et al. Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. Med J Aust. 2006;185(5):263-7.

114. Moher D, Jones A, Lepage L, for the Consort Group. Use of the consort statement and quality of reports of randomized trials: A comparative before-and-after evaluation. JAMA. 2001;285(15):1992-5.
115. Turner L, Shamseer L, Altman Douglas G, Weeks L, Peters J, Kober T, et al. Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. Cochrane Database of Systematic Reviews. 2012(11):MR000030.
116. Boutron I, Moher D, Altman DG, Schulz KF, Ravaud P, for the CG. Extending the CONSORT Statement to Randomized Trials of Nonpharmacologic Treatment: Explanation and Elaboration. Annals of Internal Medicine. 2008;148(4):295-309.
117. Braslow JT, Duan N, Starks SL, Polo A, Bromley E, Wells KB. Generalizability of studies on mental health treatment and outcomes, 1981 to 1996. Psychiatric Services. 2005;56(10):1261-8.
118. Ahmad N, Boutron I, Dechartres A, Durieux P, Ravaud P. Applicability and generalisability of the results of systematic reviews to public health practice and policy: a systematic review. Trials. 2010;11(20).
119. Eldridge S, Ashby D, Bennett C, Wakelin M, Feder G. Internal and external validity of cluster randomised trials: systematic review of recent trials. BMJ. 2008;336(7649):876-80.
120. Glasgow RE, Lichtenstein E, Marcus AC. Why Don’t We See More Translation of Health Promotion Research to Practice? Rethinking the Efficacy-to-Effectiveness Transition. Am J Public Health. 2003;93(8):1261-7.
121. Green LW, Glasgow RE. Evaluating the relevance, generalization, and applicability of research: issues in external validation and translation methodology. Eval Health Prof. 2006;29(1):126-53.
122. Steckler A, McLeroy KR. The importance of external validity. Am J Public Health. 2008;98(1):9-10.
123. Glasziou P, Meats E, Heneghan C, Shepperd S. What is missing from descriptions of treatment in trials and reviews? BMJ. 2008;336.
124. Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. American Journal of Epidemiology. 2012;172(1):107-15.
125. Roy AS. Stifling new cures: The True Cost of Lengthy Clinical Drug Trials. Project FDA Report [Internet]. 2012 Accessed 2013/26/08. Available from: http://www.manhattan-institute.org/html/fda_05.htm.
126. Collier R. Rapidly rising clinical trial costs worry researchers. CMAJ. 2009;180(3):277-8.
127. Sculpher MJ, Pang FS, Manca A, Drummond MF, Golder S, Urdahl H, et al. Generalisability in economic evaluation studies in healthcare: a review and case studies. Health Technology Assessment. 2004;8(49).
128. Kravitz RL, Duan N, Braslow J. Evidence-Based Medicine, Heterogeneity of Treatment Effects, and the Trouble with Averages. Milbank Quarterly. 2004;82(4):661-87.
129. Kraemer H, Frank E, Kupfer D. Moderators of treatment outcomes: Clinical, research, and policy importance. JAMA: The Journal of the American Medical Association. 2006;296(10):1286-9.
130. Drummond MF, Bloom BS, Carrin G, Hillman AL, Hutchings HC, Knill-jones RP, et al. Issues in the Cross-National Assessment of Health Technology. International Journal of Technology Assessment in Health Care. 1992;8(04):670-82.
131. O'Brien BJ. A tale of two (or more) cities: Geographic transferability of pharmacoeconomic data. American Journal of Managed Care. 1997;3:S33-S9.
132. Boulenger S, Nixon J, Drummond M, Ulmann P, Rice S, de Pouvourville G. Can economic evaluations be made more transferable? European Journal of Health Economics. 2005;6(4):334-46.
133. Barbieri M, Drummond M, Rutten F, Cook J, Glick HA, Lis J, et al. What Do International Pharmacoeconomic Guidelines Say about Economic Data Transferability? Value in Health. 2010;13(8):1028-37.
134. Drummond M, Barbieri M, Cook J, Glick HA, Lis J, Malik F, et al. Transferability of economic evaluations across jurisdictions: ISPOR Good Research Practices Task Force report. Value in Health. 2009;12(4):409-18.

135. Welte R, Feenstra T, Jager H, Leidl R. A decision chart for assessing and improving the transferability of economic evaluation results between countries. Pharmacoeconomics. 2004;22(13):857-76.
136. Goeree R, Burke N, O'Reilly D, Manca A, Blackhouse G, Tarride JE. Transferability of economic evaluations: approaches and factors to consider when using results from one geographic area for another. . Current Medical Research & Opinion. 2007;23(4):671-82.
137. Barbieri M, Drummond M, Willke R, Chancellor J, Jolain B, Towse A. Variability of cost-effectiveness estimates for pharmaceuticals in Western Europe: Lessons for inferring generalizability. Value in Health. 2005;8(1):10-23.
138. Ayanian JZ, Weissman JS. Teaching Hospitals and Quality of Care: A Review of the Literature. Milbank Quarterly. 2002;80(3):569-93.
139. Goeree R, He J, O'Reilly D, Tarride J-E, Xie F, Lim M, et al. Transferability of health technology assessments and economic evaluations: a systematic review of approaches for assessment and application. ClinicoEconomics and Outcomes Research. 2011;3(1):89-104.
140. Grieve R, Cairns J, Thompson SG. Improving costing methods in multicentre economic evaluation: the use of multiple imputation for unit costs. Health Economics. 2010;19(8):939-54.
141. Raikou M, Briggs A, Gray A, McGuire A. Centre-specific or average unit costs in multi-centre studies? Some theory and simulation. Health Economics. 2000;9(3):191-8.
142. Ridyard CH, Hughes DA. Methods for the Collection of Resource Use Data within Clinical Trials: A Systematic Review of Studies Funded by the UK Health Technology Assessment Program. Value in Health. 2010;13(8):867-72.
143. Kupersmith J. Quality of Care in Teaching Hospitals: A Literature Review. Academic Medicine. 2005;80(5):458-66.
144. Campbell SM, Hann M, Hacker J, Burns C, Oliver D, Thapar A, et al. Identifying predictors of high quality care in English general practice: observational study. BMJ. 2001;323.
145. Gutacker N, Bojke C, Daidone S, Devlin N, Street A. Analysing Hospital Variation in Health Outcomeat the Level of EQ-5D Dimensions: CHE Research Paper 74 2012 Accessed 11/04/2013. Available from: http://www.york.ac.uk/media/che/documents/papers/researchpapers/CHERP74_analysing_hospital_variation_health_outcome_EQ-5D.pdf.
146. Greiner W, Weijnen T, Nieuwenhuizen M, Oppe S, Badia X, Busschbach J, et al. A single European currency for EQ-5D health states. European Journal of Health Economics. 2003;4(3):222-31.
147. Knies S, Evers SAA, Candel MJM, Severens J, Ament AHA. Utilities of the EQ-5D: transferable or not? Pharmacoeconomics. 2009;27(9):767-79.
148. Oppong R, Kaambwa B, Nuttall J, Hood K, Smith R, Coast J. The impact of using different tariffs to value EQ-5D health state descriptions: an example from a study of acute cough/lower respiratory tract infections in seven countries. The European Journal of Health Economics. 2013;14(2):197-209.
149. Sakthong P, Charoenvisuthiwongs R, Shabunthom R. A comparison of EQ-5D index scores using the UK, US, and Japan preference weights in a Thai sample with type 2 diabetes. Health and Quality of Life Outcomes. 2008;6(1):71.
150. Johnson JA, Luo N, Shaw JW, Kind P, Coons SJ. Valuations of EQ-5D Health States: Are the United States and United Kingdom Different? Medical Care. 2005;43(3):221-8.
151. Heyland DK, Kernerman P, Gafni A, Cook DJ. Economic evaluations in the critical care literature: do they help us improve the efficiency of our unit? Critical Care Medicine1996;24(9):1591-8.
152. Späth H-M, Carrère M-O, Fervers B, Philip T. Analysis of the eligibility of published economic evaluations for transfer to a given health care system: Methodological approach and application to the French health care system. Health Policy. 1999;49(3):161-77.
153. Chase D, Rosten C, Turner S, Hicks NJ, Milne R. Development of a toolkit and glossary to aid in the adaptation of health technology assessment (HTA) reports for use in different contexts. Health Technology Assessment. 2009;13(59).

154. Antonanzas F, Rodriguez-Ibeas R, Juarez C, Hutter F, Lorente R, Pinillos M. Transferability indices for health economic evaluations: methods and applications. Health Economics. 2009;18(6):629-43.
155. Manca A, Sculpher MJ, Goeree R. The analysis of multinational cost-effectiveness data for reimbursement decisions: a critical appraisal of recent methodological developments. Pharmacoeconomics. 2010;28(12):1079-96.
156. Cook JR, Drummond M, Glick H, Heyse JF. Assessing the appropriateness of combining economic data from multinational clinical trials. Statistics in Medicine. 2003;22(12):1955-76.
157. Gail M, Simon R. Testing for Qualitative Interactions Between Treatment Effects and Patient Subsets. Biometrics. 1985;41(2):361-72.
158. Piantadosi S, Gail M. A comparison of the power of two tests for qualitative interactions. Statistics in Medicine. 1993;12(13):1239-48.
159. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. The Lancet. 2000;355(9209):1064-9.
160. Gunter L, Zhu J, Murphy S. Variable selection for qualitative interactions in personalized medicine while controlling the family-wise error rate. Journal of Biopharmaceutical Statistics. 2011;21(6):1063-78.
161. Coyle D, Drummond MF. Analyzing differences in the costs of treatment across centers within economic evaluations. International Journal of Technology Assessment in Health Care. 2001;17(2):155-63.
162. Willke RJ, Glick HA, Polsky D, Schulman K. Estimating country-specific cost-effectiveness from multinational clinical trials. Health Economics. 1998;7(6):481-93.
163. Willan AR, Briggs AH, Hoch JS. Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data. Health Economics. 2004;13(5):461-75.
164. Manca A, Rice N, Sculpher MJ, Briggs AH. Assessing generalisability by location in trial-based cost-effectiveness analysis: the use of multilevel models. Health Economics. 2005;14(5):471-85.
165. Hoch JS, Briggs AH, Willan AR. Something old, something new, something borrowed, something blue: a framework for the marriage of health econometrics and cost-effectiveness analysis. Health Economics. 2002;11(5):415-30.
166. Nixon RM, Thompson SG. Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. Health Economics. 2005;14(12):1217-29.
167. Manca A, Lambert PC, Sculpher M, Rice N. Cost-effectiveness analysis using data from multinational trials: the use of bivariate hierarchical modeling. Medical Decision Making. 2007;27(4):471-90.
168. Draper D, Hodges J, Mallows C, Pregibon D. Exchangeability and Data analysis. Journal of the Royal Statistical Society, series A. 1993;156:9-37.
169. Bernardo J, Smith A. Bayesian Statistics. New York: John Wiley and Sons; 1994.
170. Drummond M, Manca A, Sculpher M. Increasing the generalizability of economic evaluations: Recommendations for the design, analysis, and reporting of studies. International Journal of Technology Assessment in Health Care. 2005;21(2):165-71.
171. Street A, Maynard A. Activity based financing in England: the need for continual refinement of payment by results. Health Economics, Policy, & Law. 2007;2(Pt:4):4-27.
172. Malcomson JM. Hospital cost differences and payment by results. Health Economics, Policy, & Law. 2007;2(Pt:4):4-33.
173. Laudicella M, Olsen KR, Street A. Examining cost variation across hospital departments-a two-stage multi-level approach using patient-level data. Social Science and Medicine. 2010;71(10):November.
174. Cookson R, Laudicella M. Do the poor cost much more? The relationship between small area income deprivation and length of stay for elective hip replacement in the English NHS from 2001 to 2008. Social Science & Medicine. 2011;72(2):173-84.
175. Manca A, Willan AR. 'Lost in translation': accounting for between-country differences in the analysis of multinational cost-effectiveness data. Pharmacoeconomics. 2006;24(11):1101-19.

176. Glasziou PP, Simes RJ, Hall J, Donaldson C. Design of a cost-effectiveness study within a randomized trial: The LIPID trial for secondary prevention of IHD. Controlled Clinical Trials. 1997;18(5):464-76.

177. Parliament of the United Kingdom. Health and Social Care Act 2012 (c.7). London: The Stationery Office; 2012.

178. World Health Organization. Prevention of hospital-acquired infections: a practical guide. Geneva: WHO; 2002.

179. World Health Organization. Report on the Burden of Endemic Health Care-Associated Infection Worldwide. Geneva: WHO; 2011.

180. Smyth ETM, McIlvenny G, Enstone JE, Emmerson AM, Humphreys H, Fitzpatrick F, et al. Four country healthcare associated infection prevalence survey 2006: overview of the results. J Hosp Infect. 2008;69(3):230-48.

181. Leaper DJ, Van Goor H, Reilly J, Petrosillo N, Geiss HK, Torres AJ, et al. Surgical site infection - a European perspective of incidence and economic burden. International Wound Journal. 2004;1(4):247-73.

182. Allegranzi B, Nejad SB, Combescure C, Graafmans W, Attar H, Donaldson L, et al. Burden of endemic health-care-associated infection in developing countries: systematic review and meta-analysis. Lancet. 2011;377:228-41.

183. Gaynes RP, Culver DH, Horan TC, Edwards JR, Richards C, Tolson JS, et al. Surgical Site Infection (SSI) Rates in the United States, 1992-1998: The National Nosocomial Infections Surveillance System Basic SSI Risk Index. Clinical Infectious Diseases. 2001;33:S69-S77.

184. HELICS. Surveillance of Surgical Site Infections: Surgical Site Infections 2004 2006 [Accessed 2010/11/05]. Available from: http://www.ecdc.europa.eu/IPSE/helicshome.htm.

185. Soleto LBS, Pirard MMDMPH, Boelaert MMDP, Peredo RMD, Vargas RMD, Gianella AMD, et al. Incidence of Surgical-Site Infections and the Validity of the National Nosocomial Infections Surveillance System Risk Index in a General Surgical Ward in Santa Cruz, Bolivia • Infection Control and Hospital Epidemiology. 2003;24(1):26-30.

186. Wassef MAH, A.; Abdul Rahman, E.M.; El-Sherif, R.H. A prospective surveillance of surgical site infections: study for efficacy of preoperative antibiotic prophylaxis. African Journal of Microbiological Research. 2012;6(12):3072-8.

187. Hafez S, Saied T, Hasan E, Elnawasany M, Ahmad E, Lloyd L, et al. Incidence and modifiable risk factors of surveillance of surgical site infections in Egypt: A prospective study. American Journal of Infection Control. 2012;40(5):426-30.

188. de Oliveira AC, Ciosak SI, Ferraz EM, Grinbaum RS. Surgical site infection in patients submitted to digestive surgery: Risk prediction and the NNIS risk index. American Journal of Infection Control. 2006;34(4):201-7.

189. Mawalla B, Mshana S, Chalya P, Imirzalioglu C, Mahalu W. Predictors of surgical site infections among patients undergoing major surgery at Bugando Medical Centre in Northwestern Tanzania. BMC Surgery. 2011;11(1):21.

190. de Lissovoy G, Fraeman K, Hutchins V, Murphy D, Song D, Vaughn BB. Surgical site infection: incidence and impact on hospital utilization and treatment costs. American Journal of Infection Control. 2009;37(5):387-97.

191. Coello R, Charlett A, Wilson J, Ward V, Pearson A, Borriello P. Adverse impact of surgical site infections in English hospitals. J Hosp Infect. 2005;60(2):93-103.

192. Health Protection Agency. Surveillance of surgical site infections in NHS hospitals in England, 2010/2011. London: Health Protection Agency, 2011.

193. European Centre for Disease Prevention and Control. Annual Epidemiological Report 2011. Reporting on 2009 surveillance data and 2010 epidemic intelligence data. Stockholm: ECDC; 2011.

194. Smith RL, Bohl JK, McElearney ST, Friel CM, Barclay MM, Sawyer RG, et al. Wound infection after elective colorectal resection. Annals of surgery. 2004;239(5):599-607.

195. Tanner J, Khan D, Aplin C, Ball J, Thomas M, Bankart J. Post-discharge surveillance to identify colorectal surgical site infection rates and related costs. J Hosp Infect. 2009;72(3):243-50.

196. Blumetti J, Luu M, Sarosi G, Hartless K, McFarlin J, Parker B, et al. Surgical site infections after colorectal surgery: Do risk factors vary depending on the type of infection considered? Surgery. 2007;142(5):704-11.
197. Howard DPJ, Datta G, Cunnick G, Gatzen C, Huang A. Surgical site infection rate is lower in laparoscopic than open colorectal surgery. Colorectal Disease. 2010;12(5):423-7.
198. Serra-Aracil X, Espin-Basany E, Biondo S, Guirao X, Orrego C, Sitges-Serra A. SUrgical site infection in elective operations for colorectal cancer after the application of preventive measures. Archives of Surgery. 2011;146(5):606-12.
199. Horan TC, Gaynes RP, Martone WJ, Jarvis WR, Emori TG. CDC Definitions of Nosocomial Surgical Site Infections, 1992: A Modification of CDC Definitions of Surgical Wound Infections. Infection Control and Hospital Epidemiology. 1992;13(10):606-8.
200. Centers for Disease Control and Prevention. Guideline for Prevention of Surgical Site Infection. Infection Control and Hospital Epidemiology. 1999;20(4):247-78.
201. Garner JS, Jarvis WR, Emori TG, Horan TC, Hughes JM. CDC definitions for nosocomial infections, 1988. American Journal of Infection Control. 1988;16(3):128-40.
202. Bruce J, Russell EM, Mollison J, Krukowski ZH. The measurement and monitoring of surgical adverse events. Health Technology Assessment. 2001;5(22):1-194.
203. Glenister HM TL, Cooke EM, Bartlett CLR. A study of surveillance methods for detecting hospital infection. London: Public Health Laboratory Service, 1992.
204. Ayliffe GA CM, Cookson BD, Emmerson AM, Falkiner FR, French GL et al. National prevalence survey of hospital acquired infections: definitions. A preliminary report of the Steering Group of the Second National Prevalence Survey. J Hosp Infect. 1993;24(1):17.
205. Peel AL, Taylor EW. Proposed definitions for the audit of postoperative infection: a discussion paper. Surgical Infection Study Group. Annals of the Royal College of Surgeons of England. 1991;73(6):385-8.
206. Health Protection Agency. Protocol for the Surveillance of Surgical Site Infection 2011 Accessed 10/01/2013. Available from: http://www.hpa.org.uk/webc/HPAwebFile/HPAweb_C/1194947388966.
207. Gibbons CB, J.; Carpenter, J.; Wilson, A.P.; Wilson, J.; Pearson, A.; Lamping, D.L.; Krukowski, Z.H.; Reeves, B.C.;. Identification of risk factors by systematic review and development of risk-adjusted models for surgical site infection. Health Technology Assessment 2011;15(30).
208. Wilson AP GR, Treasure T, Sturridge MF. A clinical trial of teicoplanin compared with a combination of flucloxacillin and tobramycin as antibiotic prophylaxis for cardiac surgery: the use of a scoring method to assess the incidence of wound infection. Journal of Hospital Infection. 1986;7(Suppl A):7.
209. Bailey IS, Karran SE, Toyn K, Brough P, Ranaboldo C, Karran SJ. Community surveillance of complications after hernia surgery. BMJ. 1992;304(6825):469-71.
210. Wilson APR, Helder N, Theminimulle SK, Scott GM. Comparison of wound scoring methods for use in audit. Journal of Hospital Infection. 1998;39(2):119-26.
211. Altemeier WAC, W.R.; Hummel, R.P.;. Surgical considerations of endogenous infections--sources, types, and methods of control. Surg Clin North Am. 1968;48(1):227-40.
212. Weigelt JA, Lipsky BA, Tabak YP, Derby KG, Kim M, Gupta V. Surgical site infections: Causative pathogens and associated outcomes. American Journal of Infection Control. 2010;38(2):112-20.
213. Misteli HW, A.F.; Rosenthal, R.; Oertli, D.; Marti, W.R.; Weber, W.P. Spectrum of pathogens in surgical site infections at a Swiss university hospital. Swiss Medical Weekly. 2011;140(w13146).
214. Buggy D. Can anaesthetic management influence surgical-wound healing? The Lancet. 2000;356(9227):355-7.
215. Faraday N, Rock P, Lin EE, Perl TM, Carroll K, Stierer T, et al. Past History of Skin Infection and Risk of Surgical Site Infection After Elective Surgery. Annals of surgery. 2013;257(1):150-4.
216. Hübner M, Diana M, Zanetti G, Eisenring MC, Demartines N, Troillet N. Surgical site infections in colon surgery: The patient, the procedure, the hospital, and the surgeon. Archives of Surgery. 2011;146(11):1240-5.

217. Culver DH, Horan TC, Gaynes RP, Martone WJ, Jarvis WR, Emori TG, et al. Surgical wound infection rates by wound class, operative procedure, and patient risk index. The American Journal of Medicine. 1991;91(3, Supplement 2):S152-S7.
218. Anonymous. New classification of physical status. Anesthesiology. 1963;24(111).
219. Haley RW, Culver DH, Morgan WM, White JW, Emori TG, Hooton TM. Identifying patients at high risk of surgical wound infection. A simple multivariate index of patient susceptibility and wound contamination. American Journal of Epidemiology. 1985;121(2):206-15.
220. Leong G, Wilson J, Charlett A. Duration of operation as a risk factor for surgical site infection: comparison of English and US data. Journal of Hospital Infection. 2006;63(3):255-62.
221. Anderson DJ, Chen LF, Sexton DJ, Kaye KS. Complex surgical site infections and the devilish details of risk adjustment: Important implications for public reporting. Infection Control and Hospital Epidemiology. 2008;29(10):941-6.
222. Clements ACA, Tong ENC, Morton AP, Whitby M. Risk stratification for surgical site infections in Australia: evaluation of the US National Nosocomial Infection Surveillance risk index. Journal of Hospital Infection. 2007;66(2):148-55.
223. Horan TC, Culver DH, Gaynes RP, Jarvis WR, Edwards JR, Reid CR. Nosocomial Infections in Surgical Patients in the United States, January 1986-June 1992. Infection Control and Hospital Epidemiology. 1993;14(2):73-80.
224. Jarvis WR. Selected aspects of the socioeconomic impact of nosocomial infections: Morbidity, mortality, cost, and prevention. Infection Control and Hospital Epidemiology. 1996;17(8):552-7.
225. Astagneau P, Rioux C, Golliot F, Brucker G. Morbidity and mortality associated with surgical site infections: results from the 1997-1999 INCISO surveillance. Journal of Hospital Infection. 2001;48(4):267-74.
226. DiPiro J, Martindale R, Bakst A, Vacani P, Watson P, Miller M. Infection in surgical patients: effects on mortality, hospitalization, and postdischarge care. American Journal of Health-System Pharmacy. 1998;55(8):777-81.
227. Kirkland KB, Briggs JP, Trivette SL, Wilkinson WE, Sexton DJ. The Impact of Surgical-Site Infections in the 1990s: Attributable Mortality, Excess Length of Hospitalization, and Extra Costs •. Infection Control and Hospital Epidemiology. 1999;20(11):725-30.
228. Bayat A, McGrouther DA, Ferguson MW. Skin scarring. BMJ. 2003;326(7380):88-92.
229. Whitehouse JD, Friedman ND, Kirkland KB, Richardson WJ, Sexton DJ. The Impact of Surgical-Site Infections Following Orthopedic Surgery at a Community Hospital and a University Hospital: Adverse Quality of Life, Excess Length of Stay, and Extra Cost. Infection Control and Hospital Epidemiology. 2002;23(4):183-9.
230. Perencevich EN, Sands KE, Cosgrove SE, Guadagnoli E, Meara E, Platt R. Health and economic impact of surgical site infections diagnosed after hospital discharge. Emerging infectious diseases. 2003;9(2):196-203.
231. Andersson AE, Bergh I, Karlsson J, Nilsson K. Patients' experiences of acquiring a deep surgical site infection: An interview study. American Journal of Infection Control. 2010;38(9):711-7.
232. Weber BW, Zwahlen M, Reck S, Feder-Mengus C, Misteli H, Rosenthal R, et al. Economic Burden of Surgical Site Infections at a European University Hospital. Infection Control and Hospital Epidemiology. 2008;29(7):623-9.
233. Junker T, Mujagic E, Hoffman H, Rosenthal R, Misteli H, Zwahlen M, et al. Prevention and control of surgical site infections: review of the Basel Cohort Study. Swiss Medical Weekly. 2012;142.
234. Anderson DJ, Kaye KS, Chen LF, Schmader KE, Choi Y, Sloane R, et al. Clinical and Financial Outcomes Due to Methicillin Resistant *Staphylococcus aureus* Surgical Site Infection: A Multi-Center Matched Outcomes Study. PLoS ONE. 2009;4(12):e8305.
235. Leaper DJ. Surgical-site infection. British Journal of Surgery. 2010;97(11):1601-2.
236. Kashimura N, Kusachi S, Konishi T, Shimizu J, Kusunoki M, Oka M, et al. Impact of surgical site infection after colorectal surgery on hospital stay and medical expenditure in Japan. Surgery Today. 2012;42(7):639-45.

237. Urban JA. Cost Analysis of Surgical Site Infections. Surgical Infections. 2006;7(s1):4.
238. Reilly J, Twaddle S, McIntosh J, Kean L. An economic analysis of surgical wound infection. Journal of Hospital Infection. 2001;49(4):245-9.
239. Alfonso JL, Pereperez SB, Canoves JM, Martinez MM, Martinez IM, Martin-Moreno JM. Are we really seeing the total costs of surgical site infections? A Spanish study. Wound Repair and Regeneration. 2007;15(4):474-81.
240. Engemann JJ, Carmeli Y, Cosgrove SE, Fowler VG, Bronstein MZ, Trivette SL, et al. Adverse Clinical and Economic Outcomes Attributable to Methicillin Resistance among Patients with Staphylococcus aureus Surgical Site Infection. Clinical Infectious Diseases. 2003;36(5):592-8.
241. McGarry SA, Engemann JJ, Schmader KE, Sexton DJ, Kaye KS. Surgical-Site Infection Due to Staphylococcus aureus Among Elderly Patients: Mortality, Duration of Hospitalization, and Cost •. Infection Control and Hospital Epidemiology. 2004;25(6):461-7.
242. Umscheid CA, Mitchell MD, Doshi JA, Agarwal R, Williams K, Brennan PJ. Estimating the Proportion of Healthcare-Associated Infections That Are Reasonably Preventable and the Related Mortality and Costs. Infection Control and Hospital Epidemiology. 2011;32(2):101-14.
243. National Institute for Health and Clinical Excellence. Surgical site infection - prevention and treatment of surgical site infection. Clinical Guideline 78. London: RCOG Press; 2008.
244. Gottrup F. Prevention of surgical-wound infections. New England Journal of Medicine. 2000;342(3):202-4.
245. Pessaux P, Msika S, Atalla D, Hay JM, Flamant Y. Risk Factors for Postoperative Infectious Complications in Noncolorectal Abdominal Surgery: A Multivariate Analysis Based on a Prospective Multicenter Study of 4718 Patients. Archives of Surgery. 2003;138(3):314-24.
246. LizanGarcia M, GarciaCaballero J, AsensioVegas A. Risk factors for surgical-wound infection in general surgery: A prospective study. Infection Control and Hospital Epidemiology. 1997;18(5):310-5.
247. Biscione F, Couto R, Pedrosa T. Performance, Revision and Extension of the National Nosocomial Infections Surveillance System's Risk Index in Brazilian Hospitals. Infection Control and Hospital Epidemiology. 2012;33(2):124-34.
248. Anaya DA, Cormier JN, Xing Y, Koller P, Gaido L, Hadfield D, et al. Development and validation of a novel stratification tool for identifying cancer patients at increased risk of surgical site infection. Annals of Surgery. 2012;255(1):134-9.
249. Stockley JM, Allen RM, Thomlinson DF, Constantine CE. A district general hospital's method of post-operative infection surveillance including post-discharge follow-up, developed over a five-year period. Journal of Hospital Infection. 2001;49(1):48-54.
250. Sands K, Vineyard G, Platt R. Surgical Site Infections Occurring after Hospital Discharge. Journal of Infectious Diseases. 1996;173:963-70.
251. Petherick E, Dalton J, Moore P, Cullum N. Methods for identifying surgical wound infection after discharge from hospital: a systematic review. BMC Infectious Diseases. 2006;6(1):170.
252. Reilly J, Allardice G, Bruce J, Hill R, McCoubrey J. Procedure-Specific Surgical Site Infection Rates and Postdischarge Surveillance in Scotland •. Infection Control and Hospital Epidemiology. 2006;27(12):1318-23.
253. Rioux C, Grandbastien B, Astagneau P. Impact of a six-year control programme on surgical site infections in France: results of the INCISO surveillance. Journal of Hospital Infection. 2007;66(3):217-23.
254. Manniën J, van den Hof S, Muilwijk J, van den Broek Peterhans J, van Benthem B, Wille JC. Trends in the Incidence of Surgical Site Infection in The Netherlands •. Infection Control and Hospital Epidemiology. 2008;29(12):1132-8.
255. Gastmeier P, Sohr D, Schwab F, Behnke M, Zuschneid I, Brandt C, et al. Ten years of KISS: The most important requirements for success. Journal of Hospital Infection. 2008;70, Supplement 1(0):11-6.
256. Morton AP, Clements ACA, Doidge SR, Stackelroth J, Curtis M, Whitby M. Surveillance of healthcare-acquired infections in Queensland, Australia: Data and lessons from the first 5 years. Infection Control and Hospital Epidemiology. 2008;29(8):695-701.

257. Geubbels ELPE, Bakker HG, Houtman P, van Noort-Klaassen MA, Pelk MSJ, Sassen TM, et al. Promoting quality through surveillance of surgical site infections: Five prevention success stories. American Journal of Infection Control. 2004;32(7):424-30.
258. Mabit C, Marcheix PS, Mounier M, Dijoux P, Pestourie N, Bonnevialle P, et al. Impact of a surgical site infection (SSI) surveillance program in orthopedics and traumatology. Orthopaedics & Traumatology-Surgery & Research. 2012;98(6):690-5.
259. Hibbard JH, Stockard J, Tusler M. Hospital Performance Reports: Impact On Quality, Market Share, And Reputation. Health Affairs. 2005;24(4):1150-60.
260. Brandt C, Sohr D, Behnke M, Daschner F, Ruden H, Gastmeier P. Reduction of surgical site infection rates associated with active surveillance. Infection Control and Hospital Epidemiology. 2006;27(12):1347-51.
261. Cole WR, Bernard HR. Wound Isolation in Prevention of Postoperative Wound Infection. Surgery Gynecology and Obstetrics with International Abstracts of Surgery. 1967;125(2):257-&.
262. Applied Medical. Alexis TM Wound Retractor System 2011 Accessed 10/01/2011. Available from: http://www.appliedmedical.com/products/product_card.aspx?prodGroupID=9&catID=31&Name=Alexis%3Csup%3E%AE%3C/sup%3E+wound+retractor+system
263. 3M Healthcare. 3M Steri-Drape Wound-edge protector 2011 [Accessed 2011/11/14]. Available from: http://solutions.3m.com/wps/portal/3M/en_US/infection-prevention-solutions/home/products/?PC_7_RJH9U52308DUB0IIL8TMGN3013_nid=PTDM8B1RG4be4Q6TMNMMVPgl.
264. Raahave D. Aseptic Barriers of Plastic to Prevent Bacterial-Contamination of Operation Wounds. Acta Chirurgica Scandinavica. 1974;140(8):603-10.
265. Mohan HM, McDermott S, Fenelon L, Fearon NM, O'Connell PR, Oon SF, et al. Plastic wound retractors as bacteriological barriers in gastrointestinal surgery: a prospective multi-institutional trial. Journal of Hospital Infection. 2012;81(2):109-13.
266. Webster J, Alghamdi A. Use of plastic adhesive drapes during surgery for preventing surgical site infection. Cochrane Database of Systematic Reviews. 2007;Reviews 2007 Issue 4.
267. Maxwell JG, Ford CR, Peterson DE, Richards RC. Abdominal wound infections and plastic drape protectors. American Journal of Surgery. 1969;118(6):844-8.
268. Psaila JV, Wheeler MH, Crosby DL. The role of plastic wound drapes in the prevention of wound infection following abdominal surgery. British Journal of Surgery. 1977;64:729-32.
269. Sookhai S, Redmond HP, Deasy JM. Impervious wound-edge protector to reduce postoperative wound infection: a randomised, controlled trial. Lancet. 1999;353(9164):1585-.
270. Nystrom PO, Broome A, Hojer H, Ling L. A Controlled Trial of A Plastic Wound Ring Drape to Prevent Contamination and Infection in Colorectal Surgery. Diseases of the Colon & Rectum. 1984;27(7):451-3.
271. Horiuchi T, Tanishima H, Tamagawa K, Matsuura I, Nakai H, Shouno Y, et al. Randomized, controlled investigation of the anti-infective properties of the Alexis retractor/protector of incision sites. Journal of Trauma-Injury Infection and Critical Care. 2007;62(1):212-5.
272. Centre for Reviews and Dissemination. Systematic reviews - CRD's guidance for undertaking reviews in health care. York: CRD; 2009.
273. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ. 2009;339:332-6.
274. Armstrong EC. The well-built clinical question: the key to finding the best evidence efficiently. Wisconsin Medical Journal. 1999;98(2):25-8.
275. Gurevich I, Horan TC. Surgical Site Infections: Simplifying the Definitions. Infection Control and Hospital Epidemiology. 1995;16(11):667-8.
276. The Cochrane Collaboration. Review Manager (Rev Man) [computer program]. Version 5.0 ed. Copenhagen: The Nordic Cochrane Centre; 2008.
277. DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials. 1986;7(3):177-88.

278. Whitehead A. Meta-Analysis of Controlled Clinical Trials. Chichester: John Wiley and Sons; 2002.
279. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Statistics in Medicine. 2002;21(11):1539-58.
280. Thompson SG. Systematic Review: Why sources of heterogeneity in meta-analysis should be investigated. BMJ. 1994;309(6965):1351-5.
281. Begg CB, Mazumdar M. Operating Characteristics of a Rank Correlation Test for Publication Bias. Biometrics. 1994;50(4):1088-101.
282. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997;315(7109):629-34.
283. Alexander-Williams J, Oates GD, Brown PP, Burden DW, McCall J, Hutchison AG, et al. Abdominal wound infections and plastic wound guards. British Journal of Surgery. 1972;59(2):142-6.
284. Gamble SS, Hopton DS. Plastic ring wound drapes in elective colorectal surgery. Journal of the Royal College of Surgeons of Edinburgh. 1984;29(4):232-3.
285. Batz W, Marcus D, Rothmund M. Value of Ring Drape and Incision Drape to Prevent Wound-Infection in Colorectal Surgery - A Controlled Randomized Study. Aktuelle Chirurgie. 1987;22(4):149-52.
286. Redmond HP, Meagher PJ, Kelly CJ, Deasy JM. Use of An Impervious Wound-Edge Protector to Reduce the Postoperative Wound-Infection Rate. British Journal of Surgery. 1994;81(12):1811.
287. Brunet P, Bounoua F, Bugnon PY, Gautier-Benoit C. Interet des champs a anneau en chirurgie abdominale (The use of ring drapes in abdominal surgery). Lyon chirurgical. 1994;90(6):438-41.
288. Lee P, Waxman K, Taylor B, Yim S. Use of Wound-Protection System and Postoperative Wound-Infection Rates in Open Appendectomy A Randomized Prospective Trial. Archives of Surgery. 2009;144(9):872-5.
289. Reid K, Pockney P, Draganic B, Smith SR. Barrier Wound Protection Decreases Surgical Site Infection in Open Elective Colorectal Surgery: A Randomized Clinical Trial. Diseases of the Colon & Rectum. 2010;53(10):1374-80.
290. Harrower HW. Isolation of Incisions Into Body Cavities. American Journal of Surgery. 1968;116(6):824-6.
291. Nystrom PO, Brote L. Effects of a plastic wound drape on contamination with enterobacteria and on infection after appendicectomy. Acta Chirurgica Scandinavica. 1980;146(1):65-70.
292. Pollock AV. Prevention of wound infection by an antiseptic wound protector. Journal of the Royal Society of Medicine. 1980;73(11):831.
293. Anthony T, Murray BW, Sum-Ping JT, Lenkovsky F, Vornik VD, Parker BJ, et al. Evaluating an Evidence-Based Bundle for Preventing Surgical Site Infection: A Randomized Trial. Archives of Surgery. 2010;146(3):263-9.
294. Horiuchi T, Tanishima H, Tamagawa K, Sakaguchi S, Shono Y, Tsubakihara H, et al. A wound protector shields incision sites from bacterial invasion. Surgical Infections. 2010;11(6):501-3.
295. Nakagoe T, Sawai T, Tsuji T, Nanashima A, Jibiki Ma, Yamaguchi H, et al. Minilaparotomy Wound Edge Protector (Lap-Protector): A New Device. Surgery Today. 2001;31(9):850-2.
296. Kercher KW, Nguyen TH, Harold KL, Poplin ME, Matthews BD, Sing RF, et al. Plastic wound protectors do not affect wound infection rates following laparoscopic-assisted colectomy. Surgical Endoscopy and Other Interventional Techniques. 2004;18(1):148-51.
297. Varela JE, Wilson SE, Nguyen NT. Laparoscopic surgery significantly reduces surgical-site infections compared with open surgery. Surgical Endoscopy and Other Interventional Techniques. 2010;24(2):270-6.
298. Kiran RP, El-Gazzaz GH, Vogel JD, Remzi FH. Laparoscopic Approach Significantly Reduces Surgical Site Infections after Colorectal Surgery: Data from National Surgical Quality Improvement Program. Journal of the American College of Surgeons. 2010;211(2):232-8.
299. Shabanzadeh DM, Sørensen LT. Laparoscopic surgery compared with open surgery decreases surgical site infection in obese patients: A systematic review and meta-analysis. Annals of surgery. 2012;256(6):934-45.

300. Targarona EM, Balagué C, Knook MM, Trías M. Laparoscopic surgery and surgical infection. British Journal of Surgery. 2000;87(5):536-44.
301. Edwards JPA, Ho L.; Tee, May C.; Dixon, Elijah; Ball, Chad G.; . Wound protectors reduce surgical site infection: a meta-analysis of randomized controlled trials. Annals of surgery. 2012;256(1):6.
302. Theodoridis TD, Chatzigeorgiou KN, Zepiridis L, Papanicolaou A, Vavilis D, Tzevelekis F, et al. A prospective randomized study for evaluation of wound retractors in the prevention of incision site infection after caesarean section. Clin Exp Obstet Gynecol. 2011;38:57-9.
303. Cheng KP, Roslani AC, Sehha N, Kueh JH, Law CW, Chong HY, et al. ALEXIS O-Ring wound retractor vs conventional wound protection for the prevention of surgical site infections in colorectal resections1. Colorectal Disease. 2012;14(6):e346-e51.
304. Eborall H, Stewart M, Cunningham-Burley S, Price J, Fowkes FG. Accrual and drop out in a primary prevention randomised controlled trial: qualitative study. Trials. 2011;12(1):7.
305. Mihaljevic A, Michalski C, Erkan M, Reiser-Erkan C, Jager C, Schuster T, et al. Standard abdominal wound edge protection with surgical dressings vs coverage with a sterile circular polyethylene drape for prevention of surgical site infections (BaFO): study protocol for a randomized controlled trial. Trials. 2012;13(1):57.
306. Torrance GW, Feeny DH, Furlong WJ, Barr RD, Zhang Y, Wang Q. Multiattribute Utility Function for a Comprehensive Health Status Classification System: Health Utilities Index Mark 2. Medical Care. 1996;34(7):702-22.
307. Feeny D, Furlong W, Torrance GW, Goldsmith CH, Zhu Z, DePauw S, et al. Multiattribute and Single-Attribute Utility Functions for the Health Utilities Index Mark 3 System. Medical Care. 2002;40(2):113-28.
308. Kaplan RM, Bush JW, Berry CC. Health status: types of validity and the index of well-being. Health Services Research. 1976;11(4):478-507.
309. Ware JE, Jr., Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: Construction of Scales and Preliminary Tests of Reliability and Validity. Medical Care. 1996;34(3):220-33.
310. Ware JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Medical Care. 1992;30(6):473-83.
311. McHorney CA, Ware JE, Jr., Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and Clinical Tests of Validity in Measuring Physical and Mental Health Constructs. Medical Care. 1993;31(3):247-63.
312. Brazier J, Yang Y, Tsuchiya A, Rowen D. A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. The European Journal of Health Economics. 2010;11(2):215-25.
313. Elliott R, Weatherly H, Hawkins N, Cranny G, Chambers D, Myers L, et al. An economic model for the prevention of MRSA infections after surgery: non-glycopeptide or glycopeptide antibiotic prophylaxis? The European Journal of Health Economics.11(1):57-66.
314. Cranny G, Elliott R, Weatherly H, Chambers D, Hawkins N, Myers L, et al. A systematic review and economic model of switching from non-glycopeptide to glycopeptide antibiotic prophylaxis for surgery (Provisional abstract). Health Technology Assessment. 2008(1):1-147.
315. Brasel KJ, Borgstrom DC, Weigelt JA. Cost-utility analysis of contaminated appendectomy wounds. Journal of the American College of Surgeons. 1997;184(1):23-30.
316. Bailey RR, Stuckey DR, Norman BA, Duggan AP, Bacon KM, Connor DL, et al. Economic value of dispensing home-based preoperative chlorhexidine bathing cloths to prevent surgical site infection. Infection Control and Hospital Epidemiology. 2011(5):465-71.
317. Lee BY, Wiringa AE, Bailey RR, Lewis GJ, Feura J, Muder RR. Staphylococcus aureus vaccine for orthopedic patients: an economic model and analysis. Vaccine. 2010;28(12):2465-71.
318. Slobogean GP, O'Brien PJ, Brauer CA. Single-dose versus multiple-dose antibiotic prophylaxis for the surgical treatment of closed fractures: a cost-effectiveness analysis. Acta Orthopaedica. 2010(2):256-62.

319. Lee BY, Wiringa AE, Mitgang EA, McGlone SM, Afriyie AN, Song Y, et al. Routine pre-cesarean Staphylococcus aureus screening and decolonization: a cost-effectiveness analysis. American Journal of Managed Care. 2011;17(10):693-700.
320. Lee BY, Tsui BY, Bailey RR, Smith KJ, Muder RR, Lewis GJ, et al. Should Vascular Surgery Patients Be Screened Preoperatively for Methicillin-Resistant Staphylococcus aureus? Infection Control and Hospital Epidemiology. 2009;30(12):1158-65.
321. Thoma A, Khuthaila D, Rockwell G, Veltri K. Cost-utility analysis comparing free and pedicled TRAM flap for breast reconstruction. Microsurgery. 2003;23(4):287-95.
322. Poulsen KB, Gottschau A. Long-term prognosis of patients with surgical wound infections. World Journal of Surgery. 1997;21(8):799-804.
323. Lee BY, Wiringa AE, Bailey RR, Goyal V, Lewis GJ, Tsui BY, et al. Screening Cardiac Surgery Patients for MRSA: An Economic Computer Model. American Journal of Managed Care. 2010;16(7):e163-e73.
324. Tengs TO, Wallace A. One Thousand Health-Related Quality-of-Life Estimates. Medical Care. 2000;38(6):583-637.
325. Tsevat J, Durand-Zaleski I, Pauker SG. Cost-Effectiveness of Antibiotic Prophylaxis for Dental Procedures in Patients with Artificial Joints. American Journal of Public Health. 1989;79(6):739-43.
326. Selai C, Rosser R. Eliciting EuroQol descriptive data and utility scale values from inpatients. A feasibility study. Pharmacoeconomics. 1995;8(2):147-58.
327. Sackett DL, Torrance GW. The utility of different health states as perceived by the general public. Journal of Chronic Diseases. 1978;31(11):697-704.
328. Lawrence WF, Fleishman JA. Predicting EuroQoL EQ-5D Preference Scores from the SF-12 Health Survey in a Nationally Representative Sample. Medical Decision Making. 2004;24(2):160-9.
329. Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D, et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement. BMJ. 2013;346.
330. Jarman AF, Wray NP, Wenner DM, Ashton CM. Trials and tribulations: the professional development of surgical trialists. The American Journal of Surgery. 2012;204(3):339-46.
331. Wilson J, Charlett A, Leong G, McDougall C, Duckworth G. Rates of surgical site infection after hip replacement as a hospital performance indicator: analysis of data from the English mandatory surveillance system. Infection Control & Hospital Epidemiology. 2008;29(3):219-26.
332. Janson M, Lindholm E, Anderberg B, Haglind E. Randomized trial of health-related quality of life after open and laparoscopic surgery for colon cancer. Surgical Endoscopy. 2007;21(5):747-53.
333. Curtis L. Unit costs of health and social care 2011. Canterbury: Personal Social Services Unit, University of Kent; 2011.
334. Cosgrove SE, Qi YM, Kaye KS, Harbarth SM, Karchmer AW, Carmeli Y. The Impact of Methicillin Resistance in Staphylococcus aureus Bacteremia on Patient Outcomes: Mortality, Length of Stay, and Hospital Charges •. Infection Control and Hospital Epidemiology. 2005;26(2):166-74.
335. Department of Health. NHS Reference Costs 2009-2010. 2010 Accessed 15/09/2011. Available from: https://www.gov.uk/government/publications/nhs-reference-costs-2009-2010.
336. Curtis L. Unit costs of health and social care 2010. Canterbury: Personal Social Services Research Unit, University of Kent; 2010.
337. NHS The Information Centre. Prescription Cost Analysis 2010 2011 [Accessed 2013/09/23]. Available from: http://www.hscic.gov.uk/pubs/prescostanalysis2010.
338. Philips Z, Bojke L, Sculpher M, Claxton K, Golder S. Good Practice Guidelines for Decision-Analytic Modelling in Health Technology Assessment: A Review and Consolidation of Quality Assessment. Pharmacoeconomics. 2006;24(4).
339. Caro JJ, Briggs AH, Siebert U, Kuntz KM. Modeling Good Research Practices—Overview: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force–1. Medical Decision Making. 2012;32(5):667-77.
340. Delgado-Rodriguez M, Gomez-Ortega A, Sillero-Arenas M, Llorca J. Epidemiology of surgical-site infections diagnosed after hospital discharge: A prospective cohort study. Infection Control and Hospital Epidemiology. 2001;22(1):24-30.

341. NHS The Information Centre. Hospital Episode Statistics for England: 2009-2010 2011 Accessed 2013/09/23. Available from: http://www.hscic.gov.uk/pubs/hesadmitted1011.

342. Bojke L, Claxton K, Sculpher M, Palmer S. Characterizing Structural Uncertainty in Decision Analytic Models: A Review and Application of Methods. Value in Health. 2009;12(5):739-49.

343. Briggs AH. Handling Uncertainty in Cost-Effectiveness Models. Pharmacoeconomics. 2000;17(5):479-500.

344. Elliott RA, Weatherly HLA, Hawkins NS, Cranny G, Chambers D, Myers L, et al. An economic model for the prevention of MRSA infections after surgery: non-glycopeptide or glycopeptide antibiotic prophylaxis? European Journal of Health Economics. 2010;11(1):57-66.

345. Kind P, Hardman G, Macran S. UK population norms for EQ-5D. Centre for Health Economics Discussion Paper. 1999;172.

346. Pada SK, Ding Y, Ling ML, Hsu LY, Earnest A, Lee TE, et al. Economic and clinical impact of nosocomial meticillin-resistant Staphylococcus aureus infections in Singapore: a matched case-control study. Journal of Hospital Infection. 2011;78(1):36-40.

347. Drummond M. Introducing economic and quality of life measurements into clinical studies. Annals of Medicine. 2001;33(5):344-9.

348. Dolan P. Modeling Valuations for EuroQol Health States. Medical Care. 1997;35(11):1095-108.

349. Davey P, Lynch B, Malek M, Byrne D, Thomas P. Cost-effectiveness of single dose cefotaxime plus metronidazole compared with three doses each of cefuroxime plus metronidazole for the prevention of wound infection after colorectal surgery. Journal of Antimicrobial Chemotherapy. 1992;30(6):855-64.

350. Boyd KA, Fenwick E, Briggs A. Using an iterative approach to economic evaluation in the drug development process. Drug Development Research. 2010;71(8):470-7.

351. Torgerson DJ, Byford S. Economic modelling before clinical trials. BMJ. 2002;325.

352. Pinkney T, Bartlett D, Hawkins W, Mak T, Youssef H, Futaba K, et al. Reduction of surgical site infection using a novel intervention (ROSSINI): study protocol for a randomised controlled trial. Trials.12(1):217.

353. Dobson AJ, Barnett AG. Introduction to Generalized Linear Models. Boca Raton, FL: Chapman and Hall/CRC; 2008.

354. Pinkney TD, Calvert M, Bartlett DC, Gheorghe A, Redman V, Dowswell G, et al. Impact of wound edge protection devices on surgical site infection after laparotomy: multicentre randomised controlled trial (ROSSINI Trial). BMJ. 2013;347.

355. Joint Formulary Committee. British National Formulary. 61st edition. London: British Medical Association and Royal Pharmaceutical Society of Great Britain; 2011.

356. Manca A, Hawkins N, Sculpher MJ. Estimating mean QALYs in trial-based cost-effectiveness analysis: the importance of controlling for baseline utility. Health Economics. 2005;14(5):487-96.

357. Department of Health. NHS Reference costs 2010-2011.2011 Accessed 04/09/2013. Available from: https://www.gov.uk/government/publications/2010-11-reference-costs-publication.

358. Department of Health. NHS Reference costs 2007-2008. 2008 Accessed 04/09/2013. Available from: http://webarchive.nationalarchives.gov.uk/20130107105354/http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_098945.

359. R Development Core Team. R: A language and environment for statistical computing. 2.15.3 ed. Vienna, Austria: R Foundation for Statistical Computing; 2013.

360. Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. 2011;45(3).

361. Barie PS. Does a Well-Done Analysis of Poor-Quality Data Constitute Evidence of Benefit? Annals of surgery. 2012;255(6).

362. Ramsey SD, McIntosh M, Sullivan SD. Design issues for conducting cost-effectiveness analyses alongside clinical trials. Annual Review of Public Health. 2001;22:129-41.

363. Smyth RMD, Kirkham JJ, Jacoby A, Altman DG, Gamble C, Williamson PR. Frequency and reasons for outcome reporting bias in clinical trials: Interviews with trialists. BMJ. 2011;342:c7153(7789).
364. Al-Marzouki S, Roberts I, Evans S, Marshall T. Selective reporting in clinical trials: analysis of trial protocols accepted by The Lancet. The Lancet. 2008;372(9634):201.
365. Sandelowski M, Barroso J. Creating Metasummaries of Qualitative Findings. Nursing Research. 2003;52(4):226-33.
366. QSR International Pty Ltd. NVivo qualitative data analysis software. 8 ed2008.
367. Dwan K, G AD, Cresswell L, Blundell M, L GC, R WP. Comparison of protocols and registry entries to published reports for randomised controlled trials. Cochrane Database of Systematic Reviews. 2011(1).
368. Ross JS, Mulvey GK, Hines EM, Nissen SE, Krumholz HM. Trial Publication after Registration in ClinicalTrials.Gov: A Cross-Sectional Analysis. PLoS Med. 2009;6(9):e1000144.
369. Chan A, Hrobjartsson A, Haahr M, Gotzsche P, Altman D. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA. 2004;291:2457 - 65.
370. Chan A, Hrobjartsson A, Jorgensen K, Gotzsche P, Altman D. Discrepancies in sample size calculations and data analyses reported in randomised trials: comparison of publications with protocols. BMJ. 2008;337:a2299.
371. Blümle A, Meerpohl JJ, Rücker G, Antes G, Schumacher M, Elm Ev. Reporting of eligibility criteria of randomised trials: cohort study comparing trial protocols with subsequent articles. BMJ. 2011;342.
372. Sweetman E, Doig G. Failure to report protocol violations in clinical trials: a threat to internal validity? Trials. 2011;12(1):214.
373. Ring N, Jepson R, Ritchie K. Methods of synthesizing qualitative research studies for health technology assessment. International Journal of Technology Assessment in Health Care. 2011;27(04):384-90.
374. Limkakeng AT, de Oliveira LLH, Moreira T, Phadtare A, Garcia Rodrigues C, Hocker MB, et al. Systematic review and metasummary of attitudes toward research in emergency medical conditions. Journal of Medical Ethics. 2013.
375. Fletcher B, Gheorghe A, Moore D, Wilson S, Damery S. Improving the recruitment activity of clinicians in randomised controlled trials: a systematic review. BMJ Open. 2012;2(1).
376. Tetzlaff J, Chan A-W, Kitchen J, Sampson M, Tricco A, Moher D. Guidelines for randomized clinical trial protocol content: a systematic review. Systematic Reviews. 2012;1(1):43.
377. Treweek S, McCormack K, Abalos E, Campbell M, Ramsay C, Zwarenstein M, et al. The Trial Protocol Tool: The PRACTIHC software tool that supported the writing of protocols for pragmatic randomized controlled trials. Journal of Clinical Epidemiology. 2006;59:1127 - 33.
378. Jones A, Conroy E, Williamson P, Clarke M, Gamble C. The use of systematic reviews in the planning, design and conduct of randomised trials: a retrospective cohort of NIHR HTA funded trials. BMC Medical Research Methodology. 2013;13(1):50.
379. Liamputtong P. Focus Group Methodology: Principles and Practice. Bodmin: Sage; 2011.
380. Krueger RA, Casey MA. Focus groups: A practical guide for applied research: Sage; 2000.
381. Hsieh HF, Shannon SE. Three Approaches to Qualitative Content Analysis. Qualitative Health Research. 2005;15(9):1277-88.
382. University of Birmingham. Clinical Trials Units at Birmingham 2013 [Accessed 08/09/2013]. Available from: http://www.birmingham.ac.uk/research/activity/mds/centres/bcct/units/index.aspx.
383. McPherson G, Campbell M, Elbourne D. Use of randomisation in clinical trials: a survey of UK practice. Trials. 2012;13(1):198.
384. Kolb B, Whishaw IQ. Fundamentals of Neuropsychology 5th edition. 5th edition ed. New York: Worth Publishers; 2003.
385. Paramasivan S, Huddart R, Hall E, Lewis R, Birtle A, Donovan J. Key issues in recruitment to randomised controlled trials with very different interventions: a qualitative investigation of recruitment to the SPARE trial (CRUK/07/011). Trials. 2011;12(1):78.

386. Brady M, Stott D, Norrie J, Chalmers C, St George B, Sweeney P, et al. Developing and evaluating the implementation of a complex intervention: using mixed methods to inform the design of a randomised controlled trial of an oral healthcare intervention after stroke. Trials. 2011;12(1):168.
387. Murdoch M, McColl E, Howel D, Deverill M, Buckley B, Lucas M, et al. INVESTIGATE-I (INVasive Evaluation before Surgical Treatment of Incontinence Gives Added Therapeutic Effect?): study protocol for a mixed methods study to assess the feasibility of a future randomised controlled trial of the clinical utility of invasive urodynamic testing. Trials. 2011;12(1):169.
388. Grant A, Dreischulte T, Treweek S, Guthrie B. Study protocol of a mixed-methods evaluation of a cluster randomized trial to improve the safety of NSAID and antiplatelet prescribing: data-driven quality improvement in primary care. Trials. 2012;13(1):154.
389. Kaur G, Smyth R, Williamson P. Developing a survey of barriers and facilitators to recruitment in randomized controlled trials. Trials. 2012;13(1):218.
390. Hamm M, Scott S, Klassen T, Moher D, Hartling L. Do health care institutions value research? A mixed methods study of barriers and facilitators to methodological rigor in pediatric randomized trials. BMC Medical Research Methodology. 2012;12(1):158.
391. Djulbegovic B. The Paradox of Equipoise: The Principle That Drives and Limits Therapeutic Discoveries in Clinical Research. Cancer Control. 2009;16(4):342-7.
392. Ross S, Grant A, Counsell C, Gillespie W, Russell I, Prescott R. Barriers to Participation in Randomised Controlled Trials: A Systematic Review. Journal of Clinical Epidemiology. 1999;52(12):1143-56.
393. Beckett M, Quiter E, Ryan G, Berrebi C, Taylor S, Cho M, et al. Bridging the gap between basic science and clinical practice: The role of organizations in addressing clinician barriers. Implementation Science. 2011;6(1).
394. McDonald A, Treweek S, Shakur H, Free C, Knight R, Speed C, et al. Using a business model approach and marketing techniques for recruitment to clinical trials. Trials. 2011;12(1):74.
395. Harris J. Scientific research is a moral duty. Journal of Medical Ethics. 2005;31(4):242-8.
396. Schaefer G, Emanuel EJ, Wertheimer A. THe obligation to participate in biomedical research. JAMA. 2009;302(1):67-72.
397. Chan A-W, Tetzlaff JM, Gøtzsche PC, Altman DG, Mann H, Berlin JA, et al. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. BMJ. 2013;346.
398. Zarin DA, Tse T, Williams RJ, Califf RM, Ide NC. The ClinicalTrials.gov Results Database — Update and Key Issues. New England Journal of Medicine. 2011;364(9):852-60.
399. Gheorghe A, Roberts TE, Ives JC, Fletcher BR, Calvert M. Centre Selection for Clinical Trials and the Generalisability of Results: A Mixed Methods Study. PLoS ONE [Electronic Resource]. 2013;8(2):e56560.
400. Wei JW, Heeley EL, Jan S, Huang Y, Huang Q, Wang J-G, et al. Variations and Determinants of Hospital Costs for Acute Stroke in China. PLoS ONE. 2010;5(9):e13041.
401. Rivers PA, Fottler MD, Younis MZ. Does certificate of need really contain hospital costs in the United States? Health Education Journal. 2007;66(3):September.
402. Becker ER. National trends and determinants of hospitalization costs and lengths-of-stay for uterine fibroids procedures. Journal of Health Care Finance. 2007;33(3):1-16.
403. Saleh SS, Racz M, Hannan E. The effect of preoperative and hospital characteristics on costs for coronary artery bypass graft. Annals of surgery. 2009;249(2):335-41.
404. Sonig A, Khan IS, Wadhwa R, Thakur JD, Nanda A. The impact of comorbidities, regional trends, and hospital factors on discharge dispositions and hospital costs after acoustic neuroma microsurgery: a United States nationwide inpatient data sample study (2005-2009). Neurosurgical Focus. 2012;33(3):E3.
405. Adam T, Evans DB. Determinants of variation in the cost of inpatient stays versus outpatient visits in hospitals: A multi-country analysis. Social Science and Medicine. 2006;63(7):1700-10.
406. Bellanger MM, Or Z. What can we learn from a cross-country comparison of the costs of child delivery? Health Economics. 2008;17(SUPPL.#1):S47-57.

407. Petrinco M, Pagano E, Desideri A, Bigi R, Ghidina M, Ferrando A, et al. Information on Center Characteristics as Costs' Determinants in Multicenter Clinical Trials: Is Modeling Center Effect Worth the Effort? Value in Health. 2009;12(2):325-30.
408. Huttin C, de Pouvourville G. The impact of teaching and research on hospital costs. An empirical study in the French context. HEPAC Health Economics in Prevention and Care. 2001;2(2):47-53.
409. Levy CR, Fish R, Kramer AM. Site of death in the hospital versus nursing home of Medicare skilled nursing facility residents admitted under Medicare's Part A benefit. Journal of the American Geriatrics Society. 2004;52(8):1247-54.
410. Friedman B, Jiang HJ, Elixhauser A, Segal A. Hospital Inpatient Costs for Adults with Multiple Chronic Conditions. Medical Care Research and Review. 2006;63(3):327-46.
411. Lee KH, Yang SB, Choi M. The association between Hospital ownership and technical efficiency in a managed care environment. Journal of Medical Systems. 2009;33(4):307-15.
412. Chaikledkaew U, Pongchareonsuk P, Chaiyakunapruk N, Ongphiphadhanakul B. Factors affecting health-care costs and hospitalizations among diabetic patients in Thai public hospitals. Value in Health. 2008;11(SUPPL.#1):S69-S74.
413. McCollam PL, Lage MJ, Bala M. A comparison of total hospital costs for percutaneous coronary intervention patients receiving abciximab versus tirofiban. Catheterization and Cardiovascular Interventions. 2001;54(2):2001.
414. Stock GN, McDermott C. Operational and contextual drivers of hospital costs. Journal of Health, Organisation and Management. 2011;25(2):142-58.
415. Baker LC, Phibbs CS, Guarino C, Supina D, Reynolds JL. Within-year variation in hospital utilization and its implications for hospital costs. Journal of Health Economics. 2004;23(1):191-211.
416. Daidone S, D'Amico F. Technical efficiency, specialization and ownership form: evidences from a pooling of Italian hospitals. Journal of Productivity Analysis. 2009;32(3):203-16.
417. Gutacker N, Bojke C, Daidone S, Devlin NJ, Parkin D, Street A. Truly inefficient or providing better quality of care? Analysing the relationship between risk-adjusted hospital costs and patients' health outcomes. Health Economics. 2013;22(8):931-47.
418. Mason A, Street A, Miraldo M, Siciliani L. Should prospective payments be differentiated for public and private healthcare providers? Health Economics, Policy and Law. 2009;4(04):383-403.
419. Department of Health. PbR and the Market Forces Factor (MFF) in 2013-14. 2013 Accessed 11/07/2013. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/141395/PbR-and-the-MFF-in-2013-14.pdf.pdf.
420. Kristensen T, Laudicella M, Ejersted C, Street A. Cost variation in diabetes care delivered in English hospitals. Diabetic Medicine. 2010;27(8):August.
421. Department of Health. NHS Reference Costs 2011-2012. 2012 Accessed 11/07/2013. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/127112/2011-12-reference-costs-publication.pdf.pdf.
422. NHS The Information Centre. Hospital Estates and Facilities Statistics 2011/2012. 2013 Accessed 29/05/2013. Available from: http://www.hefs.ic.nhs.uk/.
423. Department of Health. Reference Cost Index: 2011 to 2012 Episodes. 2012 Accessed 29/05/2013. Available from: https://www.gov.uk/government/publications/nhs-reference-costs-financial-year-2011-to-2012.
424. Health & Social Care Information Centre. Hospital Episode Statistics, Admitted Patient Care - England 2011-12. 2013 Accessed 29/05/2013. Available from: http://www.hscic.gov.uk/searchcatalogue?q=title%3A%22Hospital+Episode+Statistics%2C+Admitted+patient+care+-+England%22&area=&size=10&sort=Relevance.
425. Gillam SJ, Siriwardena AN, Steel N. Pay-for-Performance in the United Kingdom: Impact of the Quality and Outcomes Framework—A Systematic Review. The Annals of Family Medicine. 2012;10(5):461-8.

426. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Statistics in Medicine. 2006;25(24):4279-92.
427. Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, et al. Evaluating non-randomised intervention studies. Health Technology Assessment. 2003;7(27).
428. Gomes M, Grieve R, Nixon R, Ng ESW, Carpenter J, Thompson SG. Methods for covariate adjustment in cost-effectiveness analysis that use cluster randomised trials. Health Economics. 2012;21(9):1101-18.
429. McCarron CE, Pullenayegum EM, Thabane L, Goeree R, Tarride J-E. Bayesian Hierarchical Models Combining Different Study Types and Adjusting for Covariate Imbalances: A Simulation Study to Assess Model Performance. PLoS ONE. 2011;6(10):e25635.
430. Grieve R, Nixon R, Thompson SG, Normand C. Using multilevel models for assessing the variability of multinational resource use and cost data. Health Economics. 2005;14(2):185-96.
431. O'Brien BJ, Briggs AH. Analysis of uncertainty in health care cost-effectiveness studies: an introduction to statistical issues and methods. Statistical Methods in Medical Research. 2002;11(6):455-68.
432. Health & Social Care Information Centre. Hospital Episode Statistics2013 Accessed 08/08/2013. Available from: http://www.hscic.gov.uk/hes.
433. World Health Organization. World Health Report 2000. Health systems: improving performance. Geneva: WHO; 2000.
434. Anand S, Ammar W, Evans T, Hasegawa T, Kissimova-Skarbek K, Langer A, et al. Report of the Scientific Peer Review Group on Health Systems Performance Assessment. In: Murray CJL, Evans DB, editors. Health Systems Performance Assessment Debates, Methods and Empiricism. Geneva: WHO; 2003.
435. Frenk J. The World Health Report 2000: Expanding the horizon of health system performance. Health, Policy and Planning. 2010;25(5):343-5.
436. McKee M. The World Health Report 2000: 10 years on. Health, Policy and Planning. 2010;25(5):346-8.
437. National Institute for Health and Clinical Excellence. Surgical site infection - Evidence Update June 2013. London: NICE; 2013.
438. Jones RS, Brown C, Opelka F. Surgeon compensation: "Pay for performance," the American College of Surgeons National Surgical Quality Improvement Program, the Surgical Care Improvement Program, and other considerations. Surgery. 2005;138(5):829-36.
439. Hawn MT, Vick CC, Richman J, Holman W, Deierhoi RJ, Graham LA, et al. Surgical site infection prevention: time to move beyond the surgical care improvement program. Annals of Surgery. 2011;254(3):494-9.
440. Edmiston CE, Spencer M, Lewis BD, Brown KR, Rossi PJ, Henen CR, et al. Reducing the risk of surgical site infections: did we really think SCIP was going to lead us to the promised land? Surgical Infections. 2011;12(3):169-77.
441. Stulberg JJ, Delaney CP, Neuhauser DV, Aron DC, Fu P, Koroukian SM. ADherence to surgical care improvement project measures and the association with postoperative infections. JAMA. 2010;303(24):2479-85.
442. Wick EC, Hobson DB, Bennett JL, Demski R, Maragakis L, Gearhart SL, et al. Implementation of a Surgical Comprehensive Unit-Based Safety Program to Reduce Surgical Site Infections. Journal of the American College of Surgeons. 2012;215(2):193-200.
443. Lawson EH, Hall B, Ko CY. Risk factors for superficial vs deep/organ-space surgical site infections: Implications for quality improvement initiatives. JAMA Surgery. 2013;148(9):849-58.
444. Pibouleau L, Boutron I, Reeves BC, Nizard R, Ravaud P. Applicability and generalisability of published results of randomised controlled trials and non-randomised studies evaluating four orthopaedic procedures: methodological systematic review. [Review] [24 refs]. BMJ. 2009;339:b4538.
445. Claxton K, Fenwick E, Palmer S, Sculpher M, Abrams KR, Sutton AJ. Building a reference case for Bayesian applications to health economics and outcomes research. CHE Technical Series [Internet]. 2004 Accessed 27/09/2013; Paper 35. Available from:

https://www.york.ac.uk/media/che/documents/papers/technicalpapers/CHE%20Technical%20Paper%2035.pdf.

446. Freemantle N, Hessel F. The Applicability and Generalizability of Findings from Clinical Trials for Health-Policy Decisions. Pharmacoeconomics. 2009;27(1):5-10.

APPENDICES

Appendix 1. PRISMA checklist for the systematic review of WEPD clinical effectiveness

Appendix 2. Search strategy for the systematic review of WEPD clinical effectiveness

Ovid SP Medline

1	wound protect*.mp.
2	wound-protect*.mp.
3	wound guard.mp.
4	wound-guard.mp.
5	wound edge protect*.mp.
6	wound-edge protect*.mp.
7	impervious wound drape.mp.
8	impervious wound protect*.mp.
9	ring drape.mp.
10	drape protect*.mp.
11	barrier protect*.mp.
12	ViDrape.mp. [mp=title, original title, abstract, name of substance word, subject heading word, unique identifier]
13	Vi Drape.mp. [mp=title, original title, abstract, name of substance word, subject heading word, unique identifier]
14	Steri Drape.mp. [mp=title, original title, abstract, name of substance word, subject heading word, unique identifier]
15	SteriDrape.mp. [mp=title, original title, abstract, name of substance word, subject heading word, unique identifier]
16	Alexis.mp. [mp=title, original title, abstract, name of substance word, subject heading word, unique identifier]
17	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16
18	exp Surgical Wound Infection/
19	surgical infection*.mp.
20	exp Wound Infection/
21	exp Postoperative Complications/
22	exp Bacterial Infections/
23	surgical site infection*.mp.
24	wound complication*.mp.
25	postoperative infection*.mp.
26	18 or 19 or 20 or 21 or 22 or 23 or 24 or 25
27	17 and 26

Ovid EMBASE Classic + Embase

1	wound protect*.mp.
2	wound-protect*.mp.
3	wound guard.mp.
4	wound-guard.mp.
5	wound edge protect*.mp.
6	wound-edge protect*.mp.
7	impervious wound drape.mp.
8	impervious wound protect*.mp.
9	ring drape.mp.

10	drape protect*.mp.
11	barrier protect*.mp.
12	ViDrape.mp. [mp=title, original title, abstract, name of substance word, subject heading word, unique identifier]
13	Vi Drape.mp. [mp=title, original title, abstract, name of substance word, subject heading word, unique identifier]
14	Steri Drape.mp. [mp=title, original title, abstract, name of substance word, subject heading word, unique identifier]
15	SteriDrape.mp. [mp=title, original title, abstract, name of substance word, subject heading word, unique identifier]
16	Alexis.mp. [mp=title, original title, abstract, name of substance word, subject heading word, unique identifier]
17	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16
18	exp surgical infection/
19	surgical wound infection*.mp.
20	exp wound infection/
21	exp postoperative complication/
22	exp bacterial infection/
23	surgical site infection*.mp.
24	exp wound complication/
25	exp postoperative infection
26	18 or 19 or 20 or 21 or 22 or 23 or 24 or 25
27	17 and 26

EBSCO CINAHL Plus

1	TX wound protect*
2	TX wound-protect*
3	TX wound guard
4	TX wound-guard
5	TX wound edge protect*
6	TX wound-edge protect*
7	TX impervious wound drape
8	TX impervious wound protect*
9	TX ring drape
10	TX drape protect*
11	TX barrier protect*
12	TX ViDrape
13	TX Vi Drape
14	TX Steri Drape
15	TX SteriDrape
16	TX Alexis
17	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16
18	TX surgical wound infection*
19	TX surgical infection*
20	TX wound infection*
21	TX wound complication*

22	TX postoperative complication*
23	TX bacterial infection*
24	TX surgical site infection*
25	TX postoperative infection*
26	18 or 19 or 20 or 21 or 22 or 23 or 24 or 25
27	17 and 26

ISI Web of Knowledge – Science Citation Index Expanded, Conference Proceedings Citation Index- Science and Conference Proceedings Citation Index- Social Science & Humanities

All Terms were searched as ‘Topic’.

1	wound protect*
2	wound-protect*
3	wound guard
4	wound-guard
5	wound edge protect
6	wound-edge protect*
7	impervious wound drape
8	impervious wound protect*
9	ring drape
10	drape protect*
11	barrier protect*
12	ViDrape
13	Vi Drape
14	Steri Drape
15	SteriDrape
16	Alexis
17	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16
18	surgical wound infection*
19	surgical site infection*
20	surgical infection*
21	wound infection*
22	bacterial infection*
23	wound complication*
24	postoperative complication*
25	postoperative infection*
26	18 or 19 or 20 or 21 or 22 or 23 or 24 or 25
27	17 and 26

Cochrane Library (all databases)

1 wound protect*:ti,ab,kw
2 wound-protect*: ti,ab,kw
3 wound guard: ti,ab,kw
4 wound-guard: ti,ab,kw
5 wound edge protect:ti,ab,kw
6 wound-edge protect*:ti,ab,kw
7 impervious wound drape:ti,ab,kw
8 impervious wound protect*:ti,ab,kw
9 ring drape:ti,ab,kw
10 drape protect*:ti,ab,kw
11 barrier protect*: ti,ab,kw
12 ViDrape: ti,ab,kw
13 Vi Drape: ti,ab,kw
14 Steri Drape: ti,ab,kw
15 SteriDrape: ti,ab,kw
16 Alexis: ti,ab,kw
17 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16
18 surgical wound infection*: ti,ab,kw
19 surgical site infection*: ti,ab,kw
20 surgical infection*: ti,ab,kw
21 wound infection*: ti,ab,kw
22 bacterial infection*: ti,ab,kw
23 wound complication*: ti,ab,kw
24 postoperative complication*: ti,ab,kw
25 postoperative infection*: ti,ab,kw
26 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25
27 17 and 26

Appendix 3. Search strategy for the systematic review of SSI utility values

OVID MEDLINE and MEDLINE-In-Process

1	exp Quality of Life/
2	Quality of life.tw.
3	Life quality.tw.
4	Hrql.tw.
5	Hrqol.tw.
6	Hql.tw.
7	Qol.tw.
8	Ql.tw.
9	Sf$.tw.
10	Short form.tw.
11	Shortform.tw.
12	Euroqol.tw.
13	Eq 5d.tw.
14	Eq5d.tw.
15	Qaly$.tw.
16	Quality adjusted life year$.tw.
17	Hye.tw.
18	Psychological general well being.tw.
19	Pgwb$.tw.
20	Health utilit$.tw.
21	Hui$.tw.
22	Quality of wellbeing.tw.
23	Quality of well being.tw.
24	Qwb$.tw.
25	General health questionnaire$.tw.
26	Ghq.tw.
27	Nottingham health profile.tw.
28	Nhp.tw.
29	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28
30	Exp Surgical Wound Infection/
31	Exp Wound Infection/
32	Surgical wound$.tw.
33	Postoperative wound infection.tw.
34	Surgical wound infection.tw.
35	(wound infection adj8 surgery).tw.
36	Wound infec$.tw.
37	30 or 31 or 32 or 33 or 34 or 35 or 36
38	29 and 37

OVID EMBASE

1 exp Quality of Life/
2 Quality of life.tw.
3 Life quality.tw.
4 Hrql.tw.
5 Hrqol.tw.
6 Hql.tw.
7 Qol.tw.
8 Ql.tw.
9 Sf$.tw.
10 Short form.tw.
11 Shortform.tw.
12 Euroqol.tw.
13 Eq 5d.tw.
14 Eq5d.tw.
15 Qaly$.tw.
16 Quality adjusted life year$.tw.
17 Hye.tw.
18 Psychological general well being.tw.
19 Pgwb$.tw.
20 Health utilit$.tw.
21 Hui$.tw.
22 Quality of wellbeing.tw.
23 Quality of well being.tw.
24 Qwb$.tw.
25 General health questionnaire$.tw.
26 Ghq.tw.
27 Nottingham health profile.tw.
28 Nhp.tw.
29 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28

30 Exp Surgical Wound Infection/
31 Exp Wound Infection/
32 Exp Surgical wound/
33 Exp Surgical infection/
34 Surgical wound infection.tw.
35 (wound infection adj8 surgery).tw.
36 Wound infec$.tw.
37 30 or 31 or 32 or 33 or 34 or 35 or 36
38 29 and 37

ISI Web of Science

1 TS=(euroqol or euro qol or eq5d or eq 5d or eq-5d)

2 TS=(hui or hui1 or hui2 or hui3)

3 TS=(sf36 or sf 36 or short form 36 or shortform 36 or sf thirtysix or sf thirty six or shortform thirtysix or shortform thirty six or short form thirtysix or short form thirty six)

4 TS=(sf6 or sf 6 or short form 6 or shortform 6 or sf six or sfsix or shortform six or short form six)

5 TS=(sf12 or sf 12 or short form 12 or shortform 12 or sf twelve or sftwelve or shortform twelve or short form twelve)

6 TS=(sf16 or sf 16 or short form 16 or shortform 16 or sf sixteen or sfsixteen or shortform sixteen or short form sixteen)

7 TS=(sf20 or sf 20 or short form 20 or shortform 20 or sf twenty or sftwenty or shortform twenty or short form twenty)

8 TS=(quality of wellbeing or qwb)

9 TS=(quality adjusted life or qaly$)

10 TS=(eortc)

11 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10

12 TS=(cost effective* or cost-effective or cost-utility or cost utility or cost-benefit or cost benefit or economic evaluation*)

13 11 or 12

14 TS=(surgical site infection*)

15 TS=(surgical infection*)

16 TS=(wound infection*)

17 TS=(surgical wound infection*)

18 TS=(wound infection adj8 surgery)

19 TS=(surgical wound*)

20 12 or 13 or 14 or 15 or 16 or 17

21 13 and 20

NHS Economic Evaluation Database (NHS EED)

1 Surgical site infect*

2 Surgical wound infect*

3 Surg* infect*

4 Wound infect*

5 SSI

6 1 or 2 or 3 or 4 or 5

Appendix 4. Screened papers not included in the systematic review of SSI utility values

Study ID	Study type	Reason for exclusion
Cooper 2002	Decision model	Contains no utility data.
Davey 1992	Decision model	Contains no utility data.
Edwards 2006	Decision model	Not relevant for SSI: utility refers to 'severe infections in ICU' and has been approximated as the utility value for the 'unconscious' state.
Elliott 2010	Decision model	Duplicate of Cranny et al (2009)
Falavigna 2011	Primary study	SF-36 instrument used to elicit HRQoL for deep wound infection after spine surgery. Time of elicitation was long and unclear, median 22 months (range 6 to 108 months).
Fedorka 2011	Primary study	SF-12 instrument used to elicit HRQoL after above the knee amputation in patients with infected total knee arthroplasty. Did not have an appropriate control group (uninfected).
Fisman 2001	Decision model	Not relevant for SSI: utility values refer to functional prosthesis and resection arthroplasty.
Ginandes 2003	Primary study	SF-36 instrument used to elicit HRQoL in a trial investigating the effect of hypnosis on wound healing. No relevant comparison reported i.e. infected vs uninfected.
Hertzman 1990	Decision model	Contains no utility data.
Immer 2005	Primary study	SF-36 instrument used to elicit HRQoL in patients with deep sternal wound infection after cardiac surgery. No relevant comparison reported i.e. infected vs uninfected.
Juricek 2010	Primary study	SF-36 instrument used to elicit HRQoL in patients with complications following spinal surgery. Time of elicitations was 6, 12 and 24 months post-operatively and reporting is unclear.
Klesius 2004	Primary study	Nottingham Health Profile used to assess HRQoL in patients undergoing sternal resection to treat deep sternal infections following cardiac surgery. No relevant comparison reported i.e. infected vs uninfected.
Klinger 2006	Primary study	SF-36 instrument to evaluate HRQoL of patients with infected total knee arthroplasty. Time of elicitation mean 4.5 years (range 2 to 11 years). No relevant comparison reported i.e. infected vs uninfected.
Kobayashi 2011	Primary study	SF-36 instrument used to elicit HRQoL in patients with deep sternal wound infection after cardiac surgery. Time of elicitation mean 47.3 months after discharge.
Laudermilch 2010	Primary study	SF-36 instrument used to elicit HRQoL in patients undergoing revision total knee arthroplasty. Time of elicitation mean 3.3 years (range 2 to 5.7 years).
Leung 2011	Primary study	SF-12 instrument used to elicit HRQoL in patients with infected total knee arthroplasty. No relevant comparison reported i.e. infected vs uninfected.
Mbah 2012	Primary study	SF-36 to assess HRQoL in patients after pancreatic surgery. No relevant comparison reported i.e. infected vs uninfected.
Meek 2004	Primary study	SF-12 to evaluate HRQoL in patients after revision knee arthroplasty. Unclear whether patient groups being

Study ID	Study type	Reason for exclusion
		compared were infected/uninfected at the time of HRQoL elicitation.
Melling 2001	Primary study	No HRQoL instrument was administered.
Mok 2009	Primary study	SF-36 instrument to evaluate HRQoL in patients after instrumented posterior spinal fusion. Time of elicitation minimum 2 years post-operatively.
Naylor 2009	Primary study	SF-36 instrument to evaluate HRQoL in patients after total hip arthroplasty vs total knee arthroplasty. No relevant comparison reported i.e. infected vs uninfected.
Nguyen 2001	Primary study	SF-36 instrument to evaluate HRQoL in patients after laparoscopic vs open gastric bypass. No relevant comparison reported i.e. infected vs uninfected.
Nguyen 2007	Primary study	VascuQol questionnaire used to assess HRQoL after infrainguinal bypass in patients with wound complications. Wound complication defined as a composite of: infection, hematoma, seroma or lymphatic leak, necrosis, dehiscence, and erythema.
Pada 2011	Primary study	EQ-5D instrument used to assess HRQoL after general surgery. Estimates are presented for soft skin/tissue infection, which includes surgical site infection, decubitus ulcer infection and other types of infections not accounted for. Not clear whether utility estimates refer to SSI patients.
Petilon 2012	Primary study	SF-36 instrument to evaluate HRQoL in patients after lumbar fusion complicated by deep wound infection. Time of elicitation 2 years post-operatively.
Poelman 2010	Primary study	SF-36 instrument to evaluate HRQoL in patients after incisional hernia. No relevant comparison reported i.e. infected vs uninfected.
Robotham 2011	Decision model	Not relevant for SSI and long time horizon: QALYs estimated using a study for the first five years after ICU discharge.
Saeed 2001	Primary study	EQ-5D instrument used to assess HRQoL after coronary artery bypass surgery. No relevant comparison reported i.e. infected vs uninfected.
Slover 2006	Decision model	Invoked disutility values do not refer to SSI patients.
Slover 2011	Decision model	Contains no utility data.
Sonnenberg 1999	Decision model	Does not refer to SSI.
Teshima 2009	Primary study	No HRQoL instrument applied.
Uyl-de Groot 2004	Primary study	SF-36 instrument to evaluate HRQoL in patients after radical vulvectomy and bilateral inguino-femoral lymphadenectomy. No relevant comparison reported i.e. infected vs uninfected.
VandenBergh 1996	Decision model	Contains no utility data.

Study ID	Study type	Reason for exclusion
Wassenberg 2011	Decision model	Contains no utility data.
Whitehouse 2002	Primary study	SF-36 instrument to evaluate HRQoL in patients after general surgery. Time of elicitation was 1 year post-operatively.
Wynne 2004	Primary study	Patient comfort assessed for three wound dressings used in patients with sternal wound infections. No relevant comparison reported i.e. infected vs uninfected.
Young 2006	Decision model	Contains no utility data.

Appendix 5. CHEERS Statement for the WEPD vs. standard care decision model

Section/item	Item No	Recommendation	Reported on page No
Title and abstract			
Title	1	Identify the study as an economic evaluation or use more specific terms such as "cost-effectiveness analysis", and describe the interventions compared.	n/a
Abstract	2	Provide a structured summary of objectives, perspective, setting, methods (including study design and inputs), results (including base case and uncertainty analyses), and conclusions.	n/a
Introduction			
Background and objectives	3	Provide an explicit statement of the broader context for the study.	133
		Present the study question and its relevance for health policy or practice decisions.	133
Methods			
Target population and subgroups	4	Describe characteristics of the base case population and subgroups analysed, including why they were chosen.	133
Setting and location	5	State relevant aspects of the system(s) in which the decision(s) need(s) to be made.	133
Study perspective	6	Describe the perspective of the study and relate this to the costs being evaluated.	133
Comparators	7	Describe the interventions or strategies being compared and state why they were chosen.	133
Time horizon	8	State the time horizon(s) over which costs and consequences are being evaluated and say why appropriate.	134
Discount rate	9	Report the choice of discount rate(s) used for costs and outcomes and say why appropriate	134
Choice of health outcomes	10	Describe what outcomes were used as the measure(s) of benefit in the evaluation and their relevance for the type of analysis performed.	135
Measurement of effectiveness	11a	Single study-based estimates: Describe fully the design features of the single effectiveness study and why the single study was a sufficient source of clinical effectiveness data.	n/a
	11b	Synthesis-based estimates: Describe fully the methods used for identification of included studies and synthesis of clinical effectiveness data.	135
Measurement and valuation of preference based outcomes	12	If applicable, describe the population and methods used to elicit preferences for outcomes.	n/a
Estimating resources and costs	13a	Single study-based economic evaluation: Describe approaches used to estimate resource use associated with the alternative interventions. Describe primary	n/a

Section/item	Item No	Recommendation	Reported on page No
		or secondary research methods for valuing each resource item in terms of its unit cost. Describe any adjustments made to approximate to opportunity costs.	
	13b	Model-based economic evaluation: Describe approaches and data sources used to estimate resource use associated with model health states. Describe primary or secondary research methods for valuing each resource item in terms of its unit cost. Describe any adjustments made to approximate to opportunity costs.	135-136
Currency, price date, and conversion	14	Report the dates of the estimated resource quantities and unit costs. Describe methods for adjusting estimated unit costs to the year of reported costs if necessary. Describe methods for converting costs into a common currency base and the exchange rate.	135
Choice of model	15	Describe and give reasons for the specific type of decision analytical model used. Providing a figure to show model structure is strongly recommended.	137-138
Assumptions	16	Describe all structural or other assumptions underpinning the decision-analytical model.	137-139
Analytical methods	17	Describe all analytical methods supporting the evaluation. This could include methods for dealing with skewed, missing, or censored data; extrapolation methods; methods for pooling data; approaches to validate or make adjustments (such as half cycle corrections) to a model; and methods for handling population heterogeneity and uncertainty.	140-143
Results			
Study parameters	18	Report the values, ranges, references, and, if used, probability distributions for all parameters. Report reasons or sources for distributions used to represent uncertainty where appropriate. Providing a table to show the input values is strongly recommended.	144-146
Incremental costs and outcomes	19	For each intervention, report mean values for the main categories of estimated costs and outcomes of interest, as well as mean differences between the comparator groups. If applicable, report incremental cost-effectiveness ratios	147-148
Characterising uncertainty	20a	Single study-based economic evaluation: Describe the effects of sampling uncertainty for the estimated incremental cost and incremental effectiveness parameters, together with the impact of methodological assumptions (such as discount rate, study perspective).	n/a
	20b	Model-based economic evaluation: Describe the	147-154

Section/item	Item No	Recommendation	Reported on page No
		effects on the results of uncertainty for all input parameters, and uncertainty related to the structure of the model and assumptions.	
Characterising heterogeneity	21	If applicable, report differences in costs, outcomes, or cost-effectiveness that can be explained by variations between subgroups of patients with different baseline characteristics or other observed variability in effects that are not reducible by more information.	148
Discussion			
Study findings, limitations, generalisability, and current knowledge	22	Summarise key study findings and describe how they support the conclusions reached. Discuss limitations and the generalisability of the findings and how the findings fit with current knowledge.	154-166
Other			
Source of funding	23	Describe how the study was funded and the role of the funder in the identification, design, conduct, and reporting of the analysis. Describe other non-monetary sources of support.	n/a
Conflicts of interest	24	Describe any potential for conflict of interest of study contributors in accordance with journal policy. In the absence of a journal policy, we recommend authors comply with International Committee of Medical Journal Editors recommendations.	n/a

Appendix 6. CHEERS Statement for the ROSSINI trial economic evaluation

Appendix 7. EQ-5D questionnaire used in the ROSSINI trial

414

Appendix 8. Case report forms used in the ROSSINI trial

415

ROSSINI – Reduction of Surgical Site Infection using a Novel Intervention
CRF 2 – Operative Data (Version 3.0 dated 14th March 2011)

417

Appendix 9. Studies included in the systematic review of RCT protocols

Study ID	Surname of Chief Investigator	HTA Reference	Project title
1	Gregory	03/46/09	Development and evaluation by a cluster randomised trial of a psychosocial intervention in children and teenagers experiencing diabetes: the DEPICTED study
2	Cockayne	05/513/02	EVerT: cryotherapy versus salicylic acid for the treatment of verrucae - a randomised controlled trial
3	Crawford	04/39/04	Group art therapy as an adjunctive treatment for people with schizophrenia: a randomised controlled trial (MATISSE)
4	Rintoul	06/302/216	Clinical effectiveness and costeffectiveness of endobronchial and endoscopic ultrasound relative to surgical staging in potentially resectable lung cancer: results from the ASTER randomised controlled trial
5	Banerjee	04/11/02	Study of the use of antidepressants for depression in dementia: the HTA -SADD trial - a multicentre, randomised, double-blind, placebo-controlled trial of the clinical effectiveness and cost-effectiveness of sertraline and mirtazapine
7	Collinson	09/22/16	Randomised Assessment of Treatment using Panel Assay of Cardiac markers - Contemporary Biomarker Evaluation (RATPAC CBE)
8	Woods	06/304/229	REMCARE: reminiscence groups for people with dementia and their family caregivers - effectiveness and cost-effectiveness pragmatic multicentre randomised trial
9	Lenney	05/503/04	Management of Asthma in School age Children On Therapy (MASCOT): a randomised, double-blind, placebo-controlled, parallel study of efficacy and safety
10	N'Dow	05/46/01	Types of urethral catheter for reducing symptomatic urinary tract infections in hospitalised adults requiring short-term catheterisation: multicentre randomised controlled trial and economic evaluation of antimicrobial- and antisepticimpregnated urethral catheters (the CATHETER trial)
11	Molassiotis	07/31/02	The effectiveness and cost-effectiveness of acupressure for the control and management of chemotherapy-related acute and delayed nausea
12	Gilbody	07/41/05	Smoking cessation for people with severe mental illness: a pilot study and definitive randomised evaluation of a bespoke service
13	Gilbody	08/19/04	Collaborative care and active surveillance for screen-positive elders with sub-clinical depression: a pilot study and definitive and randomised evaluation - the CASPER trial
14	Gilbert	08/58/02	A randomised trial to increase the uptake of smoking cessation services using personal targeted risk information and taster sessions
15	Powell	05/503/10	MAGnesium NEbuliser Trial In Children (MAGNETIC)
16	Underwood	06/02/01	Exercise for depression in care home residents: a randomised controlled trial with cost-effectiveness analysis

Study ID	Surname of Chief Investigator	HTA Reference	Project title
			(OPERA)
17	Little	05/10/01	PRImary care Streptococcal Management study (PRISM)
18	Russell	04/35/08	Folate Augmentation of Treatment - Evaluation for Depression: randomised controlled trial (FolATED)
19	Willner	08/53/34	A cluster randomised controlled trial of a manualised cognitive behavioural anger management intervention delivered by supervised lay therapists to people with intellectual disabilities
20	Wiles	06/404/02	Cognitive behavioural therapy as an adjunct to pharmacotherapy for treatment resistant depression in primary care: a randomised controlled trial
21	Nashef	07/01/34	A randomised controlled trial to investigate the clinical and cost effectiveness of adding an ablation device-based maze procedure as a routine adjunct to elective cardiac surgery for patients with pre-existing atrial fibrillation (AMAZE)
22	Macrae	05/506/03	Control of Hyperglycaemia in Paediatric intensive care trial (the CHIP trial)
23	Coulton	06/304/142	The effectiveness and cost-effectiveness of opportunistic screening and stepped care interventions for older hazardous alcohol users in primary care (AESOPS)
24	Lamb	07/32/05	SARAH: Strengthening And stretching for people with Rheumatoid Arthritis of the Hands: The clinical and cost-effectiveness of an exercise programme over and above usual care
25	Lloyd Scott	06/303/84	Randomised controlled trial of tumour-necrosis-factor inhibitors against combination intensive therapy with conventional disease modifying anti-rheumatic drugs in established rheumatoid arthritis: the TACIT trial
26	Stallard	06/37/04	A single blind randomised controlled trial to determine the effectiveness of group cognitive behaviour therapy (CBT) in the prevention of depression in high risk adolescents
27	Dumville	07/60/26	VenUS IV (Venous leg Ulcer Study IV): A randomised controlled trial of compression hosiery versus compression bandaging in the treatment of venous leg ulcers
28	Snooks	07/01/21	Care of older people who fall: evaluation of the clinical and cost effectiveness of new protocols for emergency ambulance personnel to assess and refer to appropriate community based care
29	Goyder	07/25/02	A randomised controlled trial and cost-effectiveness evaluation of "booster" interventions to sustain increases in physical activity in middle-aged adults in deprived urban neighbourhoods
30	Williams	06/78/03	COmparison of iNfliximab and ciclosporin in STeroid Resistant Ulcerative Colitis: a Trial (CONSTRUCT)
31	Goodacre	06/01/02	The 3Mg Trial: Randomised controlled trial of intravenous or nebulised magnesium sulphate or standard therapy for acute severe asthma
32	Allen	06/39/02	A multicentre, randomised, placebo controlled trial of lactic acid bacteria in prevention of antibiotic-associated diarrhoea (AAD) & Clostridium difficile diarrhoea (CDD) in patients aged 65 years & over admitted to hospital and receiving antibiotics

Study ID	Surname of Chief Investigator	HTA Reference	Project title
33	Rai	08/38/01	First trimester progesterone therapy in women with a history of unexplained recurrent miscarriages: A randomised, double-blind, placebo-controlled, multi-centre trial [The PROMISE (PROgesterone in recurrent MIScarriage) Trial]
34	Coleman	06/07/01	Double-blind, randomised, placebo-controlled trial of nicotine replacement therapy (NRT) in pregnancy - SNAP
35	Wolf	05/515/01	SLEEPS: Safety profiLe, Efficacy and Equivalence in Paediatric intensive care Sedation: a comparison of clonidine and midazolam
36	Logan	08/14/51	Getting out of the house: a multi centre trial to evaluate an outdoor mobility intervention for people who have had a stroke
37	Field	09/61/01	United Kingdom Lung Cancer Screening Trial (UKLS)
38	Chakravarthy	07/36/01	A randomised controlled trial (RCT) of alternative treatments to Inhibit VEGF in patients with Age-related choroidal Neovascularisation (IVAN)
39	Christie	06/44/05	Maximising engagement, motivation and long term change in a structured intensive education programme in diabetes for children, young people and their families: child and adolescent structured competencies approach to diabetes education
40	James	06/303/205	A randomised phase III trial of Docetaxel plus Prednisolone vs. Docetaxel with Prednisolone plus either Zoledronic acid, Strontium-89 or both agents combined (TRAPEZE)
41	Gilbert	08/13/47	CATheter Infections in Children - the CATCH trial
42	Earl	06/303/98	PERSEPHONE - duration of trastuzumab study with chemotherapy in early breast cancer: six versus twelve months
43	Mehanna	06/302/129	Positron Emission Tomography-Computerised Tomography scans (PET-CT) guided watch and wait policy versus planned neck dissection for the management of locally advanced (N2/N3) nodal metastases in patients with head and neck squamous cancer
44	Young	06/04/01	A randomised controlled trial of high frequency oscillatory ventilation in patients with acute respiratory distress syndrome (OSCAR)
45	Clark	06/404/84	A randomised controlled trial of Outpatient Polyp Treatment (OPT) for abnormal uterine bleeding
46	Tyrer	07/01/26	Cognitive-behavioural therapy for Health Anxiety in Medical Patients (CHAMP)
47	Cassell	07/43/01	The relative clinical and cost-effectiveness of three contrasting approaches to partner notification for curable sexually transmitted infections (STIs): a cluster randomised trial in primary care
48	Powell	07/37/64	Can emergency endovascular aneurysm repair (eEVAR) improves the survival from ruptured abdominal aortic aneurysm?
49	Knowles	09/104/16	CONtrol of Faecal Incontinence using Distal NeuromodulaTion (CONFIDeNT)
50	Clark	06/80/01	Does home oxygen therapy (HOT) in addition to standard care improve disease severity and symptoms in chronic heart failure?

Study ID	Surname of Chief Investigator	HTA Reference	Project title
51	Iliffe	06/36/04	Multi-centre cluster trial in primary care comparing a community group exercise programme with home based exercise and with usual care for people aged 65 and over
52	Brittended	06/45/02	Randomised controlled trial comparing foam sclerotherapy, alone or in combination with endovenous laser therapy, with conventional surgery as a treatment for varicose veins
53	Morrell	08/56/02	A randomised controlled trial of continuous positive airway pressure treatment in older people with obstructive sleep apnoea hypopnoea syndrome (PREDICT)
54	Williamson	09/01/27	An open randomised study of autoinflation in school age children (4-11 years) with otitis media with effusion (OME) in primary care
55	Anie	07/48/01	An evaluation of the effectiveness of ibuprofen and morphine for acute pain in sickle cell disease
56	Sackley	08/14/30	A cluster randomised controlled trial of an occupational therapy intervention for residents with stroke living in UK care-homes
57	Livingston	08/14/06	The START (STrAtegies for RelaTives) study: a pragmatic randomised controlled trial to determine the effectiveness of a manual based coping strategy programme in promoting the mental health of carers of people with dementia
58	Cunningham	09/91/16	Bronchiolitis of Infancy Discharge Study (BIDS)
59	Williams	06/403/51	A randomised controlled trial to compare the safety and effectiveness of doxycycline (200 mg/day) with prednisolone (0.5 mg/kg/day) for initial treatment of bullous pemphigoid
60	Priebe	07/60/43	Financial incentives to improve adherence to anti-psychotic maintenance medication in non-adherent patients - a cluster randomised controlled trial: FIAT (Financial Incentives for Adherence to Treatment)
61	Gilbody	06/43/05	The Randomised Evaluation of the Effectiveness and Acceptability of Computerised Therapy (REEACT) Trial
62	Camobell	08/53/15	The effectiveness and cost effectiveness of telephone triage of patients requesting same day consultations in general practice: a cluster randomised controlled trial comparing nurse-led and GP-led management systems. - The ESTEEM trial
63	Hillmen	07/01/38	A randomised, phase II trial in previously untreated patients with chronic lymphocytic leukaemia to compare fludarabine, cyclophosphamide and rituximab with fludarabine and cyclophosphamide, mitoxantrone and low dose rituximab. (CLL6)
64	Reeves	06/402/94	A multi-centre randomised controlled trial of Transfusion Indication Threshold Reduction on transfusion rates, morbidity and healthcare resource use following cardiac surgery (TITRe 2)
65	McClinton	08/71/01	Use of drug therapy in the management of symptomatic ureteric stones in hospitalised adults: multicentre placebo controlled randomised trial of calcium channel blockers (nifedipine) and alpha blockers (tamsulosin) - The SUSPEND trial

Study ID	Surname of Chief Investigator	HTA Reference	Project title
66	Little	09/127/19	Positive Online WEight Reduction (POWER)
67	Clarke	07/01/07	Randomised controlled trial to assess the clinical- and cost-effectiveness of physiotherapy and occupational therapy in Parkinson's disease (PD REHAB)
68	Costa	08/116/97	A randomised controlled trial of percutaneous fixation with Kirschner wires versus volar locking-plate fixation in the treatment of adult patients with a displaced fracture of the distal radius
69	Rangan	06/404/53	Pragmatic multi-centre randomised trial of surgical versus non-surgical treatment for proximal fracture of the humerus in adults
70	Mendelow	07/37/16	Surgical Trial In Traumatic intraCerebral Haemorrhage [STITCH)
71	Halliday	06/301/233	Asymptomatic Carotid Surgery Trial-2 (ACST-2): an international randomised trial to compare carotid endarterectomy with carotid artery stenting to prevent stroke
72	Rowan	07/52/03	CALORIES: A phase III, open, multicentre, randomised controlled trial comparing the clinical and cost-effectiveness of early nutritional support in critically ill patients via the parenteral versus the enteral route
73	Rowan	07/37/47	Protocolised Management In Sepsis (ProMISe): a multicentre, randomised controlled trial of the clinical and cost-effectiveness of early protocolised resuscitation for emerging septic shock
74	Kuyken	08/56/01	Preventing depressive relapse in NHS Practice through mindfulness-based cognitive therapy (MBCT)
75	Barnes	08/116/12	Amisulpride augmentation in clozapine-unresponsive schizophrenia (AMICUS)
76	Ussher	07/01/14	A pragmatic randomized controlled trial of physical activity as an aid to smoking cessation during pregnancy
77	Brocklehurst	06/38/01	A multicentre randomised controlled trial of an intelligent system to support decision making in the management of labour using the cardiotocogram (INFANT)
78	Song	09/91/36	A randomised controlled trial of self-help materials for the prevention of smoking relapse
79	Carr	05/47/02	The clinical and cost-effectiveness of arthroscopic versus open surgical repair for tears of the rotator cuff (UKUFF trial)
80	Paton	06/403/90	A randomised controlled trial of a protease inhibitor monotherapy versus continuing combination antiretroviral therapy for HIV-1 infected patients previously established on a dual nucleoside and non-nucleoside combination regimen
81	Barnes	07/83/01	Antidepressant Controlled Trial of Negative symptoms in Schizophrenia (ACTIONS)
82	Thursz	08/14/44	STeroids or Pentoxifyline for Alcoholic Hepatitis (STOPAH) Trial
83	Orrell	08/116/06	Individual Cognitive Stimulation Therapy for dementia (iCST Trial)
84	McMurran	08/53/06	Psychoeducation with problem solving (PEPS) therapy for adults with personality disorder: A community-based, randomised controlled trial
85	Gates	07/37/69	A randomised controlled trial of the LUCAS mechanical compression/decompression device for out of hospital

Study ID	Surname of Chief Investigator	HTA Reference	Project title
			cardiac arrest
86	Brocklehurst	08/22/02	A study of position during the late stages of labour in women with an epidural
87	Khan	09/22/50	Can magnetic resonance imaging scan replace, or triage the use of laparoscopy in establishing diagnosis among women presenting in secondary care with chronic pelvic pain?
88	Cottrell	07/33/01	SHIFT. Self-Harm Intervention, Family Therapy: a randomised controlled trial of family therapy vs. treatment as usual for young people seen after second or subsequent episodes of self-harm
89	Hewer	07/51/01	Torpedo-CF: Trial of optimal therapy for pseudomonas eradication in cystic fibrosis
90	Robertson	09/127/41	A randomised controlled trial evaluating the effectiveness and cost-effectiveness of Families for Health, a family-based childhood obesity management intervention delivered in a community setting for ages 7 to 11
91	Willett	07/37/61	Comparison of close contact cast (CCC) technique to open surgical reduction and internal fixation (ORIF) in the treatment of unstable ankle fractures in patients over 60 years
92	Priebe	08/116/68	Effectiveness and Cost-Effectiveness of Body Psychotherapy in the Treatment of Negative Symptoms of Schizophrenia. A multi-centre randomised controlled trial
93	Khan	09/55/38	Antiepileptic drug (AED) management in Pregnancy: An evaluation of effectiveness, cost effectiveness and acceptability of dose adjustment strategies
94	McDermott	09/55/33	A Randomised Controlled Trial In Patients With Respiratory Muscle Weakness Due to Motor Neurone Disease of the NeuRx RA/4 Diaphragm Pacing System (DiPALS)
95	Tickle	08/14/19	A randomised control trial to measure the effects and costs of a dental caries prevention regime for young children attending primary care dental services (Northern Ireland Caries Prevention In Practice Trial - NIC-PIP trial)
96	Goodyer	06/05/01	Randomised controlled trial of brief psychodynamic psychotherapy, cognitive behaviour therapy and treatment as usual in adolescents with moderate to severe depression attending routine child and adolescent mental health clinics
97	Glazener	07/60/18	Clinical and cost-effectiveness of surgical options for the management of anterior and/or posterior vaginal wall prolapse: two randomised controlled trials within a Comprehensive Cohort Study
98	Simpson	08/44/04	Weight Loss Maintenance in Adults: A 3 arm individually randomised controlled trial to evaluate the impact of a 12-month multi-component intervention and less intensive version compared to a control on weight loss maintenance in obese adults
99	Watson	08/24/02	A pragmatic multicentre randomised controlled trial comparing stapled haemorrhoidopexy to conventional excisional surgery for haemorrhoidal disease
100	Lamb	09/80/04	Physical activity programmes for community dwelling people with mild to moderate dementia (DAPA - Dementia And Physical Activity)
102	Webb	08/53/31	Long-term tapering versus standard prednisolone (steroid) therapy for the treatment of the initial episode of

Study ID	Surname of Chief Investigator	HTA Reference	Project title
			childhood nephrotic syndrome: national multicentre randomised double blind controlled trial
103	Kitchener	09/164/01	Strategies to increase cervical screening uptake at first invitation (STRATEGIC)
104	Jayne	07/89/01	FIAT (Fistula-in-ano trial) comparing Surgisis® anal fistula plug versus surgeon's preference (advancement flap, fistulotomy, cutting seton) for transsphincteric fistula-in-ano
105	Buch	08/116/75	SWITCH - Randomised- controlled trial of switching to alternative tumour necrosis factor-blocking drugs or abatacept or rituximab in patients with rheumatoid arthritis who have failed an initial TNF-blocking drug
106	Heller	08/107/01	The REPOSE (Relative Effectiveness of Pumps Over MDI and Structured Education) Trial
107	Jeffcoate	09/01/53	Evaluation of lightweight fibreglass heel casts in the management of ulcers of the heel in diabetes
108	Clarkson	09/01/45	Improving the Quality of Dentistry (IQuaD): A randomised controlled trial comparing oral hygiene advice and periodontal instrumentation for the prevention and management of periodontal disease in dentate adults attending dental primary care
109	Walsh	09/144/51	The Age of Blood Evaluation Study (ABLE)
110	Drayson	08/116/69	Tackling Early Morbidity and Mortality in Myeloma: Assessing the benefit of antibiotic prophylaxis and its effect on healthcare associated infections
111	Jayne	08/56/04	Plasma exchange and glucocorticoids in anti-neutrophil cytoplasm antibody associated systemic vasculitis: a randomized controlled trial. PEXIVAS
112	McPherson	08/53/22	FEMME trial: Randomised trial of treating Fibroids with either Embolisation or MyoMectomy to measure the Effect on quality of life
113	Beard	08/14/08	Total or Partial Knee Arthroplasty Trial (TOPKAT)
114	Barr	05/12/01	Randomised control trial of surveillance and no surveillance for patients with Barrett's oesophagus - BOSS (Barrett's Oesophagus Surveillance Study)
115	Goodacre	06/302/19	The RATPAC (Randomised Assessment of Treatment using Panel Assay of Cardiac markers) trial: a randomised controlled trial of point-of-care cardiac markers in the emergency department
116	Nelson	02/37/03	VenUS III: a randomised controlled trial of therapeutic ultrasound in the management of venous leg ulcers
117	Williams	05/16/01	A multicentre randomised controlled trial and economic evaluation of ion-exchange water softeners for the treatment of eczema in children: the Softened Water Eczema Trial (SWET)
118	Kitchener	03/04/02	MAVARIC - a comparison of automation-assisted and manual cervical screening: a randomised controlled trial
119	Rodgers	02/41/06	BoTULS: a multicentre randomised controlled trial to evaluate the clinical effectiveness and cost-effectiveness of treating upper limb spasticity due to stroke with botulinum toxin type A
120	Cross	03/13/06	A randomised controlled equivalence trial to determine the effectiveness and cost-utility of manual chest physiotherapy techniques in the management of exacerbations of chronic obstructive pulmonary disease

Study ID	Surname of Chief Investigator	HTA Reference	Project title
			(MATREX)
121	Kilby	07/01/44	The PLUTO Trial: Percutaneous shunting in Lower Urinary Tract Obstruction
122	Blair	08/14/39	Randomised controlled trial of continuous subcutaneous insulin infusion compared to multiple daily injection regimens in children and young people at diagnosis of type I diabetes mellitus
123	Torgerson	09/77/01	Randomised trial of a multifaceted podiatry intervention for fall prevention
124	Lovell	09/81/01	Obsessive Compulsive Treatment Efficacy Trial (OCTET)
125	Johnson	09/144/50	Randomised controlled trial of the clinical and cost-effectiveness of a contingency management intervention for reduction of cannabis use and of relapse in early psychosis
126	Hamilton-Shield	09/127/04	Changing eating behaviours to treat childhood obesity in the community using Mandolean: the ComMando, (Community Mandolean) randomised trial
127	Brown	09/91/21	A randomised, multi-stage phase II/III study of Sunitinib comparing Temporary cessation with Allowing continuation, at the time of maximal response, in the first-line treatment of locally advanced/metastatic Renal cell carcinoma (the STAR trial)
128	Blazeby	09/127/53	BY-BAND. Gastric BYpass or adjustable gastric BANDing surgery to treat morbid obesity: a multi-centre randomised controlled trial
129	Lewis	03/45/07	A pragmatic randomised controlled trial to evaluate the cost-effectiveness of a physical activity intervention as a treatment for depression: the treating depression with physical activity (TREAD) trial

Appendix 10. Systematic review: complete results of meta-summary

	Number (%) of included RCTs reporting each consideration													
	TOTAL		NON-PHARMACOLOGIC (n=33)			PHARMACOLOGIC (n=96)			CLUSTER (n=14)			NON-CLUSTER (n=115)		
	N	% total	N	% total	% group	N	% total	% group	N	% total	% group	N	% total	% group
CENTRE SELECTION CONSIDERATIONS	**78**	**60%**	**56**	**43%**	**58%**	**22**	**17%**	**67%**	**13**	**10%**	**93%**	**65**	**50%**	**57%**
DIVERSITY AND REPRESENTATIVENESS	**31**	**24%**	**26**	**20%**	**27%**	**5**	**4%**	**15%**	**6**	**5%**	**43%**	**25**	**19%**	**22%**
POPULATION CHARACTERISTICS	**14**	**11%**	**13**	**10%**	**14%**	**1**	**1%**	**3%**	**2**	**2%**	**14%**	**12**	**9%**	**10%**
cultural background	1	1%	1	1%	1%	0	0%	0%	0	0%	0%	1	1%	1%
ethnicity	9	7%	8	6%	8%	1	1%	3%	1	1%	7%	8	6%	7%
socio-economic status	10	8%	9	7%	9%	1	1%	3%	2	2%	14%	8	6%	7%
HEALTH SERVICE DELIVERY	**15**	**12%**	**13**	**10%**	**14%**	**2**	**2%**	**6%**	**6**	**5%**	**43%**	**9**	**7%**	**8%**
patient case-mix	2	2%	2	2%	2%	**1**	1%	3%	2	2%	14%	2	2%	2%
intervention throughput	1	1%	1	1%	1%	**0**	0%	0%	1	1%	7%	1	1%	1%
organisations or practitioners	9	7%	9	7%	9%	**1**	1%	3%	9	7%	64%	9	7%	8%
services offered	4	3%	4	3%	4%	**0**	0%	0%	4	3%	29%	4	3%	3%
CENTRE SETTING	**15**	**12%**	**12**	**9%**	**13%**	**3**	**2%**	**9%**	**2**	**2%**	**14%**	**13**	**10%**	**11%**
environment	1	1%	1	1%	1%	0	0%	0%	0	0%	0%	1	1%	1%
regions	3	2%	1	1%	1%	2	2%	6%	0	0%	0%	3	2%	3%
size	1	1%	1	1%	1%	0	0%	0%	1	1%	7%	0	0%	0%
plain setting	2	2%	2	2%	2%	0	0%	0%	2	2%	14%	2	2%	2%
type of communities	1	1%	1	1%	1%	0	0%	0%	1	1%	7%	1	1%	1%
urban vs. rural	8	6%	8	6%	8%	1	1%	3%	8	6%	57%	8	6%	7%
CENTRE CHARACTERISTICS	**57**	**44%**	**39**	**30%**	**41%**	**18**	**14%**	**55%**	**7**	**5%**	**50%**	**50**	**39%**	**43%**
CENTRE SETTING	**4**	**3%**	**3**	**2%**	**3%**	**1**	**1%**	**3%**	**2**	**2%**	**14%**	**2**	**2%**	**2%**
geographical location	2	2%	1	1%	1%	1	1%	3%	1	1%	7%	1	1%	1%
uniqueness in the region	1	1%	1	1%	1%	0	0%	0%	0	0%	0%	1	1%	1%
deprivation status	1	1%	1	1%	1%	0	0%	0%	1	1%	7%	0	0%	0%
HEALTH SERVICE DELIVERY ('RESEARCH-	**16**	**12%**	**11**	**9%**	**11%**	**5**	**4%**	**15%**	**0**	**0%**	**0%**	**16**	**12%**	**14%**

	Number (%) of included RCTs reporting each consideration													
	TOTAL		**NON-PHARMACOLOGIC (n=33)**			**PHARMACOLOGIC (n=96)**			**CLUSTER (n=14)**			**NON-CLUSTER (n=115)**		
	N	**% total**	**N**	**% total**	**% group**	**N**	**% total**	**% group**	**N**	**% total**	**% group**	**N**	**% total**	**% group**
READY')														
centre of excellence	1	1%	1	1%	1%	0	0%	0%	0	0%	0%	1	1%	1%
clinical interest	6	5%	2	2%	2%	4	3%	12%	0	0%	0%	6	5%	5%
computer systems	1	1%	0	0%	0%	1	1%	3%	0	0%	0%	1	1%	1%
Department of Health approved centre	1	1%	1	1%	1%	0	0%	0%	0	0%	0%	1	1%	1%
links with other facilities	3	2%	2	2%	2%	1	1%	3%	0	0%	0%	3	2%	3%
NHS centre	7	5%	6	5%	6%	1	1%	3%	0	0%	0%	7	5%	6%
satisfactory peer review	1	1%	1	1%	1%	0	0%	0%	0	0%	0%	1	1%	1%
INTERVENTION	**31**	**24%**	**24**	**19%**	**25%**	**7**	**5%**	**21%**	**3**	**2%**	**21%**	**28**	**22%**	**24%**
appropriate training	3	2%	3	2%	3%	0	0%	0%	0	0%	0%	3	2%	3%
suitable to implement the intervention	16	12%	12	9%	13%	4	3%	12%	0	0%	0%	16	12%	14%
experience in delivering the intervention	13	10%	10	8%	10%	3	2%	9%	0	0%	0%	13	10%	11%
not running the intervention	6	5%	6	5%	6%	0	0%	0%	3	2%	21%	3	2%	3%
performance in delivering the intervention	5	4%	4	3%	4%	1	1%	3%	0	0%	0%	5	4%	4%
RESEARCH	**19**	**15%**	**11**	**9%**	**11%**	**8**	**6%**	**24%**	**2**	**2%**	**14%**	**17**	**13%**	**15%**
able to support research	2	2%	2	2%	2%	0	0%	0%	0	0%	0%	2	2%	2%
part of a research network	10	8%	4	3%	4%	6	5%	18%	1	1%	7%	9	7%	8%
research experience	10	8%	6	5%	6%	4	3%	12%	0	0%	0%	10	8%	9%
interest in research	3	**2%**	2	2%	2%	1	1%	3%	1	1%	7%	2	2%	2%
CENTRE SIZE	**22**	**17%**	**16**	**12%**	**17%**	**6**	**5%**	**18%**	**4**	**3%**	**29%**	**18**	**14%**	**16%**
catchment area	7	5%	4	3%	4%	3	2%	9%	1	1%	7%	6	5%	5%
patient throughput	11	9%	8	6%	8%	3	2%	9%	1	1%	7%	10	8%	9%
size of centre	5	4%	4	3%	4%	1	1%	3%	2	2%	14%	3	2%	3%
TRIAL PARTICIPATION	**37**	**29%**	**23**	**18%**	**24%**	**14**	**11%**	**42%**	**8**	**6%**	**57%**	**29**	**22%**	**25%**
RECRUITMENT	**17**	**13%**	**10**	**8%**	**10%**	**7**	**5%**	**21%**	**3**	**2%**	**21%**	**14**	**11%**	**12%**
ability to recruit	10	8%	4	3%	4%	6	5%	18%	0	0%	0%	10	8%	9%
access to study population	8	6%	6	5%	6%	2	2%	6%	3	2%	21%	5	4%	4%
TRIAL CONSTRAINTS	**5**	**4%**	**5**	**4%**	**5%**	**0**	**0%**	**0%**	**4**	**3%**	**29%**	**1**	**1%**	**1%**
proximity to study site	2	2%	2	2%	2%	0	0%	0%	2	2%	14%	0	0%	0%

	Number (%) of included RCTs reporting each consideration													
	TOTAL		**NON-PHARMACOLOGIC (n=33)**			**PHARMACOLOGIC (n=96)**			**CLUSTER (n=14)**			**NON-CLUSTER (n=115)**		
	N	**% total**	**N**	**% total**	**% group**	**N**	**% total**	**% group**	**N**	**% total**	**% group**	**N**	**% total**	**% group**
costs to trial	2	2%	2	2%	2%	0	0%	0%	1	1%	7%	1	1%	1%
time frame of trial	1	1%	1	1%	1%	0	0%	0%	1	1%	7%	0	0%	0%
ENSURING TRIAL PROCESSES AND REQ.	**24**	**19%**	**13**	**10%**	**14%**	**11**	**9%**	**33%**	**3**	**2%**	**21%**	**21**	**16%**	**18%**
arrange follow-up	1	1%	1	1%	1%	0	0%	0%	0	0%	0%	1	1%	1%
compliance with trial procedures and regulatory requirements	17	13%	8	6%	8%	9	7%	27%	2	2%	14%	15	12%	13%
ensuring communication	6	5%	1	1%	1%	5	4%	15%	0	0%	0%	6	5%	5%
identify champions	1	1%	1	1%	1%	0	0%	0%	0	0%	0%	1	1%	1%
required time, staff, facilities	16	12%	7	5%	7%	9	7%	27%	2	2%	14%	14	11%	12%
SUPPORT	**7**	**5%**	**6**	**5%**	**6%**	**1**	**1%**	**3%**	**3**	**2%**	**21%**	**4**	**3%**	**3%**
support from centre management	4	3%	4	3%	4%	0	0%	0%	2	2%	14%	2	2%	2%
support from funding bodies	0	0%	0	0%	0%	0	0%	0%	0	0%	0%	0	0%	0%
support from local commissioners	1	1%	0	0%	0%	1	1%	3%	0	0%	0%	1	1%	1%
support from local stakeholders	1	1%	1	1%	1%	0	0%	0%	0	0%	0%	1	1%	1%
support from research network	1	1%	1	1%	1%	0	0%	0%	1	1%	7%	0	0%	0%
WILLINGNESS	**9**	**7%**	**7**	**5%**	**7%**	**2**	**2%**	**6%**	**1**	**1%**	**7%**	**8**	**6%**	**7%**
willing to randomise	4	3%	3	2%	3%	1	1%	3%	0	0%	0%	4	3%	3%
willingness to perform the intervention	4	3%	3	2%	3%	1	1%	3%	0	0%	0%	4	3%	3%
willing to participate	3	2%	3	2%	3%	0	0%	0%	1	1%	7%	2	2%	2%

Appendix 11. Invitation to participate in the focus groups

Investigating the rationale for centre selection in randomised controlled trials coupled with economic evaluations

Dear Ms YYYY,

You are **invited** to participate in a **focus group** that will explore the rationale for centre selection in randomised controlled trials coupled with economic evaluations (RCT-EEs). The aim of the study is to elicit a comprehensive list of both methodological and pragmatic considerations that should be taken into account when considering the enrolment of a centre (e.g. hospital, GP practice, school, community as a whole etc.).

If you agree to **participate**, you will be invited to participate in one **focus group** session on the topic explained above. Prior to expressing the intention to attend you will have the opportunity to contact the researchers and ask any questions you may have on the conduct of the study. Several weeks after attending the focus group you will receive an **electronic questionnaire** on the rationale for centre selection in RCT-EEs, which you will be invited to provide feedback to. You will not be asked to fill in the questionnaire, but simply to **comment** on its format and content.

This research is carried out as part of a PhD studentship in health economics and is based in Primary Care Clinical Sciences at the University of Birmingham. The study is sponsored by the University of Birmingham.

Your participation is very important for the successful completion of this PhD studentship and will be very much appreciated.

To register your interest and ask any questions, please contact

Dr. Melanie Calvert (main investigator)
Primary Care Clinical Sciences
University of Birmingham
[redacted]

Adrian Gheorghe (studentship holder)
Primary Care Clinical Sciences
University of Birmingham
[redacted]

Thank you.

Appendix 12. Informed consent form for participation in the focus groups

Investigating the rationale for centre selection in randomised controlled trials coupled with economic evaluations

Informed Consent form Version 1.0 26th July 2011

		Participant: Please initial each section
1	I confirm that I have read and understood the Participant Information Sheet for this study (version 2.0 26th July 2011) and have had the opportunity to ask questions.	□
2	I understand that a copy of my Informed Consent form, information about me and my progress will be supplied in confidence to the study researchers at the University of Birmingham (Primary Care Clinical Sciences).	□
3	I understand that my participation in this study is voluntary and that I may withdraw participation at any time, without giving a reason, and without my legal rights being affected.	□
4	I understand that the information that I will provide during the focus group will be collected only with audio digital equipment, further transcribed and analysed by the study researchers.	□
5	I understand that, if I decide to withdraw participation from the study, the information I had provided in the focus group discussion prior to my withdrawal cannot be eliminated and will be analysed by the researchers.	□
6	I understand that due to the nature of the research, anonymity cannot be ensured, but the confidentiality of my data will be strictly protected by the researchers during every stage of the research, as detailed in the Participant Information Sheet (version 2.0 26th July 2011).	□
7	I understand that after I have taken part in the focus group I will be sent a questionnaire and asked to provide feedback on its content and structure. I give permission for this questionnaire to be sent to me, and I understand that I do not have to send comments back if I do not want to. I understand that researchers may contact me by email or phone to remind me to send my comments.	□
8	I agree to protect the confidentiality of data collected from the other participants in the focus group.	□
9	I agree to take part in this study.	□

______________________________ ______________________________
Name of Participant (please print) Name of Researcher (please print)

______________________________ ______________________________
Signature of Participant Signature of Researcher

Date today __ __/__ __/__ __ __ __ Date today __ __/__ __/__ __ __ __

Appendix 13. Focus group topic guide

Investigating the rationale for centre selection in randomised controlled trials coupled with economic evaluations

Focus group topic guide

Scenario [as it will be presented to participants at the beginning of the focus group]

The discussion will consider a hypothetical multi-centre parallel randomised controlled trial (RCT) coupled with an economic evaluation. Resource use data and outcome data (e.g. clinical markers, health-related quality of life information) are being collected alongside the RCT. A centre can range from a GP surgery to an entire community, depending on the nature of the investigation.

Topic guide

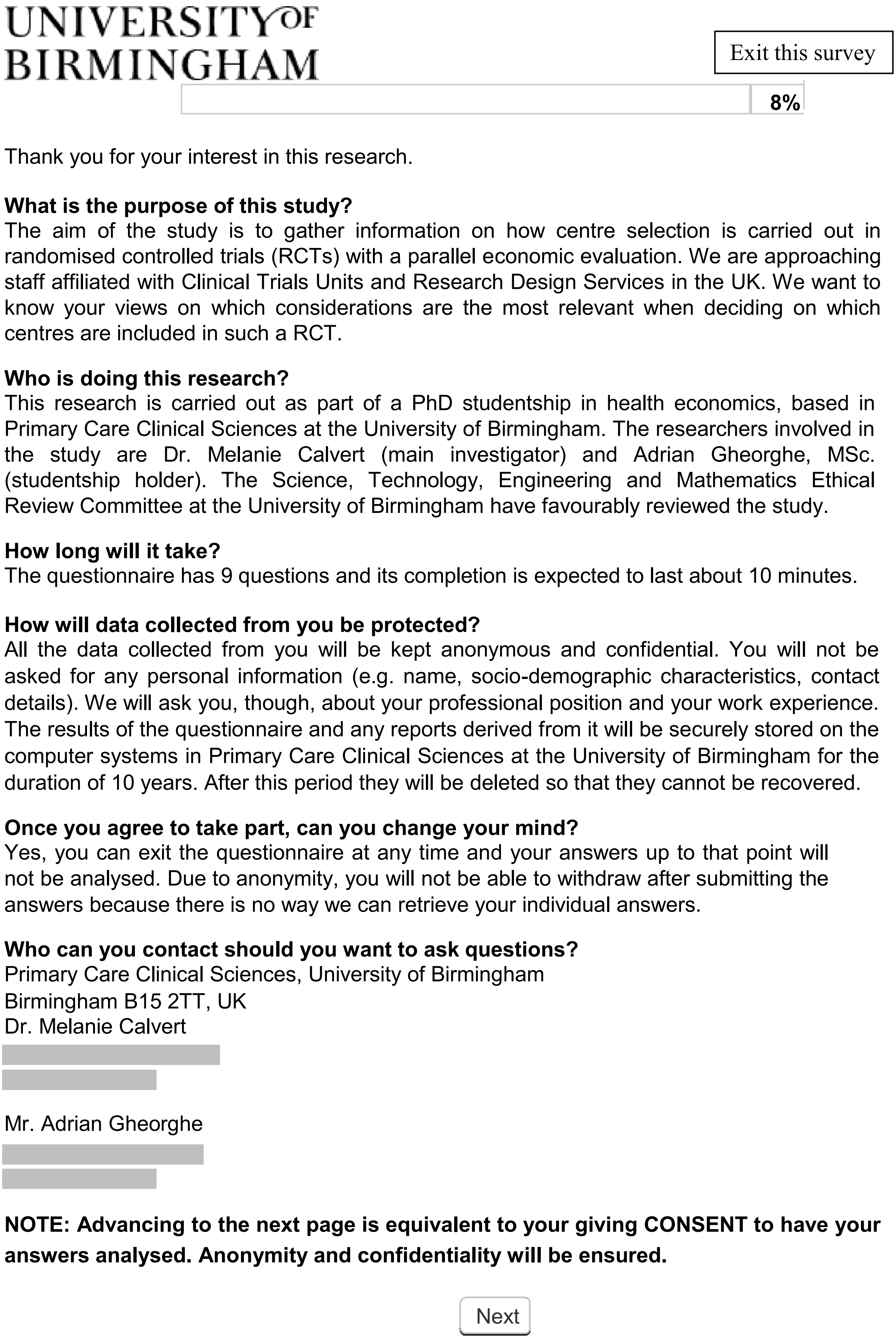
1. What sorts of things might you consider when selecting centres for this kind of trial?

    a.Which things are the most important?

    i)At what stage in the trial should these things be considered?

2. Who should be involved in making the decision about centre selection?

Appendix 14. Full version of the online survey



Thank you for your interest in this research.

What is the purpose of this study?
The aim of the study is to gather information on how centre selection is carried out in randomised controlled trials (RCTs) with a parallel economic evaluation. We are approaching staff affiliated with Clinical Trials Units and Research Design Services in the UK. We want to know your views on which considerations are the most relevant when deciding on which centres are included in such a RCT.

Who is doing this research?
This research is carried out as part of a PhD studentship in health economics, based in Primary Care Clinical Sciences at the University of Birmingham. The researchers involved in the study are Dr. Melanie Calvert (main investigator) and Adrian Gheorghe, MSc. (studentship holder). The Science, Technology, Engineering and Mathematics Ethical Review Committee at the University of Birmingham have favourably reviewed the study.

How long will it take?
The questionnaire has 9 questions and its completion is expected to last about 10 minutes.

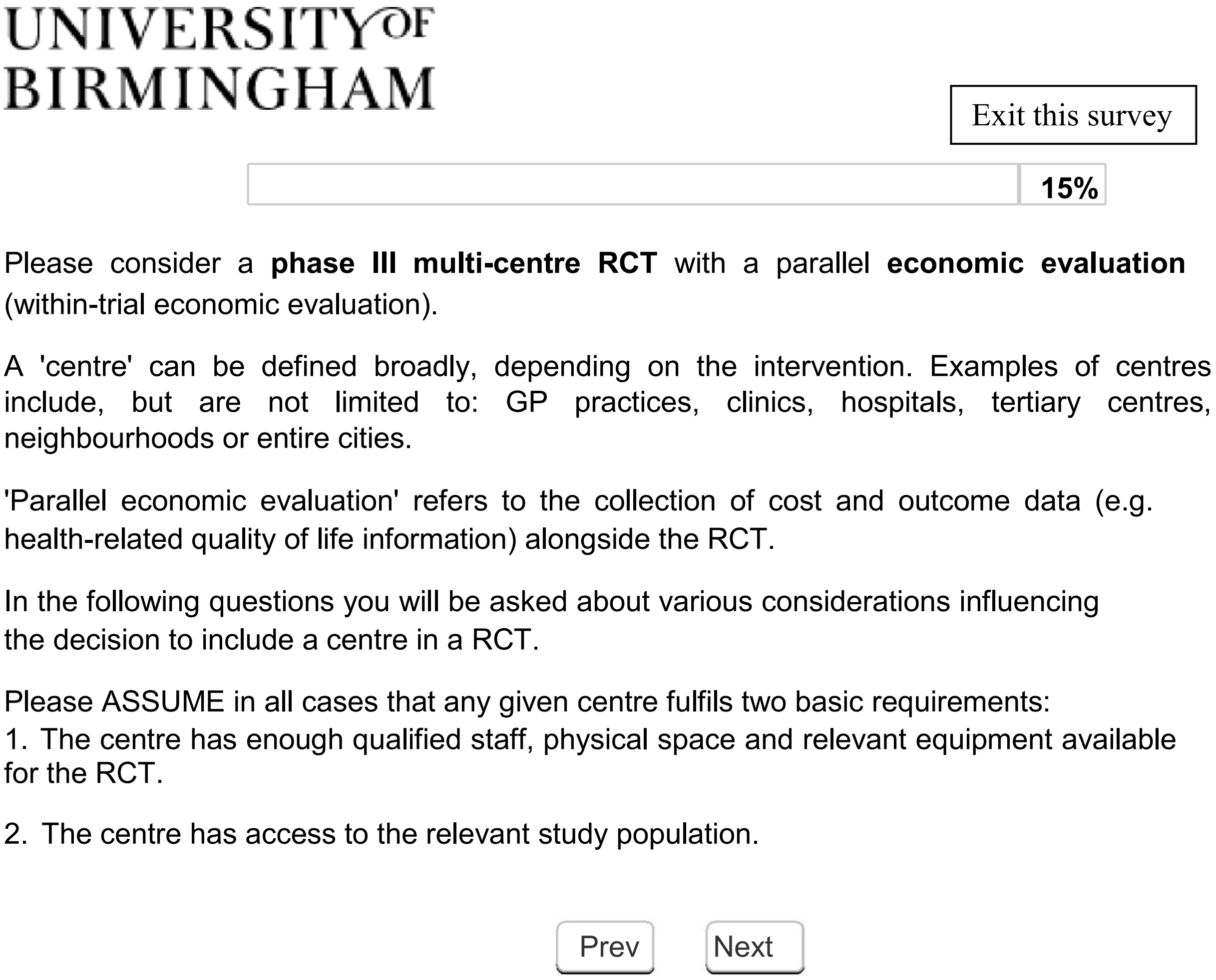
How will data collected from you be protected?
All the data collected from you will be kept anonymous and confidential. You will not be asked for any personal information (e.g. name, socio-demographic characteristics, contact details). We will ask you, though, about your professional position and your work experience. The results of the questionnaire and any reports derived from it will be securely stored on the computer systems in Primary Care Clinical Sciences at the University of Birmingham for the duration of 10 years. After this period they will be deleted so that they cannot be recovered.

Once you agree to take part, can you change your mind?
Yes, you can exit the questionnaire at any time and your answers up to that point will not be analysed. Due to anonymity, you will not be able to withdraw after submitting the answers because there is no way we can retrieve your individual answers.

Who can you contact should you want to ask questions?
Primary Care Clinical Sciences, University of Birmingham
Birmingham B15 2TT, UK
Dr. Melanie Calvert
[redacted]

Mr. Adrian Gheorghe
[redacted]

NOTE: Advancing to the next page is equivalent to your giving CONSENT to have your answers analysed. Anonymity and confidentiality will be ensured.



Please consider a **phase III multi-centre RCT** with a parallel **economic evaluation** (within-trial economic evaluation).

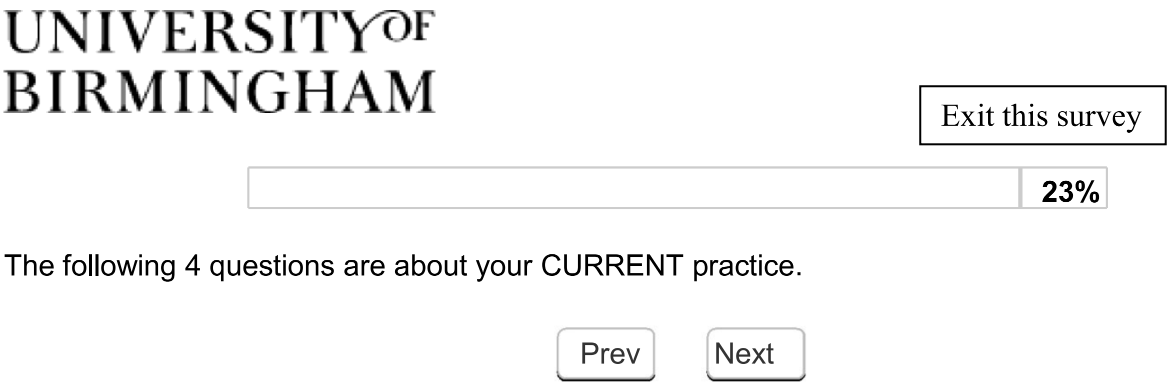
A 'centre' can be defined broadly, depending on the intervention. Examples of centres include, but are not limited to: GP practices, clinics, hospitals, tertiary centres, neighbourhoods or entire cities.

'Parallel economic evaluation' refers to the collection of cost and outcome data (e.g. health-related quality of life information) alongside the RCT.

In the following questions you will be asked about various considerations influencing the decision to include a centre in a RCT.

Please ASSUME in all cases that any given centre fulfils two basic requirements:
1. The centre has enough qualified staff, physical space and relevant equipment available for the RCT.

2. The centre has access to the relevant study population.



The following 4 questions are about your CURRENT practice.

UNIVERSITY OF
BIRMINGHAM

Exit this survey

31%

1. YOUR CURRENT PRACTICE: When considering the inclusion of a centre in a RCT with a parallel economic evaluation, which centre characteristics do you usually look for?

Please choose the most important characteristics from the list below. No explicit ranking is required. Please select a MINIMUM of 3 and a MAXIMUM of 5 answers.

Local clinical staff understand the methodological underpinnings of RCTs.

The geographical location of the centre is convenient for logistical reasons.

The centre belongs to a relevant research network.

Local staff have had experience with conducting RCTs in the past.

The centre's computer systems are compatible with the trial centre's computer systems.

The centre retains or contributes to generalisability in terms of clinical practice.

The centre has support from local commissioners to participate in the RCT.

The centre retains or contributes to generalisability in terms of population characteristics. There is a good communication relationship between the trials unit and centre staff.

The centre is able to obtain necessary approvals (including R&D) in a timely manner. The centre is able to recruit the desired number of patients in a timely manner.

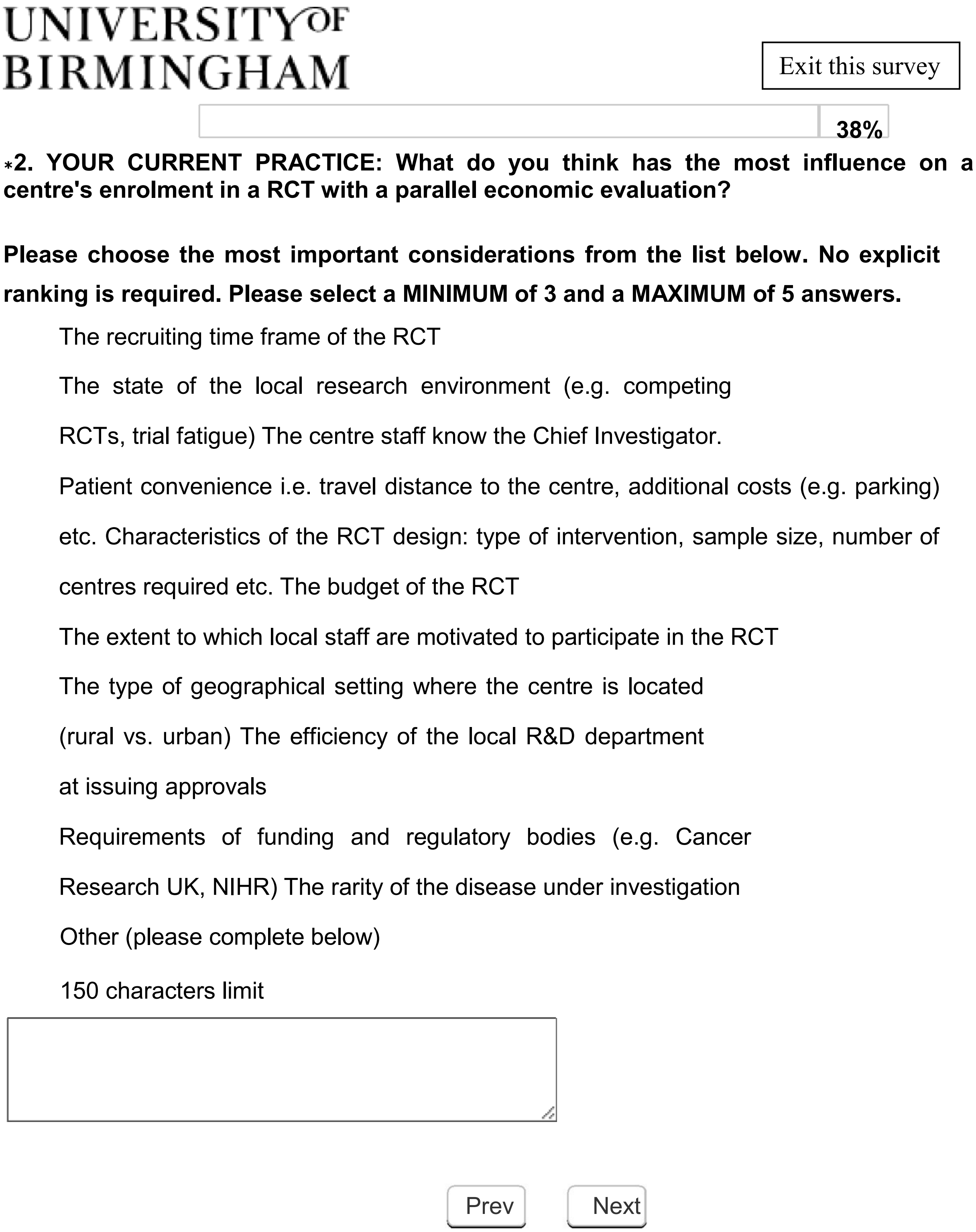
The centre staff show a degree of interest in the RCT.

The centre retains or contributes to generalisability in terms of economic evaluation results.

Other (please complete below).

150 characters limit

Prev Next



*2. YOUR CURRENT PRACTICE: What do you think has the most influence on a centre's enrolment in a RCT with a parallel economic evaluation?**

Please choose the most important considerations from the list below. No explicit ranking is required. Please select a MINIMUM of 3 and a MAXIMUM of 5 answers.

The recruiting time frame of the RCT

The state of the local research environment (e.g. competing

RCTs, trial fatigue) The centre staff know the Chief Investigator.

Patient convenience i.e. travel distance to the centre, additional costs (e.g. parking)

etc. Characteristics of the RCT design: type of intervention, sample size, number of

centres required etc. The budget of the RCT

The extent to which local staff are motivated to participate in the RCT

The type of geographical setting where the centre is located

(rural vs. urban) The efficiency of the local R&D department

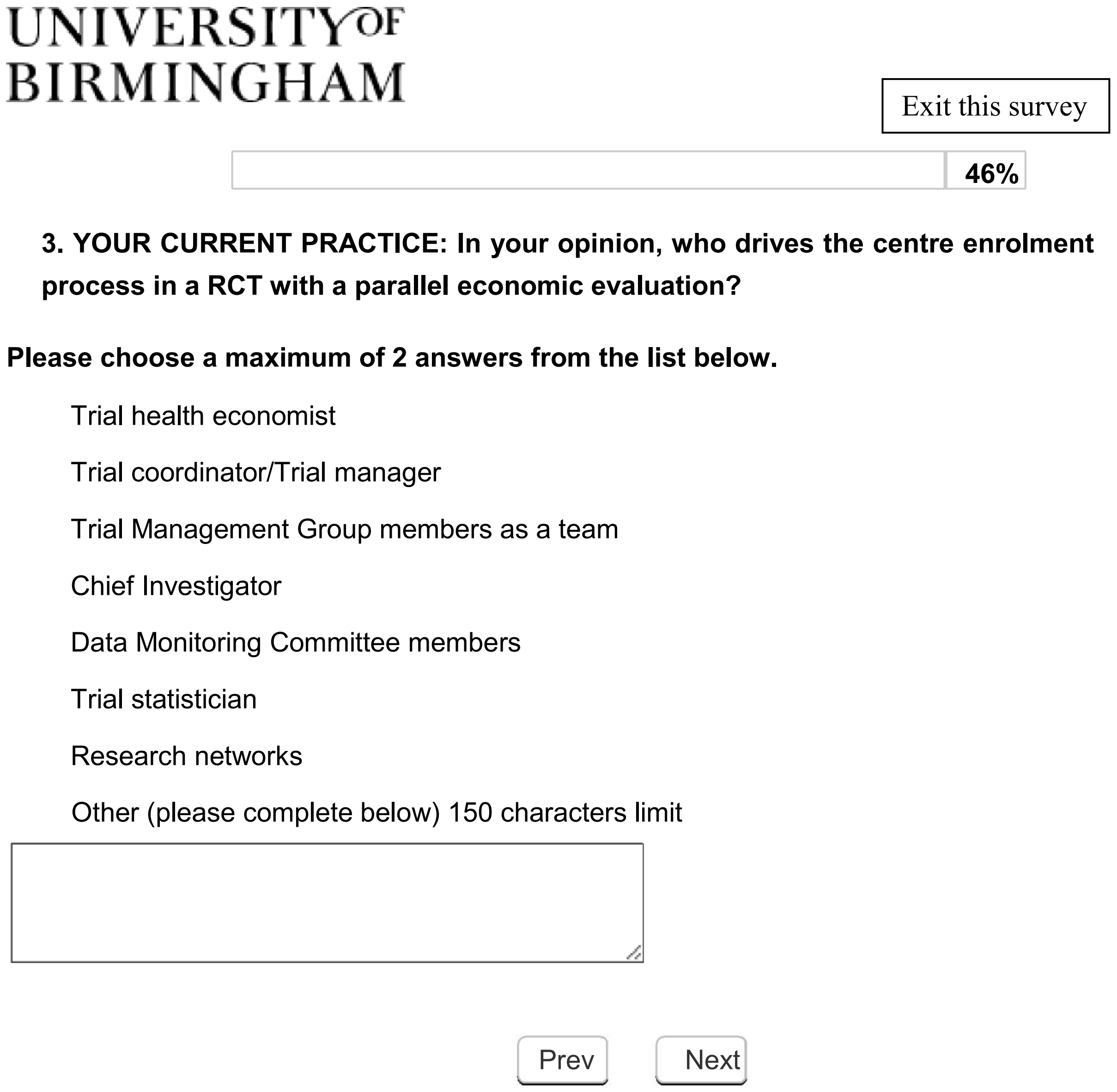
at issuing approvals

Requirements of funding and regulatory bodies (e.g. Cancer

Research UK, NIHR) The rarity of the disease under investigation

Other (please complete below)

150 characters limit



3. YOUR CURRENT PRACTICE: In your opinion, who drives the centre enrolment process in a RCT with a parallel economic evaluation?

Please choose a maximum of 2 answers from the list below.

Trial health economist

Trial coordinator/Trial manager

Trial Management Group members as a team

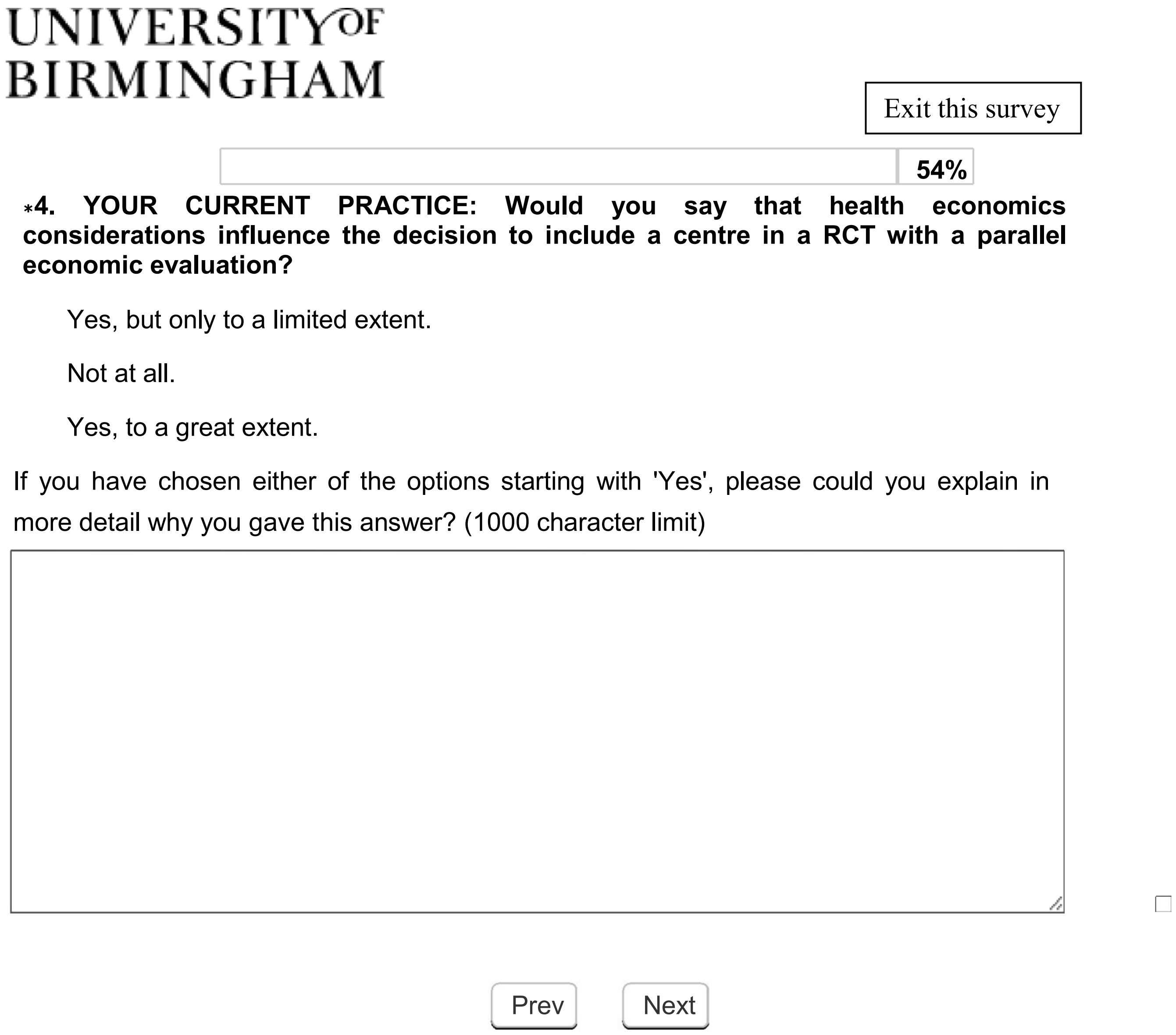
Chief Investigator

Data Monitoring Committee members

Trial statistician

Research networks

Other (please complete below) 150 characters limit



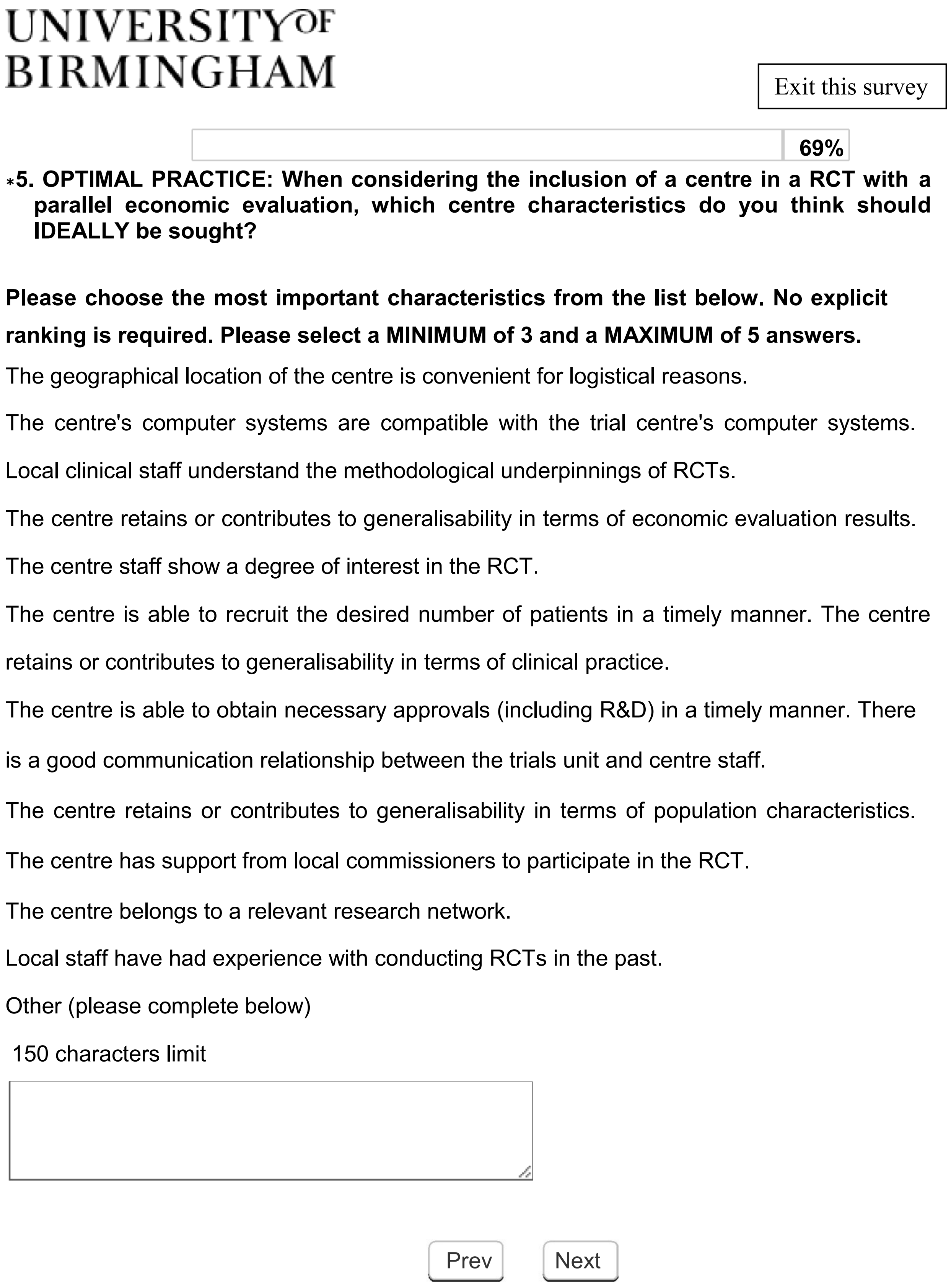
***4. YOUR CURRENT PRACTICE: Would you say that health economics considerations influence the decision to include a centre in a RCT with a parallel economic evaluation?**

Yes, but only to a limited extent.

Not at all.

Yes, to a great extent.

If you have chosen either of the options starting with 'Yes', please could you explain in more detail why you gave this answer? (1000 character limit)



*5. OPTIMAL PRACTICE: When considering the inclusion of a centre in a RCT with a parallel economic evaluation, which centre characteristics do you think should IDEALLY be sought?

Please choose the most important characteristics from the list below. No explicit ranking is required. Please select a MINIMUM of 3 and a MAXIMUM of 5 answers.

The geographical location of the centre is convenient for logistical reasons.

The centre's computer systems are compatible with the trial centre's computer systems.

Local clinical staff understand the methodological underpinnings of RCTs.

The centre retains or contributes to generalisability in terms of economic evaluation results.

The centre staff show a degree of interest in the RCT.

The centre is able to recruit the desired number of patients in a timely manner. The centre retains or contributes to generalisability in terms of clinical practice.

The centre is able to obtain necessary approvals (including R&D) in a timely manner. There is a good communication relationship between the trials unit and centre staff.

The centre retains or contributes to generalisability in terms of population characteristics.

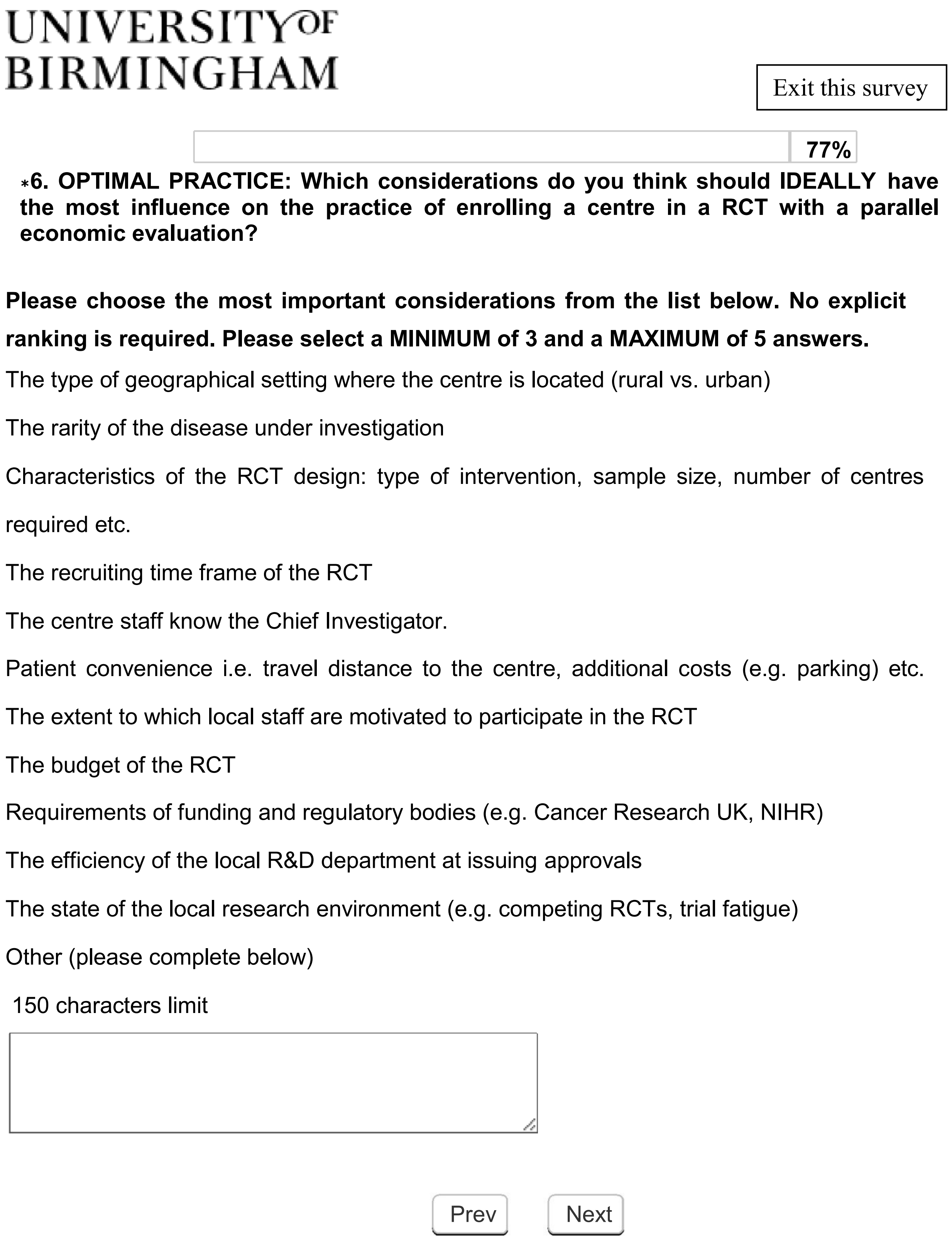
The centre has support from local commissioners to participate in the RCT.

The centre belongs to a relevant research network.

Local staff have had experience with conducting RCTs in the past.

Other (please complete below)

150 characters limit



*6. OPTIMAL PRACTICE: Which considerations do you think should IDEALLY have the most influence on the practice of enrolling a centre in a RCT with a parallel economic evaluation?**

Please choose the most important considerations from the list below. No explicit ranking is required. Please select a MINIMUM of 3 and a MAXIMUM of 5 answers.

The type of geographical setting where the centre is located (rural vs. urban)

The rarity of the disease under investigation

Characteristics of the RCT design: type of intervention, sample size, number of centres required etc.

The recruiting time frame of the RCT

The centre staff know the Chief Investigator.

Patient convenience i.e. travel distance to the centre, additional costs (e.g. parking) etc.

The extent to which local staff are motivated to participate in the RCT

The budget of the RCT

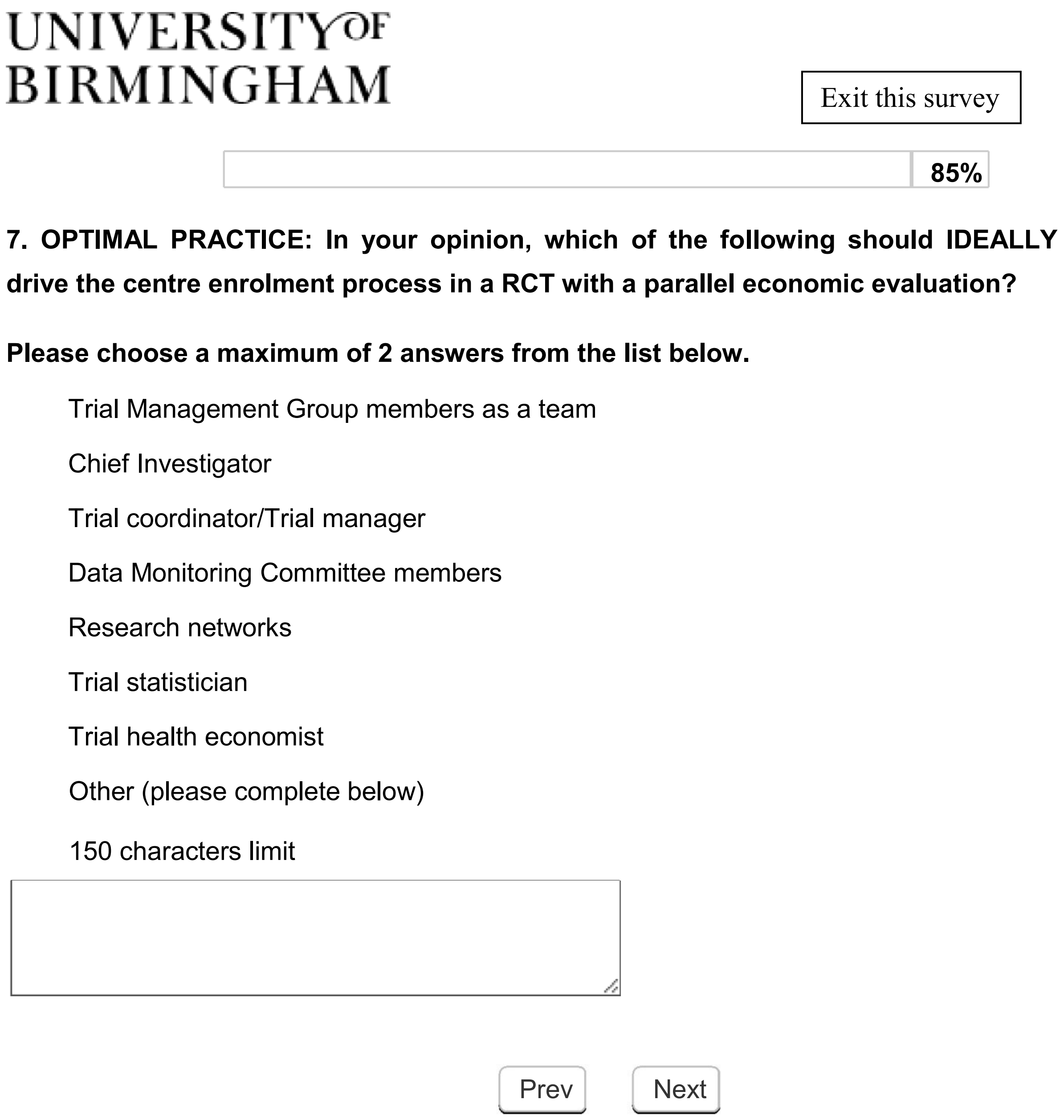
Requirements of funding and regulatory bodies (e.g. Cancer Research UK, NIHR)

The efficiency of the local R&D department at issuing approvals

The state of the local research environment (e.g. competing RCTs, trial fatigue)

Other (please complete below)

150 characters limit



7. OPTIMAL PRACTICE: In your opinion, which of the following should IDEALLY drive the centre enrolment process in a RCT with a parallel economic evaluation?

Please choose a maximum of 2 answers from the list below.

Trial Management Group members as a team

Chief Investigator

Trial coordinator/Trial manager

Data Monitoring Committee members

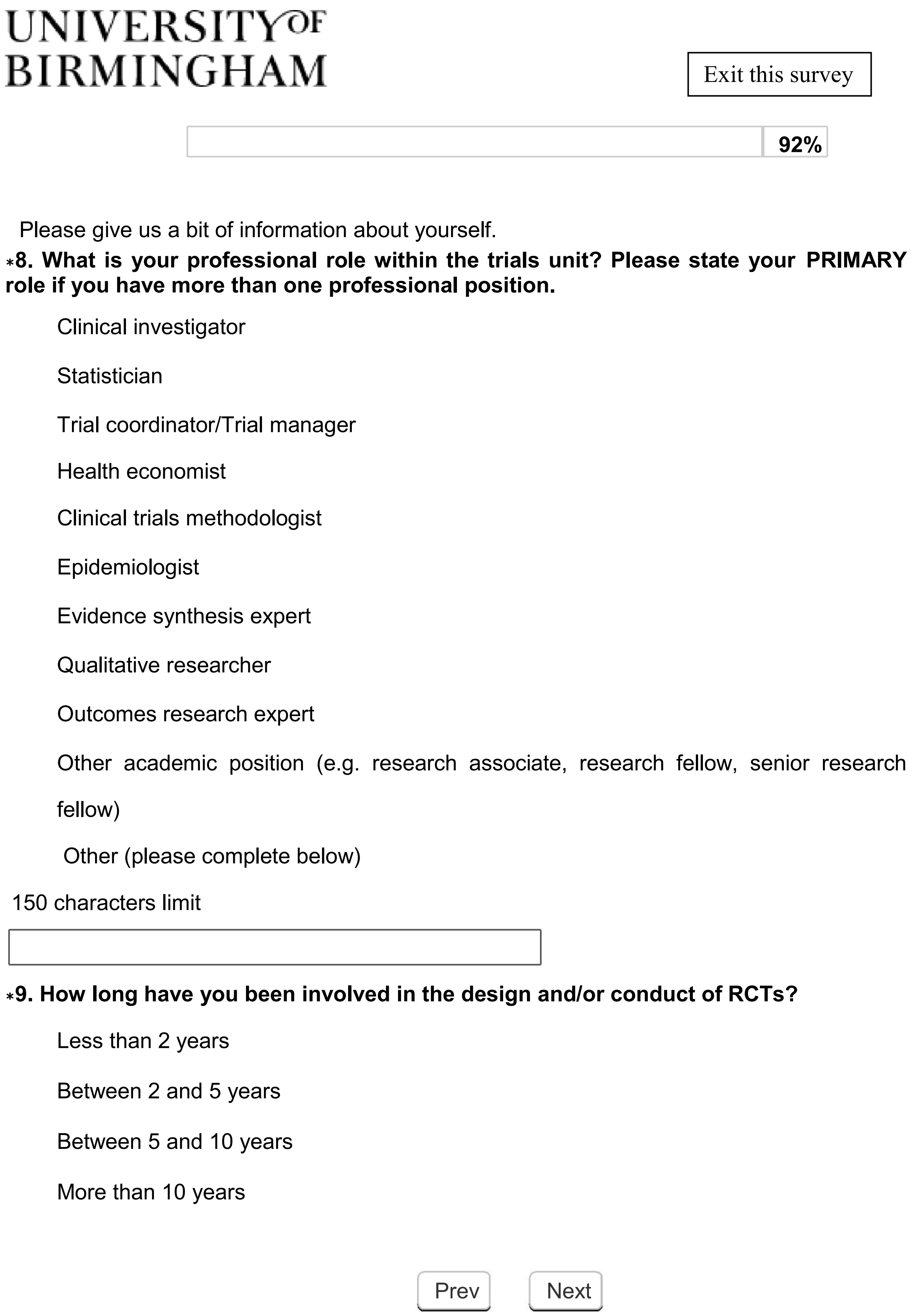
Research networks

Trial statistician

Trial health economist

Other (please complete below)

150 characters limit



Please give us a bit of information about yourself.

***8. What is your professional role within the trials unit? Please state your PRIMARY role if you have more than one professional position.**

Clinical investigator

Statistician

Trial coordinator/Trial manager

Health economist

Clinical trials methodologist

Epidemiologist

Evidence synthesis expert

Qualitative researcher

Outcomes research expert

Other academic position (e.g. research associate, research fellow, senior research fellow)

Other (please complete below)

150 characters limit

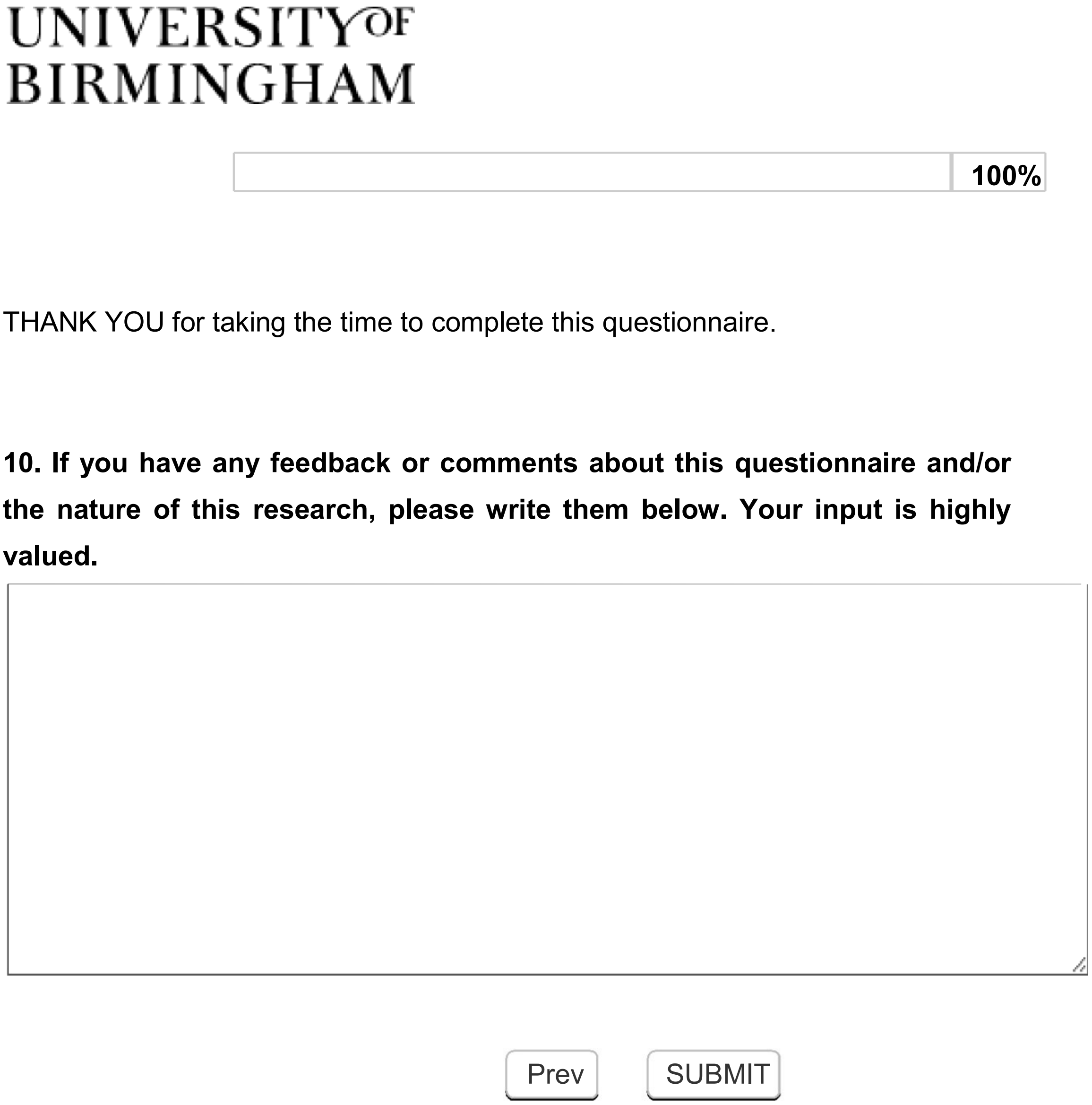
***9. How long have you been involved in the design and/or conduct of RCTs?**

Less than 2 years

Between 2 and 5 years

Between 5 and 10 years

More than 10 years



THANK YOU for taking the time to complete this questionnaire.

10. If you have any feedback or comments about this questionnaire and/or the nature of this research, please write them below. Your input is highly valued.

Appendix 15. Invitation to participate in the online survey

Investigating the rationale for centre selection in randomised controlled trials

with parallel economic evaluations

Dear Prof. YYYY,

You are invited to participate in a study which aims to gather information on how centre selection is carried out in randomised controlled trials with a parallel economic evaluation (RCT-EEs). For this purpose we have devised an electronic questionnaire which is being circulated to all 48 UK Clinical Research Collaborative (UKCRC) Clinical Trials Units and 10 Research Design Services (RDS) in the UK. The questionnaire has 9 multiple-choice questions and its completion should last less than 10 minutes.

If you agree to participate, you are invited to **complete** the online questionnaire (link below) and also to **circulate** it within Barts and the London Pragmatic Clinical Trials Unit for completion by staff involved in the design and conduct of RCT-EEs. We are interested in the views of the following professionals: clinical investigators, trial coordinators/trial managers, statisticians, health economists and any other academic position (e.g. research associate, research fellow).

https://surveymonkey.com/s/J9R7FKK

This research is carried out as part of a PhD studentship in health economics and is based in Primary Care Clinical Sciences at the University of Birmingham. The study is sponsored by the University of Birmingham.

Your participation is very important for the successful completion of this PhD studentship and will be very much appreciated.

For any questions you may have, please contact

Dr. Melanie Calvert (main investigator)
Primary Care Clinical Sciences
University of Birmingham
[redacted]

Adrian Gheorghe (studentship holder)
Primary Care Clinical Sciences
University of Birmingham
[redacted]

Thank you.

Appendix 16. Survey: Free text responses in the 'Other' field, by question

Q1. YOUR CURRENT PRACTICE: When considering the inclusion of a centre in a RCT with a parallel economic evaluation, which centre characteristics do you usually look for?

"Track record in delivering recruitment and high quality data"

"Local knowledge, experience and expertise in the disease area/intervention under investigation"

"These choices are often made by PI/TM rather than stats teams members but these would be my choices"

"The only important thing for me is that the staff are interested and engaged"

Q2. YOUR CURRENT PRACTICE: What do you think has the most influence on a centre's enrolment in a RCT with a parallel economic evaluation?

"It is difficult to say what the centre perspective is, but critical issues are whether the research question is clinically relevant and how practical (i.e. easy) the enrolment pathway is (i.e. are patients easily identified)"

"I presume you mean from centre's perspective?"

"A PI who is keen and actively encourages staff to recruit to the RCT."

"Feeling the research question is of importance to them and the populations they serve"

"If they think the trial is addressing a really important clinical question that they can relate to and want the answer to. The support they receive to participate in the trial."

"Relationship with study CI"

"I can't say that I see much relevance in most of the answers suggested."

"Promotion / raising the profile of the service they provide"

Q3. YOUR CURRENT PRACTICE: In your opinion, who drives the centre enrolment process in a RCT with a parallel economic evaluation?

"In the CTU environment we have a senior trialist with responsibility for operational delivery. This is a key oversight role and really pushed the Trial Manager and CI. Also sometime we have a clinical co-ordinator - they will take an active role in centre identification. In cluster trials, the lead stats methodologist also has oversight, as centre characteristics are key."

"Sponsors, in particular commercial sponsors"

"LOCAL INVESTIGATOR OR NURSE"

> *"I don't think this is any different for a trial without parallel economic evaluation"*

> *"Principal investigators, availability of research nurses either from the local R&D or one of the research networks or the CLRN."*

Q4. YOUR CURRENT PRACTICE: Would you say that health economics considerations influence the decision to include a centre in a RCT with a parallel economic evaluation?

> *"Efforts are usual made to recruit centres that serve different socio-economic backgrounds."*

> *"In my experience, the health economic component is only considered in cluster trials"*

> *"We would consider generalisability and that includes economics, but not as a separate criteria."*

> *"I have no idea - it is not in my experience. Maybe I have been included as a survey participant in error?"*

> *"It's surely obvious that if including a centre would vitiate a health economic component of a study in which that component is quite essential, then the centre would not be used. As long as including an otherwise good centre wouldn't damage the study, I guess the economic side would play a pretty minor role, since the scientific integrity of the study would have to come first."*

Q5. OPTIMAL PRACTICE: When considering the inclusion of a centre in a RCT with a parallel economic evaluation, which centre characteristics do you think should IDEALLY be sought?

> *"I have problems with this question. All of these should be ideally sought. My perfect centre would meet all of these (although the IT system is irrelevant). The next question asks for the prioritisation of these which is more useful."*

> *"Able to collect additional resource use data"*

> *"As before - track record in delivery patient recruitment and high quality data"*

> *"Local staff have an interest, experience and expertise in the disease area or intervention of interest, adequate facilities in place and good working relationships between research staff and local service providers e.g. labs, haematology, R&D etc."*

> *"I think health economics should NOT be done in parallel. This is a hopeless questionnaire"*

Q6. OPTIMAL PRACTICE: Which considerations do you think should IDEALLY have the most influence on the practice of enrolling a centre in a RCT with a parallel economic evaluation?

> *"Again I have a problem here as you have dropped items in this question which were previously important - all the ones about generalisability."*

“As before - question and whether clinically important is critical and ease of recruitment and follow up (does the research and clinical pathways facilitate the research process and/or where there is a mismatch, are there adequate resources)”

“Importance of being involved in research and the benefits it could potentially have for patients”

“This is a terrible questionnaire!”

“Having a motivated PI”

“Relevance to them and their population”

“The importance of the clinical question”

“The potential benefit to the NHS and its patients.”

“The answer "know the CI" is nearly right, but gives an unfortunate "chummy" impression. It's essential for there to be a good working relationship between the principal scientific staff.”

Q7. OPTIMAL PRACTICE: In your opinion, which of the following should IDEALLY drive the centre enrolment process in a RCT with a parallel economic evaluation?

“Local lead investigators”

“Existing expertise across the UK in the disease area/intervention of interest”

“LOCAL INVESITGATOR OR NURSE”

“Not sure about this one, it all depends on what perspective you take. I am a CI and PI, ideally I would like a generalisable (nationally and locally) set of centres for my trials, but you rarely have all the information about centres that you would want to make selection of centres as informed by data as one would want.”

“See answer to previous question.”

“Principal investigator”

Q10. Final comments

“As indicated in my "other response" boxes I have problems with some of these questions as the options available changed. What does "ideal" mean? In an "ideal" world we would so much trouble doing R&D approvals, all staff would have time to recruit patients - I got the feeling that you are probing how the challenges of recruitment, logistics and bureaucracy stop us doing the most scientifically valid trials, where issues of generalisability might lead us to choose centres which would be more representative, rather than those where we can get through the approvals process and recruit lots of

patients quickly. If that was you aim, I think that the questions could be better put, by defining "optimally" more clearly."

"Many of Qs difficult to understand so my answers will reflect that. sorry"

"Not easy to choose between the options given, and if I answered it again I am not sure I would give exactly the same answers!"

"Good luck! I would be interested in the findings of this work."

"Good luck with your study"

"The questions are tricky to be 'black and white' in responses. In the ideal world, we would carry out national audits / surveys before doing a RCT, and we would then benchmark local / national potential centres for our trial against the national picture, in order to be fully informed about generalisability of those centres - in terms of current clinical practice, skill mix of teams, population being served and current health economic information. In reality we rarely have that type of full information, and we hope that randomisation sorts out at least some of these problems and the trial is then focusing only on the between group comparisons. Clearly larger trials with many centres have a better chance of being nationally generalisable."

"My initial role in trials was as a trial health economist, but I am now more of a trial methodologist. Despite my health economic background I don't take into account very much whether or not a centre enhances economic generalisability. This is mainly because it is so difficult to recruit that the overwhelming objective is to get the numbers into the study. Furthermore, those centres that recruit are not so different in their general characteristics from those that do not. Consequently, it seems to me that generalisability is high whether or not one seeks a generalisable sample or not."

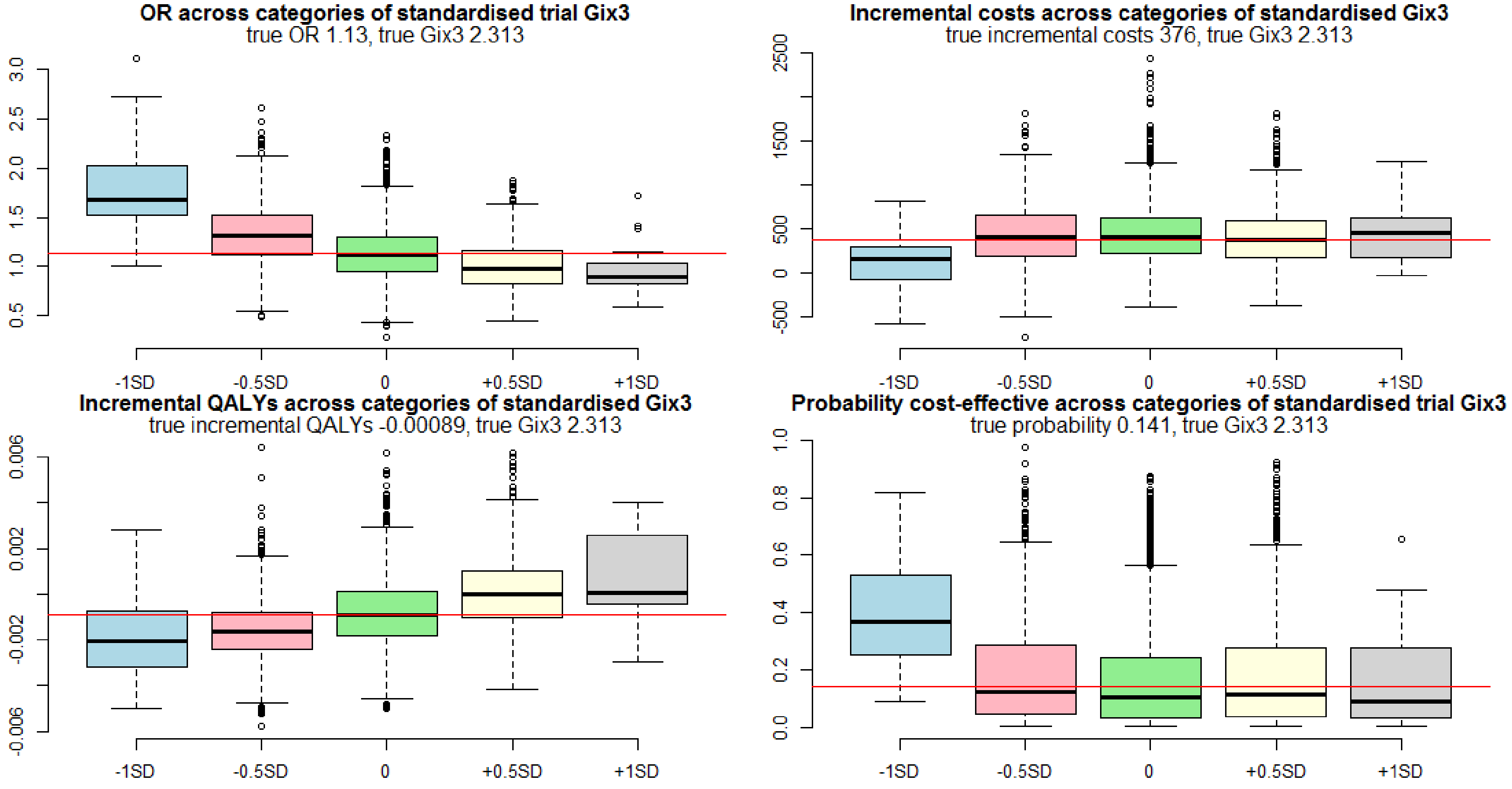
"Interesting survey"

"The questions themselves are not so clear. There are two perspectives that they could be approached 1: study design e.g. ideally all those centres selected should participate, 2: the practicalities of conducting a study, which may contradict the requirements for good study design."

"This is a very well-organised way of conducting a survey. It makes it easy for people to participate. Although I have limited experience of 'running' an RCT, I have many years working with those who do. I didn't really have the opportunity to show that clearly in the body of the survey."

Appendix 17. Generalisability index sensitivity analyses

Figure A17.1 Clinical and cost-effectiveness estimates in simulated RCTs across categories of standardised trial-Gix3



OR across categories of standardised trial Gix3
true OR 1.13, true Gix3 2.313

Incremental costs across categories of standardised Gix3
true incremental costs 376, true Gix3 2.313

Incremental QALYs across categories of standardised Gix3
true incremental QALYs -0.00089, true Gix3 2.313

Probability cost-effective across categories of standardised trial Gix3
true probability 0.141, true Gix3 2.313

Figure A17.2 Clinical and cost-effectiveness estimates in simulated RCTs across categories of standardised trial-Gix4a

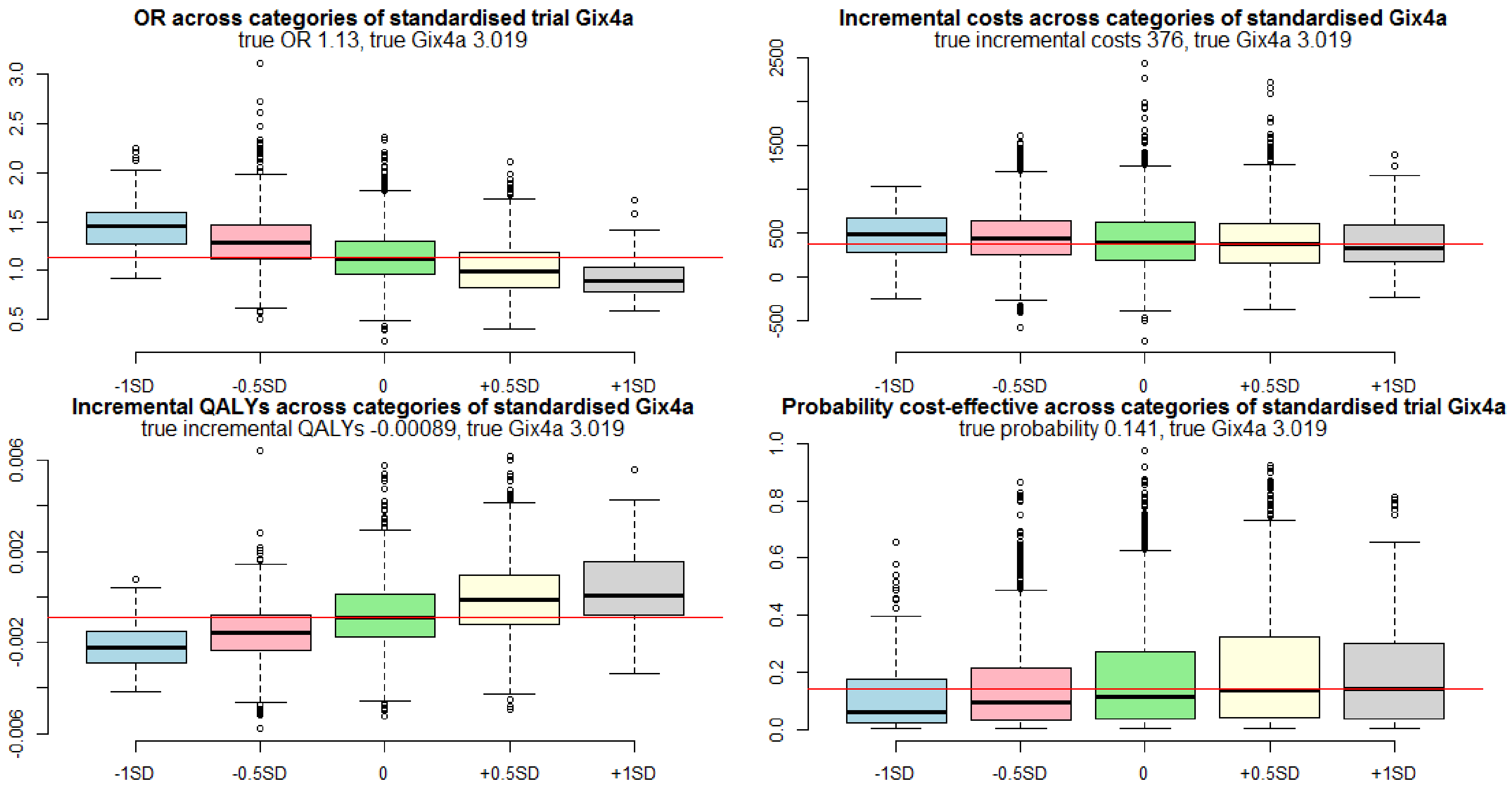


Figure A17.3 Clinical and cost-effectiveness estimates in simulated RCTs across categories of standardised trial-Gix4b

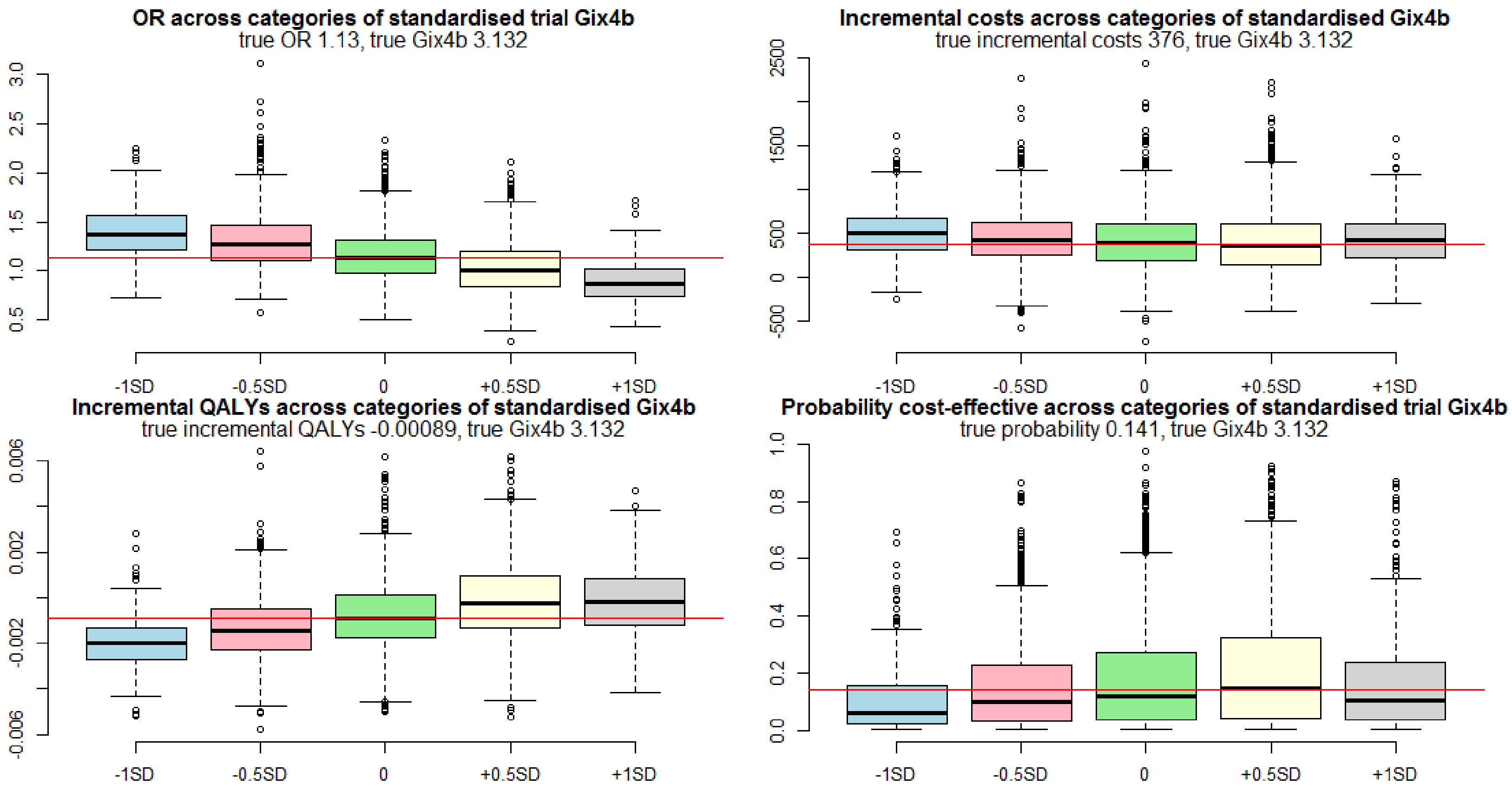


Figure A17.4 Clinical and cost-effectiveness estimates in simulated RCTs across categories of standardised trial-Gix4z

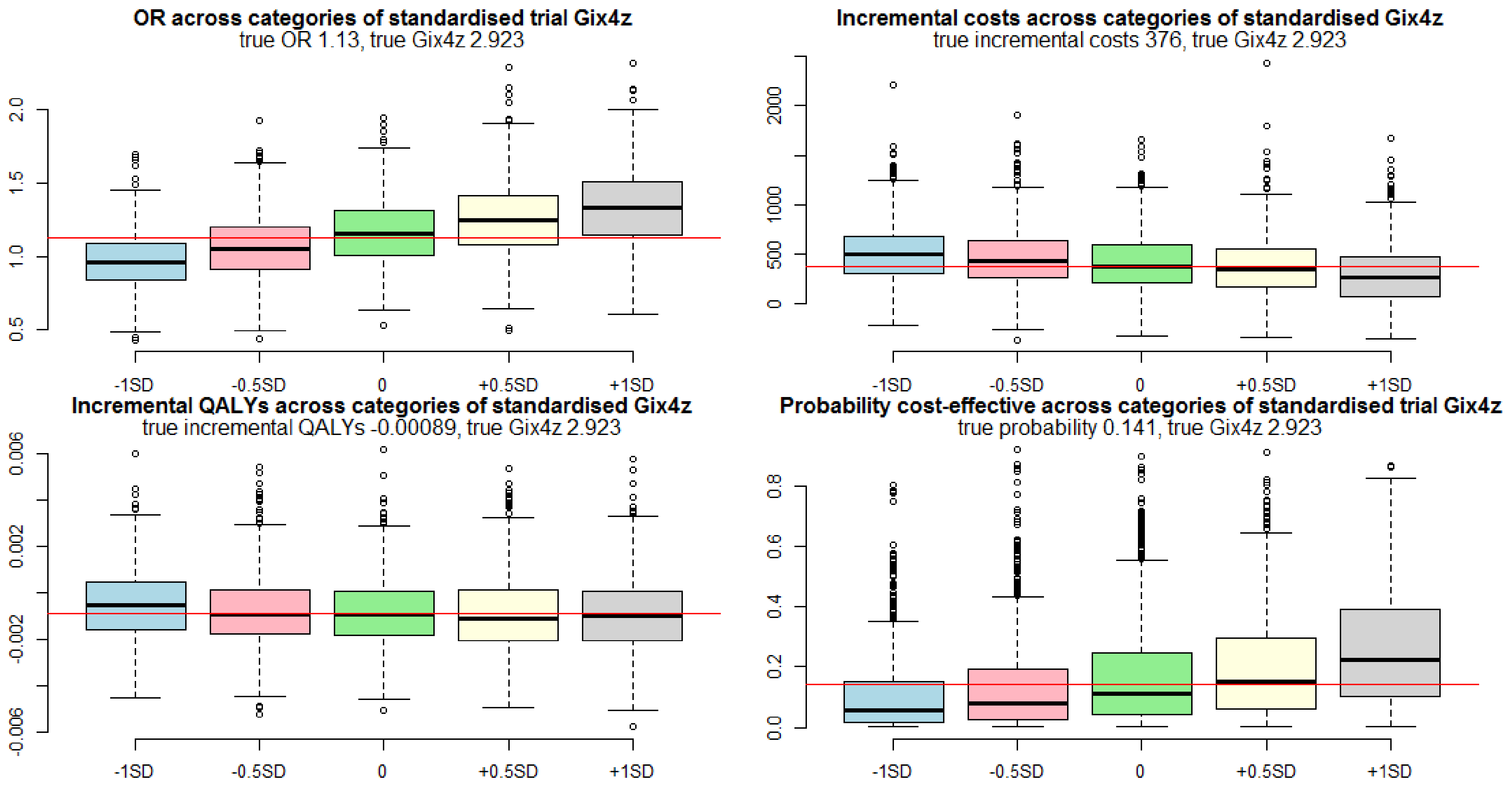


Figure A17.5 Clinical and cost-effectiveness estimates in simulated RCTs across categories of standardised trial-Gix90

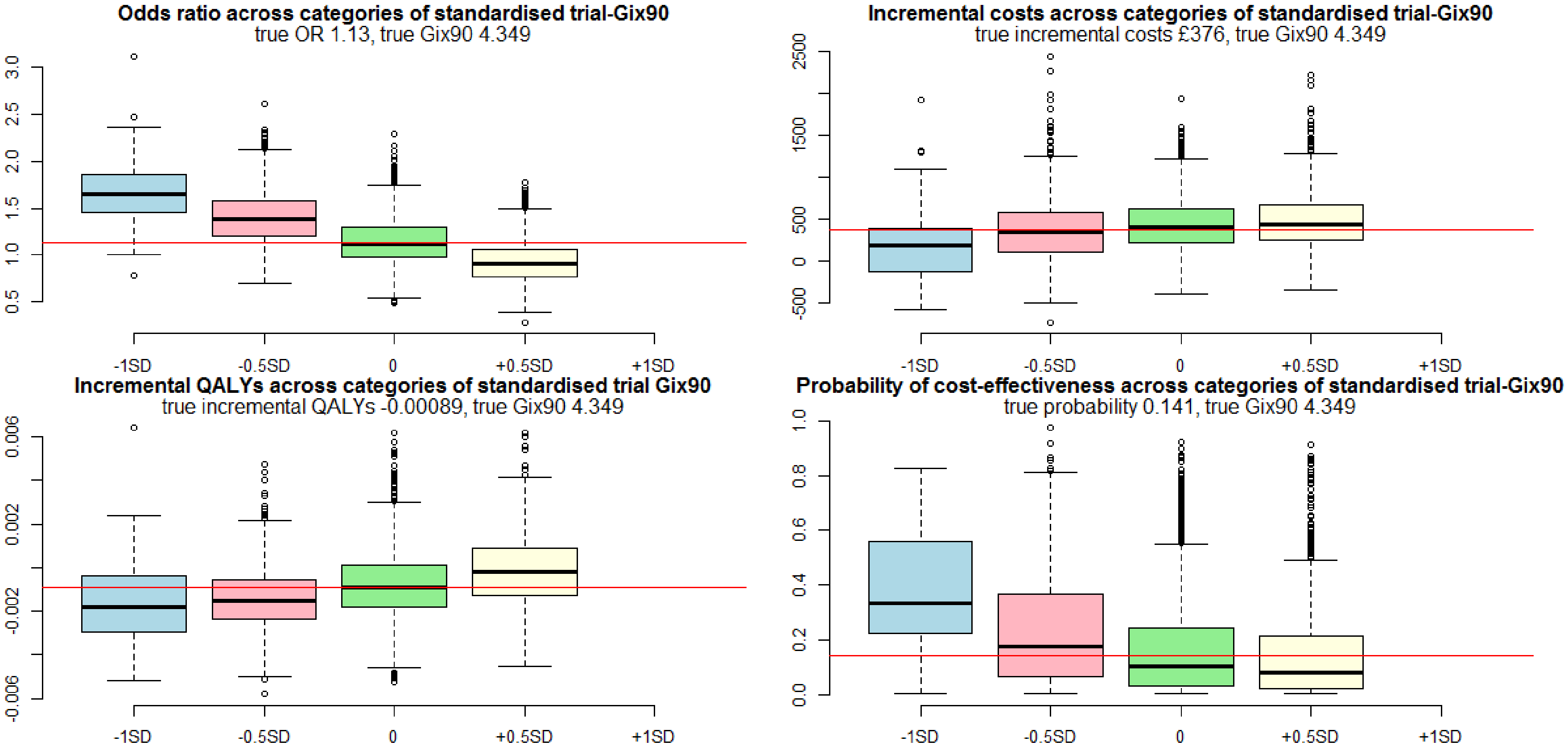
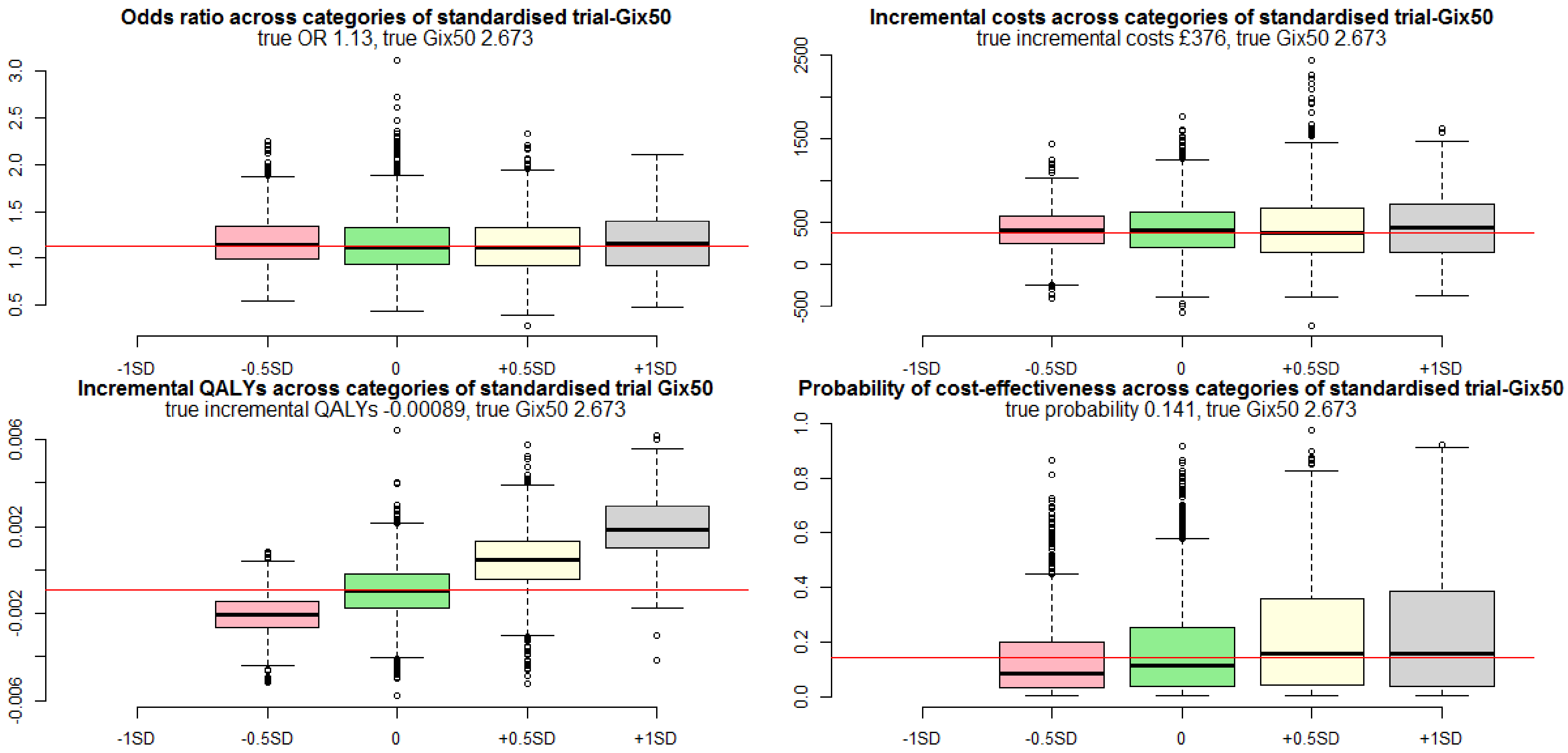


Figure A17.6 Clinical and cost-effectiveness estimates in simulated RCTs across categories of standardised trial-Gix50



Odds ratio across categories of standardised trial-Gix50
true OR 1.13, true Gix50 2.673

Incremental costs across categories of standardised trial-Gix50
true incremental costs £376, true Gix50 2.673

Incremental QALYs across categories of standardised trial Gix50
true incremental QALYs -0.00089, true Gix50 2.673

Probability of cost-effectiveness across categories of standardised trial-Gix50
true probability 0.141, true Gix50 2.673