EXTERNAL QUALITY ASSESSMENT IN CLINICAL CHEMISTRY - AN EXAMINATION OF REQUIREMENTS, APPLICATIONS AND BENEFITS

by

DAVID GRAHAME BULLOCK

A thesis submitted to the
Faculty of Science
of the
University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

Department of Clinical Chemistry
University of Birmingham
Wolfson Research Laboratories
Queen Elizabeth Medical Centre
Edgbaston
Birmingham B15 2TH
England

September 1987

SYNOPSIS

This thesis describes studies of the requirements, applications and benefits of external quality assessment (EQA) of clinical chemistry laboratories.

The involvement of EQA in assessing the analytical quality of results from clinical chemistry laboratories is well-established. This thesis addresses the contribution of EQA in improving clinical chemistry practice and therefore patient care rather than as a method of 'policing' laboratory performance which is the objective of many national EQA schemes. The use of EQA in the assessment of interlaboratory agreement, of analytical methods, of individual laboratory performance, and of quality control and calibration materials is discussed.

Surveys are examined as a means to assess the prevailing standard of performance and determine priorities for further EQA to improve performance. The contribution of scoring systems to scheme success by making the information more intelligible to participants is described. The importance of reliable target values is shown, and the reproducibility and accuracy of consensus values in such schemes have been studied. EQA data are shown to be invaluable in providing information on the relative performance of analytical procedures, and on factors such as analyte concentration and laboratory workload which affect performance. The stepwise interpretation of the Variance Index scoring system, and the use of graphical presentations in assisting the assessment of laboratory performance are described. Finally, the use of EQA data in the study of the suitability of quality control materials is examined, with particular reference to their commutability, their use in calibration, and the effects of manufacturing procedures upon their properties.

This thesis illustrates the importance of EQA to clinical chemistry practice and to patient care.

ACKNOWLEDGEMENTS

I thank Professor T P Whitehead for providing the opportunity to study for the PhD degree, and for his advice and encouragement in the preparation of this thesis. I also thank Professor J G Ratcliffe and Dr L J Kricka for their advice and constructive criticism.

I am grateful to the members of the Steering Committee on External Quality Assessment for General Clinical Chemistry, particularly Professor D W Moss, Professor J T Whicher, Dr C E Wilde, Mr P M G Broughton and Mr D M Browning, and to Dr R A Bacchus, Dr S S Brown, Dr R E Chambers, Mrs A Green, Dr J B Holton, Dr I Smith, Mr N J Smith, Mr P White, Dr K Wiener and Dr E Worthy, for their expert advice and assistance in many aspects of scheme and survey design, operation and interpretation. I thank the Department of Health and Social Security for funding the UK External Quality Assessment Schemes; the World Health Organization for funding the International EQAS; the Riyadh Al-Kharj Programme for funding the Middle East EQAS; and various manufacturers for providing data and donating materials for distribution through the schemes.

I am indebted to the staff of the UK External Quality Assessment Scheme for General Clinical Chemistry: Nina Frazer, Gwen James, Diana Chester, Helene Bird, Helen Hurley and the Medical Laboratory Scientific Officers, for their expert and unstinting assistance and support in the operation of the schemes; to Margaret Peters, Ian Clark, Ann Gregory, Jagdish Parekh and their colleagues for providing the computing support essential for these schemes; to the scheme participants; to Gary Thorpe and Stephen Smith for reading sections of the manuscript; to Robin Chambers for the statistics in section 13.2; and most of all to my wife Patricia for her forbearance and support.

ABBREVIATIONS USED

A1-AT	$alpha_1$-antitrypsin
AA I	AutoAnalyzer I
AA II	AutoAnalyzer II
Ab	antibody
ACB	Association of Clinical Biochemists
ALP	alkaline phosphatase (EC 3.1.3.1)
ALT	alanine transaminase (alanine aminotransferase; EC 2.6.1.2)
ALTM	all-laboratory trimmed mean
AST	aspartate transaminase (aspartate aminotransferase; EC 2.6.1.1)
BIS	Bias Index Score
C3	complement component 3
C4	complement component 4
CAP	College of American Pathologists
CCV	Chosen Coefficient of Variation
CDC	Centers for Disease Control
CK	creatine kinase (EC 2.7.3.2)
CPC	o-cresolphthalein complexone
CPDA	citrate/phosphate/dextrose/adenine
CV	coefficient of variation
DGKC	Deutsche Gesellschaft für klinische Chemie
DV	designated value
ECCLS	European Committee for Clinical Laboratory Standards
EDTA	ethylene diamine tetra-acetic acid
EQA	external quality assessment
EQAS	external quality assessment scheme
GCMS	gas chromatography-mass spectrometry
GFR	German Federal Republic
GGT	gamma-glutamyl transferase (EC 2.3.2.2)
GOD	glucose oxidase (EC 1.1.3.4)
HB_SAg	hepatitis B surface antigen
HDL	high density lipoprotein
Hep	heparin/Mn^{2+}
HIQC	UKEQAS Human Serum for Intralaboratory Quality Control
HIV	human immunodeficiency virus (HTLV-III; LAV)
IEQAS	International External Quality Assessment Scheme
IFCC	International Federation of Clinical Chemistry
IgA	immunoglobulin A
IgG	immunoglobulin G
IgM	immunoglobulin M
IQC	internal quality control
IRMA	immunoradiometric assay
ISE	ion-selective electrode
LD	lactate dehydrogenase (EC 1.1.1.27)
MEEQAS	Middle East External Quality Assessment Scheme
MRBIS	Mean Running Bias Index Score
MRVIS	Mean Running Variance Index Score
MTB	methylthymol blue
NAC	N-acetyl cysteine
NEQAS	national external quality assessment scheme
NHS	National Health Service
OMRVIS	Overall Mean Running Variance Index Score

PEG	polyethylene glycol
PheA	phenylalanine
PhT	phosphotungstate/Mg^{2+}
PKU	phenylketonuria
QC	quality control
QCM	quality control material
RIA	radioimmunoassay
RID	radial immunodiffusion
SCE	Committee on Enzymes of the Scandinavian Society for Clinical Chemistry
SD	standard deviation
SDBIS	Standard Deviation of the Bias Index Score
SDD	standard deviation difference
SMA	Sequential Multi-channel Analyzer
SMAC	Sequential Multi-channel Analyzer with Computer
SPS-01	Supraregional Protein Reference Units calibration material for immunochemical serum protein assays
SSA	sulphosalicylic acid
T3	triiodothyronine
T4	thyroxine
TBG	thyroxine-binding globulin
TIBC	total iron-binding capacity
TLC	thin layer chromatography
TPTZ	tripyridyl-S-triazine
Tris	tris(hydroxymethyl)-aminomethane
TSH	thyroid stimulating hormone (thyrotrophin)
UKEQAS	UK External Quality Assessment Scheme
VAR	variability of the bias
VI	Variance Index
VIS	Variance Index Score
WIIS	WRL International Intercomparison Scheme
WHO	World Health Organization
WRL	Wolfson Research Laboratories

CONTENTS

Chapter 1 INTRODUCTION

1.1 Quality assurance in clinical chemistry 1
1.1.1 Quality assurance 2
1.1.2 Internal quality control 2
1.1.3 External quality assessment 2
1.2 Terminology in quality assurance 2
1.3 Early external quality assessment surveys 3
1.4 External quality assessment through occasional surveys 5
1.5 The development of internal quality control 7
1.6 External quality assessment 9
1.6.1 The need for external quality assessment 9
1.6.2 Interlaboratory surveys 10
1.7 The development of external quality assessment schemes 11
1.8 Evolution of the UKEQAS for General Clinical Chemistry 13
1.9 Potential applications of data from external quality assessment 14
1.10 Consideration of the requirements, applications and benefits of external quality assessment 16

ASSESSMENT OF INTERLABORATORY AGREEMENT

Chapter 2 EXTERNAL QUALITY ASSESSMENT SURVEYS IN ASSESSING THE STATE OF THE ART

2.1 Introduction 17
2.2 The relationship of patient care to reliability of laboratory results 19
2.2.1 Sensitivity, specificity and efficiency 20
2.2.2 Analytical goals 23
2.3 The quality of extra-laboratory assays in clinical chemistry 24
2.3.1 Exploratory survey - Survey 1 26
2.3.2 First national survey - Survey 2 29
2.3.3 Second national survey - Survey 3 32
2.3.4 Conclusions from the surveys 35
2.4 Laboratory examples 37
2.4.1 Salicylate and paracetamol 38
2.4.2 Specific proteins in serum 40
2.4.3 Serum bilirubin in paediatrics 42
2.4.4 Urinary total protein 44
2.4.5 Diagnosis of phenylketonuria 46
2.5 Summary 49

Chapter 3 FUNDAMENTAL REQUIREMENTS OF SCHEME AND SURVEY DESIGN

3.1 Introduction 51
3.2 Organisation of the scheme 51
3.2.1 Scheme administration 52
3.2.2 Frequency of distributions 52
3.2.3 Number of specimens in each distribution 56
3.2.4 Period between assay and receipt of report 62

3.3 Validity of the specimens 63
3.3.1 Species of origin 65
3.3.2 Specimen presentation 67
3.4 Assessment of performance 70
3.4.1 Report presentation 72
3.4.2 Scoring system for performance assessment 74
3.4.3 Source of target values 74
3.5 Summary 79

Chapter 4 SCORING SYSTEMS IN EXTERNAL QUALITY ASSESSMENT

4.1 Introduction - the need for scoring systems 80
4.2 Classification of scoring systems 80
4.2.1 Pass/fail systems 80
4.2.2 Semi-quantitative systems 81
4.2.3 Quantitative systems 81
4.2.4 'Hybrid' systems 82
4.3 Scoring as a stimulus to improvement 82
4.4 Scoring in assessment of the state of the art 82
4.5 Assessment of progress - comparisons over time 84
4.6 Comparisons over geography 87
4.7 Appraisal of scoring systems 89
4.7.1 Assessment of imprecision 89
4.7.2 SD differences 91
4.7.3 The Variance Index (VI) system 91
4.7.4 BIAS and VAR 92
4.8 Selection of Chosen Coefficients of Variation (CCVs) 93
4.9 Summary 97

Chapter 5 THE VALIDITY OF CONSENSUS VALUES

5.1 Introduction 99
5.2 Reproducibility of consensus values 102
5.2.1 Studies using UKEQASs 102
5.2.2 Studies in the International and Middle East EQASs 106
5.3 Ad hoc comparisons of consensus values 106
5.3.1 National EQASs 106
5.3.2 Commercial EQASs 108
5.4 Comparisons with reference and definitive methods 108
5.4.1 Definitive methods 113
5.4.2 DGKC reference methods 116
5.5 Systematic comparisons between schemes 116
5.5.1 International EQAS 118
5.5.2 Middle East EQAS 118
5.6 The WRL International Intercomparison Scheme (WIIS) 118
5.6.1 Principles and establishment 122
5.6.2 Experience with scheme operation 123
5.6.3 Comparison of results 125
5.6.4 Appraisal of the scheme 134
5.7 Summary 137

Chapter 6 THE EFFECTS OF ANALYTE LEVEL ON INTERLABORATORY AGREEMENT

6.1 Introduction 139
6.2 Studies in the UKEQAS for General Clinical Chemistry 140
6.2.1 Study design 140
6.2.2 Overall data 142
6.2.3 Method-related data 148
6.3 Studies in other schemes 149
6.3.1 UKEQAS for Salicylate and Paracetamol 149
6.3.2 UKEQAS for Lead in Blood 154
6.3.3 UKEQAS Enzyme Surveys 154
6.4 Summary 157

Chapter 7 EXTERNAL QUALITY ASSESSMENT OF NON-QUANTITATIVE TESTS: AMINOACID INVESTIGATIONS

7.1 Introduction 159
7.1.1 Specimen provision 160
7.1.2 Survey design 161
7.1.3 Performance criteria 161
7.1.4 Scoring systems 162
7.1.5 EQA of non-quantitative screening assays 163
7.2 Development of the UKEQAS for Phenylketonuria (PKU) Screening 163
7.2.1 Specimen provision 163
7.2.2 Scheme design 165
7.2.3 Performance assessment and scoring systems 166
7.3 Establishment of UKEQAS Surveys of Urinary Aminoacid Investigations 173
7.3.1 Specimen provision 173
7.3.2 Scheme design 175
7.3.3 Performance assessment 177
7.3.4 Scoring systems 182
7.4 Summary 184

ASSESSMENT OF ANALYTICAL METHODS

Chapter 8 EXTERNAL QUALITY ASSESSMENT IN THE CHOICE OF ANALYTICAL PROCEDURES

8.1 Introduction 186
8.1.1 Reagent and instrument manufacturers 187
8.1.2 Scientific literature 187
8.1.3 Evaluations 188
8.1.4 Information from colleagues 189
8.2 External quality assessment data in method selection 189
8.3 The selection of a method for serum calcium assay 192
8.3.1 Assessment of EQAS data 192
8.3.2 Appraisal of the conclusions 196
8.4 Assessment of factors other than interlaboratory agreement 201
8.4.1 Accuracy 203
8.4.2 Turnround time 206
8.5 Summary 210

ASSESSMENT OF INDIVIDUAL LABORATORY PERFORMANCE

Chapter 9 SCORING AS A STIMULUS TO IMPROVED LABORATORY PERFORMANCE

9.1 Introduction - the assessment of performance 212
9.2 Hierarchical interpretation of scoring 214
9.2.1 The OMRVIS and assessment of overall performance 215
9.2.2 Assessment of performance for individual analytes 218
9.2.3 Appraisal of contributory factors 221
9.2.4 Method assessment 223
9.3 Assessment of performance relative to other laboratories 227
9.3.1 Competition in EQA 230
9.4 Assessment of improvements in performance 231
9.4.1 VI scoring in assessing changes 232
9.4.2 Cumulation period 237
9.5 The detection of unsatisfactory performance 239
9.5.1 Application to UKEQASs 240
9.6 The basis of assessment - state of the art or clinical requirements? 242
9.7 Summary 245

Chapter 10 THE USE OF GRAPHICAL PRESENTATIONS OF EXTERNAL QUALITY ASSESSMENT DATA

10.1 Introduction 247
10.2 Youden plots 247
10.3 Frequency distributions 249
10.3.1 The distribution of results 249
10.3.2 The distribution of scores 251
10.4 Assessment of performance over time 253
10.4.1 Overall performance 254
10.4.2 Performance for individual analytes 254
10.5 Assessment of results for an individual analyte 259
10.5.1 Linearly-related specimens 259
10.5.2 Cumulation procedures 259
10.6 Use of graphical presentations by EQASs 267
10.6.1 Application 267
10.6.2 Resource implications 268
10.7 Summary 271

Chapter 11 STUDIES OF FACTORS AFFECTING PERFORMANCE

11.1 Introduction 273
11.1.1 The Nuffield survey of factors affecting analytical performance in clinical chemistry laboratories 273
11.2 The influence of laboratory workload on performance 275
11.2.1 Effects on overall performance 276
11.2.2 Effects for single analytes 276
11.3 The influence of laboratory type on performance 283
11.4 Summary 285

Chapter 12 THE SUITABILITY OF QUALITY CONTROL MATERIALS FOR EXTERNAL QUALITY ASSESSMENT

12.1 Introduction 286
12.1.1 Precision and bias control in IQC 287
12.1.2 Bias control in EQA 287
12.1.3 Calibration of assays 288
12.1.4 Method comparison 289
12.2 High density lipoprotein (HDL) cholesterol assay 289
12.2.1 Investigation of commercial QC sera 291
12.2.2 Use of lyophilised sera in an EQA survey 295
12.3 Sodium and potassium assay using direct-reading ion-selective electrodes (ISEs) 297
12.3.1 UKEQAS for General Clinical Chemistry 302
12.4 Summary 307

Chapter 13 THE EFFECTS OF CALIBRATION ON INTERLABORATORY AGREEMENT

13.1 Introduction 309
13.2 Specific proteins in serum 312
13.2.1 SPS-01 calibration study 312
13.2.2 Potential confounding factors 317
13.2.3 Period following SPS-01 study 320
13.3 Pregnancy oestrogens in urine 320
13.3.1 Commutability between specimens and calibrant 321
13.3.2 Relationship to interlaboratory agreement 323
13.4 Total urinary protein 324
13.5 Assays of enzyme activity in serum 326
13.5.1 Calibration studies 326
13.5.2 Commutability considerations - amylase assay 333
13.6 Summary 335

Chapter 14 THE EFFECTS OF SPECIES OF ORIGIN AND MANUFACTURING TECHNIQUE ON QUALITY CONTROL MATERIAL BEHAVIOUR

14.1 Introduction 337
14.2 Study of the effects of species and of material manufacturer 339
14.3 The effects of species of origin 341
14.3.1 Examination of overall data 343
14.3.2 Examination of method-related data 343
14.3.3 Appraisal of the results 353
14.4 The effects of material manufacturer 358
14.4.1 Examination of overall data 358
14.4.2 Examination of method-related data 359
14.4.3 Appraisal of the results 364
14.5 Effect of addition of ethylene glycol 365
14.5.1 UKEQAS for General Clinical Chemistry 365
14.5.2 UKEQAS Enzyme Surveys 367
14.6 Summary 369

Chapter 15 **ASSESSMENT OF THE EFFECTS OF HEAT-TREATMENT ON THE BEHAVIOUR OF LYOPHILISED HUMAN SERUM**

15.1 Introduction 371
15.2 Effects of heat-treatment 372
15.2.1 Evaluation data 373
15.2.2 Inorganic, organic and enzyme constituents 373
15.2.3 Thyroid-related and steroid hormones 377
15.2.4 Appraisal of the results 381
15.3 Studies on lot HIQC/13 382
15.4 Summary 384

Chapter 16 **GENERAL DISCUSSION**

16.1 The contribution to patient care of external quality assessment in clinical chemistry 385
16.1.1 The relationship of analytical quality to patient care 385
16.1.2 Consideration of the requirements, applications and benefits of external quality assessment 386
16.2 EQA in the assessment of interlaboratory agreement 387
16.2.1 EQA in assessing the state of the art 387
16.2.2 Fundamental requirements of scheme design 387
16.2.3 The application of scoring systems in EQA 388
16.2.4 Consensus values in EQA 390
16.2.5 The influence of analyte concentration on interlaboratory agreement 391
16.2.6 EQA of non-quantitative investigations 392
16.3 EQA in the assessment of analytical methods 392
16.4 EQA in the assessment of individual laboratory performance 393
16.4.1 Scoring systems for performance assessment 394
16.4.2 Use of graphical data presentations in EQA 395
16.4.3 Investigation of factors affecting performance 395
16.5 EQA in the assessment of quality control material suitability 396
16.5.1 Suitability of QC materials for EQA 396
16.5.2 Calibration effects in EQA 397
16.5.3 Effects of material source and processing 398
16.5.4 Effects of heat-treatment 398
16.6 Conclusion 399

REFERENCES 400

APPENDICES

Appendix I **EXTERNAL QUALITY ASSESSMENT SCHEMES AND SURVEYS**

I.1 General aspects of EQA scheme and survey design 417
I.2 UK External Quality Assessment Schemes (UKEQASs) 418
I.2.1 UKEQAS for General Clinical Chemistry 418
I.2.2 UKEQAS Enzyme Surveys 418

I.2.3 UKEQAS for Lead in Blood 424
I.2.4 UKEQAS for Urinary Pregnancy Oestrogens 424
I.2.5 UKEQAS for Specific Proteins 425
I.2.6 UKEQAS for Salicylate and Paracetamol 425
I.2.7 UKEQAS for PKU Screening 425
I.3 Other EQASs 426
I.3.1 International EQAS (IEQAS) 426
I.3.2 Middle East EQAS (MEEQAS) 426
I.3.3 Intensive EQASs 426
I.4 WRL International Intercomparison Scheme (WIIS) 427
I.5 EQA surveys 428
I.5.1 HDL cholesterol 428
I.5.2 Urinary total protein 428
I.5.3 Extra-laboratory assays 428
I.5.4 Urinary aminoacids 429

Appendix II VARIANCE INDEX SCORING SYSTEM

II.1 General aspects of Variance Index (VI) scoring 430
II.2 VI scoring in UKEQASs 430
II.2.1 Bias Index Score (BIS) 430
II.2.2 Designated Value (DV) 430
II.2.3 Chosen Coefficient of Variation (CCV) 430
II.2.4 Variance Index Score (VIS) 431
II.2.5 Mean Running VIS (MRVIS) 431
II.2.6 Mean Running BIS (MRBIS) 431
II.2.7 Standard Deviation of the BIS (SDBIS) 431
II.2.8 Overall Mean Running VIS (OMRVIS) 431
II.3 VI scoring in other EQASs 431
II.3.1 International EQAS and Middle East EQAS 431
II.3.2 Intensive EQASs 431
II.4 Graphical presentation of VIS data 431

Appendix III STUDY DESIGNS

III.1 Studies on the validity of consensus values 434
III.1.1 Reproducibility 434
III.1.2 Ad hoc comparisons between EQASs 434
III.1.3 Comparisons with values assigned by reference and definitive methods 434
III.1.4 Comparisons between EQASs 434
III.1.5 Comparisons among NEQASs in the WIIS 434
III.2 Studies of interlaboratory agreement 435
III.2.1 Relationship with analyte level 435
III.2.2 Relationship with species of origin 438
III.2.3 Relationship with manufacturer 438
III.3 Studies in the UKEQAS for PKU Screening 439
III.3.1 Relationship with workload 439
III.3.2 Relationship with analytical method 439
III.4 Calibration studies 439
III.4.1 UKEQAS Enzyme Surveys 439
III.4.2 UKEQAS for Urinary Pregnancy Oestrogens 440
III.4.3 Urine protein surveys 440
III.4.4 UKEQAS for Specific Proteins 440
III.5 Studies on HDL cholesterol assay 441
III.5.1 Materials and methods 441
III.5.2 Study protocol 443

III.6 Studies on heat-treated human serum 443
III.6.1 Preparation of materials 443
III.6.2 Evaluation at WRL 443
III.6.3 UKEQAS distributions 444

Appendix IV UKEQAS HUMAN SERUM FOR INTRALABORATORY QUALITY CONTROL

IV.1 Serum processing and production procedures 445
IV.2 Value assignment procedures 445

Appendix V SUPPLIERS OF QUALITY CONTROL MATERIALS AND REAGENTS 449

LIST OF FIGURES

Figure 2.1 Hypothetical distributions of results from populations of normal and diseased individuals, assuming equal contributions from biological and analytical variability (SD) 21

Figure 2.2 Effect on distributions of results from populations of normal and diseased individuals of doubling (A) and halving (B) the analytical variability 22

Figure 2.3 Comparison of glucose results in the second WRL survey of extra-laboratory assays, April 1986, with those from UKEQAS distribution of the same serum 31

Figure 2.4 Distributions of glucose results in the third WRL survey of extra-laboratory assays, November 1986 34

Figure 3.1 Average performance (MRVIS and SDBIS) for all participants in the UKEQAS for Urinary Pregnancy Oestrogens, 1980-1984 54

Figure 3.2 Graphs for phosphate and calcium of laboratory result against designated value for two participants in the Middle East EQAS 57

Figure 3.3 Changes in average VIS for participants in an 'intensive' EQAS 61

Figure 4.1 Relationship of interlaboratory agreement (recalculated CV and average VIS) with analyte concentration (mmol/L) for urea in UKEQAS for General Clinical Chemistry, 1980-1982 83

Figure 4.2 Improvement in interlaboratory agreement (average OMRVIS) in UKEQAS for General Clinical Chemistry, 1972-1986 85

Figure 4.3 Improvement in interlaboratory agreement (average MRVIS) in UKEQAS for Lead in Blood, 1979-1985 86

Figure 5.1 Schematic design for WRL International Intercomparison Scheme (WIIS), with several repesentative laboratories from each national EQAS participating 121

Figure 5.2 Stability of cumulative average percentage deviations for calcium in WIIS 127

Figure 6.1 Relationship with the recalculated mean of recalculated CV for chloride and of average CV for iron in study period 143

Figure 6.2 Relationship with recalculated CV and average VIS for bilirubin in study period 144

Figure 6.3 Relationship with average VIS for glucose in study and validation periods 145

Figure 6.4 Relationship with percentages of VISs >200, >300 and 400 and of numbers of results excluded for creatinine 147

Figure 6.5 Relationship with recalculated CV for creatinine by AAII/SMA and by Other 150

Figure 6.6 Relationship with percentage difference from recalculated mean for glucose by Beckman Glucose Analyzer and by YSI Glucose Analyzer 151

Figure 6.7 Relationship with average VIS for urea by Manual urease and by AAII/SMA 152

Figure 6.8 Relationship between interlaboratory agreement (CV; average VIS) and mean salicylate concentration in UKEQAS for Salicylate & Paracetamol, 1985-1986 153

Figure 6.9 Relationship between interlaboratory agreement (CV) and mean in UKEQAS for Lead in Blood in 1973-1974 (■), 1978 (o) and 1983 (●) 155

Figure 6.10 Relationship between interlaboratory agreement and mean activity for AST and CK in UKEQAS Enzyme Surveys and UKEQAS for General Clinical Chemistry, 1985-1986 156

Figure 7.1 Example of report format in UKEQAS for PKU Screening: Summary of coded actions and PheA concentrations returned by participants 168

Figure 7.2 Example of report format in UKEQAS for PKU Screening. A: Summary of classifications by participants, with corresponding scores. B: Summary of total scores in and average scores for most recent surveys 172

Figure 7.3 Format of results document in UKEQAS Surveys of Urinary Aminoacid Investigation 178

Figure 7.4 Format of report in UKEQAS Surveys of Urinary Aminoacid Investigation: Specimen description and summary of responses by participants 179

Figure 7.5 Format of report in UKEQAS Surveys of Urinary Aminoacid Investigation: Summary of spot test results 180

Figure 8.1 Graphs of laboratory result against method mean for calcium by methylthymol blue in UKEQAS for General Clinical Chemistry, 1986, for participants with (A) good performance, and (B) poor performance 200

Figure 8.2 Relationship between percentage recovery and paracetamol added for enzymic and Glynn & Kendal method groups in UKEQAS for Salicylate and Paracetamol, 1985-1986 204

Figure 8.3 Relationship between interlaboratory agreement (CV) and paracetamol added for enzymic and Glynn & Kendal method groups in UKEQAS for Salicylate and Paracetamol, 1985-1986 205

Figure 8.4 Relationship between analytical method and average score for Surveys 6-19 in UKEQAS for PKU Screening 207

Figure 8.5 Relationship between analytical method and average turnround time (delay between specimen receipt and analysis; working days) for Surveys 17-20 in UKEQAS for PKU Screening 208

Figure 8.6 Relationship between 1983 workload and average turnround time (delay between specimen receipt and analysis; working days) for Surveys 17-20 in UKEQAS for PKU Screening 209

Figure 9.1 Primary page of report for participant laboratory in UKEQAS for General Clinical Chemistry 216

Figure 9.2 Example secondary page of report for participant laboratory in UKEQAS for General Clinical Chemistry 217

Figure 9.3 Demonstration of the effects of bias and imprecision (inconsistent bias) upon the relationship between laboratory result (y axis) and designated value (x axis) 222

Figure 9.4 Relationship between laboratory result (y axis) and designated value (method mean; x axis) for amylase and urea in UKEQAS for General Clinical Chemistry 224

Figure 9.5 Relationship between laboratory result (y axis) and designated value (method mean; x axis) for CK and potassium in UKEQAS for General Clinical Chemistry 225

Figure 9.6 'Hybrid' histogram format used in International EQAS 229

Figure 9.7 Graph of OMRVIS against time for a participant with deteriorating performance in UKEQAS for General Clinical Chemistry 233

Figure 9.8 Graphs of MRVIS against time for participants with (A) improving and (B) deteriorating performance in UKEQAS for Lead in Blood 234

Figure 9.9 Graphs against time of (A) average OMRVIS for all participants and (B) OMRVIS for an individual participant in Middle East EQAS 235

Figure 9.10 Graph of running scores for paracetamol against time for a participant in the UKEQAS for Salicylate & Paracetamol, 1984-1987, showing effect of changing from chemical to enzymic assay procedure 238

Figure 10.1 Youden plot of results in participants' reports from DGKC Ringversuch 4/87 in GFR 248

Figure 10.2 Histogram presentation of results for lead and cadmium in UKEQAS for Lead in Blood 250

Figure 10.3 Presentation of participants' BIAS against VAR for TSH in UKEQAS for Thyroid-related Hormones 252

Figure 10.4 Changes in OMRVIS for participants with (A) good and (B) improving performance in UKEQAS for General Clinical Chemistry 255

Figure 10.5 Changes in MRVIS for calcium for a participant in UKEQAS for General Clinical Chemistry 256

Figure 10.6 Changes in MRVIS for a participant in UKEQAS for Lead in Blood 257

Figure 10.7 Changes in MRVIS, MRBIS and SDBIS for a participant in UKEQAS for Urinary Pregnancy Oestrogens 258

Figure 10.8 Example plots of laboratory result against analyte concentration for a set of linearly-related specimens 260

Figure 10.9 Graphs of laboratory result against designated value for participants in an 'intensive' EQA scheme, showing (A) bias and (B) imprecision 261

Figure 10.10 Graphical displays of results against designated values for sodium and amylase for participants in UKEQAS for General Clinical Chemistry 263

Figure 10.11 Tabular displays of results and designated values for sodium and amylase for participants in UKEQAS for General Clinical Chemistry 265

Figure 10.12 Tabular display of results and ratio to designated value for TSH for a participant in UKEQAS for Thyroid-related Hormones 266

Figure 10.13 Eaxample of mixed text and graphical presentation of data in participant's report from Netherlands national EQAS 270

Figure 11.1 Average OMRVIS for groups of participants in UKEQAS for General Clinical Chemistry, 1984-1986 277

Figure 11.2 Relationship to annual workload of (A) average score for Surveys 15-19 and (B) turnround time in UKEQAS for PKU Screening 281

Figure 11.3 Relationship to (A) 1984 annual workload and (B) 1985 batch freqency of performance (MRVIS at June 1985) in UKEQAS for Urinary Pregnancy Oestrogens 282

Figure 12.1 Differences between the PhT and Hep precipitation procedures for the 25 sera studied, in (A) mean results and (B) within-batch SD 294

Figure 12.2 Relationship with HDL (A) and total (B) cholesterol concentration of between-laboratory agreement (CV) for liquid and lyophilised sera 300

Figure 12.3 Relationship with total protein of difference for sodium between direct ISE and overall mean in UKEQAS for General Clinical Chemistry, 1984 305

Figure 13.1 Interlaboratory agreement (CV) before, with and after SPS-01 for IgG, IgM and C3 in UKEQAS for Specific Proteins 315

Figure 13.2 Relationship between interlaboratory agreement (CV) and concentration for IgM and C3 in UKEQAS for Specific Proteins 318

Figure 13.3 Effect of 'calibration' for AST in UKEQAS Enzyme Survey 17, October 1984 329

Figure 13.4 Effect of 'calibration' for ALP in UKEQAS Enzyme Survey 17, October 1984 330

Figure 13.5 Effect of 'calibration' for amylase in UKEQAS Enzyme Survey 17, October 1984 331

Figure 14.1 Relationship with recalculated CV for glucose and with average VIS for total protein, classified by species of origin 344

Figure 14.2 Relationship with average VIS for bilirubin, classified by species of origin, in study and validation periods 345

Figure 14.3 Relationship with average VIS for iron, classified by species of origin, in study and validation periods 346

Figure 14.4 Relationship with recalculated CV for albumin by Manual BCG, classified by species of origin 347

Figure 14.5 Relationship with recalculated CV and percentage difference from recalculated mean for albumin by BCP, classified by species of origin 348

Figure 14.6 Relationship with percentage difference from recalculated mean for sodium by Corning/EEL 430/450 and Ion-selective electrode, classified by species of origin 350

Figure 14.7 Relationship with percentage difference from recalculated mean for sodium by Indirect ion-selective electrode and Direct ion-selective electrode, classified by species of origin 351

Figure 14.8 Relationship with average VIS for phosphate by Manual/discrete analyser colorimetric, classified by species of origin, in study and validation periods 352

Figure 14.9 Relationship with percentage difference from recalculated mean for total protein by AAII/SMA and Vickers M300/D300, classified by species of origin 354

Figure 14.10 Relationship with percentage difference from recalculated mean for total protein by Continuous flow blanked biuret and Manual/discrete analyser unblanked biuret, classified by species of origin 355

Figure 14.11 Relationship with average VIS and percentage of VISs >200 for bilirubin, classified by manufacturer 360

Figure 14.12 Relationship with recalculated CV and percentage of VISs >200 for magnesium, classified by manufacturer 361

Figure 14.13 Relationship with average VIS for sodium by Continuous flow flame and Ion-selective electrode, classified by manufacturer 362

Figure 14.14 Relationship with recalculated CV for creatinine by Manual/discrete analyser endpoint and Other, classified by manufacturer 363

Figure IV.1 Package insert for lot HIQC/12 - description of purpose, preparation and value assignment procedures 446

Figure IV.2 Package insert for lot HIQC/12 - assigned values 447

LIST OF TABLES

Table 2.1 Comparison of analytical goals for commonly-determined analytes in serum with median within-laboratory imprecision from the Wellcome QC Programme in 1971 and 1986 25

Table 2.2 Comparison of the results from the first WRL survey of extra-laboratory assays, October 1985, with UKEQAS distribution of the same lyophilised serum 28

Table 2.3 Comparison of the results from the second WRL survey of extra-laboratory assays, April 1986, with UKEQAS distribution of the same lyophilised serum 30

Table 2.4 Results from the third WRL survey of extra-laboratory assays, November 1986 33

Table 2.5 Interlaboratory agreement in the first three UKEQAS surveys of salicylate and paracetamol assay, November 1983 - August 1984 39

Table 2.6 Average interlaboratory agreement for immunoglobulins for the four specimens in the first UKEQAS survey of specific protein assays, September 1980 41

Table 2.7 Interlaboratory agreement in the first UKEQAS survey of paediatric bilirubin assay, April 1984 43

Table 2.8 Interlaboratory agreement in the first UKEQAS survey of urinary total protein assay, April 1985 45

Table 2.9 Quantitative results and detection limits reported for specimens 8 (urea and salts in distilled water) and 9 (normal human urine) in UKEQAS urinary total protein surveys 3 and 4, June and November 1986 47

Table 2.10 Interlaboratory agreement in UKEQAS surveys of phenylketonuria screening and quantitative phenylalanine assay, June and November 1978 48

Table 3.1 Considerations in the selection of QC materials based on human or animal serum 68

Table 3.2 Considerations in the selection of lyophilised or liquid QC materials 71

Table 4.1 Relative performance of participants in the UKEQAS for Lead in Blood, classified according to principal component(s) of workload and laboratory type 88

Table 4.2 Average running scores (OMRVIS) in EQASs administered from Wolfson Research Laboratories, 1985 90

Table 4.3 Chosen Coefficients of Variation (CCVs) and average Variance Index Scores (VISs) for all participants during 1986 in the UKEQAS for General Clinical Chemistry 94

Table 4.4 Ratio of average VISs obtained in International EQAS and Middle East EQAS to those for distribution of the same material in the UKEQAS for General Clinical Chemistry, 1986 96

Table 5.1 Reproducibility of consensus values in the UKEQAS for General Clinical Chemistry, 1981-1987 104

Table 5.2 Reproducibility of consensus values in the UKEQAS for Urinary Pregnancy Oestrogens, 1980-1981 (24 distributions) 105

Table 5.3 Reproducibility of consensus values in the International EQAS (11 distributions) and Middle East EQAS (8 distributions), 1985-1987 107

Table 5.4 Comparison of consensus values in the UKEQAS for General Clinical Chemistry with those in the Netherlands EQAS for Armtrol bovine serum, lots 488 and 489, 1980 109

Table 5.5 Comparison of consensus values in the UKEQAS for General Clinical Chemistry with those in the Norwegian (Seronorm, lot 158) and South African (UKEQAS human serum, lot L4/80) EQASs 110

Table 5.6 Comparison of consensus values in the UKEQAS for General Clinical Chemistry with those in the Merz+Dade QAP, 1984 (Levels I and II) and 1986 (Level II) 111

Table 5.7 Comparison of consensus values in the UKEQAS for General Clinical Chemistry with those in the Ortho QC Program (1979-1980) and the Wellcome Group QA Programme (1985) 112

Table 5.8 Comparison of consensus values in the UKEQAS for General Clinical Chemistry with NBS definitive method values for WHO Experimental Reference Serum, lot 4976, 1977 114

Table 5.9 Comparison of consensus values in the UKEQAS for General Clinical Chemistry obtained on Seronorm, lot 158, with definitive method values and with values transferred from NBS SRM909 115

Table 5.10 Comparison of consensus values in the UKEQAS for general Clinical Chemistry with reference method values assigned by the DGKC to Roche Control Sera N and P, 1977 - 1986 117

Table 5.11 Comparison of consensus values in the International EQAS with those in the UKEQAS for General Clinical Chemistry, 1985-1987 119

Table 5.12 Comparison of consensus values in the Middle East EQAS with those in the UKEQAS for General Clinical Chemistry, September 1983 - May 1985 120

Table 5.13 Countries with NEQASs participating in the WRL International Intercomparison Scheme 124

Table 5.14 Cumulative average percentage deviation from WIIS consensus in Trials 101-112 126

Table 5.15 Overall average cumulative average percentage deviation from WIIS consensus in Trials 101-112, with corresponding data and average percentage deviation of national EQAS from WIIS consensus (Table 5.18) only for countries providing information on national performance 129

Table 5.16 Percentage deviations of WIIS participants from their national EQAS 131

Table 5.17 Comparison of performance in the UKEQAS for General Clinical Chemistry and the WIIS 132

Table 5.18 Percentage deviation of national EQAS from WIIS consensus 133

Table 5.19 Average VISs in the WIIS for Trials 101-112, with ratio to those in UKEQAS for General Clinical Chemistry 136

Table 6.1 Indicators of performance used in study of relationships between analyte level and interlaboratory agreement 141

Table 7.1 Coded alternative laboratory actions which would normally be taken following the first analysis of a specimen by the routine screening procedure, used in the UKEQAS for PKU Screening 167

Table 7.2 First scoring system in UKEQAS for PKU Screening, used in Survey 4, November 1981 170

Table 7.3 Turnround times for UKEQAS Surveys 2 and 3 of Urinary Aminoacid Investigation, October 1986 and February 1987 176

Table 7.4 Summary of tentative categorisation of participants' responses in UKEQAS Surveys 1-3 of Urinary Aminoacid Investigation 183

Table 8.1 Average interlaboratory agreement for serum calcium assay in UKEQAS for General Clinical Chemistry, January-June 1980 193

Table 8.2 Average interlaboratory agreement for serum calcium assay in UKEQAS for General Clinical Chemistry, January-June 1986 (n = 12 distributions), with average running scores at June 1986 197

Table 8.3 Average and range of running scores for serum calcium assay by methylthymol blue method group in UKEQAS for General Clinical Chemistry, 1983 and 1984 199

Table 8.4 Average MRVIS and recovery and repeatability studies in UKEQAS for Lead in Blood, 1980, with ranges for dithizone method group 202

Table 9.1 Comparison, in terms of VIS, of performance attained in UKEQAS for General Clinical Chemistry with analytical goals 220

Table 9.2 Average running scores for method groups and subgroups for sodium assay in UKEQAS for General Clinical Chemistry, 1987 226

Table 11.1 Performance (average OMRVIS) for groups of participants in UKEQAS for General Clinical Chemistry, December 1986 278

Table 11.2 Performance (1981) of participants in the UKEQAS for Lead in Blood classified according to annual blood lead assay workload (1980) and interval between batches 279

Table 12.1 Mean and within-batch precision (SD) of HDL cholesterol assay by PhT and Hep procedures on pooled patients' sera and 25 QC sera 293

Table 12.2 Comparison of within-batch and between-day mean and precision (CV) of HDL cholesterol assay for 6 sera 296

Table 12.3 Intralaboratory precision, mean results and between-laboratory agreement for HDL and total cholesterol in Survey 3, 1980 298

Table 12.4 Summary of intralaboratory precision, mean results and between-laboratory agreement for HDL and total cholesterol in Surveys 2 (1979), 3 (1980), and 4 or 5 (1981) 299

Table 12.5 Comparison of performance (average running scores) for sodium and potassium assay by direct ISE instruments in UKEQAS for General Clinical Chemistry, 1984 303

Table 13.1 Mean between-laboratory CV for immunoglobulins before, with and after SPS-01 313

Table 13.2 Mean between-laboratory CV for C3, C4 and A1-AT before, with and after SPS-01 314

Table 13.3 Average and range of overall means before, with and after SPS-01 319

Table 13.4 Effect of 'calibration' on between-laboratory agreement for urinary pregnancy oestrogen assay 322

Table 13.5 Effect of 'calibration' on between-laboratory agreement for urinary total protein assay in Surveys 2 and 3 325

Table 13.6 Overall statistics, irrespective of method, after 'calibration' (after exclusion of results more than 2 SD from the untrimmed mean) in UKEQAS Enzyme Surveys 14-18 327

Table 13.7 Effect of 'calibration' for CK in Survey 17, October 1984 328

Table 13.8 Effect of 'calibration' for amylase in Surveys 14-18 332

Table 14.1 Classification of materials studied by species of origin and manufacturer 342

Table 14.2 Interlaboratory agreement obtained for human serum stabilised with ethylene glycol in the UKEQAS for General Clinical Chemistry, April 1982, compared with that for lyophilised serum 366

Table 14.3 Interlaboratory agreement obtained for human serum stabilised with ethylene glycol in UKEQAS Enzyme Survey 13, 1983, compared with that in Enzyme Survey 14 for 'reliable' method groups 368

Table 15.1 Comparison of pH, turbidity and vial-to-vial variability for heat-treated batch and lots HIQC/9 to HIQC/12 374

Table 15.2 Comparison of interlaboratory agreement (CV) for heat-treated batch and lots HIQC/9 to HIQC/12, with mean value for heat-treated batch 375

Table 15.3 Comparison of average Variance Index Scores (VISs) for heat-treated batch and lots HIQC/9 to HIQC/12 376

Table 15.4 Comparison of method-related data for general clinical chemistry analytes for heat-treated batch and lot HIQC/12 378

Table 15.5 Comparison of method-related data for hormones for heat-treated batch and lot HIQC/12 379

Table 15.6 Comparison of method-related data for steroid hormones for heat-treated batch and lot HIQC/12 380

Table 15.7 Total and free thyroxine content for heat-treated batch and lots HIQC/9 to HIQC/12, and lot HIQC/13 (also heat-treated) 383

Table I.1 Description of UKEQASs at January 1987 419

Table I.2 Materials distributed in UKEQAS for General Clinical Chemistry, 1978-1987 420

Table II.1 Chosen Coefficients of Variation for analytes not in UKEQAS for General Clinical Chemistry 432

Table III.1 Materials distributed in International EQAS, 1984-1987 435

Table III.2 Materials distributed in Middle East EQAS, 1983-1987 436

Table III.3 Identity of the QC materials assessed for HDL cholesterol assay 442

Chapter 1:

INTRODUCTION

1.1 Quality assurance in clinical chemistry

The primary role of any clinical laboratory is to assist in the diagnosis, monitoring and prevention of disease and the monitoring of therapy, through provision of quantitative or qualitative data on specimens from patients or subjects. These data can only be of optimal assistance if they are completely reliable, both within and among laboratories. Unreliable laboratory reports may lead to inaccurate diagnoses or inappropriate treatment, and increase suffering and healthcare costs thereby or by necessitating repeated tests.

All measurements, including clinical chemical analyses, are subject to variance: 'variance' here denotes discord and discrepancy in the results of measuring the same quantity on the same material (Whitehead, 1977) rather than its statistical usage as the square of the standard deviation (SD). This variance was largely overlooked in the early development of clinical chemistry. Clinical chemists only became aware of the importance of quality assurance in their professional activity when external quality assessment (EQA) surveys, such as those of Belk and Sunderman (1947) in the USA and later of Wootton and King (1953) in the UK, revealed large variations in the the results obtained on the same specimen in different laboratories.

The terms quality assurance, internal quality control (IQC) and external quality assessment (EQA) are used here as defined by the World Health Organization (WHO; WHO, 1981) and the European Committee for Clinical Laboratory Standards (ECCLS; Leblanc et al, 1985a):

1.1.1 Quality assurance

All measures taken to increase within-laboratory reproducibility and between-laboratory comparability, and to ensure the usefulness of laboratory investigations generally.

1.1.2 Internal quality control

The set of procedures undertaken by the staff of a laboratory for the continual evaluation of the reliability of the work of the laboratory and its emergent results, in order to decide whether they are reliable enough to be released.

1.1.3 External quality assessment

A system of retrospectively and objectively comparing results from different laboratories by means of an external agency.

1.2 Terminology in quality assurance

The terminology used in quality assurance, both within laboratories and in the published literature, was confused until 1981.

Originally "quality control" was the main term used to denote the general field of quality assurance, and this was qualified to designate the aspects of analytical QC as internal quality control and external quality control. This usage was confirmed by the recommendations of the IFCC Expert Panel on Quality Control (Büttner et al, 1979a), developed during the 1970s.

"Internal quality control" was widely and almost universally used, but terms such as "preventive measures" (Whitehead, 1977) and "good laboratory practice" were evolved to cover much of quality assurance. In the field of EQA a much greater diversity in nomenclature had arisen, with "external quality control" being

used only occasionally. Thus, there were "roundrobins", "interlaboratory surveys", "proficiency testing surveys", "quality assurance programs" and "clinical laboratory improvement programs"; many systems, eg the National Quality Control Scheme in the UK (Whitehead et al, 1973), were simply designated as "quality control schemes".

Confusion over the objectives of the various aspects of quality assurance was almost inevitable in such a situation, with IQC and external QC being considered to be as interchangeable as this nomenclature implied. The WHO Working Group considered that a change in terminology was essential, both to emphasise the complementary nature of IQC and EQA and to bring out the objectives of EQA besides the 'control' of individual laboratory performance.

The terms "quality assurance", "internal quality control" and "external quality assessment" were therefore selected, as best conveying the intent and content of these aspects of activity (WHO, 1981). This terminology, defined above, has also been endorsed by ECCLS (Leblanc et al, 1985a) and the International Committee for Standardization in Haematology (Lewis, 1984).

1.3 Early external quality assessment surveys

The earliest published EQA surveys of clinical chemical analyses were conducted in the USA by Belk and Sunderman (1947), with the primary objective of checking accuracy. These surveys, involving 59 pathologists' laboratories in Pennsylvania, used aqueous solutions of salts, glucose or urea, serum (for protein assays) and preserved whole blood (for haemoglobin assay). Essentially arbitrary limits of acceptability were set, of $\pm$10% for most analytes but based on the authors' opinion of satisfactory

laboratory practice. Belk and Sunderman expressed surprise at the scatter and consequent unreliability of the results obtained. For example, calcium results ranged from 3 to 13 and from 7 to 15 mg/dL (0.75 - 3.25 and 1.75 - 3.75 mmol/L) for solutions prepared to contain 6.6 and 12.6 mg/dL (1.65 and 3.15 mmol/L) respectively, and more than 60% of participants failed to obtain 'acceptable' results for at least half of the specimens. There seemed, however, to be no evidence of an overall bias relative to weighed-in values.

In further enquiries, most of the participating pathologists attributed the unreliability of results to the inadequate numbers and poor training of their technical staff. Some also blamed poor equipment and lack of space, and lack of understanding between the pathologist and his staff; the authors felt this latter reflected poor communication between hospital wards and the laboratory as a whole.

Other groups in the USA also carried out local (statewide) surveys around this time (Snavely and Golden, 1949 and 1951). The spread of results was broadly similar to that shown by Belk and Sunderman, with about one third of results being classified as unsatisfactory. These authors attributed much of the poor performance to inadequate staffing, training and supervision, but also commented on the deficiencies of pre-calibrated instruments (Snavely and Golden, 1949) and advocated the trial of standardised methods in further surveys (Snavely and Golden, 1951). Surveys were also undertaken in other laboratory disciplines, with similar findings (Marsters, 1949; Hardy, 1952); there was over 10 years' experience of EQA in syphilis serology testing (Cumming et al, 1935). Again these authors

stressed the need for improvement through provision of advice and assistance and through standardisation of procedures, with EQA providing an educational stimulus for improvement.

A few years later, the first survey in the UK was undertaken by Wootton and King (1953). These workers had derived reference intervals (then termed "normal ranges") for common laboratory investigations, and wished to determine their applicability to assays carried out in other laboratories. The initial study involved determination by 21 laboratories of 6 analytes in a specimen of preserved whole blood. The wide scatter of results (eg from 45% to 170% of the mean for urea) prompted a further survey of 9 assays among 36 laboratories, this time using an aqueous solution of pure inorganic and simple organic compounds at known concentrations. Though there was no evidence of consistent bias from the weighed-in value in most cases (apart from an average overestimation of creatinine by about 40%), a considerable part of the variability persisted. The authors asserted that the use of different analytical methods did not explain the observed lack of agreement.

Wootton and King emphasised the need to investigate further the causes of the between-laboratory variance revealed by their survey, and proposed the institution of a system of regular checks on a national scale.

1.4 External quality assessment through occasional surveys

Through the years following the initial surveys described above a number of workers pursued the study of between-laboratory variability in results. Many of these reported the results of 'one-off' local or national EQA surveys (eg Tonks and Allen,

1955; Holtz, 1959; Tonks, 1963; Desmond, 1964; Gowenlock, 1969), and in one case an international survey (Wootton, 1956). Other publications reviewed the results of regular but essentially local EQA schemes (eg Shuey and Cebel, 1949; Campbell, 1962; Merritt et al, 1965; Evans et al, 1966).

These largely confirmed the findings of the earlier surveys. Shuey and Cebel (1949), reporting several years' experience in surveys of military laboratories, postulated that the efficacy of EQA was directly related to the survey frequency and to the number of specimens included. Interestingly, Merritt et al (1965) reported only insignificant overall improvement in laboratories provided with additional advice and assistance.

The 1954 international survey, sponsored by IFCC, showed that the unsatisfactory between-laboratory agreement was not confined to individual countries, and demonstrated for the first time that laboratories could compare two solutions more accurately than determine the absolute concentration of an analyte (Wootton, 1956). Overall, there appeared to be a need for more satisfactory analytical methods and procedures for their control (eg Wootton, 1957).

The EQA surveys conducted during the 1950s and early 1960s were limited in scope. The facilities available to their organisers were also limited. Thus the specimens may have been less than ideal, the surveys often covered a few analytes only, and the intervals between the survey itself and the appearance of reports for participants and/or publication in the scientific literature were prolonged. Surveys other than those mentioned above were also carried out, but the results were not published.

1.5 The development of internal quality control

Publication and disussion of the findings of the initial EQA surveys led to a growing awareness of the importance of quality assurance in clinical laboratory medicine. In particular the possibilities of developing IQC procedures based on the successful applications of quality control (QC) in manufacturing industry (eg Shewhart, 1931) were explored by authors such as Levey and Jennings (1950). Such procedures had already been implemented in analytical chemistry laboratories (Wernimont, 1946; Mitchell, 1947).

One problem, however, is that while industrial QC employs measured characteristics of the manufactured item itself to monitor the production process chemical analyses yield only an analytical result, which should differ for each specimen. Thus most IQC procedures in clinical chemistry use the reference sample technique, whereby a control specimen (or specimens) is analysed periodically among the clinical specimens (eg Archibald, 1950; Levey and Jennings, 1950; Henry and Segalove, 1952; Benenson et al, 1955; Henry, 1959). Clinical and control specimens are assumed to be affected identically by the analytical procedure, though this may not be true in all cases if commercial QC materials are used. In the earlier literature there is confusion between the use of "standards" for calibration of the assay and/or for its control; Archibald (1950) was the first to advocate the inclusion of a pooled serum specimen with every set of determinations, to be carried through all steps of the analysis (including calculation), for control purposes.

In addition to the reference sample approach, other workers

advocated the use in IQC of the results on clinical specimens, through the calculation of 'daily means' (Waid and Hoffmann, 1955; Hoffmann et al, 1961; Whitehead and Morris, 1969). This approach yields information on performance which is not subject to errors due to the 'artificial' nature of some QCMs. It does, however, require additional calculations and is prone to influences from variation in the population from which the specimens originate, necessitating extra precautions to avoid erroneous conclusions (eg Dixon and Northam, 1970).

IQC procedures have been continually improved over subsequent years, and increasingly sophisticated systems have been devised (eg Whitby et al, 1967; Whitehead, 1976 and 1977; Grannis and Caragher, 1977; Stamm, 1981; Büttner et al, 1983a), some endorsed by international organisations such as WHO and IFCC. Westgard and his colleagues have made a notable contribution in validating statistically the effectiveness of IQC systems (eg Westgard et al, 1979; Westgard and Groth, 1981), and the so-called 'Westgard rules' (Westgard et al, 1981) have been widely accepted. More recently, Westgard and Groth (1983) have proposed models based on the predictive value approach previously used to assess and optimise the interpretation of diagnostic procedures (Galen and Gambino, 1975). Fraser (1983) has also advocated use of analytical goals based on biological variability in place of statistical considerations (eg $\pm$2 SD) in setting control limits.

In parallel with the development of IQC programmes came the realisation of the importance of quality assurance, covering all activities from test selection, through specimen collection and transportation, laboratory analysis and report generation, to report interpretation. Many aspects of quality assurance

represent 'good laboratory practice', and were previously termed 'preventive measures' by Whitehead (1977). There have also been improvements in education and awareness of the need for quality in laboratory investigation (Büttner et al, 1980b).

IQC is now applied almost universally in the developed world as an integral part of laboratory practice. These procedures are a major cost in both time and materials (Tydeman et al, 1982), but this is accepted as essential to the production of reliable laboratory results; as in industry, the price of poor quality is greater than that of good quality (Price, 1984).

1.6 External quality assessment

However sophisticated a laboratory's IQC programme may be, it was apparent by the 1960s that EQA was essential to attain and maintain comparability of results among laboratories. The reasons for this were later categorised (WHO, 1981; Büttner et al, 1983b).

1.6.1 The need for external quality assessment

These groups pointed out that comparability of results among laboratories (transferability with respect to geography) is essential to ensure uniformity of interpretation with regard to:

- common reference intervals
- mobility of medical staff between healthcare facilities (eg hospitals)
- movement of patients between healthcare facilities
- analyses provided by different laboratories at different stages of investigation or treatment
- clinical or epidemiological multicentre studies
- application of legal provisions (eg Health and Safety

Commission, 1980)

The emphasis of IQC is on control of batch-to-batch imprecision, and its basic question is whether the batch is sufficiently similar to preceding batches to permit release of results from the laboratory. Such control is normally prospective relative to reporting procedures, and unsatisfactory batches of results can therefore be identified and eliminated. The primary objective of EQA, however, is the assessment of accuracy. This is more usually expressed as bias relative to a designated or target value, since the true value for many analytes in biological specimens is unknown and the designated value ("correct value"; Whitehead, 1977) must be used as an approximation to this (WHO, 1981).

Since its outcome cannot influence release of results, EQA is essentially a retrospective activity and is complementary to IQC. These factors prompted the change in terminology (see section 1.2 above; WHO, 1981), since the previous use of "external quality control" had misled some into thinking that IQC and EQA were interchangeable. The term "external quality assessment" also emphasises the aspects (assessment of interlaboratory agreement, of analytical methods, and of quality control materials) of EQA other than 'control' of individual laboratories' performance, which are discussed in section 1.9 below and which form the subject of this thesis.

1.6.2 Interlaboratory surveys

EQA may be considered to include procedures such as specimen exchange (eg Whitehead, 1977), but usually takes the form of interlaboratory surveys. In essence, an interlaboratory survey (WHO, 1981; Büttner et al, 1983b) includes the following stages:

- distribution of a specimen or specimens from the organising laboratory to participant laboratories
- assay of the specimen(s) by participants according to the instructions provided
- return of results to the organising centre
- data processing, usually using computer facilities, of these results to yield information on interlaboratory agreement and other aspects of performance
- distribution to participants of a report based on the processed data, which may include interpretive comments (general, and/or specific to individual laboratories such as those returning discrepant results)

EQA surveys may be proficiency surveys conducted for the purpose of licensing the participant laboratories, ie determining whether their results are sufficiently reliable to provide satisfactory patient care, or be primarily educational in intent.

1.7 The development of external quality assessment schemes

During the 1960s the potential benefits of established EQASs on a national scale became increasingly apparent, echoing the original conclusions of King and Wootton (1953). Thus, in his Presidential address to the Association of Clinical Pathologists Jordan (1965) emphasised the need for EQA in addition to IQC and hoped the Association would support the establishment of national EQASs for the UK, which need not be restricted to clinical chemistry.

This changing attitude, towards acceptance by many clinical chemists that participation in EQA was not only desirable but essential for improving the reliability of laboratory investigations, was important since it ensured an enthusiastic initial group of participating laboratories. At the same time

developments in electronic data processing in laboratories enabled not only the easier and more efficient use of IQC procedures, but also the rapid processing of the large numbers of results regularly required to operate a national EQAS (NEQAS).

Following this professional pressure, several proposed national schemes secured funding, primarily from governmental sources. In the UK, National Quality Control Schemes (Whitehead and Woodford, 1981) covering firstly clinical chemistry and later haematology were initiated by groups in Birmingham (Whitehead et al, 1973) and London (Lewis and Burgess, 1969; Ward and Lewis, 1975), to be joined by a commercial scheme operated by Wellcome Diagnostics (then Wellcome Reagents Ltd). Activities were also started in other countries, including Canada (Ley and Ezer, 1974) the German Federal Republic (GFR; Bundesärztekammer, 1971; Stamm, 1975) and the USA (Gilbert, 1975a; Grannis and Caragher, 1977). In some cases the development of EQASs was stimulated by legislation, such as the Calibration Law in GFR and the Clinical Laboratories Improvement Act in the USA, whereas in others professional societies provided the main impetus.

These schemes evolved over the succeeding years, and the scope of EQA was broadened through the introduction of NEQASs covering more specialised aspects of clinical chemistry such as radioimmunoassays (RIAs) for hormones (eg Hunter and McKenzie, 1979; Röhle and Breuer, 1978; Groom, 1985a) and other clinical laboratory disciplines (see Whitehead and Woodford, 1981). Other countries, such as Belgium (de Leenheer et al, 1983), France, and Holland (Jansen et al, 1977), also initiated NEQASs.

A different EQA design, with interlaboratory comparison of

results from IQC procedures using the same material, was proposed by Limonard (1979) and implemented in Holland (Jansen and Jansen, 1980). This was based on the principles of 'regional quality control' administered by the College of American Pathologists (CAP; Lawson et al, 1980) and the commercial schemes operated by companies such as American Hospital Supply, and its deficiencies have been discussed by Tonks (1982).

By the mid-1980s, NEQASs had been established in most countries of the developed world and many in the developing world. There was common ground in many aspects of scheme design (WHO, 1981), but differences in details have persisted. In particular the derivation of designated values from assays by reference laboratories (eg Hansert and Stamm, 1980) or from consensus values (eg Gilbert, 1976; Grannis, 1976) has continued to generate considerable dispute. This and the other main factors are discussed in detail in Chapters 3 to 6.

1.8 Evolution of the UKEQAS for General Clinical Chemistry

This scheme was instituted in 1969 as the National Quality Control Scheme, supported by Department of Health and Social Security (DHSS) research and development funds. The main objectives (Whitehead et al, 1973) were:

- frequent distributions (every 14 days) of reliable specimens, prepared from human serum
- rapid return of results by participants
- rapid processing of these results, to make a report available within 10 days of specimen receipt
- results presentation to enable participants to assess their performance

- voluntary and confidential participation, open to all UK clinical laboratories
- assessment of any improvements in precision or accuracy, and of any effects of workload or analytical procedures

This design appeared to be well-accepted by the participants, which grew in numbers to about 350 from the initial 200 laboratories. By 1971 each analyte had been surveyed up to 21 times but there was no evidence of improving overall performance during this period. To provide a greater stimulus for improvement, a system for scoring performance, the Variance Index (VI), was then introduced and thereafter between-laboratory agreement was seen to improve (Whitehead et al, 1973). The authors postulated that a scoring system assisted in both assessing and monitoring changes in participants' performance; the VI system was later modified to increase its utility (Whitehead, 1977; Bullock and Wilde, 1985).

Modifications, some described in Whitehead (1977) and many of which are discussed in this thesis, were made to the basic scheme design in succeeding years. In particular the scope was broadened by occasional surveys (eg Bold and Browning, 1975; McSweeney et al, 1979) and by the introduction of sub-schemes for more specialised assays, the first of which was the UKEQAS for Lead in Blood (Bullock et al, 1986c). The current situation has recently been reviewed briefly (Bullock, 1985; DHSS, 1986b).

1.9 Potential applications of data from external quality assessment

Most clinical chemists think of EQA purely in terms of monitoring the performance of individual laboratories. Though many EQASs are established primarily with this objective, the true purposes of

EQA are much wider, as emphasised by use of the term external quality assessment (WHO, 1981).

Any EQA survey or scheme will yield information about the degree of agreement among results from the various participating laboratories. This interlaboratory concordance, often referred to as 'the state of the art', can be useful in comparisons with medical requirements and judging whether improvements are required. One means for such improvement is EQA, and the information can be used to determine priorities for EQAS establishment (WHO, 1981).

This assessment of the state of the art can be extended to individual analytical procedures. Thus the performance characteristics of method principles and instrumentation can be assessed (WHO, 1981; Jansen et al, 1981).

The main objective of most EQASs is the assessment of the performance of individual participant laboratories. A number of determinants of performance can be examined, covering aspects such as turnround time for assays in addition to analytical characteristics such as accuracy and precision (WHO, 1981). The aim is not only to stimulate and monitor improvements in performance, but also to identify those laboratories experiencing problems and in particular need of assistance (eg Browning, 1984; Walker, 1985).

Finally, interlaboratory agreement in EQA is intimately related to the behaviour of the materials distributed. Thus EQA data provide a means for study of the properties and quality of QC materials (eg Jansen, 1980). A related aspect is the ability to assess the effects of calibration practices (eg Jansen and

Jansen, 1983).

1.10 Consideration of the requirements, applications and benefits of external quality assessment

As an important element of quality assurance, external quality assessment is a determinant of the quality of patient care. The fundamental aspects of EQAS design do not seem, however, to be based on clear logical principles. Though their role in monitoring laboratory performance is obvious, the potential roles of EQA schemes in the assessment of interlaboratory agreement, in the assessment of analytical methods, in the assessment of individual laboratory performance and in the assessment of quality control materials, appear not to have been widely appreciated.

In this thesis these applications of external quality assessment schemes and their benefits are examined in turn, with consideration of the basic requirements of scheme design. This thesis is concerned with the following question: **What has external quality assessment to contribute to the scientific development of clinical chemistry and to patient care, besides the 'policing' of laboratory performance?**

Chapter 2:

EXTERNAL QUALITY ASSESSMENT SURVEYS IN ASSESSING THE STATE OF THE ART

2.1 Introduction

From the very earliest surveys EQA has been used to assess the state of the art (Cumming et al, 1935; Belk and Sunderman, 1947; Wootton and King, 1953). The survey organisers, concerned for the quality of patient care decisions based on the results of laboratory investigations, sought information on the degree of interlaboratory agreement and within-laboratory precision, and EQA was the tool best suited to this objective.

Reasons (more than one of which may motivate surveys in many cases) for desiring such information include:

- curiosity about the situation
- determining the factors contributing to good or poor performance
- monitoring the effects on interlaboratory agreeement of other activities (eg the introduction of recommended methods)
- assessment of the need for continued survey and for establishment of an EQA scheme

Consider for example an assay for which a number of analytical methods are available and which has only recently come into widespread use as a 'routine' analysis. Though the methods performed well in their originators' laboratories they are known not to be robust and to be difficult to control; current performance under routine conditions may therefore be less

satisfactory with regard to both precision and accuracy. The calibration materials are believed to be unstable, and since serum-based secondary calibrants are needed differences in the values assigned by various manufacturers are suspected, thus compounding the problems.

What then should be done about the situation? An EQAS could be initiated, but it is conceivable that the problems suspected on scientific grounds may be groundless: the EQAS would then be superfluous, and resources be wasted thereby. Alternatively, the situation may prove to be grossly unsatisfactory and a major initiative to provide better methods or improved means for their control be required: continued frequent and regular survey would again be inappropriate. Finally the specimens chosen for distribution may be unstable or otherwise unsuitable, requiring identification or development of more appropriate alternative materials.

The inevitable conclusion is that it is unwise to initiate an EQAS without first obtaining information on the current state of the art and on the design criteria for a scheme which is likely to succeed in its objectives (discussed in Chapter 3 below). The logical course of action would therefore be to carry out a pilot survey, or a short series of surveys, to assess the former.

The information generated will then be of prime assistance in determining the appropriate further action(s), if any. Thus, if the situation is deemed satisfactory then action may be unnecessary (or at least not urgent), and if it is less so then an EQAS to stimulate improvement may be indicated. If it is totally unsatisfactory then immediate actions by way of method selection and recommendation or by implementation of effective

IQC procedures will be required.

Numerous such situations exist within clinical chemistry. The continually-widening test repertoire has provided a major challenge to EQAS organisers in selecting which analytes require regular schemes, and even in allocating priorities for exploratory survey since resources are limited. Clinical relevance, within-laboratory performance data and evidence from limited EQA activities all lend some assistance in this latter, which should then yield the quantitative information on which rational policy decisions may be based.

Extra-laboratory investigations, the quality of which is of great clinical importance to the provision of reliable patient care, provide a good example of the application of EQA surveys. These therefore form the basis of the consideration below, other aspects then being discussed in relation to laboratory determinations.

Before embarking on any consideration of the application of EQA, however, the question of whether patient care in fact benefits from improved within- and between-laboratory agreement must first be addressed.

2.2 The relationship of patient care to reliability of laboratory results

The aim of quality assurance is to increase the reliability of results leaving the laboratory, but is this relevant to patient care? This question may be addressed using the the concepts of the sensitivity, specificity and predictive value of laboratory tests developed by Galen and Gambino (1975), and the determination of analytical goals in terms of their relationship

to biological variation pioneered by Harris and his co-workers (Subcommittee on Analytical Goals, 1979; Harris, 1979; reviewed by Fraser, 1983).

2.2.1 Sensitivity, specificity and efficiency

Consider the application of a laboratory test to differentiate between populations of 'normal' and a 'diseased' individuals (Figure 2.1). As with many investigations there is incomplete separation of the two groups. The test thus gives false positive and false negative results, due to failings in its specificity and sensitivity respectively, and its predictive value and efficiency are consequently suboptimal.

What then is the effect of analytical variance on these characteristics? The variation in each group is composed of both analytical and biological variation, so any change in analytical performance will be reflected in the overall spread of results. For example, Figure 2.2 shows the consequences of doubling and halving the analytical variation (here assumed initially to be equal to the inherent biological variance; Figure 2.1) on the incidence of misclassification as false positives and false negatives.

It is clear from such simulations that the clinical discrimination of the test is related directly to analytical performance. The position of the cut-off can be varied to optimise sensitivity, specificity, predictive value or efficiency according to the application (Galen and Gambino, 1975), but for any given position improvements in analytical performance will improve all these characteristics contributing to the reliability of healthcare.

Figure 2.1 Hypothetical distributions of results from populations of normal and diseased individuals, assuming equal contributions from biological and analytical variability (SD). Arrow denotes position of cut-off; FP and FN are false positive and false negative results

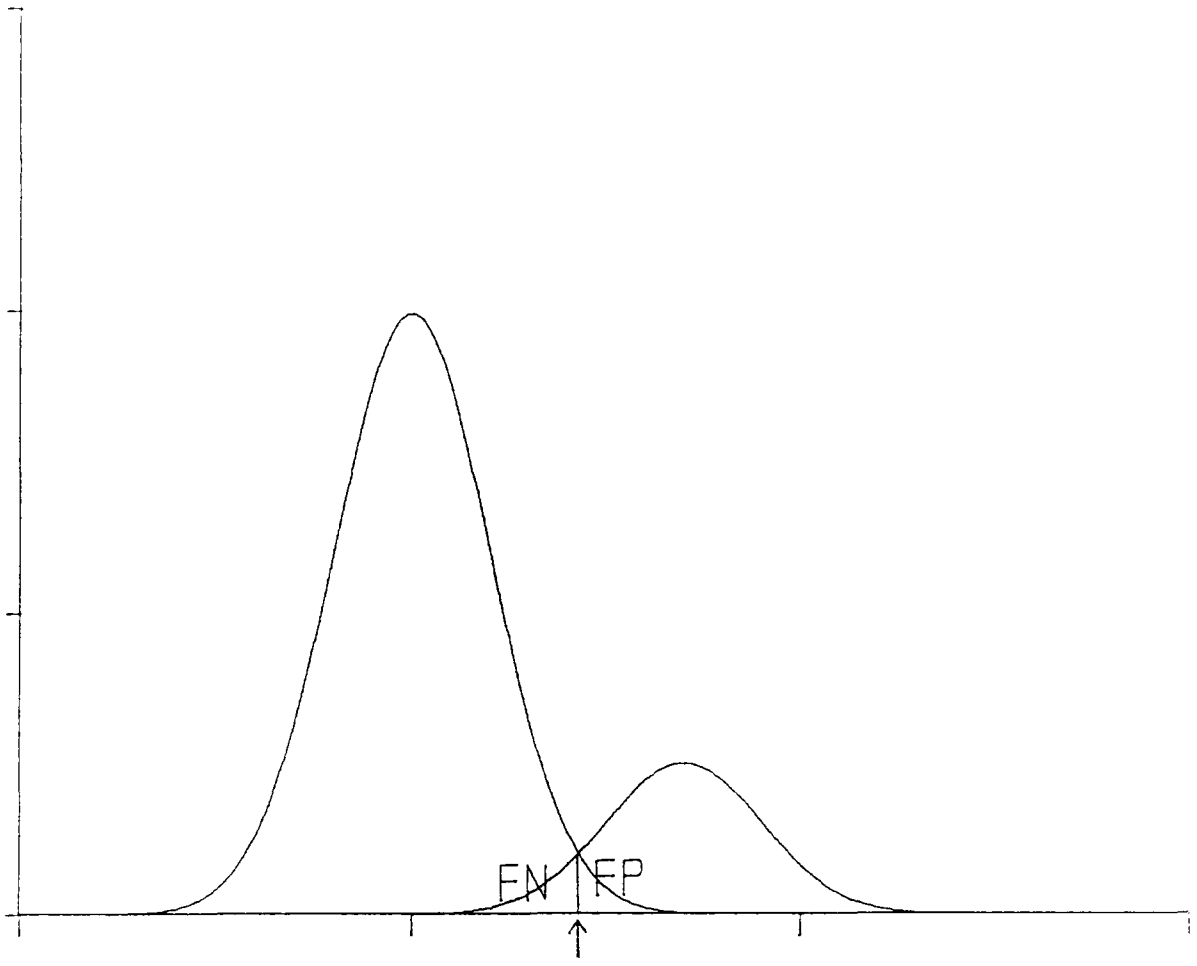
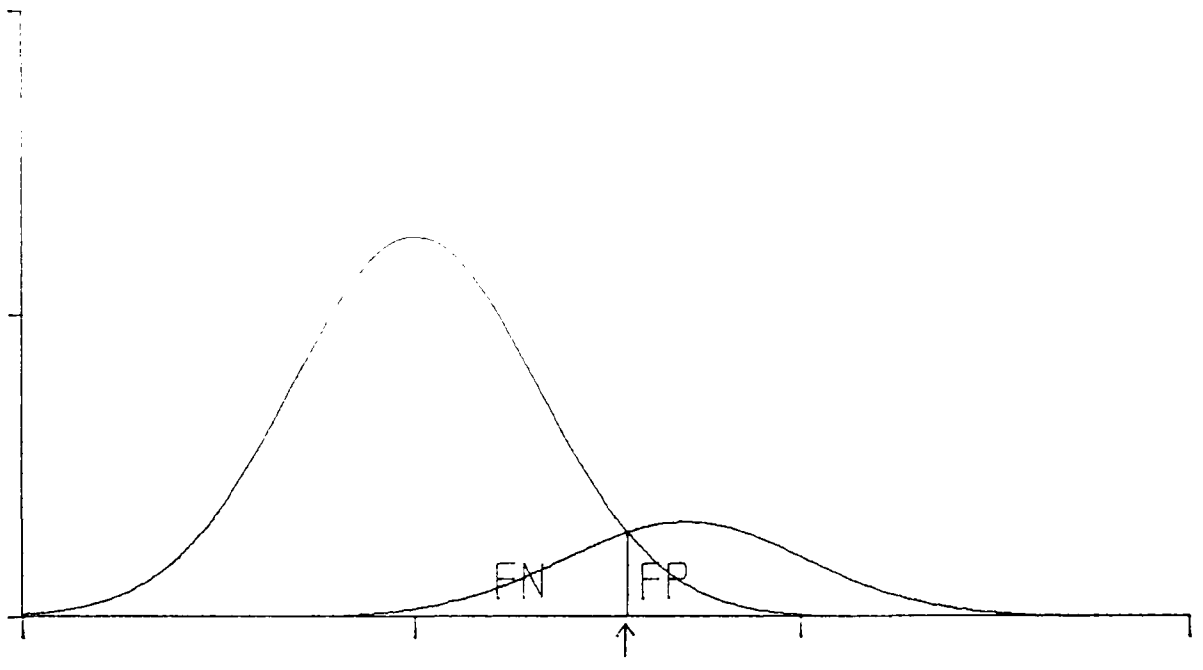
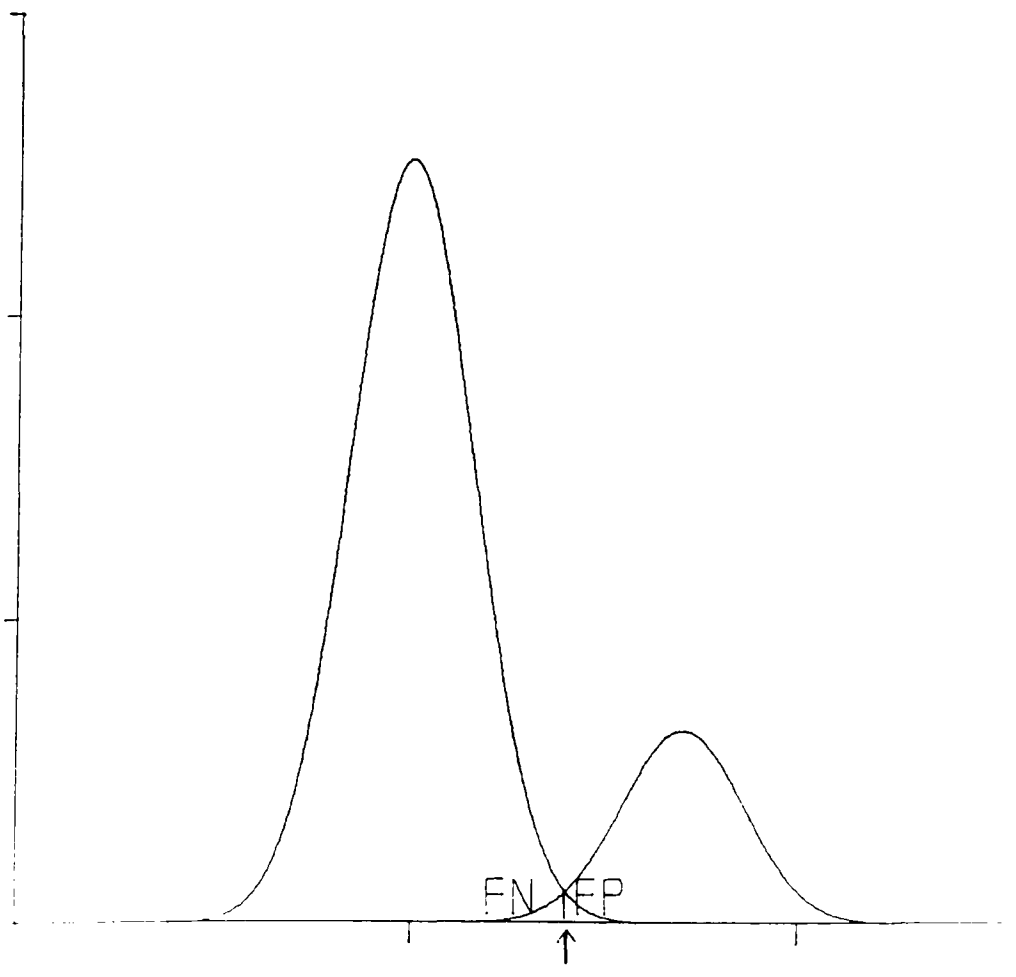


Figure 2.2 Effect on distributions of results from populations of normal and diseased individuals of doubling (A) and halving (B) the analytical variability. For explanation see Figure 2.1





These conclusions will not, however, apply to situations where there is complete separation of the groups unless deterioration in performance leads to the appearance of overlap, but such 'ideal' tests are extremely uncommon. Analogous arguments apply to the other applications of laboratory investigations, ie discrimination between several 'diseased' populations, comparison of results with externally-derived reference values, and the assessment of changes within a patient (Gräsbeck et al, 1979).

2.2.2 Analytical goals

Given the assumption that laboratory investigations do benefit patient care, one would expect both intuitively and from the analysis above that the better the analytical quality the better would be patient care. As discussed by Harris (1979) and Fraser (1983), however, analysis takes place against a background of biological variation in analyte concentrations, both within and between individuals. Thus a goal for analytical variation may be elaborated, improvement beyond which will not yield significant improvement in the variance of measured results. Conventionally (since serial monitoring of patients is the most demanding application), intra-individual variation is considered and the goal for analytical variance is taken as one quarter of the biological variance (Subcommittee on Analytical Goals, 1979):

$$(SD_{analytical})^2 < \frac{(SD_{biological})^2}{4}$$

$$\text{ie} \quad CV_{analytical} < \frac{CV_{biological}}{2}$$

Such goals may then be compared with the state of the art to determine whether improvement is necessary (eg Fraser, 1986).

Data from EQA surveys or schemes should be integral to such comparisons, which have almost without exception revealed a need for improvement beyond the current state of the art to attain these goals. For example, a study of the information available on glucose assay demonstrated generally suboptimal performance, though some of the techniques in use were capable of surpassing the goal of 2.2% CV (Fraser, 1986). Similar comparisons for a wider range of analytes, such as those summarised in Table 2.1 (Fraser CG, personal communication), confirm that this is still generally true, despite the tremendous improvements in performance since the introduction of QA into clinical chemistry.

This approach can assist in the appraisal of EQA survey data, but it must be remembered that these goals are <u>not</u> objective despite their derivation in terms of biological variance. Thus the selection of the fraction of one quarter is essentially arbitrary, dependent upon a judgement that below this level analytical variation no longer makes a significant contribution to total variance. If there are changes in the application of laboratory investigations, then these goals must be revised.

At present, however, such goals are generally not met and consequently there is a need for improvement of the reliability of most assays. Through surveys and schemes, EQA provides the means both to assess and to stimulate improvements in the quality of laboratories' analytical results and hence the reliability of patient care.

2.3 The quality of extra-laboratory assays in clinical chemistry

Extra-laboratory or 'decentralised' assays (those carried out outwith conventional laboratories by non-laboratory staff, such

Table 2.1 Comparison of analytical goals for commonly-determined analytes in serum with median within-laboratory imprecision from the Wellcome QC Programme in 1971 and 1986 (Fraser CG, personal communication; Stevens and Hjelm, 1986) Data expressed as SD at mid-point of adult reference interval, in mmol/L unless specified

	Analytical goal	Median imprecision	
		1971	1986
Sodium	0.6	2.0	1.4
Potassium	0.11	0.14	0.07
Chloride	0.8	1.9	1.6
Urea	0.30	0.71	0.25
Glucose	0.14	0.50	0.15
Calcium	0.02	0.09	0.06
Phosphate	0.04	0.11	0.04
Iron (umol/L)	2.4	3.0	1.6
Urate	0.010	0.030	0.014
Creatinine (umol/L)	2.2	15.4	6.5
Bilirubin (umol/L)	1.1	6.5	2.3
Total protein (g/L)	1.0	2.2	1.4
Cholesterol	0.16	0.39	0.12
Magnesium	0.01	0.07	0.06

as medical and nursing staff) are becoming increasingly important in patient care. Though other factors are also involved, the primary justification for such assays is to enable better clinical management through the availability of test results with greater speed and/or convenience than the laboratory could provide (Watson, 1980; Marks, 1983; Belsey et al, 1986). The scale of this activity in the UK in 1982 was revealed by a questionnaire (Browning et al, 1984), indicating that extra-laboratory assays were carried out in at least 40% of hospitals; there was collaboration with the clinical chemistry laboratory in only half of these. Such assays are likely to increase, with further devolution into community health centres and patients' homes.

Quality assurance in this situation is at least as important as within laboratories (Whitehead and Garvey, 1985), but there has been concern over the quality of such assays. Though some had endeavoured to assess quality within hospitals (eg Andrews et al, 1983; Drucker et al, 1983; Smith, 1983) or in the community (eg Petranyi et al, 1984; Burrin et al, 1985), no information was available on the situation nationally in the UK. A programme of surveys to investigate the national situation and provide the basis for rational action was therefore undertaken by the Wolfson Research Laboratories (WRL; Browning and Bullock, 1987).

2.3.1 Exploratory survey - Survey 1

It was decided that these surveys should be conducted initially through hospital laboratories. Contact with laboratories was readily available, laboratory expertise would be necessary to ensure satisfactory reconstitution of lyophilised materials, and distribution from a centralised hospital site would permit

economy of specimen usage. The disadvantage of incomplete national coverage was a secondary consideration at this stage. Sodium, potassium, glucose and bilirubin, the most widely-determined serum analytes, were selected for survey. This survey (Appendix I.5.3) therefore comprised distribution of a specimen of lyophilised bovine serum to the 210 laboratories responding to the 1982 questionnaire (Browning et al, 1984). Recipients were asked to reconstitute the specimen and provide aliquots for assay to extra-laboratory sites.

The results returned were then assessed by comparison with those obtained on distribution of the same serum though the UKEQAS for General Clinical Chemistry. This (Table 2.2) demonstrated firstly much greater variability of the extra-laboratory sodium and potassium results, with performance apparently about 50% worse than that for laboratory assays. There were few results, however, and further study on a larger scale would be required to confirm this conclusion.

The variance of extra-laboratory glucose assays appeared substantially worse. When assessed by CV and by average VIS, it was at least double that among laboratories, and the range of results obtained appeared clinically significant. The small number of results in each group made more detailed assessment according to the manufacturer and model of reflectance meter less reliable, but no major differences in performance were apparent. A national survey of glucose assay thus appeared essential.

The results also confirmed the unsuitability of bovine serum for spectrophotometric bilirubin assays, due to the presence of interfering chromophores. Survey of bilirubin was therefore discontinued, to be combined with assessment of paediatric

Table 2.2 Comparison of the results from the first WRL survey of extra-laboratory assays, October 1985, with UKEQAS distribution of the same lyophilised serum. All data recalculated after exclusion of results more than 2SD from the untrimmed mean; VISs calculated relative to the method mean.

	n	Mean	CV	Range	Average VIS
Sodium (mmol/L)					
Extra-lab	34	152.4	2.0%	147 - 164	88
UKEQAS	429	150.7	1.1%	144 - 156	57
Potassium (mmol/L)					
Extra-lab	45	7.26	3.4%	2.0 - 7.8	91
UKEQAS	438	7.23	1.9%	6.7 - 8.8	55
Glucose (mmol/L)					
Extra-lab	152	19.03	12.3%	5.6 - 23.0	183
UKEQAS	434	17.15	3.4%	14.9 - 20.4	40
Bilirubin (umol/L)					
Extra-lab*	33	123.3	17.2%	67 - 320	80
UKEQAS	380	92.9	7.3%	30 - 154	37
*Diazo	4	86.8	21.3%	67 - 110	
*Spectro	27	128.2	12.7%	81 - 320	89

bilirubin assays within laboratories (see section 2.4.3 below).

2.3.2 First national survey - Survey 2

Survey 2 was designed to assess the applicability of these findings on a national scale. Specimens were therefore sent to all UKEQAS participants; the other aspects were as in Survey 1 (Appendix I.5.3).

The variance of sodium and potassium assays (Table 2.3) was confirmed as being 1.5 times that seen in the UKEQAS for laboratory assays. Though this is a cause for concern, the ranges of results obtained did not appear likely to cause major errors in the clinical management of patients. Thus these analytes could no longer be considered as priorities for survey.

A similar relationship held for glucose using 'laboratory' instruments such as glucose oxidase (GOD)/oxygen electrode procedures (Table 2.3). When glucose assays using reagent 'sticks' and reflectance meters were considered, however, the situation was again much worse, the average scores being about four times those obtained by UKEQAS laboratories on the same specimen. The ranges of results returned also confirmed a potential for gross errors in clinical management if diagnoses or therapy were based on these results. This is emphasised by the distributions of extra-laboratory and UKEQAS results shown in Figure 2.3.

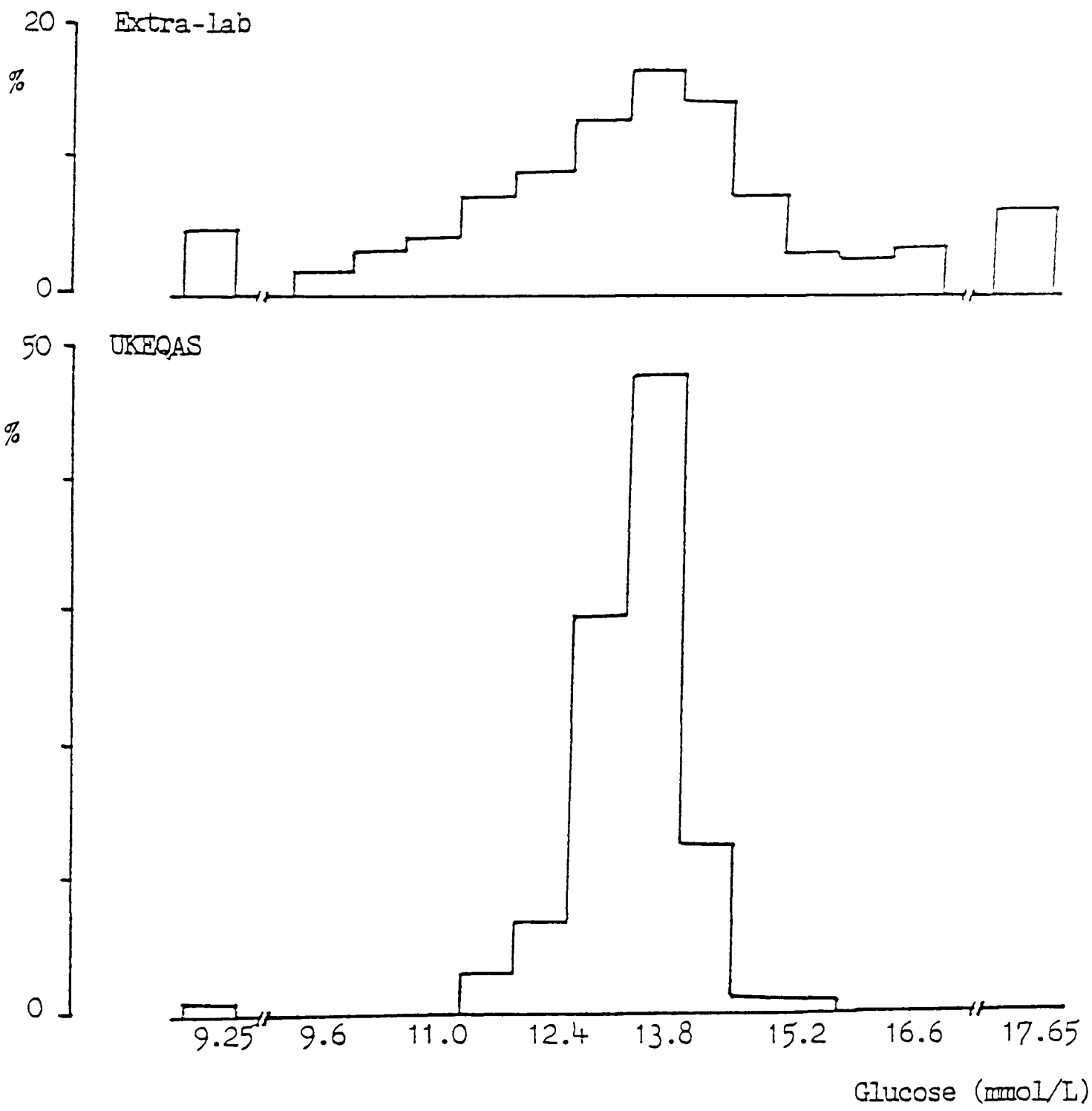
Glucose assay under extra-laboratory conditions thus appeared to be in need of greatest improvement and of further survey. The two initial surveys were susceptible to several criticisms:

- the specimens were of bovine rather than human origin

Table 2.3 Comparison of the results from the second WRL survey of extra-laboratory assays, April 1986, with UKEQAS distribution of the same lyophilised serum. All data recalculated after exclusion of results more than 2SD from the untrimmed mean; VISs calculated relative to the method mean.

	n	Mean	CV	Range	Average VIS
Sodium (mmol/L)					
Extra-lab	69	152.8	1.6%	146 - 166	85
UKEQAS	412	151.3	1.2%	141 - 165	62
Potassium (mmol/L)					
Extra-lab	71	6.17	3.0%	4.7 - 7.2	91
UKEQAS	422	6.15	1.9%	4.5 - 7.8	57
Glucose (mmol/L)					
Extra-lab*	492	13.6	14.6%	1.5 - 22.0	150
UKEQAS	418	13.4	3.6%	6.2 - 16.1	41
*GOD	31	12.7	6.6%	9.4 - 14.6	70

Figure 2.3 Comparison of glucose results in the second WRL survey of extra-laboratory assays, April 1986, with those from UKEQAS distribution of the same serum



- the lyophilised specimens were reconstituted in and distributed from laboratories, casting doubt on specimen stability
- the specimens were of serum rather than whole blood, for which the procedures are designed
- performance was assessed at a single level only, of uncertain clinical relevance
- the coverage was unlikely to be complete, due to distribution through laboratories and to self-selection of sites returning results

Though the last potential objection must be accepted in any such survey, it proved possible to resolve all others apart from the first for glucose assay by using a newly-available material in Survey 3.

2.3.3 Second national survey - Survey 3

Survey 3 (Appendix I.5.3) therefore used two liquid blood-based specimens ("Sugar-Chex") at clinically high and low levels. Other aspects were as before, though no UKEQAS results were available for comparison. Studies by the suppliers confirmed this material to be stable, have negligible vial-to-vial variability and show minimal effect of mixing efficiency (some participants reported difficulty in resuspending settled erythrocytes).

The results of this survey (Table 2.4) confirmed for glucose the findings of the earlier surveys. The ranges of results for the two specimens were clinically significant (Figure 2.4), interpretation of the extreme results of 2.1 and 22.2 mmol/L obtained for specimen 2 being obviously different.

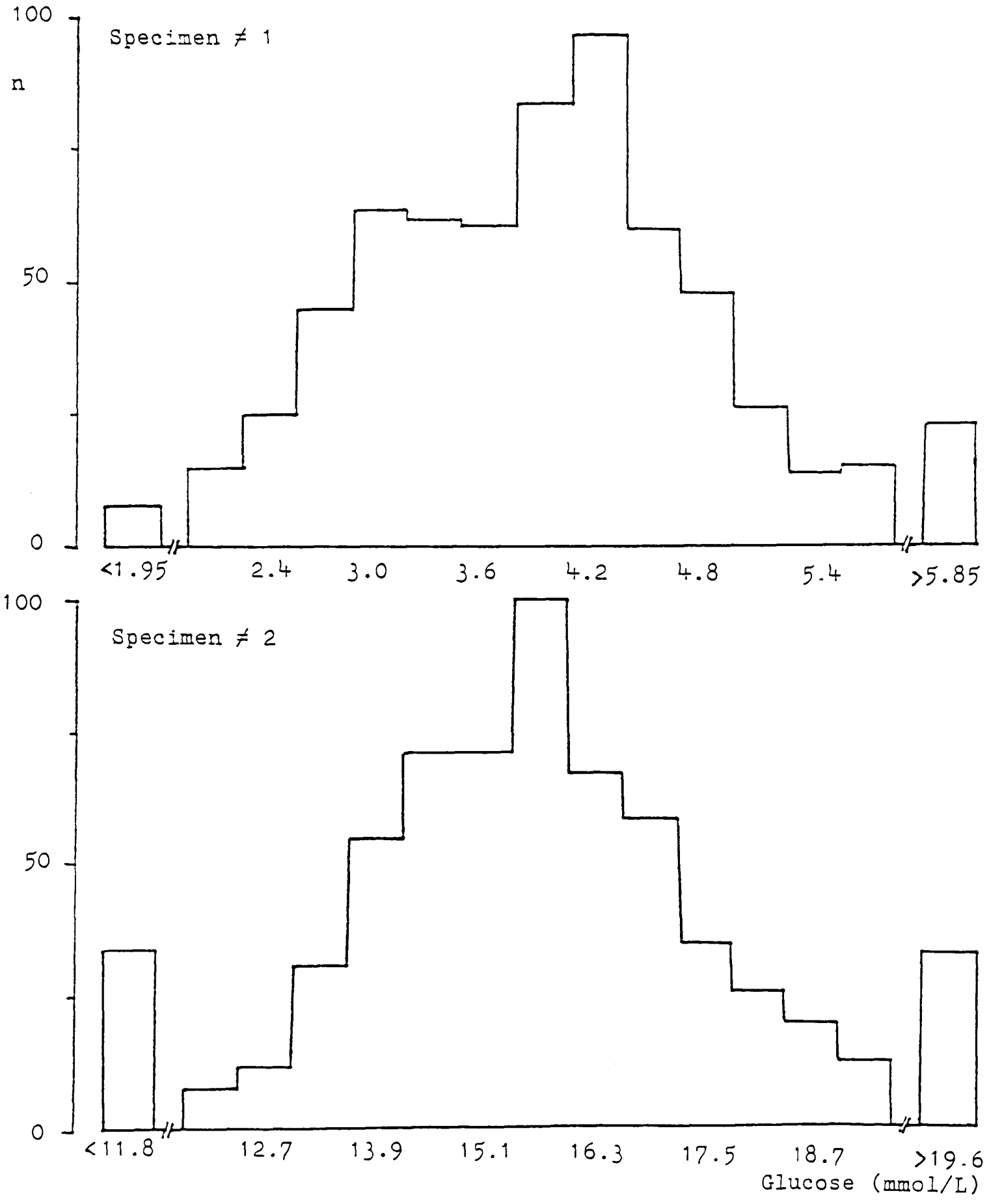
Glucose results obtained by visual reading of reagent sticks had

Table 2.4 Results from the third WRL survey of extra-laboratory assays, November 1986. All data recalculated after exclusion of results more than 2SD from the untrimmed mean; VISs calculated relative to the method mean.

	n	Mean	CV	Range	Average VIS
Glucose (mmol/L) - results from quantitative assay					
Specimen 1	623	3.84	22.4%	1.1 - 10.3	196
Specimen 2	592	15.74	11.5%	2.1 - 22.2	128
Glucose (mmol/L) - results by visual reading					
Specimen 1	192	4.68	22.1%	1 - 17	157
Specimen 2	162	16.40	8.2%	5 - 44	111

	n	Average VIS
Overall	646	162
GOD 'automated'	24	88
BCL meter	350	136
Ames meter	178	192
Hypocount meter	64	216
Other meter	30	237

Figure 2.4 Distributions of glucose results in the third WRL survey of extra-laboratory assays, November 1986



been reported by some participants in Survey 2. These were, as expected, even more variable than those from reflectance meters: the trimmed CV for 170 results was 26.7%, with a range of 2.2-44 mmol/L. In Survey 3 such results were also requested, and 192 were received. The performance (trimmed CVs 2.1 and 8.2%, average VISs 157 and 111), however, appeared similar to that using reflectance meters though the ranges of results (1-17 and 5-44 mmol/L) were greater.

2.3.4 Conclusions from the surveys

These surveys demonstrate clearly that the standard of performance of extra-laboratory assays in the UK in 1986 was not fully satisfactory. Variability appeared to be 50% greater than that of laboratory assays for sodium and potassium and for glucose using laboratory instruments, presumably reflecting the lesser experience of the operators and their lower awareness of quality assurance. The variability was about three times greater for glucose using reflectance meters, with potentially adverse clinical consequences of diagnoses or therapy based on such results. The information on visually-read glucose assays was not fully consistent, though variability appeared to be greater than for assays using meters.

How reliable are these conclusions? Survey 1 was intended as a feasibility study only, and the primary outcome was to indicate that the design could be applied on a larger scale. It also suggested that problems did exist with extra-laboratory assays, and that bovine serum was not useful for survey of bilirubin assays.

Survey 2 confirmed the existence of greater variance in extra-laboratory assays than that seen in the UKEQAS for laboratory

determinations. The variability for glucose was by far the greatest, indicating this analyte as the priority for further survey, whereas that for sodium and potassium was of doubtful clinical significance.

As described above (section 2.3.2), the design used in Surveys 1 and 2 was susceptible to several criticisms. Most were obviated by the amended design chosen for Survey 3, of glucose alone. Thus two specimens were used, at levels with clinically significance; they were stable and did not need laboratory reconstitution. The specimens were also blood-based, so their performance should have more closely reflected that obtained for clinical specimens as well appearing more realistic to the operators. The species of origin was again bovine, but differences in behaviour between materials including bovine and human erythrocytes are most unlikely to be sufficiently great to influence the conclusions drawn from Survey 3.

The final point cannot be answered at this time, since there is no centralised source of information on sites where extra-laboratory equipment is used. Indeed, many laboratories are unaware of the situation within their own hospital. The surveys may thus not have been fully representative of the situation throughout the UK: the coverage was almost certainly incomplete, and the participants all had some contact with laboratories. The information available indicates that collaboration with laboratory staff can improve performance (Smith, 1983; Whitehead and Garvey, 1985), so it is unlikely that these surveys overestimated the true variability.

The data on visual glucose readings from Survey 3 illustrate an

important point. The agreement, apparently better than that using meters, may be explained by the high proportion of results at a single value (64% and 70% of 4.5 and 17.0 mmol/L respectively), grossly distorting the distribution from Gaussian and making unreliable any judgements from parametric estimators of dispersion such as the CVs. The average VISs were also distorted, due primarily to the effect on the method mean (the average scores increased to 262 and 151 if the reflectance meter means were used as designated values).

Thus the conclusions from these surveys appear to be valid in indicating that the performance of at least glucose assays requires urgent improvement if extra-laboratory assays are to benefit patient care.

What steps should be taken to bring about such improvement? One obvious and often-repeated suggestion is to establish an NEQAS to cover these assays. There are practical problems in implementing such a scheme, relating mainly to ascertainment of participant sites, to selection of suitable and stable specimens, and not least to the sheer scale of the undertaking. Moreover, experience suggests this would be unlikely to stimulate great improvement in the absence of adequate internal quality control procedures and of quality assurance in general. The laboratory is the obvious source for the necessary expertise, and collaboration with extra-laboratory sites appears essential in the interests of safe and effective patient care. This will constitute a major challenge for clinical chemists over the coming years, and continuing EQA surveys offer the most convenient means to monitor the situation.

2.4 Laboratory examples

Other examples drawn from laboratory assays demonstrate

additional aspects of the application of EQA surveys in assessing the need for institution of an EQAS.

2.4.1 Salicylate and paracetamol

Assays of salicylate and paracetamol in serum are of great clinical importance in the diagnosis and treatment of suspected overdose, and accurate and precise determinations are essential for reliable patient care. Many of the methods in use, however, were considered to be imprecise or non-specific, as suggested by the two local EQASs in the UK (Epton, 1979; Wiener, 1980), requests for participation in which were increasingly coming from outside the respective Regions. Thus there appeared to be both a demand for regular EQA of these determinations and evidence that this was necessary, and UKEQAS surveys were therefore initiated in 1983, building on the experience of the Regional schemes.

The first survey used three specimens of lyophilised human serum to which pure drugs had been added (Appendix I.2.6). To avoid potential complications due to salicylate interference in some paracetamol assays each specimen contained only one drug (Wiener, 1980). The results (Table 2.5) showed unsatisfactory agreement. Many participants had, however, complained that the specimens were turbid and that 1 mL of serum was insufficient for assay.

Further surveys were thus necessary to confirm the conclusions. These specimens were liquid to reduce the turbidity, based on sterile equine serum (since no matrix effects due to species were expected) to which pure drugs were added, and provided in larger volume. Gentamicin was used as an antibacterial agent since azide causes spectral interference in the Trinder procedure for salicylate assay (Davies KW, personal communication). The results

Table 2.5 Interlaboratory agreement in the first three UKEQAS surveys of salicylate and paracetamol assay, November 1983 - August 1984

	n	Mean (mg/L)	CV (%)	Recovery (%)
Salicylate				
Survey 1	272	379.3	11.1	-
Survey 2	271	350.3	7.2	101.5
		489.0	7.1	101.2
Survey 3	268	556.3	6.7	101.1
Paracetamol				
Survey 1	256	118.7	16.1	-
		161.4	10.5	-
Survey 2	259	133.9	10.6	98.5
Survey 3	259	176.5	8.3	98.1
		61.5	22.6	102.5

(Table 2.5) confirmed the need for improvement, and a UKEQAS was therefore established. This continued with a similar design (Appendix I.2.6), though specimens later contained both drugs and a further change to single-specimen distributions was made. Data from the scheme on method performance are given in Chapter 8 and on performance improvements in Chapter 9.

2.4.2 Specific proteins in serum

Measurement of proteins became more widespread during the 1970s, with increased use of nephelometric and turbidimetric assays and less reliance on radial immunodiffusion (RID) procedures. This expansion, coupled with intermethod differences in values obtained for calibrants and the improved between-laboratory agreement with regular EQA documented in the USA (eg Taylor and Fulford, 1981) suggested that an NEQAS for the UK could be beneficial. Surveys were therefore undertaken, starting in 1980. Immunoglobulins G, A and M (IgG, IgA and IgM) were selected for the initial surveys (Appendix I.2.5) primarily to validate the specimens (comprising liquid human serum), since wide variations were not expected. The results, however, showed considerable variation (Table 2.6), much of which appeared to be related to the calibration materials used (Chambers et al, 1984). Further surveys were therefore undertaken, with a method grouping including classification according to calibrant.

In the initial stages distributions comprised multiple specimens, including pairs related by dilution which enabled assessment of intralaboratory precision (Chambers et al, 1984). Additional studies (see section 13.2; Chambers et al, 1987) examined the potential improvement which might be gained from use of a common calibration material. Later distributions were reduced to a

Table 2.6 Average interlaboratory agreement for immunoglobulins for the four specimens in the first UKEQAS survey of specific protein assays, September 1980. All data recalculated after exclusion of results more than 2SD from the untrimmed mean.

	n	Average CV (%)		
		IgG	IgA	IgM
Electroimmunoassay	4	31.9	14.7	36.4
AIP	12	7.6	8.7	14.9
Turbidimetry	20	11.4	17.4	17.4
Nephelometry (Hyland calibrant)	19	12.7	23.0	12.5
Nephelometry (other calibrant)	9	12.7	14.2	8.4
RID (Hyland calibrant)	5	10.5	17.6	32.5
RID (other calibrant)	118	13.6	16.0	15.5

single specimen as the UKEQAS became established and Variance Index (VI) scoring was introduced; an increasing number of proteins were incorporated during the scheme's evolution.

2.4.3 Serum bilirubin in paediatrics

The assay of bilirubin is critical in the clinical management of neonatal jaundice, with therapy based primarily on the assay results (Mollison and Cutbush, 1954; Isherwood and Fletcher, 1985). The results of overseas surveys (Schreiner and Glick, 1982; Watkinson et al, 1982; Blijenberg et al, 1984), supported by anecdotal evidence from Regional surveys within the UK, indicated between-laboratory and between-method differences. Surveys to ascertain the national state of the art in the UK therefore appeared essential, and were instituted in 1984.

The survey used a commercial lyophilised human-based QCM, since use of pooled clinical specimens on a national scale was infeasible; such QCMs are used widely for IQC with no apparent problems. All UK laboratories received the specimen, and only results obtained within laboratories were processed. These (Table 2.7) showed that there was variability among laboratories, and confirmed the existence of accuracy differences between diazotisation and direct spectrophotometric methods.

Comparability was thus not fully satisfactory, and further surveys to investigate this in more detail and to stimulate improvement would be justified. The situation was, however, not as poor as had been anticipated and this activity was therefore allocated a lower priority. An additional factor delaying further surveys was the need (DHSS, 1986a) to ensure that the human-based materials to be distributed were negative for

Table 2.7 Interlaboratory agreement in the first UKEQAS survey of paediatric bilirubin assay, April 1984

	n	Mean (umol/L)	CV (%)
Overall	315	376.7	8.3
Diazotisation methods:			
Caffeine/benzoate	62	387.7	6.0
Diphylline	9	374.7	5.8
Other accelerator	26	392.7	8.3
Dichlorophenyldiazo salt	20	385.2	7.1
Spectrophotometric methods:			
Bilirubinometer - direct	86	359.0	7.9
Bilirubinometer - diluted	17	377.8	8.1
Dilution & spectrophotometry	78	379.2	8.4

antibody to human immunodeficiency virus (HIV) at the individual donation stage. These surveys would also attempt to assess the quality of extra-laboratory assays (section 2.3.1 above).

2.4.4 Urinary total protein

Urinary protein assays are important in the diagnosis and monitoring of renal disease, and form part of the medical examination of apparently healthy subjects for insurance purposes. Thus reliability is essential, especially at low protein concentrations. Published data from the USA and Australia (Glenn, 1980; Shephard et al, 1983) and from UK Regional surveys (Legg and Hurrell, 1984) indicated that between-laboratory agreement was poor, and a national survey in the UK was therefore undertaken in 1985. The same published data suggested that the situation for other common urinary assays (glucose, electrolytes, urea and creatinine) was relatively satisfactory, so total protein only was surveyed.

The initial survey (Appendix I.5.2) confirmed that variance was high (Table 2.8), indicating the need for further activity. Institution of a full EQAS was felt, however, not to be essential, and a series of surveys was planned. These confirmed that the most widely-used method, turbidimetry with sulphosalicylic acid (SSA), appeared to give inferior performance, and that several laboratories continued to use the unsatisfactory direct biuret procedure. The surveys included more than one specimen and enabled study of the effects of using a common calibration material, discussed fully in Chapter 13. These studies (see section 13.4) suggested, however, that problems with the materials distributed might have contributed to the variability seen, and caution should therefore be exercised in

Table 2.8 Interlaboratory agreement in the first UKEQAS survey of urinary total protein assay, April 1985

		Specimen		
		1	2	3
Overall	n	326	320	335
	Mean (g/L)	5.14	6.67	0.27
	CV	38.9%	22.8%	57.1%
Turbidimetry - SSA		7.52	6.84	0.22
		41.7%	29.7%	53.6%
Turbidimetry - SSA/Na_2SO_4		4.36	6.22	0.23
		27.6%	21.5%	32.2%
Turbidimetry - TCA		4.24	7.37	0.29
		19.7%	18.9%	30.6%
Precipitation & dye binding		4.59	6.23	0.27
		16.6%	21.3%	21.0%
Precipitation & biuret		4.56	6.86	0.29
		19.1%	16.5%	42.6%
Direct biuret		4.55	7.45	0.81
		17.0%	10.1%	71.9%
Direct Coomassie blue		4.80	6.17	0.33
		16.0%	19.7%	91.7%
Benzethonium chloride		4.29	6.52	0.28
		23.8%	12.7%	15.7%

interpreting the survey results.

Another important performance characteristic studied through these surveys was the detection limit. Two specimens without detectable protein content were distributed, one of urea and sodium and potassium chlorides in distilled water and the other a normal urine. For both specimens many laboratories reported a finite protein content (Table 2.9); only a few represented transposition errors or results reported in mg/L. When results (both quantitative results and those reported as below the detection limit) were analysed according to participants' detection limit (Table 2.9), an appreciable number were shown to be inappropriate when considered in relation to the generally-accepted reference interval of 0.15 g/24h, roughly equivalent to 0.1 g/L.

2.4.5 Diagnosis of phenylketonuria

Phenylketonuria (PKU; Knox, 1972) is an inherited metabolic disease, the effects of which on mental development can be minimised by early diagnosis and dietary restriction. The UK incidence makes screening of the newborn population cost-effective, and such a programme has operated since the 1960s (Medical Research Council, 1981). A variety of method principles are in use for the assay of phenylalanine in blood specimens, and there was evidence that cases had been missed due to analytical as well as administrative errors (eg Smith, 1985; Holtzman et al, 1986). Such failures undermine the efficacy of the screening programme, and UKEQAS surveys were undertaken at the request of the DHSS/Medical Research Council Steering Committee for the PKU Register.

One survey was of the screening procedures (using dried blood

Table 2.9 Quantitative results and detection limits reported for specimens 8 (urea and salts in distilled water) and 9 (normal human urine) in UKEQAS urinary total protein surveys 3 and 4, June and November 1986. Means recalculated after exclusion of results more than 2SD from the untrimmed mean.

	Specimen	
	8	9
Quantitative results:		
n	99	145
Mean (g/L)	0.092	0.083
Range (g/L)	0.002 - 12.23	0.01 - 2.96
Results >0.1 g/L	22	23
Detection limits:		
0.5	2	-
0.3	-	3
0.2	5	6
0.1	58	63
0.05	27	33
0.02	-	4
0.01	8	4
0.005	-	2

Table 2.10 Interlaboratory agreement in UKEQAS surveys of phenylketonuria screening and quantitative phenylalanine assay, June and November 1978. All data recalculated after exclusion of results more than 2SD from the untrimmed mean.

	n	Mean (umol/L)	CV (%)
PKU screening:			
Overall	18	256	31.1
Guthrie	9	225	36.3
Fluorimetry	5	291	24.2
Chromatography	4	283	26.7
Phenylalanine assay:			
Overall, excluding Guthrie	44	483	11.2
Guthrie	8	289	31.5
Fluorimetry	24	488	12.6
Aminoacid analyser	4	497	4.7
Other method	9	495	16.9

spots or liquid whole blood specimens) and one of quantitative phenylalanine assay (using lyophilised bovine serum). The results (Table 2.10) showed the interlaboratory agreement of phenylalanine assay to be relatively satisfactory at about 11% CV, with even better agreement being obtained within the more reliable method groups. In contrast, the variability of blood phenylalanine results by the screening procedures was very poor at around 20-30% CV.

Priority for further EQA activity clearly lay with the screening procedures, since it would be of little benefit to improve slightly further the reliability of diagnostic and monitoring assays if the initial screening process was unreliable. Further surveys of screening assays then evolved into a UKEQAS, as described in Chapter 7, though the emphasis was placed on consideration of the action taken in response to screening of the specimen rather than the quantitative results obtained by these screening procedures.

2.5 Summary

In almost all cases patient care can be improved through improvement of the quality of analytical performance. The rare exceptions are when the test is so efficient that there is complete separation of the populations of interest, and when performance is already so good that analytical variance is negligible relative to biological variation.

EQA surveys provide an assessment of the state of the art for the analyte of interest, on a national or local basis and without necessarily committing the organisers to establishing a regular EQAS. This information is invaluable in determining the most

appropriate further action, and in conjunction with the clinical importance of the assays in indicating the priorities for such action.

If the situation proves relatively satisfactory an EQAS may be unnecessary, but otherwise establishment of an EQAS would be indicated. A grossly unsatisfactory situation may require other measures, such as method standardisation or the introduction of adequate QA and IQC procedures, before an EQAS might be expected to improve agreement; continued surveys would, however, enable monitoring of overall performance.

Appropriately-designed surveys will also yield valuable information on individual aspects of assay performance, eg within-laboratory imprecision, assay detection limits and the efficacy of calibration procedures. Such studies can be incorporated into an EQAS.

Chapter 3:

FUNDAMENTAL REQUIREMENTS OF SCHEME AND SURVEY DESIGN

3.1 Introduction

The aim of an EQAS is to provide participants with a reliable and objective reflection of their laboratory's routine performance. Reports from the scheme will then indicate any need for improvement, and ideally identify where the deficiencies lie and provide assistance in remedying them. EQASs can only be successful in influencing clinical chemists' actions when the participants have confidence in the design and operation of the scheme.

Three major elements of scheme design are important in providing this confidence:

- organisation of the scheme
- validity of the specimens
- assessment of performance

There is no consensus on the ideal design, though a number of the issues involved were addressed by the WHO Working Group (WHO, 1981) and guidelines were suggested. The many EQASs in operation embody different choices in design elements, and there has been no scientific study of many factors. Indeed no truly objective study of the value of EQA in improving laboratory performance is possible: all evidence on these points is circumstantial, and much is anecdotal.

3.2 Organisation of the scheme

This element of confidence relates not only to the structural design of the scheme, predominantly in terms of distribution

schedules, but also to the administrative procedures.

3.2.1 Scheme administration

It is essential for participants to perceive the organising centre as efficient as well as scientifically competent. Procedures must be evolved to ensure that, for instance, results are processed promptly and accurately, reports are correctly addressed, and queries and complaints are dealt with courteously and within a reasonable time. Any errors arising within the organising centre must be acknowledged and rapidly corrected.

All these factors contribute to the scheme's public image and may be likened to the general sense of quality assurance, ie to 'good EQA practice'.

3.2.2 Frequency of distributions

A minimum amount of information, with a certain degree of confidence, is required before anyone will give credence to conclusions drawn from the data and act upon these conclusions. With a constant number of specimens in each distribution (see section 3.2.3 below), more frequent distribution will provide this minimum more quickly.

It is therefore widely held (eg Whitehead et al, 1973; WHO, 1981) that distributions must be frequent for an EQAS to be effective, as was suggested first by Shuey and Cebel (1949). The minimum amount of information required will vary among participants, due in part to their varying degrees of confidence, but the overall conclusion remains.

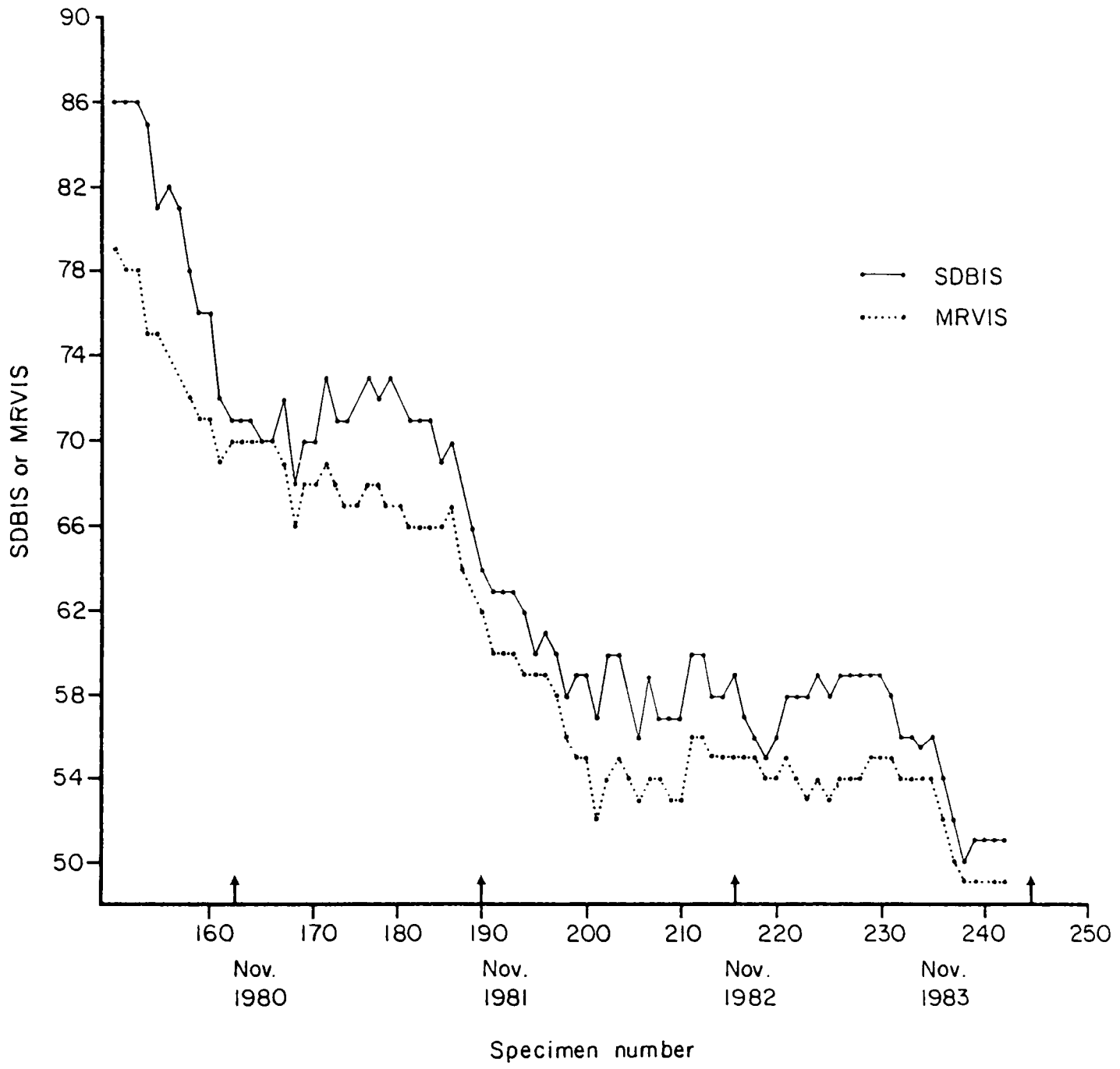
The information considered must, however, be current; no-one should take action now on the basis of performance data obtained

in past years. The more frequent the distributions, the more recent is the period to which the information pertains. There is also greater confidence that the data reflect current performance, since the two factors are intimately and probably inextricably linked.

Consider, for instance, a clinical chemist receiving EQAS reports, the first of which shows a positive bias of 7%. No action would be taken on this single figure, which may be subject to many influences. The second report shows a bias of +4%; the clinical chemist is relieved, assuming this to reflect improvement. The third, however, shows a positive bias of 11%. The average bias of +7.3% (or the change in bias of 7%), taken in conjunction with IQC data, may then be considered significant and investigative or remedial action be taken. With a two-weekly distribution schedule only one month would have elapsed, whereas with six-monthly distributions it would take one year to reach this point; confidence in the data could also have been so eroded that no action might be taken even then.

Though this proposition has not been subjected to formal study, incidental findings in the UKEQAS for Urinary Pregnancy Oestrogens are relevant. The scheme design was changed in 1980, when organisation of this scheme was transferred to WRL (Bullock and Wilde, 1985; Appendix I.2.4). There had been little evidence of improved interlaboratory agreement over the scheme's six-year previous operation with three-weekly distributions (Oakey, 1980). Following a change to two-weekly distribution of the same type of specimens, agreement improved dramatically as reflected by the average MRVIS and SDBIS for all participants (Figure 3.1). Though other aspects, discussed below, were also changed (reports

Figure 3.1 Average performance (MRVIS and SDBIS) for all participants in the UKEQAS for Urinary Pregnancy Oestrogens, 1980-1984



now gave a much more detailed analysis of participants' performance; the VI system with running scores updated each distribution replaced a scoring system applied only 9-monthly and based solely on within-laboratory reproducibility), it is likely that the increased frequency at least contributed to the improvement. Later this very success, with the continuing decrease in the number of laboratories carrying out the assay, led to redeployment of resources by reducing the distribution frequency to four-weekly.

Though frequent distributions are necessary to encourage action by participants, distributions may also be too frequent and lead some participants to feel that EQA on such a frequent basis can obviate, or at least reduce, the need for IQC. By ignoring the retrospective nature of EQA, and effectively removing the control over each analytical batch normally provided by IQC, this will lead to a loss of reliability and hence worse patient care. 'Hybrid' IQC/EQA schemes, whether commercially or professionally organised, offer an extreme example of this potential confusion. These (eg Limonard, 1979; Jansen and Jansen, 1980; Lawson et al, 1980) use the same material to provide both an IQC programme and a retrospective element of EQA; they also suffer from unconscious bias due to continual use of the same specimen, and are most appropriately classed as an IQC procedure.

In addition, it can be argued that EQA would be more cost-effective if an interval is left between distributions to allow participants time to take any remedial action necessary; this is a feature of, for example, Whitehead's 'intensive' scheme design, discussed in section 3.2.3 below.

3.2.3 Number of specimens in each distribution

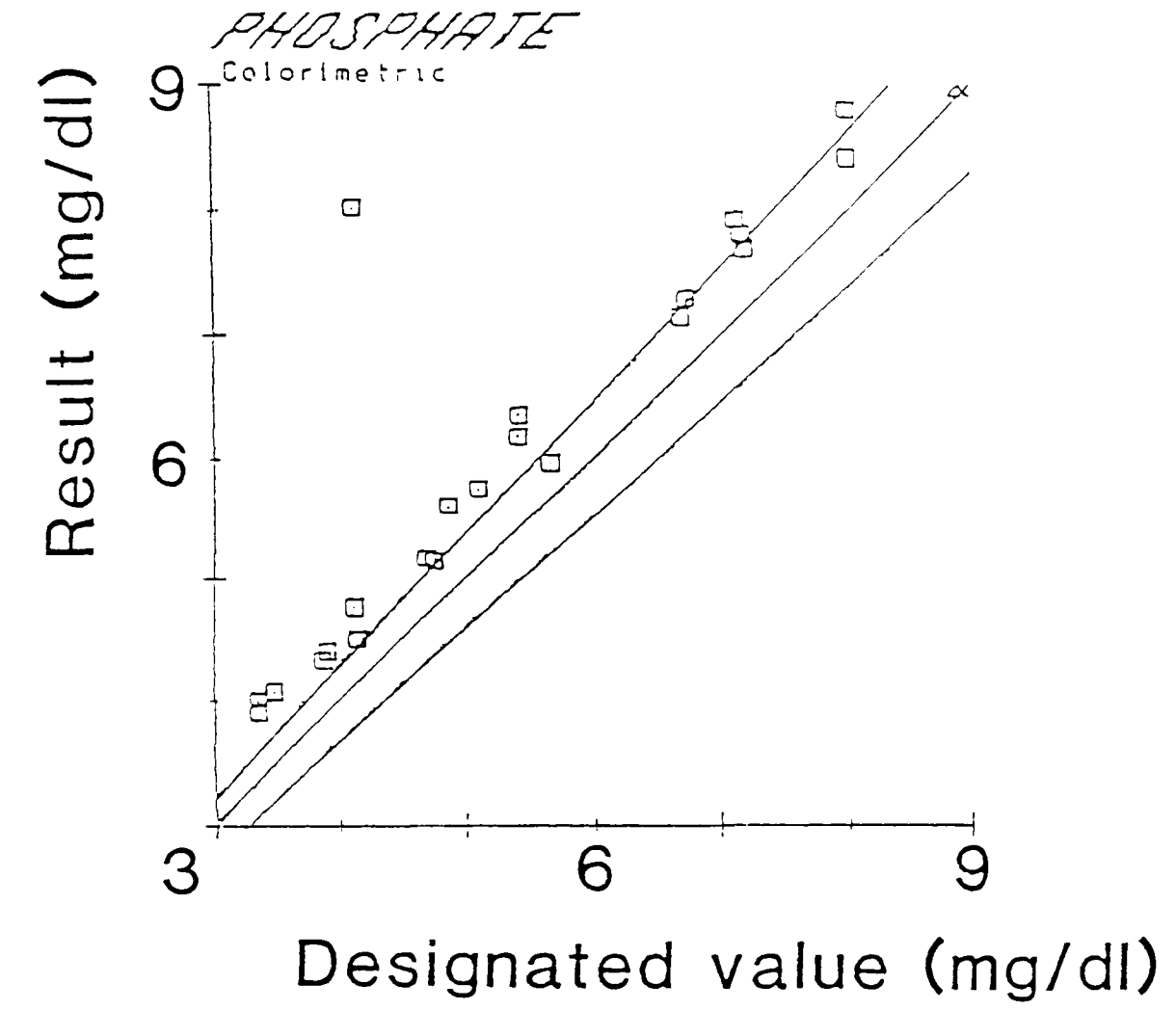
Use of a single specimen is the most widely-used and most economical approach. Distributions can be more frequent at the same specimen cost; scheme administration, in terms of specimen despatch, data processing and report despatch, is also simplified. There is no possibility of the EQAS specimens (or their results) being transposed, making the performance assessment slightly more reliable.

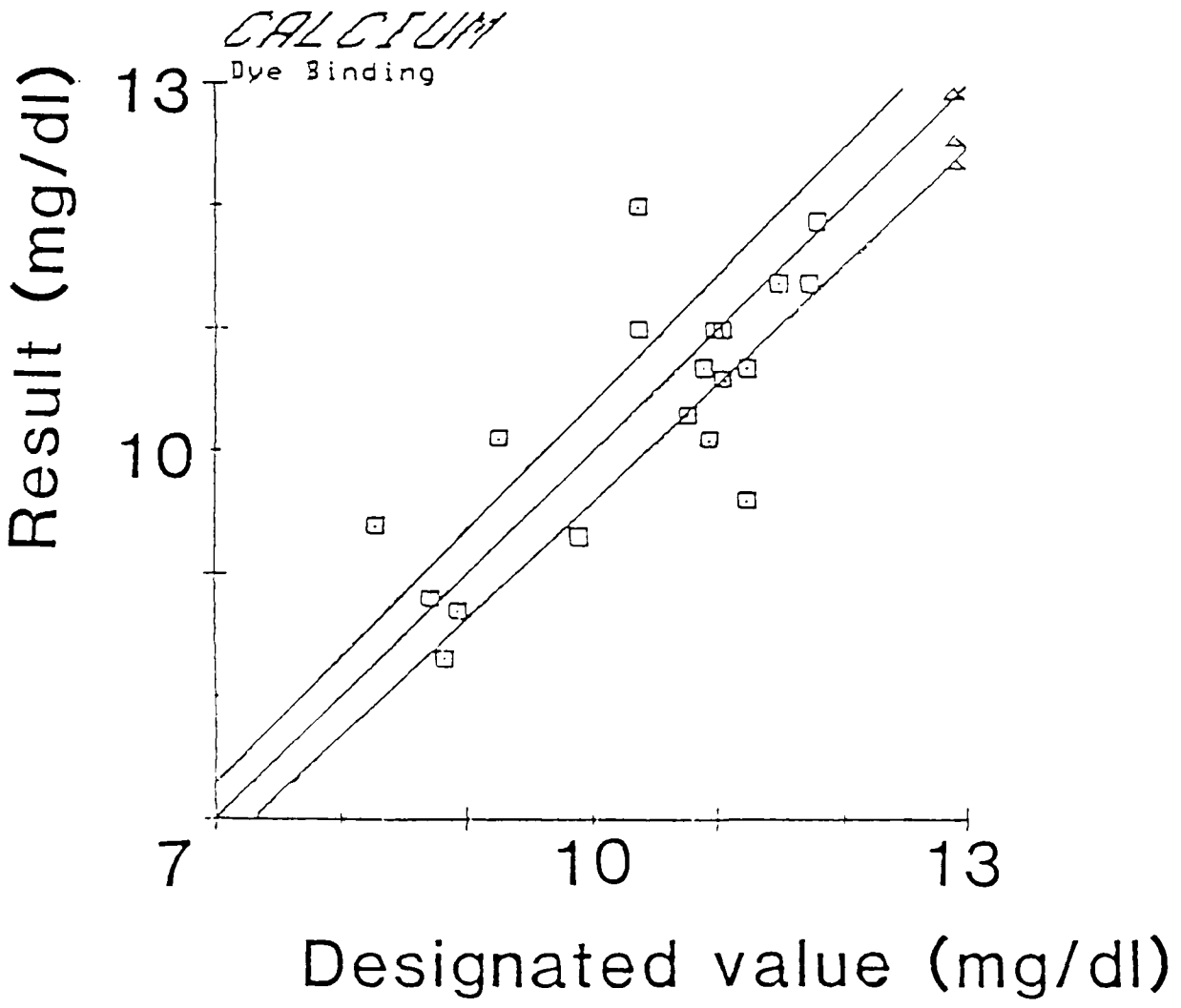
The apparent disadvantages (see below) of using a single specimen can effectively be removed by cumulation of data over a period. For example, the relationship between a laboratory's performance and analyte level can be examined by plotting their results against the designated (target) values for a number of distributions (Whitehead, 1977; Lever et al, 1981; Bullock and Wilde, 1985).

Figure 3.2 shows such cumulations for two participants in the Middle East EQAS (MEEQAS); from this the laboratories' positive bias for phosphate and imprecision for calcium are clear. One potential disadvantage is that performance may have changed over the period covered, and it is therefore essential that the temporal relationship is also examined; the issues are discussed in more detail in Chapter 10.

Distribution of a pair of specimens (eg, Stamm, 1975; Jansen et al, 1977) enables a rudimentary assessment of bias and imprecision at each distribution. This has some advantage in using more data to yield a more reliable estimate of participants' performance, important in making judgements in licensing schemes such as that in GFR (Bundesärztekammer, 1971). The Dutch scheme (Jansen et al, 1977), however, discards much of

Figure 3.2 Graphs for phosphate and calcium of laboratory result against designated value for two participants in the Middle East EQAS.





this by scoring performance solely on the worse of the two results for each analyte.

Any assessment of bias and imprecision based on two points is, however, inherently unreliable. It cannot for instance distinguish between imprecision and a linear relationship removed from or with slope dissimilar to that of the line of identity (Lever et al, 1981). Recovery studies, with distribution of spiked and unspiked specimens, can also be undertaken, which is useful in validating consensus means as well as in assessing individual participants' performance (Hunter and McKenzie, 1979).

Paired specimens may also be applied in 'ratio reporting' for interlaboratory assessment of analytes for which many numerically discordant methods are in use. A historical example is assay of enzyme activities, eg acid phosphatase (Rosalki, 1972), enabling the performance of different methods to be assessed. This approach, however, is based on intralaboratory imprecision only since any contribution from biases among laboratories is eliminated in taking the ratio. The procedure thus reduces to an interlaboratory assessment of a performance characteristic better dealt with through IQC. A more reliable estimate of intralaboratory precision may, if necessary, be derived from distribution of replicates or a pair of specimens related by dilution (eg Chambers et al, 1984) since these should be commutable.

A more useful application is in studies of calibration (eg Jansen and Jansen, 1983; Bullock and Wilde, 1985; Bullock et al, 1986b), discussed fully in Chapter 13. In essence the distribution of two specimens for assay together enables assessment of the

performance that would be obtained if all participants used a common calibration material. Such use is only occasional, and would not be a regular element of scheme design.

Simultaneous distribution of a set of specimens increases the rate at which performance data can be accumulated (Hunter and McKenzie, 1979; Groom, 1985b), with the specimens treated as independent distributions. These sets may include recovery and calibration studies, as described above for paired specimens.

If the specimens within a set are related in a known manner, however, more reliable judgements become feasible (Caragher and Grannis, 1978). Such linearly-related specimens have been used in CAP surveys (Grannis and Miller, 1976; Caragher and Grannis, 1978), and on a regular basis by some regional schemes (eg Davies KW, personal communication). Regression analysis and graphical presentation of the results describe participants' assay performance in terms of bias, imprecision or nonlinearity, though interpretations differ (Grannis, 1979; Lever and Munster, 1979).

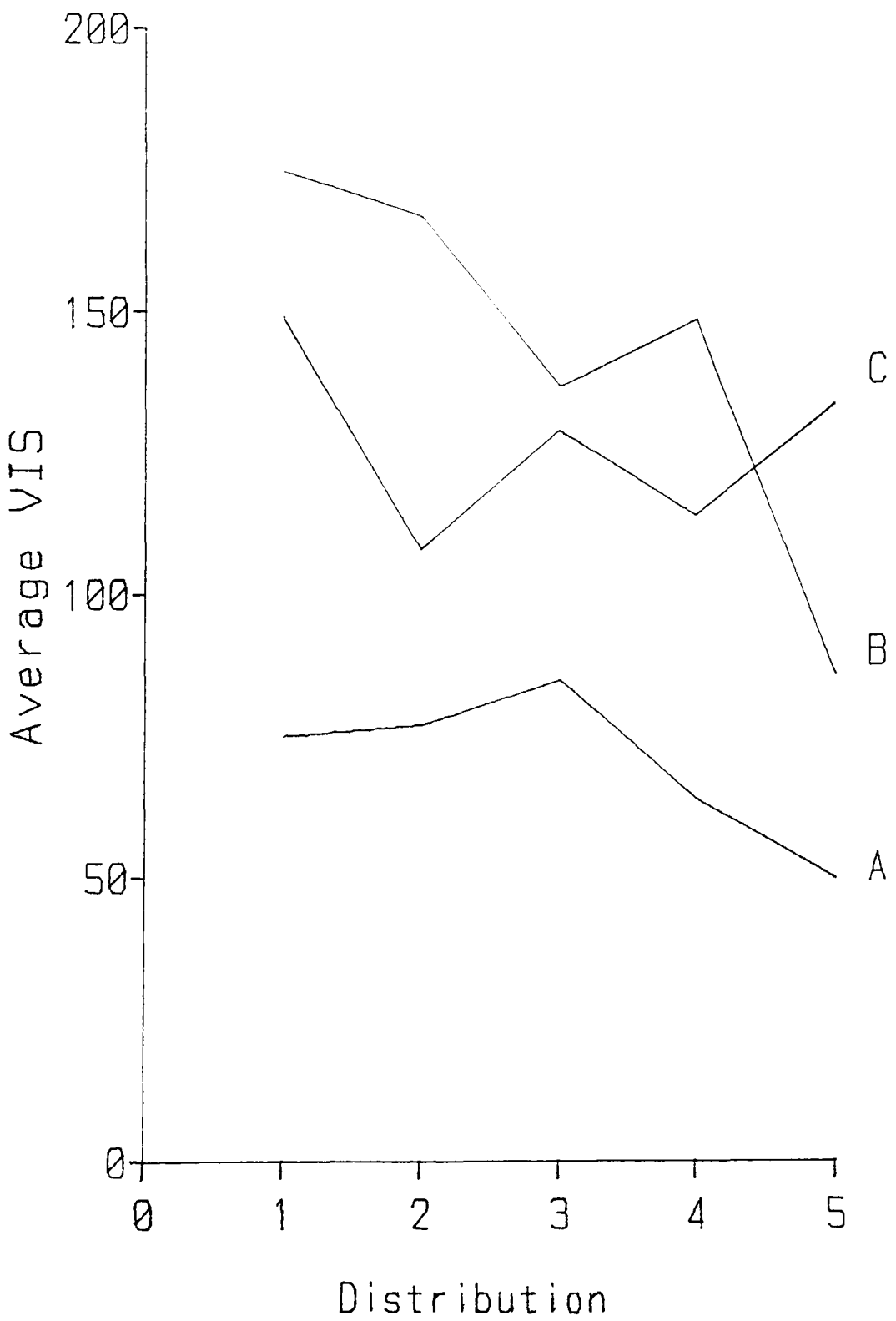
To provide such information at least 5 or 6 specimens are needed, prepared specially with analyte concentrations spread evenly over the range studied. The costs of materials and the more complex data processing are thus high. Obtaining such detailed information on performance distribution after distribution is hard to justify, and it is inappropriate to use this scheme design exclusively. Such a scheme would be extremely expensive, and may yield more information than can really be used effectively; frequent presentation of detailed information can lead to familiarity and blindness to its import.

In the 'intensive' scheme design, originated by Whitehead

(Whitehead TP, personal communication), participants receive a set of three specimens for assay in the same analytical batch. Reports give not only results and VI scores (discussed in section 3.4.2 and Chapters 4 and 9) but also a graph for each analyte relating their results to the designated (target) values for the specimens, with interpretive comments on their performance. Participants also receive a further vial of each specimen to check the efficacy of their remedial action before they receive the next distribution. With three specimens, the graphical presentation offers a reasonable chance of identifying the underlying deficiencies in participants' performance, as discussed further in Chapter 10. Provision of the second set of specimens in each distribution is expensive, however, and it may be argued that use of 6 specimens together might be more cost-effective, though Whitehead's premise is that the effectiveness of participants' actions can be improved by giving them increased confidence that they have succeeded.

Though this design is much more expensive than a conventional single-specimen system, distributions are less frequent (to allow time for corrective action and verification) and such schemes are not intended to be permanent. Rather, they should be short-term undertakings to bring the participants' results roughly into consensus so that a more conventional EQAS design may then be used for subsequent monitoring and 'fine tuning' of performance. Thus Figure 3.3 gives examples of acceptable initial performance showing some improvement, of initially unsatisfactory results which improved, and of unsatisfactory results showing no improvement. The first and second groups could benefit from transfer to a conventional scheme, whereas it is unlikely that the third would be stimulated to improve more than by the

Figure 3.3 Changes in average VIS for participants in an 'intensive' EQAS. Patterns of (A) acceptable initial performance showing some improvement,
(B) initially unsatisfactory performance which improved,
(C) unsatisfactory performance with no improvement



intensive scheme design. Depending on circumstances, this last group may then be classified as either not deserving attention or requiring considerable further guidance.

It must be remembered that the number of specimens in a distribution is linked with the distribution frequency, since the same rate of data accumulation can be attained by distributing many specimens infrequently or single specimens frequently. This choice is dictated largely by considerations of cost and organisation: if specimens are expensive and distribution costs are low the single-specimen design is favoured. Single-specimen distributions also reduce data processing requirements, and may facilitate faster return of reports to participants.

3.2.4 Period between assay and receipt of report

Participants' confidence can be increased by rapid return to them of EQAS reports, in much the same way as for distribution frequency (section 3.2.2). Clinical chemists are more likely to take note of a report referring to assays carried out the preceding week than to those a month or two before. It is also more likely that an explanation can be found for any aberrant results, since some laboratory records and analysts' recollections are of limited span.

The period between specimen assay and receipt of report by a participant is limited by communication between the participants and the organising centre, by the time required for data processing and report printing, and by the necessity to allow sufficient time for participants to assay the specimen and return results. An ideal situation, provided that designated values were available in advance, could be electronic communication of

results to the organising centre with provision of a report on performance by similar electronic means immediately after. Such a system, though becoming technically feasible, would further increase the risk of confusion between IQC and EQA and may be undesirable on these grounds. It would, however, be excellent for the 'external' control of individual users of extra-laboratory equipment.

At present postal communication is the usual means for transmission of results and reports. This, with the use by many schemes of consensus values as designated values, dictates a longer timescale than that envisaged above. Electronic data processing is essential to provide rapid turnround as well as to store and manage the resultant data.

The frequency with which assays are carried out in participants' laboratories is also relevant. If batches are infrequent, a longer period between specimen despatch and the deadline for receipt of results must be allowed. This then entails a correspondingly greater delay between assay and receipt of reports by those participants analysing specimens soon after receipt. The effects of infrequent assay batches on performance are discussed in Chapter 11.

3.3 Validity of the specimens

Effective EQA demands that the specimens distributed are reliable and have properties such that they reflect faithfully the behaviour of clinical specimens - "fidelity", as defined by Fasce et al (1973).

Clinical chemistry is concerned primarily with determinations on specimens from human patients and subjects. To be most effective,

therefore, quality assurance procedures should ideally use serum from these patients or subjects as the QCM. Such QCMs are indeed used in IQC in the daily mean, in analysis of 'carryover' specimens from the previous batch, and in the use of pooled sera. These benefit from the ready availability of material at clinically relevant concentrations, as do specimen exchange procedures. Some attempts have been made (eg Becktel et al, 1973; Leclercq, 1975) to conduct EQA through comparison of daily means or reference intervals, but the authors concluded that such procedures were impractical and insensitive for EQA though of undoubted use in IQC.

There are problems, however, associated with the instability, infectivity (DHSS, 1986a) and restricted availability of authentic clinical material, particularly for widespread EQA. Commercial products, especially those stabilised by freeze-drying or other means, are convenient and relatively free from such problems.

In EQA (other categories of QCM use are discussed in Chapters 12 to 14), the QCM must be stable, and the precision for the QCM should be identical to that for clinical specimens. If the designated value against which a participant's result is assessed is derived using an identical analytical procedure there is no further requirement on QCM properties (though an animal-based QCM would not be ideal for assessing, for example, albumin assay). If, however, the designated value is for all methods, or for a group of methods, the QCM's reponse to differing analytical procedures within the grouping used should be the same as the response of clinical specimens. This requirement is for "commutability", defined originally by Fasce et al (1973) in the

context of enzyme activity assays but for which a more general definition is the consistency of the relationship between results obtained by different analytical methods for the QCM and for freshly-drawn sera from patients (Bullock et al, 1980b; Broughton et al, 1981).

Commutability is of vital importance in EQA if information on laboratory and method performance is to be reliable. The QCMs distributed must therefore be commutable or the scheme design be such as to render negligible the effects of any lack of commutability, eg by use of method means as designated values. If not, participants will not have confidence in the information generated by the EQAS and the scheme will fail in its objectives.

The main choices are with respect to origin and presentation of the specimens, lying between human and animal sources and between liquid and lyophilised materials.

3.3.1 Species of origin

As mentioned above, a human base might be thought essential for EQA specimens and some schemes have therefore used human-based materials exclusively. Indeed one of the original objectives of the UKEQAS (Whitehead et al, 1973) was to distribute liquid pooled clinical specimens, to avoid any criticism that a participant's poor results could be due to the animal basis or lyophilised nature of the specimens.

Such specimens are cheap and relatively simple to prepare (Bowers et al, 1975; Kenny and Eaton, 1981), provided sufficient base serum is available following routine analysis of specimens from patients or through blood transfusion services (from donors whose blood is unsuitable for therapeutic purposes). The resulting

material, however, needs to be sterilised (usually by filtration) and liquid specimens have frequently proved to have limited stability in the EQAS situation. Furthermore the problems of potential infectivity of human-based materials (eg Compton et al, 1979) must be faced, not only with regard to hepatitis viruses but also to human immunodeficiency virus (HIV) and other as yet unknown agents. Recent UK guidelines (Advisory Committee on Dangerous Pathogens, 1986; DHSS, 1986a) on testing for antibody to HIV at the individual donation stage virtually preclude the general use of clinical specimens, unless heating (see Chapter 15) or chemical treatments can be applied to inactivate the virus.

In the case of proteins, such as albumin or peptide hormones, with species-related differences in structure, human serum is essential for reliable judgements of performance through EQA, though within a laboratory animal-based materials may suffice for calibration and precision control in IQC. Such differences are likely to be much less marked in total protein assay, where the analyte is already a heterogeneous mixture. For steroid and thyroid hormone assays, though the hormones do not differ in structure both their relative concentrations and the structures of their binding proteins do. Interspecies differences in behaviour can then be expected, both from these factors and from the additional effects of the different protein milieu on some (eg double antibody) procedures for separation of bound and free antigen.

For some analytes, eg enzyme activity assays, the (kinetic) properties of the added isoenzyme are more important than its species of origin in determining the specimen's properties (Moss

et al, 1985). For example alkaline phosphatase (ALP) from bovine liver is preferable to that from human placenta since placental ALP differs greatly in behaviour from the liver and bone isoenzymes which commonly comprise the activity in clinical specimens. For other analytes the stabilisers, buffers and other additives encountered in commercial products, or even the process of lyophilisation itself, may have a greater effect on the material's properties than does the origin of the base serum. Empirical studies of the effect of base serum on properties, as described and discussed in Chapter 14, may thus be the only way to resolve the question of suitability of animal-based materials.

Table 3.1 summarises the main factors influencing this choice, which has also been considered by others (eg Kenny and Eaton, 1981; Fraser and Peake, 1980). As a general rule, animal-based materials should not be used exclusively, but a human base is not essential for all analytes. Since manufacturing procedures can influence QCM properties, it is prudent to minimise the impact of matrix effects by using specimens from a wide variety of sources.

3.3.2 Specimen presentation

The primary choice is between lyophilised and liquid specimens, which may in turn contain a preservative.

Lyophilised materials are undoubtedly more stable (Lawson et al, 1982), though earlier data from some within-laboratory studies reviewed by Fraser and Peake (1980) showed only small differences in the between-day precision obtained, some in favour of liquid specimens. The outcome of such studies reflects a balance between the amount of thermal and other stress applied (greater in interlaboratory surveys than within individual laboratories), the analyte stability, the vial-to-vial inhomogeneity (including

Table 3.1 Considerations in the selection of QC materials based on human or animal serum

	Human	Animal
Availability	Limited	Effectively unlimited
Infectivity	Hepatitis HIV	Brucellosis
Cost	Expensive	Cheap
Problems	Moral - therapeutic use	Religious - in some societies
Suitability	Universal	Not for immunoassays, etc
Properties	Often related more closely to additives	

filling imprecision for lyophilised materials) and the analytical imprecision.

Most commercial lyophilised materials available in the 1960s and early 1970s were not fully satisfactory. There were problems both with imprecise vial filling and with regard to their properties since experience in optimising the procedures to cause as little disturbance as possible was then extremely limited. Many were also based on animal rather than human serum, and clinical chemists were then unsure whether there were significant differences in behaviour from that of clinical specimens (see section 3.3.1 above). Because use of such specimens might cause participants to blame discrepant results on the material distributed rather than their own laboratory's perfomance, EQAS organisers tended to favour a liquid presentation to engender confidence in the scheme.

As the numbers of participants increased, however, the difficulties of providing specimens sufficient in quality and quantity led many schemes to reconsider their initial decision. With improvements in filling and freeze drying procedures, the balance changed and now almost all EQASs in developed countries distribute commercial or commissioned lyophilised serum specimens with no apparent lack of confidence by participants.

In the tropical climate of many developing countries the need for specimen stability is even greater, but adequate lyophilisation technology is rarely available or economic to operate. Here there has therefore been a considerable problem which now appears to be potentially soluble through the use of ethylene glycol or similar chemical preservatives (Maurukas, 1975). These materials are

stable (Hartmann et al, 1981), convenient to use and have been advocated by some (eg Pope et al, 1979) for linearity evaluation. A particular advantage of ethylene glycol is that it not only prevents microbiological growth and minimises freezing damage to serum constituents by enabling storage at -20^{o}C in the liquid state, but also appears to 'sterilise' so that even infected pooled serum may be used successfully. The preparations are more viscous than serum, however, with a large dependence on temperature; storage conditions and thermal equilibration before use are therefore important in obtaining correct results, and some authors (eg Pope et al, 1979) have reported unusual intermethod biases, for example due to differences in dialyser design between AAI and AAII/SMA systems. Such effects can create problems in EQA due to non-commutability with clinical specimens. Recent endorsement by WHO (Browning et al, 1986) should nevertheless lead to more widespread use for both IQC and EQA in developing countries as a stable low-cost alternative to liquid or lyophilised materials.

Their greater stability makes lyophilised materials more popular, whereas liquid materials are cheaper to prepare and may be more convenient. Specimens incorporating ethylene glycol are becoming more popular, but can present difficulties in determination of designated values due to their lack of commutability (disussed further in Chapters 12 to 14). Table 3.2 reviews the relative advantages of these presentations, each of which may have application in EQA with the choice being dependent upon circumstances.

3.4 Assessment of performance

The experience of most scheme organisers is that the participants

Table 3.2 Considerations in the selection of lyophilised or liquid QC materials

	Lyophilised	Liquid	
		Frozen*	Ethylene glycol
Stability	Excellent	Limited	Good
Convenience	Limited	Good	Excellent
Preparation	Reconstitution	Thawing	None
Vial-to-vial variability	Dependent upon manufacture	Excellent	Mixing required
Clarity	Turbid	Good	Good
Physical properties	Little changed	Unchanged	High viscosity
Cost	High	Low	Intermediate

*with or without preservative, eg azide, antibiotics

most in need of taking note of and acting on EQAS reports are those with the least time, inclination and ability to do so. Performance assessment is therefore of paramount importance for EQASs to gain the confidence of participants and to stimulate improvement.

Elements of importance in producing a reliable assessment of performance include:

- clear and concise report presentation
- scoring system for performance assessment
- valid target values

3.4.1 Report presentation

Reports must be clear, consistent and intelligible, and above all be capable of ready interpretation.

The ideal report might thus be an individually-produced text report commenting on the current set of results, their relation to the designated values, to the results from other participants and to the laboratory's past results, with an exposition of where problems lie and suggestions for their resolution. Such a report would, however, entail considerable effort and investigation, particularly since the organiser would not be fully aware of all the relevant factors peculiar to the individual laboratory. Multiplied by the number of participants and the frequency of distributions (eg 520 x 22 = 11440 reports/year in the UKEQAS for General Clinical Chemistry) such a task becomes completely impracticable.

What then can be done to provide the basis of this information? The format of the report can be simplified in a number of ways. Preprinting and highlighting (by boxing or shading sections, or

printing in a different typeface or colour) makes the text easier to read and important sections easier to identify. Redundant or unhelpful information, or information of very limited interest such as the mean of untrimmed results, should be eliminated from the report; though the mean, the SD and the CV are determined by any two of the values it is in practice advisable to give all three since SDs are more meaningful to some clinical chemists whereas others prefer CVs.

The designated values must be given, and their derivation should be clear; a preprinted explanation, perhaps on the reverse of the report, can be useful here. The participant's results must also be shown, so that their correct attribution may be verified and they may be compared with the designated values.

The results from other participants should be indicated, in tabular, statistical or histogram form. Statistical evaluations are the most common and most widely comprehended, and should include some classification according to analytical procedure. These data are useful in method and instrument selection, as discussed in Chapter 8.

An assessment of performance in terms of a scoring system is also helpful, as outlined below and discussed in Chapter 4 and 9. Again, this information should be given in a clear and simple manner. Reports should facilitate a logical progression through a hierarchy of information in examining firstly whether analytical problems exist and secondly their nature, as outlined in Chapter 9. Some participants, however, demand more and more information in reports, and overenthusiastic scheme organisers may also be tempted to incorporate far too much, often in an indigestible form.

Presentation of any information in graphic form can improve comprehension, and this aspect is discussed in Chapter 10.

3.4.2 Scoring system for performance assessment

In any form of communication, data reduction is a major aid to the recipient's comprehension of the content. Many EQAS organisers have felt the need to use some form of scoring system as a means of data reduction and of stimulating improvements (WHO, 1981). The requirements, applications and benefits are discussed fully in Chapters 4 and 9.

3.4.3 Source of target values

Two main types of procedure have been used to derive designated (target) values for use in both EQA and IQC, involving the use of either reference laboratories or consensus values.

Reference laboratories have been used mostly in EQASs with laboratory licensing as their primary objective, eg in GFR (Hansert and Stamm, 1980) and the USA (Boone, 1984). In essence, one or more reference laboratories are selected by some means and the material is then analysed repeatedly over a period, by routine or reference procedures, and a designated value derived from the results obtained. Problems may arise at any of these stages.

Firstly, how should reference laboratories be selected? Possible criteria include the laboratories' performance, determined from EQA or other data, their reputation or the possession of particular facilities or expertise. In GFR, a reference laboratory is defined as one of which the director is well-qualified; obviously, though the laboratory may therefore have

better than average performance, there is no guarantee of this. Selection based on performance appears better on intuitive grounds, but the initial standard may not be maintained. Some system of continual reassessment is therefore advisable, and a good procedure for this is through the scheme itself. An example is the use of a reference laboratory group within the scheme, with selection criteria based on continuous demonstration of baseline security and quantitative recovery of added analyte, in the early years of the UKEQAS for Human Growth Hormone prior to a change to consensus values after their validation (Hunter and McKenzie, 1979). Where reference or definitive methods are required, the choice of reference laboratories may be dictated by their possession of the necessary equipment (eg for mass spectrometry) or willingness to undertake such work. An additional problem is whether the identity of reference laboratories should be known, either to participants or to the laboratories themselves.

The second problem lies in the analytical technique(s) used, which may be either routine or reference procedures. If the laboratory's routine method is used, eg in their normal participation in the scheme (Hunter and McKenzie, 1979), no complications arise but if a technique must be established specifically for this purpose its IQC can present great difficulties. If reference methods are employed then their validity and comparability among centres cannot be taken for granted (Eldjarn and Broughton, 1985) and must be established (eg Gaskell et al, 1984). Furthermore, if the method used differs from those used by the participants the specimen distributed must be commutable with clinical specimens, as discussed in section 3.3 above and Chapters 12 and 14. This is the main drawback of

the "new concept" proposed by Stamm (1982), whereby reference methods are used to provide a single designated value for each analyte against which all participants' results are assessed. Extensive theoretical justifications of the system and studies of the individual methods (eg Siekmann and Breuer, 1982; Külpmann et al, 1985) have been offered, but there is a lack of evidence that the intermethod relationships for the EQA specimens mirror those for clinical specimens.

Finally, the combination of the results to derive a designated value leads to controversy. The results obtained comprise several dependent sets, and though each set may be normally distributed the overall data are unlikely to be. The initial approach in GFR of using the mean as designated value and the SD to derive limits of acceptability (Stamm, 1975; Hansert and Stamm, 1980) was therefore not statistically valid, and later studies developed procedures using non-parametric statistics (Passing, 1981a and 1981b; Passing et al, 1981). Even these were not universally accepted in GFR, and Haug et al (1978) argued strongly that the disagreements observed between the reference laboratories made the whole procedure suspect and that consensus values should therefore be more reliable. Indeed, reanalysis of the data presented by Stamm (1975) yields designated values for total protein which vary by 3% (from 61.0 to 62.8 g/L) if results from only four of the five reference laboratories are considered.

Overall, the main advantage of using reference laboratory assays lies where the material is such that a true definitive or reference method will yield valid results. Their use can also be helpful where the procedure must to be rigorously defined for possible explanation in a court of law, as in the case of

'licensing' schemes, or for assigning values to reference materials (Marchandise and Colinet, 1983). Interestingly, the licensing scheme organised by CDC in the USA used a combination of criteria derived from consensus and reference laboratories in determining the acceptability of participants' performance, but CDC recently determined that examination of performance data from the CAP programme (using consensus values) now obviates the need to organise their own scheme.

Consensus values (the mean of results for all participants or for those using a particular method) have been used as designated values in many schemes (eg Whitehead et al, 1973; Jansen et al, 1977; de Leenheer et al, 1983) over a long period. Consensus values are convenient for scheme operation. They are inexpensive, requiring no more than calculation, and are available when required. In some circumstances having to wait until all results are available before assessments can be made may be a disadvantage, but is quite satisfactory in most EQASs.

Though consensus values were used initially purely on the grounds of their convenience and for lack of any viable alternative, as experience was gained they were found to be reproducible, both within a scheme (eg Gilbert, 1976) and between countries (eg Whitehead and Woodford, 1981). There was also increasing confidence that consensus values were indeed close to the true analyte concentrations in specimens of liquid human serum, though there is no theoretical reason why consensus values should be accurate. Thus Whitehead and Woodford (1981) commented on the close agreement for calcium among the UKEQAS consensus values for methods employing different analytical principles and with definitive values determined by the US National Bureau of

Standards (NBS), with similar findings for other analytes. Good agreement of CAP survey consensus values with NBS definitive methods was also documented in the USA for calcium (Gilbert, 1975b) and for other analytes (Grannis, 1976). Further studies on the validity of consensus values, primarily those from UKEQASs, are presented and discussed in Chapter 5 below.

A potential problem with consensus values is the possibility that they may 'drift'. Consider the consequences of a supplier of calibration materials assigning an incorrect value to these materials. Users will then have biased results, and if these materials are widely used there will be some effect on EQAS consensus values. Other laboratories will then note from the scheme that their performance is biased and may therefore adjust their procedures to be in consensus again. The bias generated by a single manufacturer could thus be incorporated into the scheme and affect the goal of all participants; results obtained throughout the country would then be inaccurate. There is so far no evidence, however, of such a drift from truth occurring, as discussed further in Chapter 5.

Consensus values have been used predominantly by EQASs which are primarily 'educational' in intent, such as those in Belgium, Holland, the UK and the CAP programme in the USA; they are also used in most commercial schemes. This choice has been dictated largely by their convenience and negligible cost, in the absence of any evidence to contraindicate their use. Participants now have confidence in the assessment of their performance against consensus values.

3.5 Summary

The primary consideration in EQAS design is to ensure the confidence of participants in the scheme's assessment of their performance, and thus increase the likelihood of their taking action on these conclusions.

The following elements and factors in interlaboratory survey design are important in engendering this confidence:

- scheme organisation
    - reliable administration
    - frequent distributions
    - rapid return of reports on each distribution
- specimens that are stable and have properties resembling closely those of clinical specimens
- performance assessment
    - reports that are informative, yet readily comprehended
    - a scoring system for performance assessment
    - reliable designated values, of described derivation

Though many schemes have broad similarities in design, there are individual differences. These lie mainly in the numbers of specimens constituting a distribution, in the scoring system used, and in the derivation of designated values from reference laboratories or from consensus values. Scoring systems are discussed further in Chapters 4 and 9, the validity of consensus values in Chapter 5, and the properties of QC materials in Chapters 12 to 14.

Chapter 4:

SCORING SYSTEMS IN EXTERNAL QUALITY ASSESSMENT

4.1 Introduction - the need for scoring systems

The objective of any EQA scheme is to stimulate interlaboratory concordance of numerical results. Thus, as outlined in Chapter 3, participants require a clear demonstration of whether their results are in consensus, ie whether or not corrective actions are needed. Many participants experience difficulty in comparing their results with the 'target' data provided by the scheme, whether it be in the form of designated values, histograms, statistical parameters classified according to method, etc. In addition, experience indicates that it is usually just those laboratories which have least time, inclination, and ability to devote to this task which have most need to. Some form of data reduction is therefore essential if participants are to derive maximum benefit from the scheme; this most usually takes the form of a scoring system.

4.2 Classification of scoring systems

Scheme organisers have devised a plethora of scoring systems. There are several basic types, however:

- 'pass/fail' systems
- semi-quantitative systems
- quantitative systems

4.2.1 Pass/fail systems

Such systems comprise assessment of each result against some criterion of acceptability. This criterion may be derived as a multiple of the observed SD or of the SD obtained by reference

laboratories (Bundesärztekammer, 1971; Stamm, 1975), or from an estimate of clinical requirements in the form of analytical goals (Bowyer et al, 1981) or medical need (Stamm, 1982).

4.2.2 Semi-quantitative systems

In these systems results are classified on a semi-quantitative scale, eg the allocation of 'points' according to how close the result approaches the designated value (DV). This gradation may be in terms of observed SDs in the initial definition of Variance Index (VI; Whitehead et al, 1973) or of an arbitrary scale of clinical needs (eg Jansen et al, 1977).

Later such systems were applied to other and less quantitative laboratory investigations. Examples here are the system adopted by the UKEQAS for Microbiology, which classifies participants' returns as fully correct, partly correct, wrong and badly wrong on a scale of +2 to -1 (Leblanc et al, 1985c; section 7.1.4), and the similar system adopted initially in the UKEQAS for PKU Screening, described in section 7.2.3.

4.2.3 Quantitative systems

In the final type of procedure results are scored on a continuous scale. The main applications are SD differences (SDDs or 'Z scores': eg Merritt et al, 1965; Ley and Ezer, 1974; Wellcome Diagnostics, 1984), the Variance Index system (Whitehead, 1977) and systems based on cumulated estimates of bias and precision or consistency of bias (Bacon et al, 1983; Groom, 1985d; Bullock and Wilde, 1985). In some of these applications logarithmic transformation or other statistical manoeuvres are applied to the results before and/or after scoring (Healy, 1979).

4.2.4 'Hybrid' systems

Several schemes, eg that operated in the USA by CDC (Boone, 1984), use a combination of these procedures.

4.3 Scoring as a stimulus to improvement

Scoring systems are a potent means of data reduction, to assist participants in assessing their performance relative either to other laboratories or to a defined or arbitrary standard. This objective applies to the individual laboratory situation, and is therefore discussed in detail in Chapter 9 below.

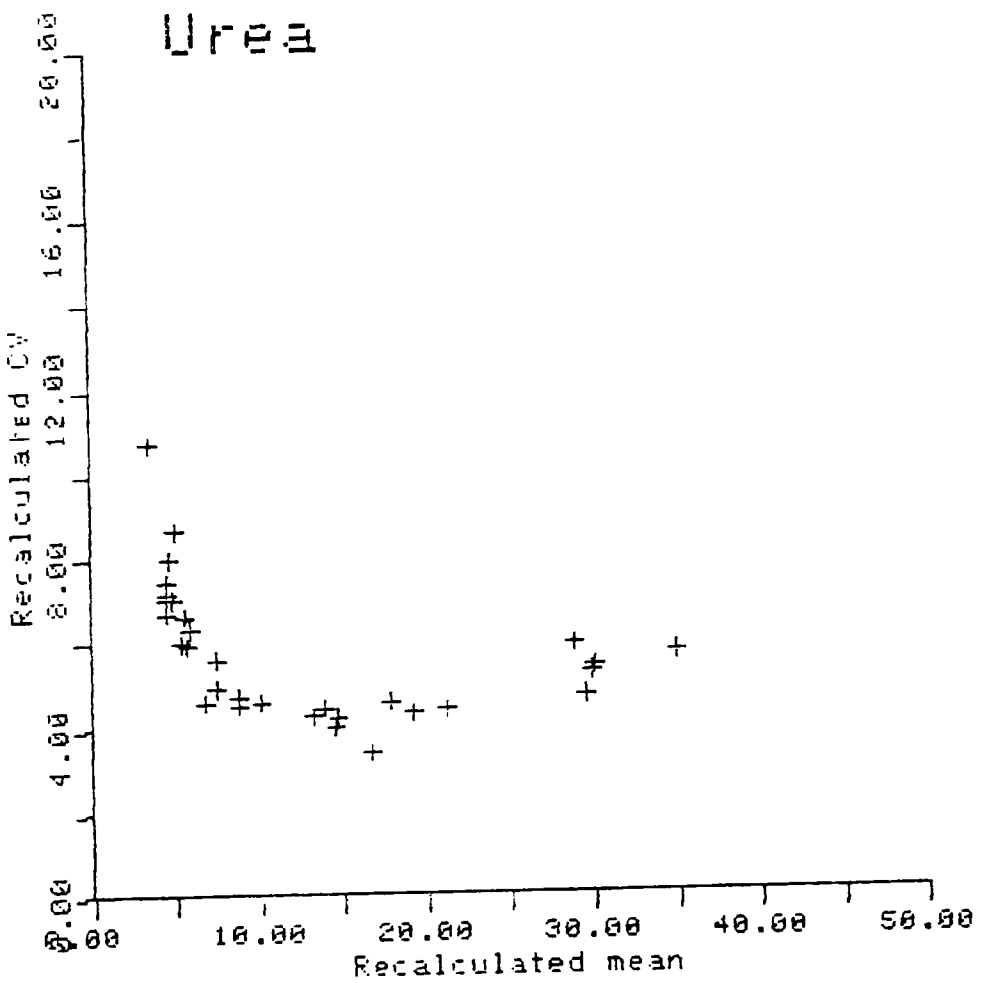
4.4 Scoring in assessment of the state of the art

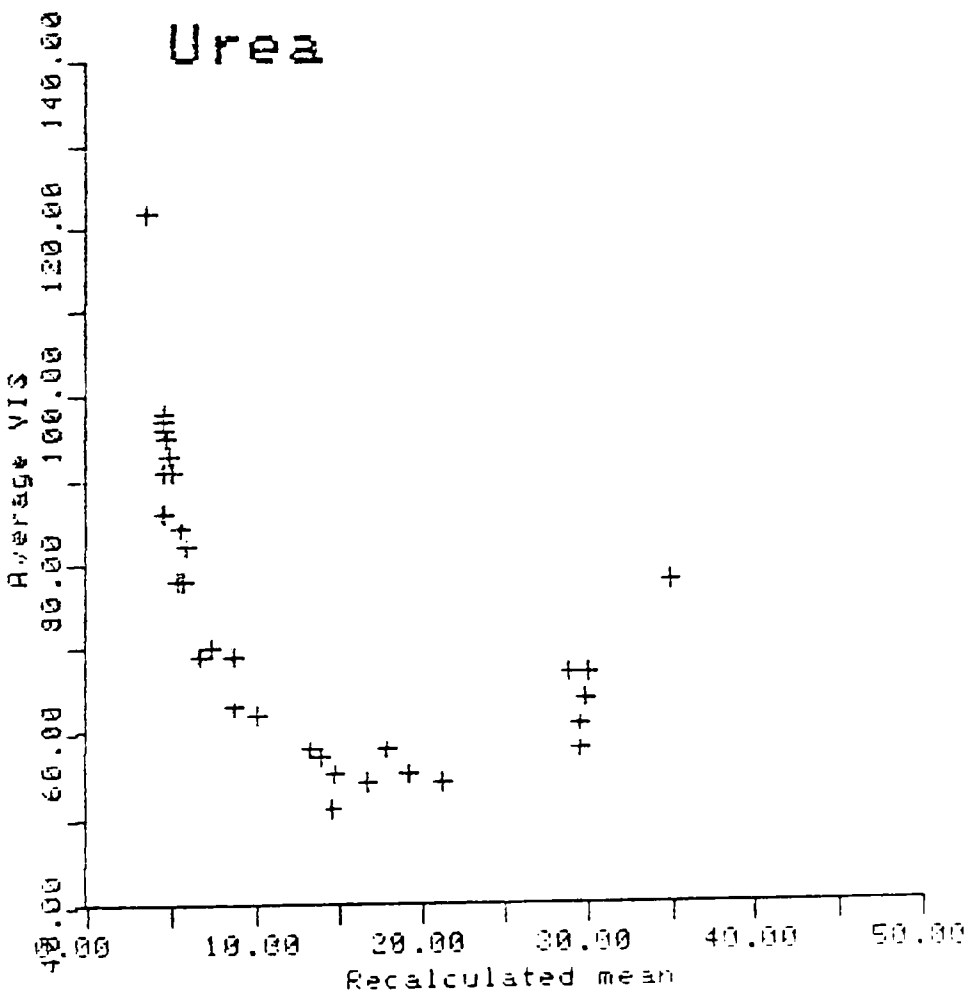
Scoring systems also simplify the assessment of overall concordance, though the primary objective of most is to stimulate improvements in the performance of individual participant laboratories.

For example, interlaboratory CVs provide a measure of the dispersion of results, but they can be influenced unduly by the proportion of results lying in the 'tails' of the distribution. If the distribution is not in fact Gaussian then the CV is not an accurate reflection of dispersion and non-parametric statistics should be considered instead (Passing, 1981b; Röhle et al 1986). Average scores provide a measure derived from all results and which is therefore effectively distribution-free. This measure is also convenient to obtain and use, and likely to be more robust, ie less variable. An additional advantage is that an average score based on use of the method mean as designated value will compensate for any intermethod differences.

Figure 4.1 illustrates this advantage in terms of variability, using data for urea from the UKEQAS for General Clinical

Figure 4.1 Relationship of interlaboratory agreement (recalculated CV and average VIS) with analyte concentration (mmol/L) for urea in UKEQAS for General Clinical Chemistry, 1980-1982





Chemistry. The relationship between interlaboratory agreement and concentration (discussed in detail in Chapter 6 below) is demonstrated more clearly by the average VISs than by the CVs.

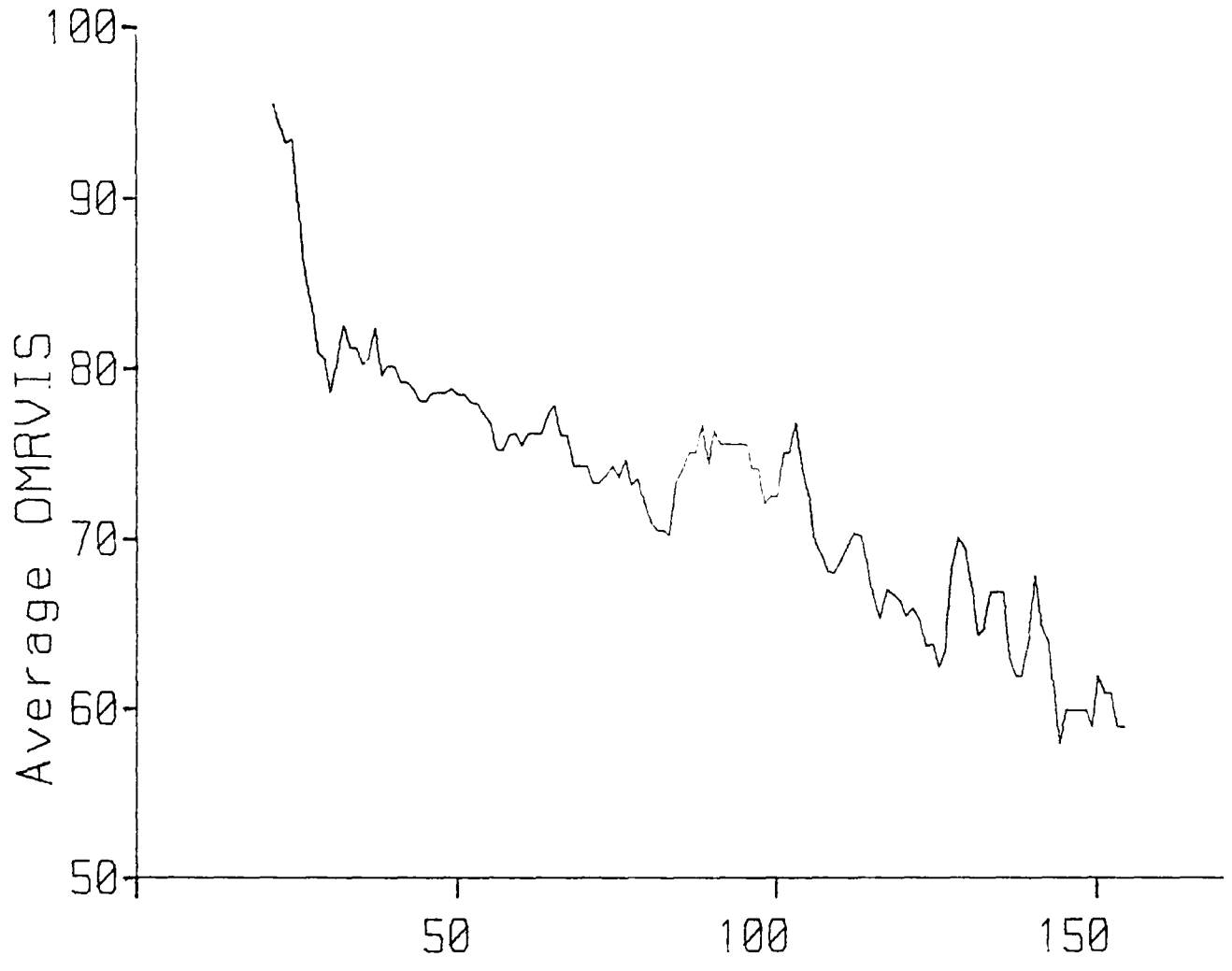
Scores also simplify the assessment of methods and other factors affecting performance (discussed in Chapters 8 and 11) through their cumulation of information over time. The resultant estimate is more representative in that it includes contributions not only from variations over time but also from differences due to analyte concentration and the material distributed. Such an application is demonstrated in section 8.3.2 below (Table 8.2) for calcium in the UKEQAS for General Clinical Chemistry.

4.5 Assessment of progress - comparisons over time

A well-designed scoring system should delineate changes in the performance of an individual laboratory. The main requirement for this, as discussed in sections 4.6 and 9.3 below, is that the score should be independent of other participants' performance. Such scoring systems then also enable assessment of trends in interlaboratory agreement, and provide one means to judge the overall success of the scheme in stimulating improvement.

Such application can be demonstrated for many schemes, as shown in Figures 4.2, 4.3 and 3.1 for the UKEQASs for General Clinical Chemistry (OMRVIS), Lead in Blood and Urinary Pregnancy Oestrogens (both MRVIS). In all cases the average scores show an improvement in interlaboratory agreement following the introduction of VI scoring into participants' reports, confirming the stimulus to performance improvement afforded by scoring. In some cases little improvement had been noted over several years of operation prior to the introduction of scoring (Whitehead et al, 1973; Oakey, 1980).

Figure 4.2 Improvement in interlaboratory agreement (average OMRVIS) in UKEQAS for General Clinical Chemistry, 1972-1986. Calculation changed in 1979, with VIs <50 no longer giving zero VIS



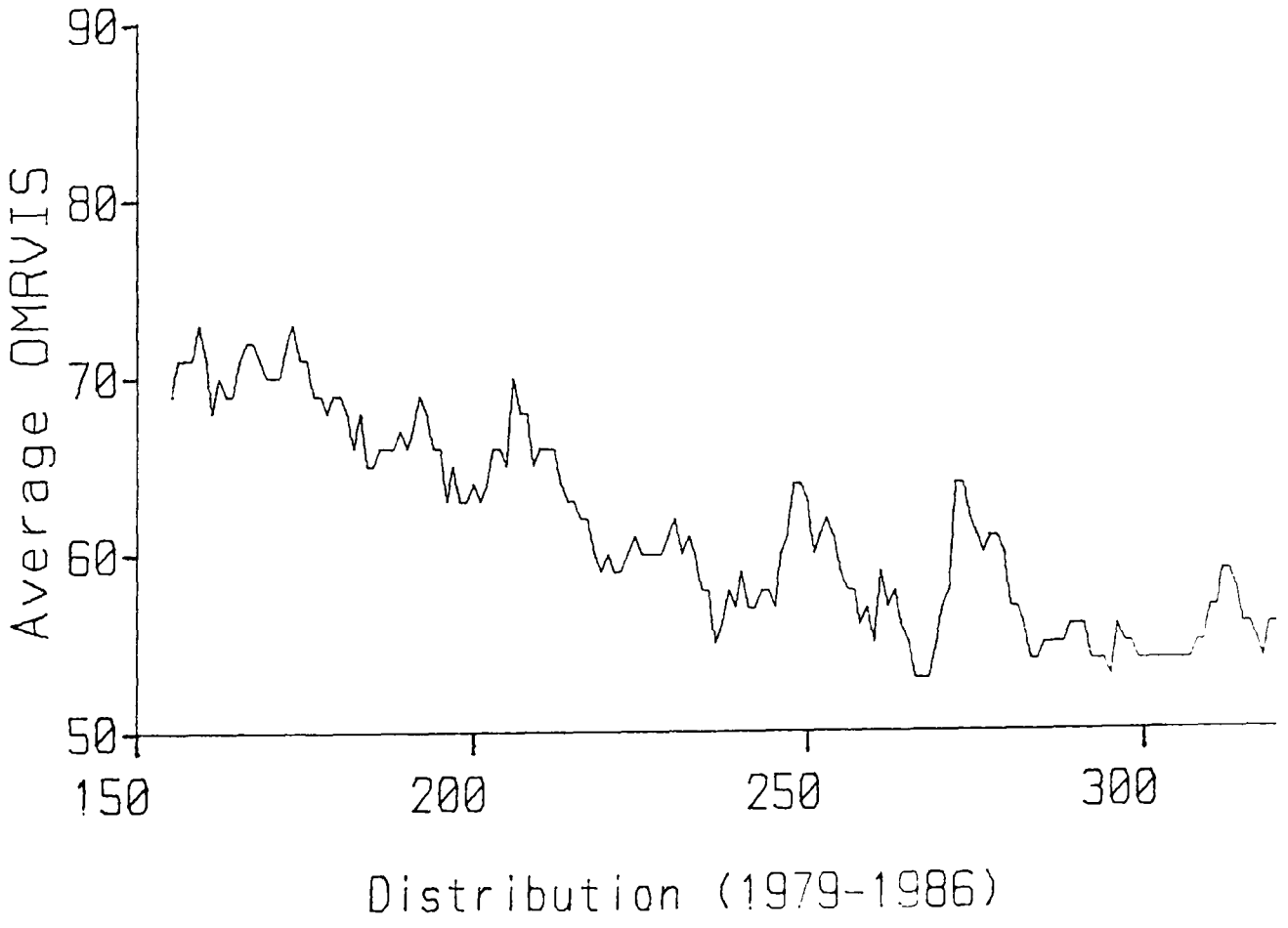
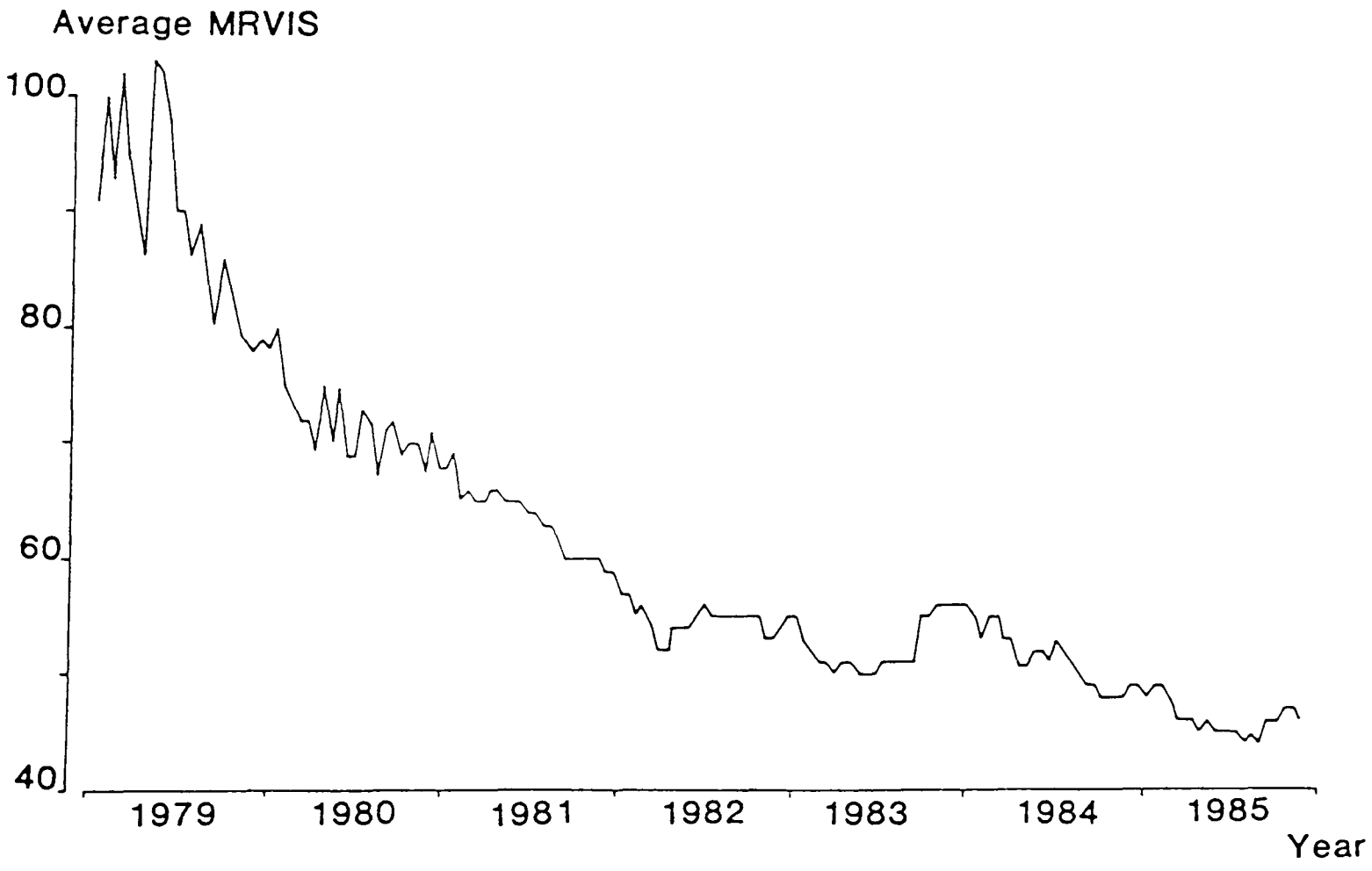


Figure 4.3 Improvement in interlaboratory agreement (average MRVIS) in UKEQAS for Lead in Blood, 1979-1985



Though in some cases other elements of scheme design were changed at the same time (distribution frequency and report format for urinary pregnancy oestrogens; report format for blood lead), and associations between design changes and performance are purely circumstantial, demonstration of the same apparent effect in three schemes is strong evidence of the effectiveness of scoring as a stimulus to improvement.

Scoring systems such as the VI system can also assist in differentiation among the factors contributing to variance. Figure 3.1 illustrates that in the UKEQAS for Urinary Pregnancy Oestrogens there was improvement in within-laboratory consistency of bias (represented by SDBIS, and equivalent in this case to imprecision) as well as in interlaboratory agreement (represented by MRVIS).

4.6 Comparisons over geography

Average scores provide a convenient and reliable assessment of performance in different groups of laboratories. These may be groups within a scheme or be the participants in different schemes. For example Table 4.1 demonstrates the relative performance of occupational monitoring, clinical, and environmental monitoring laboratories in the case of blood lead assay, where the latter group seemed to have superior performance (Bullock et al, 1986c)

If participants in different schemes are to be compared, the scoring system must again be one which is independent of the performance of other participants. Otherwise the comparison is meaningless, since for example the average SDD in any scheme will be approximately the same and does not reflect the standard of

Table 4.1 Relative performance of participants in the UKEQAS for Lead in Blood, classified according to principal component(s) of workload and laboratory type. *"Other" includes governmental, university nonclinical, and commercial laboratories

	n	MRVIS Average	MRVIS SD
Source of workload:			
Occupational	49	62	36
Clinical	26	68	36
Environmental	14	45	18
Type of laboratory:			
Health service	42	64	37
Industrial	20	72	36
Other*	27	52	28
All participants	89	62	35

performance within the scheme. The Variance Index system is such a system, and Table 4.2 shows the average OMRVISs in a number of national and regional schemes administered from WRL in 1984/1985. The calculation of the OMRVIS in the UKEQAS differs slightly from that in the other schemes (see Appendix II.2), but scores should be comparable as the majority of the materials distributed were identical. These figures confirm that the interlaboratory variance seen in the International EQAS was 2-2.5 times that in the UKEQAS for General Clinical Chemistry, a conclusion which could also but less conveniently be drawn from the average CVs.

4.7 Appraisal of scoring systems

Having reviewed the uses of scoring, how well do the various systems which have been devised fulfil these objectives?

4.7.1 Assessment of imprecision

Though bias is the prime concern of EQA, with imprecision assessed primarily through IQC procedures, it may be advantageous to assess both aspects of performance since laboratories may not have (or may not take sufficient note of) an effective IQC programme to assist in interpretation of EQA data. In addition, IQC data is by definition internal to the participant laboratory and it would otherwise be impossible in an EQAS to interpret reliably an assessment of bias without also having some indication of within-laboratory precision as an estimate of confidence in the bias assessment.

Some EQA schemes include a formal assessment of imprecision, through the replicate analysis of specimens. This may be by 'open' replicate analysis of each individual specimen, which is obviously open to manipulation by participants, as in the

Table 4.2 Average running scores (OMRVIS) in EQASs administered from Wolfson Research Laboratories, 1985. OMRVIS (Appendix II) derived from 30 VISs (40 VISs in UKEQAS); data for China EQAS obtained in 1982

	Average OMRVIS
UKEQAS	55
China EQAS	154
International EQAS	122
Middle East EQAS	92
Thailand EQAS	105
Colombia intensive EQAS	200
Mexico intensive EQAS	124

'hybrid' IQC/EQA schemes (Limonard, 1979; Jansen and Jansen, 1980; Lawson et al, 1980). The alternative is repeated distribution of the same specimen as part of the scheme design, to afford an estimate of within-laboratory precision (eg Oakey, 1980, Bacon et al, 1983; Wellcome Diagnostics, 1984; Groom, 1985d). Some of these schemes (Oakey, 1980; Wellcome Diagnostics, 1984) use this estimate as their only or primary assessment of laboratory performance. This is merely a duplication of what the laboratory's IQC program should be providing and is hence a waste of resources; if the semi-blind EQA replicate data conflict with the IQC estimate then the more favourable figure will tend to be believed.

4.7.2 SD differences

More complex algorithms, such as the standard deviation difference (SDD) or 'Z score' (eg Merritt et al, 1965; Wellcome Diagnostics, 1984), may be employed as indicators of laboratory performance. The SDD is calculated by dividing the observed SD into the difference of the participant's result from the mean. This statistic, however, can only be used to compare performance of laboratories with others in the same scheme at the same time. SDDs can **not** be compared over time since the performance of the other laboratories (and hence the observed SD) may change, nor in different schemes since the observed SD will almost certainly differ. The average SDD in any scheme at any time will always have a similar value, and changes will only represent a change in the shape of the frequency distribution of results.

4.7.3 The Variance Index (VI) system

Consideration of these factors led to the devising and subsequent refinement of the Variance Index (VI) scoring system

(Whitehead et al, 1973 and 1975; Whitehead, 1977; Bullock and Wilde, 1985). The formal definitions are given in Appendix II.2.

Variance Indices offer a convenient system for comparing assessments of performance over time and over geography. Cumulation over analytes and over time yields an Overall Mean Running VIS (OMRVIS) as an empirically-useful assessment of overall performance. Its application in comparisons of the state of the art is demonstrated in sections 4.5 and 4.6 above, and its utility in the assessment of individual laboratory performance is discussed in Chapter 9.

The VIS is a mixed index, responsive to failures both of accuracy and of precision, but the MRBIS and SDBIS (Bullock and Wilde, 1985) afford a means of separating the effects of these aspects of performance. Their interpretation is discussed in detail in Chapter 9. They thus correspond closely in intent to the BIAS and VAR estimators used in UKEQASs for hormone assay (Groom, 1985a).

4.7.4 BIAS and VAR

These parameters (Bacon et al, 1983; Groom, 1985b) provide an assessment of bias and its variability, both in percentage terms. Interpretation is similar to that of the MRBIS and SDBIS, but there are several differences.

Firstly, the scores are **not** in a 'common currency'. Thus the scores for analytes cannot be combined to give an estimate of overall performance. Interpretation must also be individual for each analyte, whereas in the VI system similar performance relative to the state of the art yields similar scores. Clinical relevance may be more readily discerned from the percentage error presentation, but VISs can be transformed simply into such terms

(eg a BIS of -85 for calcium indicates a negative bias of 3.4%).

Secondly the derivation of BIAS and VAR includes an outlier elimination procedure. Any discrepant individual biases are classified separately as 'blunders' and therefore not lost entirely from consideration, but do not contribute to the performance assessment used most commonly in these schemes. In the VI system the effects of grossly discrepant individual results are mitigated by truncation of the VIS at a maximum of 400, but such scores do still contribute.

4.8 Selection of Chosen Coefficients of Variation (CCVs)

The CCVs (Table 4.3) were originally determined as the best average CVs obtained in the UKEQAS for General Clinical Chemistry during 1972 (Whitehead et al, 1973). This gave a balance according to the state of the art existing then, and combination of scores for the various analytes would be valid. Such a derivation is independent of the clinical requirements, however, and the relative merits of assessment with respect to the state of the art and to medical needs are discussed in section 9.6.

For other analytes in this scheme the CCVs were determined from an assessment of the interlaboratory agreement obtained over a period of at least a year and the average OMRVIS at the time. This procedure was employed for the relatively 'mature' assays lithium, magnesium and osmolality. For enzyme activity assays, however, the interlaboratory agreement was unsatisfactory and short-term improvement was sought, so the 'calibrated' CVs for reliable method groups were considered (Bullock et al, 1986b).

A similar procedure using 'calibrated' data was adopted for immunoglobulin assays (Chambers et al, 1987). For single-analyte

Table 4.3 Chosen Coefficients of Variation (CCVs) and average Variance Index Scores (VISs) for all participants during 1986 in the UKEQAS for General Clinical Chemistry

	CCV	Average VIS
Sodium	1.6	61
Potassium	2.9	55
Chloride	2.2	77
Urea	5.7	55
Glucose	7.7	44
Calcium	4.0	64
Phosphate	7.8	54
Iron	15.0	46
Urate	7.7	64
Creatinine	8.9	45
Bilirubin	19.2	37
Total protein	3.9	66
Albumin	7.5	53
Cholesterol	7.6	48
Lithium	11.0	34
Magnesium	10.0	55
Osmolality	2.9	46
AST	12.5	59
ALT	17.3	52
LD	13.2	58
CK	18.5	62
ALP	15.5	57
Amylase	11.5	80

schemes no combination of scores for different analytes is required and the choice is consequently less critical. CCVs were therefore selected to give approximately similar average scores; those for salicylate and paracetamol were continued from one of the Regional schemes which preceded the UKEQAS (Epton, 1979).

Inspection of the CCVs (Table 4.3) suggests that there would be more scope for improvement in the analytes such as iron, bilirubin and creatinine which showed worst agreement initially than in for example sodium. Has the balance among the CCVs been maintained, and do they then still give VISs in a common currency?

Table 4.3 also gives the average scores from 1986 in the UK, which show a considerable range (34-80). Considering only those analytes (sodium to cholesterol) for which CCVs were established in 1972 the range is reduced (37-77) but still appreciable. Indeed the scores for iron, bilirubin and creatinine are among the lowest, and that for sodium one of the highest, but the relation with CCV is not simple. For example glucose gives the second lowest score and chloride the highest. Its clinical importance has presumably stimulated the progress for glucose, whereas the decreasing interest and workload for chloride has led to a relative deterioration in performance. The lack of major improvement for calcium despite its critical clinical application is disappointing.

Should the CCVs therefore be reallocated to bring them back into alignment? This would certainly provide a fairer assessment of performance, in line with the current relative state of the art in the UK. The system is also applied to other schemes, however, and as might be anticipated the balance differs somewhat from

Table 4.4 Ratio of average VISs obtained in International EQAS and Middle East EQAS to those for distribution of the same material in the UKEQAS for General Clinical Chemistry, 1986.
UKEQAS distribution 305 (heat-treated human serum)

	International EQAS	Middle East EQAS
Sodium	2.12	1.59
Potassium	1.88	2.23
Chloride	1.87	1.83
Urea	2.32	1.91
Glucose	2.69	1.69
Calcium	2.36	1.69
Phosphate	2.01	1.76
Iron	2.06	1.53
Urate	1.84	1.04
Creatinine	2.21	2.02
Bilirubin	2.35	1.43
Total protein	1.64	1.62
Albumin	2.22	1.68
Cholesterol	1.84	2.01

scheme to scheme (Table 4.4). Should the advantages of between-scheme comparison and immediate availability of a performance assesment then be sacrificed to give a slightly more even balance within each individual scheme?

The main argument against readjusting the balance by changing CCVs, however, is the consequent loss of continuity. Thus 15 years' performance data from the UKEQAS for General Clinical Chemistry are on file, and comparability with these past data would be forfeited through a change in CCVs. Minor changes, such as that caused by the inclusion of VISs below 50 in 1979, have been assimilated, however, and a readjustment for all schemes merits consideration.

4.9 Summary

Scoring systems make a major contribution to the effectiveness of EQA in performance assessment. This applies not only to the assessment of performance for individual participants (detailed in Chapter 9), but also to the assessment of:

- the state of the art
- the effects of analytical procedure
- progress over time
- relative performance in different schemes and countries

Many systems have been devised, and all have disadvantages.

Simple 'pass/fail' systems are useful in the licensing situation, but have no fine discrimination of performance standards. They also pose the problem of determining the criterion of acceptability. Semi-quantitative systems also have limited discrimination, and are applicable primarily to non-quantitative assays, as discussed in Chapter 7.

Among quantitative (continuously discriminating) systems, SDD scores suffer from lack of a consistent baseline; each laboratory's scores are also dependent upon other participants' performance. They cannot then be used for assessment of progress, nor for comparison of performance among schemes.

The BIAS/VAR system gives good distinction between bias and other factors contributing to errors, but is not scaled to give scores in a 'common currency' for all analytes. More importantly, there is no single estimator of total error, and performance estimates for different analytes cannot be combined.

The Variance Index system appears to overcome most of these problems, yielding an estimator of total error (VIS). VISs can be cumulated over time to give performance indicators for individual analytes (MRVIS) and also over analytes for overall performance (OMRVIS) which are robust and readily interpretable in terms of the state of the art. Similar calculations will also yield estimators of bias (MRBIS) and consistency of bias (SDBIS) for each analyte. The hierarchical interpretation of these indices is discussed in Chapter 9.

A potential problem in the VI system arises if the relative state of the art changes, ie there is a shift in performance for some analytes. Examination of data from the UK confirms there has been some movement, and a readjustment of Chosen Coefficients of Variation to restore the common currency must be considered despite the practical difficulties entailed.

Chapter 5:

THE VALIDITY OF CONSENSUS VALUES

5.1 Introduction

A consensus value is the mean, usually following some form of trimming procedure to remove 'outliers' and thus increase its reliability, of results returned by participants in an EQAS. Dependent upon circumstances the overall mean (ie including results from all participants) or the mean of results from participants using an individual method, or group of methods, may be used. The main purposes of calculating consensus values are to obtain:

- a target value (designated value) for assessment of participants' performance, as discussed in Section 3.4 and Chapter 4
- an assigned value for the material
- a value for comparison with some externally-derived designated value in the assessment of overall accuracy.

Their variability (expressed in SD or CV terms) may also be used, in deriving confidence intervals for the assigned values, in the assessment of the state of the art (see Chapters 2 and 6), in the assessment of individual participants' performance in terms of 'SD differences' ('Z scores', as discussed in Chapter 4), or in identifying some results as statistical outliers (Healy and Whitehead, 1980; Bacon et al, 1983; Groom, 1985b). In this case the procedure used for elimination can be important, and truncation (Büttner et al, 1983b), proportional censoring (Healy, 1979) and non-parametric procedures (Passing et al, 1981) each have their advocates. In most other situations,

however, the exact procedure adopted is largely irrelevant since the purpose of outlier elimination is to give a more robust estimate of the analyte concentration by removal of grossly discrepant results.

Consensus values were used in EQASs originally because they were convenient to obtain and use. The first objective of such schemes was also to attain numerical concordance of results within a country rather than strive for absolute accuracy, which was then probably not accessible within the constraints of the reference system at the time. Though there is no scientific reason why consensus values should be accurate, as experience with the developing EQASs increased it became apparent that these values had additional properties (eg Whitehead et al, 1973; Gilbert, 1976; Grannis, 1976; Whitehead, 1977).

Thus, there appeared to be consistent agreement among the means for methods with completely different analytical principles, eg atomic absorption photometric, titrimetric and dye-binding colorimetric procedures for serum calcium assay. Furthermore such method means also seemed to agree with values obtained by the newly-developing reference and definitive methods. Consensus values also appeared to be reproducible on repeated distribution of the same specimen, provided stability could be assured.

These studies, however, were conducted using pooled liquid human sera, ie the specimens distributed resembled very closely the specimens of fresh human serum to which the methods were applied and for which they had been optimised. As EQASs developed and the numbers of participants increased it became impractical to use such specimens. Improvements in manufacturing technique at this

time (around 1970-1975), yielding lyophilised products which no longer suffered from such deficiencies as high vial-to-vial variability due to imprecise filling, therefore made such materials increasingly attractive as an alternative source of specimens. Most of these materials, however, were still based on animal rather than human serum, and the full range of their properties and the effects of manufacturing procedures were unknown. Such factors are best studied empirically, as described in Chapters 14 and 15.

It was then possible that the good agreement obtained for liquid human sera would no longer be obtained with these materials. It was then necessary for EQASs to use method-related consensus values for the assessment of performance. Otherwise, if the material proved not to be commutable (see sections 3.3, 12.1 and 13.5.2), participants could be penalised for deficiencies in the specimen distributed rather than in their analytical performance on clinical specimens: this could rapidly lead to erosion of confidence in a scheme's validity. In addition, EQAS results could no longer provide a reliable comparison with reference or definitive methods since the materials distributed might not be commutable with fresh human serum. In this situation it would be impossible to relate the analytical performance attained in countries with NEQASs relying on consensus values for assessment of participants. Drifts away from accuracy might then occur, with no means for their detection, and international concordance be lost through the continuing introduction of more national or regional schemes.

International comparisons were thus essential to assess whether such drift, the potential for which appears to be inherent in the

use of consensus values, was in fact occurring. Such comparisons were initially conducted on an ad hoc basis. Usually the schemes involved were well-established, with one or both using consensus values for assessment. Comparisons arose incidentally, when both schemes distributed the same material, or deliberately, with one scheme providing material to the other or the manufacturer providing it to both. In other cases schemes used only materials distributed previously through another, longer-established scheme, often using consensus values from the latter as designated values.

Though such studies largely indicated that consensus values were indeed reproducible among as well as within schemes, more systematic study of interrelationships between NEQASs was needed. The Wolfson Research Laboratories (WRL) therefore instituted the WRL International Intercomparison Scheme (WIIS) in 1984 as a feasibility study for such a system, involving as many established schemes as practicable.

5.2 Reproducibility of consensus values

Before any comparison of consensus values between EQASs, their reproducibility within schemes must first be assessed. Some scheme designs include repeated distribution of specimens (eg for assessment of reproducibility of participants' results; Oakey, 1980) and facilitate such studies, but in others a second distribution of the same material is only occasional.

5.2.1 Studies using UKEQASs

The UKEQAS for General Clinical Chemistry (Appendix I.2.1) is of the latter type, though the scheme organisers have endeavoured to carry out such repeats about 6 months after the material's first distribution. Experience with 13 paired distributions in a

7-year period (Appendix III.1.1) is summarised in Table 5.1; the presentation in terms of percentage differences is necessitated by the wide range of analyte concentrations. From these data it is clear that the consensus values for 'mature' assays with largely satisfactory interlaboratory agreement obtained in a scheme with many participants are highly reproducible. The differences for glucose, bilirubin and enzymes represent decreases, and may therefore in part reflect material instability; the variability for iron is unexplained.

Is this also true for other cases? The UKEQAS for Lead in Blood (Appendix I.2.3) offers an example of a smaller scheme with a less mature assay. Here assessment (Bullock et al, 1986c) showed a CV of 1.2% for consensus values on 11 pairs of duplicate specimens distributed sequentially with lead concentrations in the range 1.56-3.73 umol/L, based on 47-91 results (average 68).

The UKEQAS for Urinary Pregnancy Oestrogens (Appendix I.2.4) is also a small scheme, and its design (Oakey, 1980; Bullock and Wilde, 1985) incorporates replicate distribution of a set of linearly-related specimens over one to two years. Table 5.2 summarises experience with a set of specimens for the Lever, Brombacher and RIA method groups (an overall mean is inappropriate here, since assays have differing specificities, as demonstrated by the regression slopes and discussed in section 13.3). Here again the consensus values are highly reproducible, bearing in mind the relatively low numbers of results contributing; the average VISs are less reproducible, since interlaboratory agreement was improving during this period (Figure 3.1).

Table 5.1 Reproducibility of consensus values in the UKEQAS for General Clinical Chemistry, 1981-1987. Percentage differences between overall means (SCE 37°C for AST and ALP) for 13 distributions

	n	Percentage difference	
		Average	SD
Sodium	13	0.12	0.08
Potassium	11	0.09	0.10
Chloride	7	0.17	0.14
Urea	13	0.21	0.15
Glucose	13	1.44	1.55
Calcium	13	0.21	0.20
Phosphate	11	0.28	0.27
Iron	9	1.03	0.88
Urate	13	0.70	0.83
Creatinine	13	0.54	0.42
Bilirubin	9	1.80	1.25
Total protein	13	0.17	0.12
Albumin	7	0.32	0.20
Cholesterol	11	0.47	0.49
Lithium	10	0.39	0.29
Magnesium	7	0.43	0.29
Osmolality	3	0.11	0.10
AST	1	1.95	-
ALP	2	1.98	-

Table 5.2 Reproducibility of consensus values in the UKEQAS for Urinary Pregnancy Oestrogens, 1980-1981 (24 distributions). n=48 laboratories for Lever method group, 47 for Brombacher/Hainsworth & Hall, 14 for RIA (total 123 participants)

Average VIS	Mean oestrogen (umol/2L)		
	Lever	**Brombacher**	**RIA**
92.5	36.2	36.6	28.5
80.1	36.1	36.3	27.0
71.2	35.6	37.1	28.6
72.7	34.5	35.5	27.3
71.5	46.7	47.9	38.1
70.6	46.8	47.4	35.2
76.4	44.9	47.4	37.9
64.5	44.6	47.2	34.8
64.4	61.0	65.0	49.4
68.9	61.4	65.4	49.4
70.5	62.8	63.5	49.6
63.7	60.9	65.1	46.8
74.5	81.1	86.5	66.2
58.2	82.6	85.3	68.9
68.3	82.2	86.8	69.6
56.1	82.9	85.0	62.2
59.2	113.2	118.5	92.4
52.2	116.2	116.0	87.5
57.4	112.8	117.9	89.0
57.5	114.0	118.8	90.6
63.5	140.2	146.6	108.4
60.9	140.7	143.8	107.6
67.4	139.8	145.1	107.4
55.6	140.2	145.4	108.4

5.2.2 Studies in the International and Middle East EQASs

Experience from organising EQASs in developing countries indicates that the interlaboratory agreement is substantially worse than that seen in UKEQASs, with SDs and CVs typically double the UK values (eg Table 4.4). The additional variance makes consensus values less robust, leading to preference for UKEQAS data as designated values for such schemes. Is this then reflected in the reproducibility of their consensus values?

Table 5.3 summarises the percentage differences observed in the International EQAS (Appendix I.3.1) and the Middle East EQAS (Appendix I.3.2), for comparison with the corresponding data from the UKEQAS for General Clinical Chemistry above (Table 5.1). The pattern is similar to that in the UK scheme, and the consensus values in these schemes are also reproducible. The greater variability results from the worse interlaboratory agreement and consequently larger standard errors.

5.3 Ad hoc comparisons of consensus values

Some materials distributed in the UKEQAS for General Clinical Chemistry (Appendix I.2.1) were also distributed either in other NEQASs or in commercial EQASs. The NEQASs for which data were available were those in the Netherlands (Jansen et al, 1977), Norway and South Africa (Georges, 1985). The commercial schemes were organised by Merz+Dade, Ortho Diagnostics (both 'hybrid' IQC/EQA schemes, discussed in section 3.2) and Wellcome Diagnostics. The overall consensus means were compared, since the balance between use of different method principles between the schemes appeared small.

5.3.1 National EQASs

These comparisons (Appendix III.1.2) are summarised in Tables

Table 5.3 Reproducibility of consensus values in the International EQAS (11 distributions) and Middle East EQAS (8 distributions), 1985-1987. Percentage difference between overall means

	IEQAS		MEEQAS	
	Average	**SD**	**Average**	**SD**
Sodium	0.4	0.2	0.5	0.3
Potassium	0.8	0.6	0.5	0.5
Chloride	0.6	0.3	0.4	0.3
Urea	1.4	0.8	1.0	0.8
Glucose	1.2	1.1	1.1	0.6
Calcium	1.2	1.4	0.9	0.4
Phosphate	1.6	1.2	1.7	1.3
Iron	5.2	4.5	2.8	2.1
Urate	1.4	1.0	1.2	1.0
Creatinine	2.5	2.0	1.2	0.9
Bilirubin	1.9	1.6	1.9	1.5
Total protein	1.1	1.1	1.1	0.6
Albumin	2.1	-	0.3	-
Cholesterol	1.7	1.2	2.3	1.6

5.4 and 5.5, and demonstrate excellent agreement in most cases. Similar agreement was obtained for a further three materials distributed in the UK and Holland during 1982 and 1983. The isolated instances where values diverged, eg creatinine and South Africa in Table 5.5, could not readily be explained; an independent comparison in South Africa showed good agreement with values obtained on prior distribution in the Wellcome Diagnostics scheme (Georges, 1985).

5.3.2 Commercial EQASs

Tables 5.6 and 5.7 give similar summary data from these comparisons (Appendix III.1.2). Again there is good agreement, with no consistent differences. This might be expected since some of the participants in the commercial schemes are UK laboratories, though only for the Wellcome programme (with about 25% of participants in the UK) is this likely to have been an appreciable factor. The tendency for slightly higher urate values in the UKEQAS may have been due to greater use of chemical procedures.

5.4 Comparisons with reference and definitive methods

Comparisons with values assigned to materials using definitive or reference methods serve two main purposes:

- validation of the accuracy of consensus values
- assessment of any 'drift' in consensus values

For the first the properties of the specimens considered must be very close to those of clinical specimens; liquid specimens of human serum are necessary for this. In the second application, however, it is only necessary that the composition of the specimens is consistent over time, and materials from the same manufacturer should suffice.

Table 5.4 Comparison of consensus values in the UKEQAS for General Clinical Chemistry with those in the Netherlands EQAS for Armtrol bovine serum, lots 488 and 489, 1980. Units as in Table 5.5

	Lot 488		Lot 489	
	UK	Holland	UK	Holland
Sodium	150.3	150.9	141.4	141.9
Potassium	4.95	4.95	4.38	4.40
Chloride	110.2	110.9	101.1	101.1
Glucose	9.30	10.86	5.02	5.57
Calcium	2.79	2.71	2.44	2.39
Phosphate	1.98	2.03	1.64	1.67
Iron	39.8	39.8	32.2	32.2
Urate	0.569	0.556	0.297	0.296
Creatinine	247	240	119	114
Bilirubin	138	135	23.6	22.5
Total protein	75.0	75.2	70.8	71.1
Cholesterol	4.89	5.08	4.79	4.99

Table 5.5 Comparison of consensus values in the UKEQAS for General Clinical Chemistry with those in the Norwegian (Seronorm, lot 158) and South African (UKEQAS human serum, lot L4/80) EQASs. All results in mmol/L unless specified

	UK	Norway	UK	S Africa
Sodium	137.5	137.1	141.0	140.9
Potassium	4.56	4.56	4.97	4.95
Chloride	107.7	107.1	105.2	105.8
Urea	8.6	8.6	4.59	4.76
Glucose	5.00	5.01	5.38	5.51
Calcium	2.71	2.73	2.23	2.28
Phosphate	1.06	1.07	1.32	1.34
Iron (umol/L)	29.7	29.7	-	-
Urate	0.464	0.463	0.266	0.266
Creatinine(umol/L)	145	147	87	100
Bilirubin (umol/L)	22.4	21.9	-	-
Total protein (g/L)	64.5	63.9	65.4	66.7
Albumin (g/L)	-	-	40.7	39.6
Cholesterol	2.29	2.26	4.45	4.36
Magnesium	0.88	0.87	-	-

Table 5.6 Comparison of consensus values in the UKEQAS for General Clinical Chemistry with those in the Merz+Dade QAP, 1984 (Levels I and II) and 1986 (Level II). Units as in Table 5.5; enzyme data (U/L) for SCE optimised 37°C, except amylase (Phadebas 37°C)

	Level I		Level II		Level II	
	UK	QAP	UK	QAP	UK	QAP
Sodium	140.0	140.0	122.0	122.0	111.5	112.4
Potassium	5.96	5.95	3.80	3.80	6.76	6.78
Urea	6.62	6.63	17.85	17.55	18.23	18.00
Glucose	4.72	4.68	11.10	11.05	13.21	13.14
Calcium	1.93	1.94	2.81	2.80	3.13	3.12
Phosphate	1.57	1.56	2.56	2.57	-	-
Iron	36.2	34.3	16.0	16.2	40.5	40.4
Urate	0.299	0.294	0.561	0.543	0.506	0.486
Creatinine	138	138	505	491	512	507
Bilirubin	25.2	25.2	87.4	86.3	60.8	59.5
Total protein	67.4	67.6	53.8	54.2	46.0	46.4
Albumin	42.7	43.1	36.5	36.6	28.4	28.5
Cholesterol	6.15	6.03	2.57	2.60	2.37	2.36
Magnesium	1.68	1.71	0.70	0.73	1.88	2.00
Osmolality	-	-	-	-	264	265
ALT	-	-	-	-	132	133
CK	-	-	-	-	585	577
ALP	-	-	-	-	481	461
Amylase	-	-	-	-	1106	1115

Table 5.7 Comparison of consensus values in the UKEQAS for General Clinical Chemistry with those in the Ortho QC Program (1979-1980) and the Wellcome Group QA Programme (1985). Units as in Table 5.5

	Level I		Level II			
	UK	Ortho	UK	Ortho	UK	Wellcome
Sodium	139.6	139.6	152.8	152.5	150.3	150.3
Potassium	4.2	4.21	7.25	7.26	4.48	4.48
Chloride	102.4	102.0	118.6	118.1	109.7	110.0
Urea	5.5	5.42	19.17	18.61	9.26	9.23
Glucose	4.3	4.53	17.46	17.45	9.46	9.57
Calcium	2.33	2.35	3.27	3.27	2.08	2.09
Phosphate	1.12	1.23	2.38	2.49	1.71	1.70
Iron	20.3	20.2	49.1	47.5	34.5	33.8
Urate	0.273	0.271	0.591	0.563	0.298	0.298
Creatinine	85	89	809	796	212	212
Bilirubin	8.1	8.5	103.0	96.3	39.4	39.8
Total protein	60.6	61.4	54.8	55.8	75.7	75.5
Albumin	37.7	37.4	33.9	33.1	-	-
Cholesterol	4.0	3.98	3.86	3.67	3.97	3.96
Lithium	-	-	-	-	0.90	0.90
Magnesium	-	-	-	-	1.13	1.13
Osmolality	-	-	-	-	313	313

5.4.1 Definitive methods

Several constituents of a tentative reference material prepared at CDC for WHO were analysed by definitive methods at the US National Bureau of Standards (NBS), and the assessment of the material also included UKEQAS distribution in 1977 (Appendix III.1.3). The values obtained are summarised in Table 5.8. Close agreement is demonstrated between NBS definitive method values and the consensus values for reliable method principles, supporting their validity for these simple constituents.

The reliable comparative data (ie those obtained using liquid human sera) for the UKEQAS for General Clinical Chemistry are now historical, since lyophilised materials have been distributed since the mid-1970s. A more recent comparison with definitive methods was, however, undertaken on a specimen of lyophilised equine serum (Appendix III.1.3) as part of a study of value assignment procedures (Eldjarn and Broughton, 1985). The results, summarised in Table 5.9, showed generally good agreement but anomalies were noted for sodium and cholesterol. The authors suggested that the data should be interpreted with caution, since the definitive method assays were carried out in one laboratory only and without external validation. More detailed examination showed a range of transferred values for sodium of 134.6 - 138.5 mmol/L (Eldjarn and Broughton, 1985), and the cholesterol concentration is low with the absolute deviation being only 0.14 mmol/L.

Studies on other analytes have revealed great discrepancies between definitive method values and EQAS consensus values, and steroid hormone assays (Siekmann and Breuer, 1982; Gaskell et al,

Table 5.8 Comparison of consensus values in the UKEQAS for General Clinical Chemistry with NBS definitive method values for WHO Experimental Reference Serum, lot 4976, 1977. Units as in Table 5.5

	Definitive	UKEQAS		
		Overall	Method-related	
Sodium	142.2	142.2	142.4	AutoAnalyzer flame
			142.3	IL flame
			141.9	Corning-EEL flame
Potassium	4.79	4.83	4.85	AutoAnalyzer flame
			4.82	IL flame
			4.81	Corning-EEL flame
Chloride	100.2	99.5	99.7	AutoAnalyzer
			99.4	Coulometry
			99.1	Titrimetry
Urea	5.76	5.96	5.82	AutoAnalyzer DAM
			6.43	Manual urease
Glucose	5.40	5.58	5.56	AutoAnalyzer GOD
			5.45	Manual GOD
			5.30	Beckman
			5.50	o-Toluidine
			5.93	AutoAnalyzer reduction
Calcium	2.27	2.27	2.29	AutoAnalyzer CPC
			2.25	Atomic absorption
			2.26	Vickers 300 CPC
			2.24	MTB
			2.23	EDTA titration
Urate	0.273	0.290	0.294	AutoAnalyzer chemical
			0.283	Manual chemical
			0.277	Manual uricase

Table 5.9 Comparison of consensus values in the UKEQAS for General Clinical Chemistry obtained on Seronorm, lot 158, with definitive method values and with values transferred from NBS SRM909. Units as in Table 5.5; transferred values defined by Broughton and Eldjarn (1985)

	UKEQAS	Definitive	Transferred
Sodium	137.5	135.8	-
Potassium	4.56	4.50	-
Chloride	107.7	108.0	108.5
Glucose	5.00	4.96	4.93
Calcium	2.71	2.76	2.74
Urate	0.464	-	0.462
Cholesterol	2.29	2.15	2.15
Magnesium	0.88	0.86	0.88

1984; Groom, 1985b) provide an excellent example of this situation. Here the underlying problem is one of method non-specificity, so that the mean of results obtained using relatively unsatisfactory methods cannot be expected to approach the true analyte concentration in specimens, as estimated by reference or definitive method values. Provision of such information in reports, rather than its application in scoring of performance, can stimulate method improvements and a consequent drift in consensus values towards this reference point, ie towards 'truth'.

5.4.2 DGKC reference methods

Materials for use in GFR have values assigned by the Central Reference Institution of the DGKC; such values are also used for performance assessment in the NEQAS in GFR (Bundesärztztekammer, 1971; Stamm, 1975). Values are obes are also used for performance assessment in the NEQAS in FRG (Bundesärztztekammer, 1971; Stamm, 1975). Values are obtained using approved reference methods, and in combination with a single type of QCM provide a stable baseline for assessment of any drift in consensus values.

Batches of Roche Diagnostica Control Sera N and P have been distributed regularly through the UKEQAS in addition to having values assigned by DGKC. Table 5.10 summarises the comparison of these results over 10 years (Appendix III.1.3). Minor deviations are shown, eg for iron, bilirubin and total protein, reflecting differences in the method principles used in combination with possible deficiencies in the QCMs. The relationships show no consistent trends or changes, apart from iron and creatinine (for which the reference method principle was changed) in 1986.

5.5 Systematic comparisons between schemes

Table 5.10 Comparison of consensus values in the UKEQAS for General Clinical Chemistry with reference method values assigned by the DGKC to Roche Control Sera N and P, 1977 - 1986. Average difference UKEQAS - DGKC (* denotes difference for one batch only); units as in Table 5.5

	Average difference									
	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986
n (batches)	4	2	2	2	0	4	4	3	5	5
Sodium	-0.2	-0.8	1.8	1.4		-0.7	0.1	0	0.8	0.5
Potassium	0.01	-0.14	-0.02	-0.05		-0.08	0	0.02	-0.30	-0.05
Chloride	0.1	-1.0	-1.0	2.25		0.2	0.8	0	1.1	1.1
Urea	0.12	-0.11	-0.25	0.19		-0.34	-0.47	-0.02	0.01	-0.31
Glucose	-0.06	0.12	0.07	0.30		0.10	0.19	0.15	0.29	0.46
Calcium	0	0.08	0.21	0.08		0.03	0	0.04	0.04	0.08
Phosphate	-0.03	-0.16	-0.15	-0.04		-0.03	-0.03	-0.07	0.01	0.01
Iron	3.0	1.7	2.1	5.0		5.5	5.7	6.0	3.6	-2.6
Urate	0.035	0.029	0.016	0.013		0.032	0.014	0.007	0.019	0.003
Creatinine	-1.5	6.0	5.0	-0.1		0.8	1.3	1.2	4.1	37
Bilirubin	3.7	4.4	3.2	2.1		-0.3	4.0	2.9	3.3	0.9
Total protein	-0.5	0.5	-0.6	1.7		0.8	-1.6	-1.3	-3.0	0.7
Cholesterol	0.12	-0.10	0.20	0.24		0.16		0.02	-0.02	
Lithium						-0.07	-0.08*	-0.05	-0.06	
Magnesium						0.02*	-0.10	-0.03	-0.04	

In these cases the schemes distributed specimens which had previously been distributed in the UKEQAS for General Clinical Chemistry (Appendix I.2.1), and consensus values could thus be compared for all distributions (Appendix III.1.4). In such situations, with participants striving to attain the UKEQAS consensus values used for performance assessment, good agreement might be anticipated but would not be automatic in view of the wide spread of results (see section 5.2.2 above).

5.5.1 International EQAS

Comparison based on 20 distributions yielded good agreement, summarised in Table 5.11. The lower values for glucose and bilirubin suggest possible deterioration in transit or during handling within participant laboratories. The variability in relationships appears to reflect the greater variability of consensus values in the IEQAS (Table 5.3).

5.5.2 Middle East EQAS

Table 5.12 summarises the good agreement also seen for most analytes in the 20 distributions studied (Bacchus et al, 1987). Here features similar to those described above for the IEQAS were also seen. Slight differences were seen for phosphate, iron, bilirubin and enzymic cholesterol, though there were no significant relationships with analyte concentration.

5.6 The WRL International Intercomparison Scheme (WIIS)

This scheme was established in 1984 with the aim of obtaining a more objective assessment of whether biases did exist between established national EQASs, and of the magnitude of any such biases. The principle of the scheme was to use a small number of participants from each country to reflect the accuracy base of

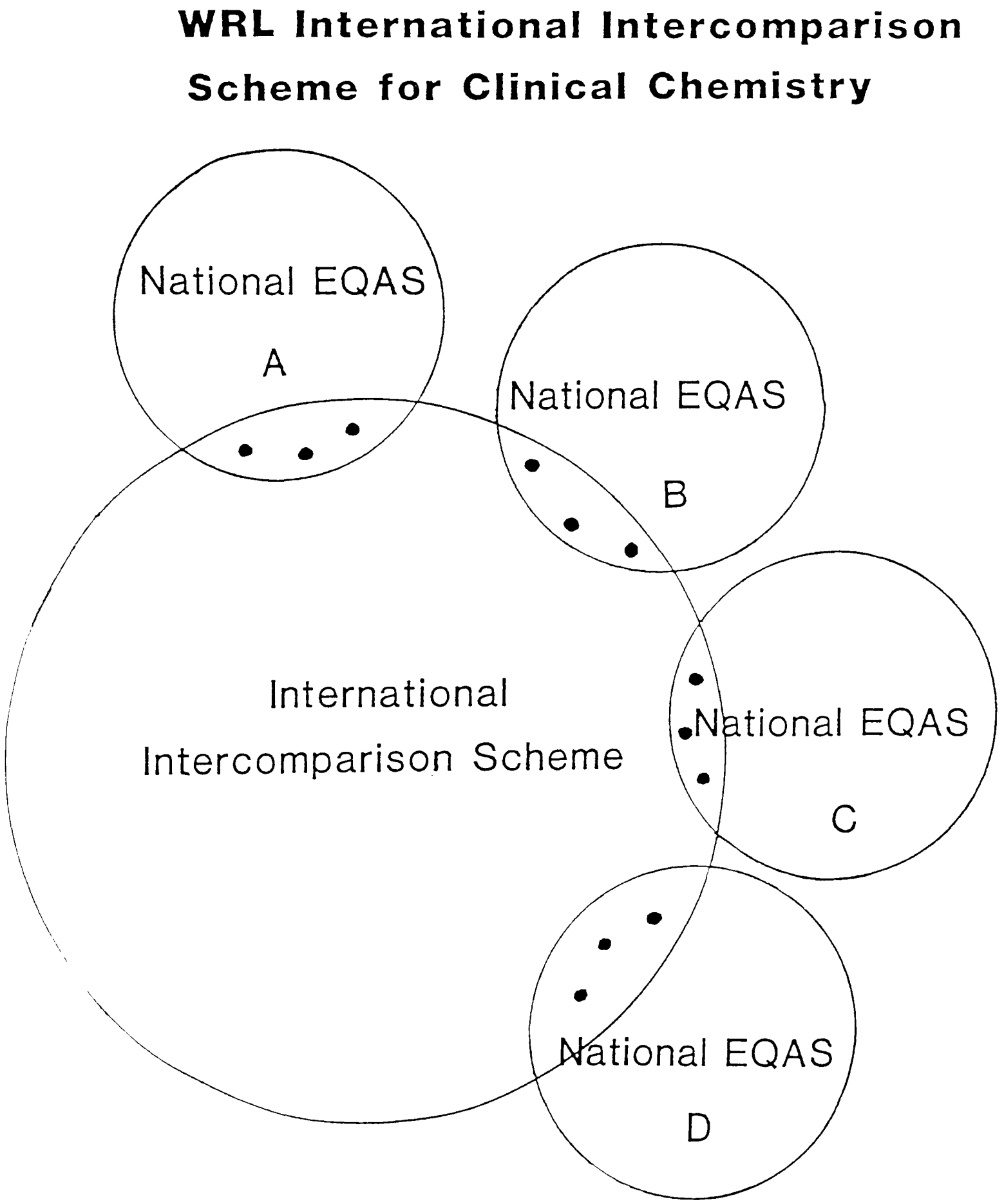
Table 5.11 Comparison of consensus values in the International EQAS with those in the UKEQAS for General Clinical Chemistry, 1985 - 1987. n=20 distributions unless specified

	Ratio IEQAS/UKEQAS (%)		
	Average	**SD**	**Range**
Sodium	99.4	0.5	98.8-100.2
Potassium	99.0	0.9	97.3-100.0
Chloride (n=18)	100.2	1.1	98.7-103.1
Urea (n=16)	99.2	4.0	89.2-104.1
Glucose	95.8	2.8	88.9-102.4
Calcium (n=19)	98.7	2.0	95.5-103.9
Phosphate (n=19)	98.8	2.5	94.2-102.7
Iron (n=19)	98.3	5.6	92.1-108.9
Urate (n=16)	99.2	2.4	96.3-105.2
Creatinine	98.5	3.4	93.4-105.8
Bilirubin	95.5	3.8	89.3-108.2
Total protein	101.0	1.3	98.8-102.8
Albumin (n=5)	99.4	2.1	96.9-101.4
Cholesterol	102.4	2.1	99.2-107.8
AST (n=6)	104.8	5.9	97.6-114.4
ALP (n=2)	103.1	-	98.1;108.0

Table 5.12 Comparison of consensus values in the Middle East EQAS with those in the UKEQAS for General Clinical Chemistry, September 1983 - May 1985. n=20 distributions unless specified

	Ratio MEEQAS/UKEQAS (%)		
	Average	**SD**	**Range**
Sodium			
Flame photometric	99.4	1.0	97.1-102.1
Indirect ISE	100.1	0.4	99.2-100.5
Potassium (n = 19)			
Flame photometric	99.1	0.7	97.3-100.3
Indirect ISE	99.6	0.8	97.5-101.0
Chloride (n = 18)			
Colorimetric	101.1	0.6	100.2-102.3
Electrometric	100.3	1.0	98.4-102.4
Urea			
Diacetylmonoxime(n=19)	97.9	3.5	93.8-110.2
Urease (n=18)	98.3	1.1	96.5-100.4
Glucose			
Enzymic	97.9	1.8	95.4-101.9
Reduction (n=15)	93.4	4.8	78.6-101.2
Calcium			
Dye binding	99.7	0.8	98.1-100.9
Phosphate			
Colorimetric	103.5	1.9	99.3-108.1
Iron (n = 18)			
Dye binding	95.6	3.3	88.9-99.9
Urate			
Colorimetric	98.3	3.8	86.6-101.5
Uricase	99.6	1.7	97.0-101.8
Creatinine			
Jaffe (non-kinetic)	99.4	2.9	94.9-107.1
Bilirubin (n = 17)			
Diazotisation	96.6	4.4	84.9-102.6
Total Protein			
Biuret	99.3	2.3	93.1-104.3
Albumin			
BCG (n = 5)	101.5	1.3	100.3-103.5
BCP (n = 3)	106.7	-	103.6-110.7
Cholesterol			
Enzymic (n=18)	96.8	1.6	93.6- 99.0
Chemical (n=17)	99.5	5.7	90.8-110.1

Figure 5.1 Schematic design for WRL International Intercomparison Scheme (WIIS), with several repesentative laboratories from each national EQAS participating. See section 5.6.1 for explanation



their national scheme (Figure 5.1).

5.6.1 Principles and establishment

These laboratories would be selected on the basis of regular participation in their NEQAS (to ensure complete returns in the WIIS), and of consistent performance. Thus, they should be among the 'best' participants in their national scheme over a period of years, and ideally with small bias; to facilitate communication these would normally include the scheme organiser's laboratory. Depending on scheme size, between 3 and 5 laboratories per scheme was felt to be appropriate for the initial stages of the WIIS, which would in effect be a feasibility study (Figure 5.1). Of course, with such a small number of laboratories they could not be assumed to be fully representative of their national EQAS for all analytes. It would therefore be necessary to obtain data regarding biases relative to their own scheme to draw reliable conclusions.

A monthly distribution frequency was chosen. To minimise postal delays sufficient materials for a six-month cycle were despatched together, as in the IEQAS and MEEQAS. Participants would be requested to report the mean of replicate assays, preferably from separate analytical batches, to reduce the effects of within-laboratory imprecision.

The design depended on the cooperation of national EQAS organisers in selecting suitable laboratories from their scheme, in redistributing specimens to them, and in subsequently providing performance data on these laboratories. Initially 33 scheme organisers were contacted, of which 28 agreed to participate from the inception of the WIIS. A further two participated later; most of these represented situations where

responsibility for scheme organisation had been transferred, though no response was obtained from the three others. Participation in the scheme was anonymous, and each country was therefore identified only by a country code; countries are listed geographically in Table 5.13.

5.6.2 Experience with scheme operation

The return rate from almost all laboratories was satisfactory. Any results received after the deadline for the main computer run were assessed and the files updated before results for the following distribution were processed.

The reports to participants showed the overall and recalulated (after truncation at $\pm$3 SD) statistics for all laboratories, and the percentage deviations from the recalculated mean for the laboratories in their own country. The individual results returned were not given, in an attempt to reduce the tendency for the participants to 'adjust' their methods to agree with the international consensus obtained through the WIIS. The purpose of this scheme thus differed fundamentally fom that of most EQASs.

Following pages showed the average percentage deviation for each analyte for the laboratories in each country, and the cumulative average percentage deviation for each analyte for each country (Appendix I.4). This latter figure was the average of all the individual percentage deviations obtained by laboratories from that country over the four most recent distributions; ie with three laboratories the maximum number of contributory percentage deviations would be 12.

The criteria for assessment were similar to those used for VIS calculation in the UKEQAS. Thus, no percentage deviations were

Table 5.13 Countries with NEQASs participating in the WRL International Intercomparison Scheme

Africa and Middle East

Saudi Arabia
South Africa

Americas

Canada
United States of America (CDC)
United States of America (CAP)
United States of America (New York State)

Asia

Burma
China (Peoples Republic)
India (Chandigarh)
India (Vellore)
Indonesia
Japan
South Korea
Thailand

Australasia

Australia
New Zealand

Eastern Europe

Bulgaria
Czechoslovakia (Czech SSR)
Czechoslovakia (Slovak SSR)
Germany (Democratic Republic)
Poland
Union of Soviet Socialist Republics

Western Europe

Belgium
Denmark
France
Germany (Federal Republic)
Netherlands
Spain
Switzerland
United Kingdom

calculated for very high or very low analyte concentrations, nor where the specimen would not necessarily yield a reliable assessment, eg for albumin on specimens of non-human origin or for urea on specimens containing Tris buffer, a urease inhibitor causing low results in some urease-based procedures. In the initial stages of scheme operation no account was taken of the methods used by individual participants. Provision was, however, made for its subsequent inclusion, and details of the methods and instrumentation used were obtained from participants.

Though facilities were offered for laboratories to return results in either SI or 'conventional' (mg/dL) unit sets, with reports in SI units, some participants initially returned grossly discrepant results. This was particularly marked for magnesium and calcium, due possibly to the use of mEq/L by some laboratories. The return rate was also initially disappointing. For these reasons the results of the first two distributions were omitted from data analyses.

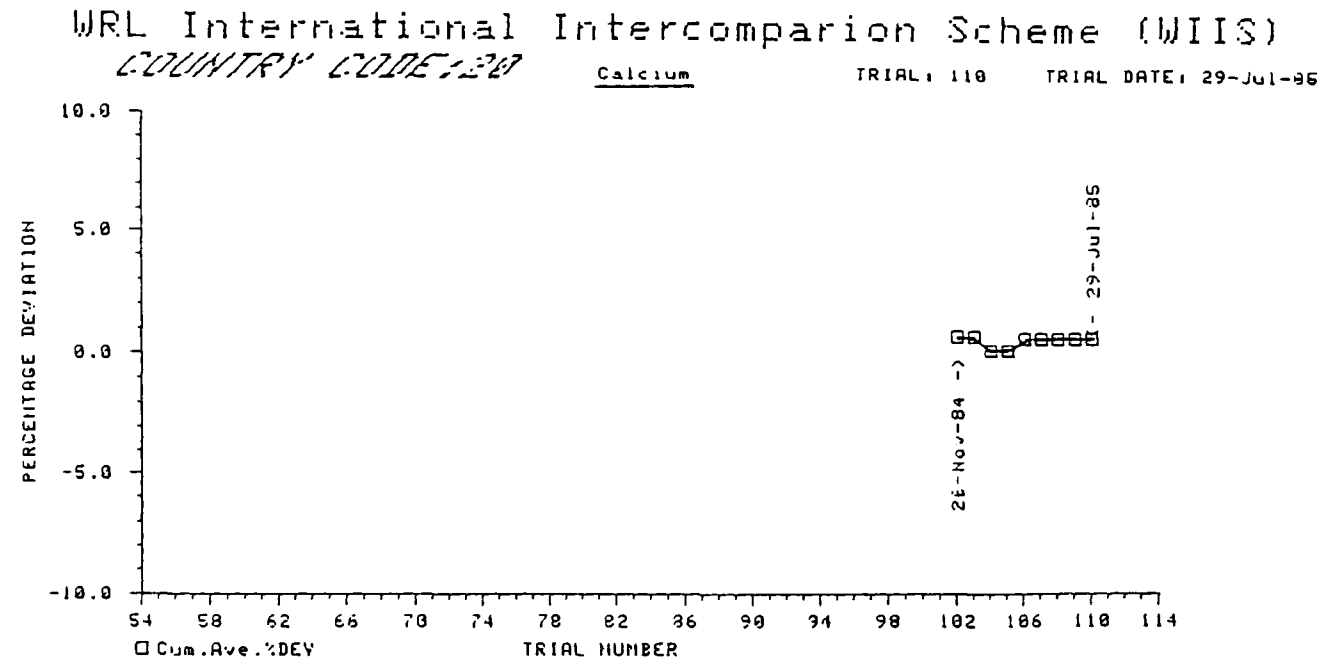
5.6.3 Comparison of results

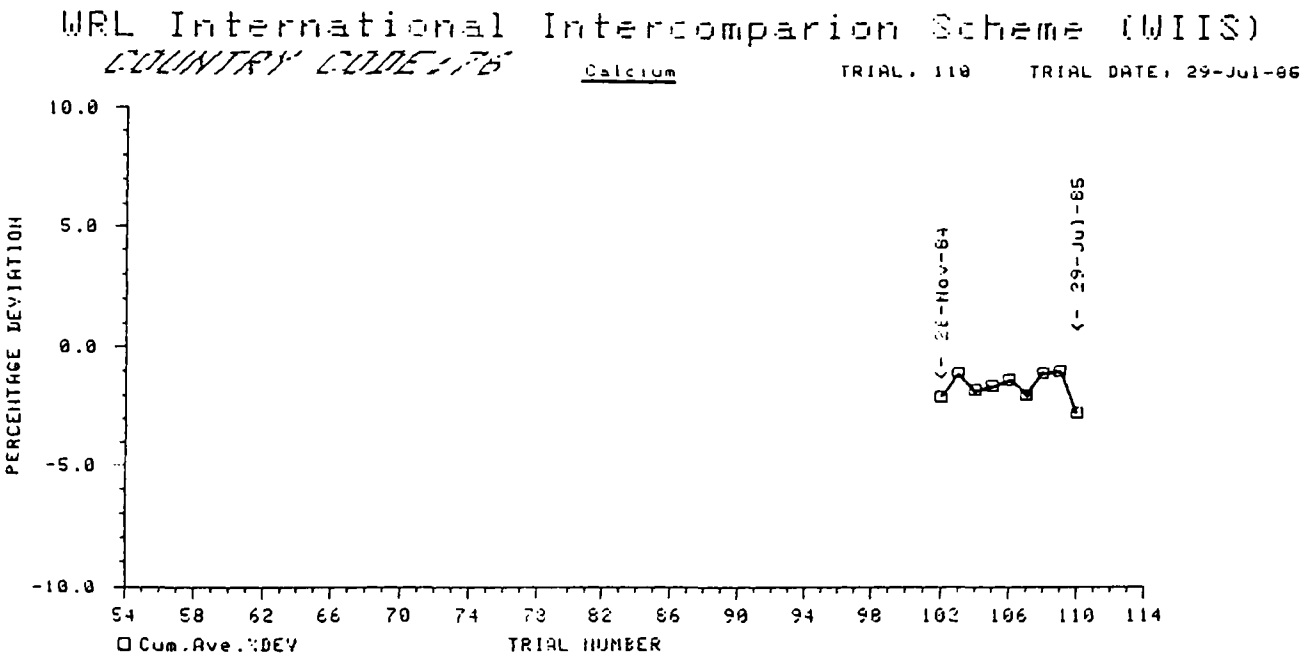
Thus distributions 101-112, being the first reliable year's data (September 1984 to August 1985), formed the basis of assessment. Table 5.14 gives the average percentage deviations for each country and analyte over these distributions. These were derived by taking the mean of the cumulative average percentage deviations at distributions 104, 108 and 112; each figure would be derived from 36 individual determinations for a country with three participants returning results reliably. As shown for calcium and countries 20 and 76 in Figure 5.2, these average percentage deviations were stable in most cases.

Table 5.14 Cumulative average percentage deviation from WIIS consensus in Trials 101-112. Asterisks indicate data from countries using predominantly chemical procedures (for urea, glucose, urate and cholesterol)

Country	Na	K	Cl	Urea	Gluc	Ca	Phos	Fe	Urate	Creat	Bili	TP	Alb	Chol	Li	Mg
20	0.2	0.37	-0.2	0.93*	2.22	0.09	-1.88	2.0	4.17*	1.2	1.07	-0.17	-1.23	-2.83	2.33	-1.67
21	0	0.8	-1.4	-0.27	-1.14	-1.34	-0.57		-3.23	-3.5	0.17	-1.97	-3.97	-3.03	0.23	-1.27
22																
23	0.57	0.53	-1.63	-3.37	-0.7	0	-2.0	0.6	0.43	-2.73	-5.83	1.0	12.87	-2.0	0.6	2.77
24	-0.37	-0.85	-2.87	-6.0	-0.77	0.05	1.0	0.4	-6.77	-3.73	-2.53	5.87	-3.7	-3.0	-0.03	-0.7
25	0.73	1.5	-0.13	0.23*	0.73	-0.5	-1.03	3.53	-5.67	-3.93	-5.17	0.13	1.53	-1.73	-3.37	8.2
26	0.73	0.8	-0.6	1.23*	1.93	0.97	-0.97	0	-5.33	-2.2	0.67	-1.6	5.03	-3.27	-2.4	-1.83
27	-0.1	-0.7	-0.5	-0.7	-0.9	-0.8	-0.1	-5.6	2.1	-2.5	4.6	-2.9	-6.7	-5.5	-3.6	3.8
28	-0.8	0.27	0.87	1.03	-2.83	0.27	1.37	-4.65	1.67	-0.83	-5.93	0.3	2.3	-2.37	-5.87	-3.43
50	0.07	0.6	0.4	-3.13	-2.73	-0.07	2.57	-0.9	-2.5	-2.17	0.3	-1.93	2.5	7.9*	-2.5	4.13
51	-0.37	-0.07	-0.17	-2.77*	-2.87	0.47	1.13	5.1	1.13	-3.6	-3.13	1.0	-0.50	3.37*	-1.03	-5.57
52	-0.3	-0.1	-1.53	-1.4	2.13*	-2.37	3.4	8.27	-2.9*	0.87	1.07	1.77	1.07	4.03*	-5.63	11.47
53	-0.77	0.033	0.5	4.93	-4.47	7.73	7.63	-2.33	-1.2*	14.1	2.1	0.5	6.7	5.63	4.1	0.93
54	0.17	0.23	0.13	0.5	1.1	-0.53	3.53	5.13	-1.2*	0.77	2.37	2.0	-0.9	4.97*	-2.5	5.1
55	0.47	-0.1	-2.7	1.6*	1.53*	-2.53	1.8	10.73	2.0*	-4.8	-1.4	-2.6	-0.57	12.87*		-45.4
70	0.43	-0.53	0.83	0.3*	1.07	-2.87	-0.7	21.63	7.27*	-0.3	3.73	0.23	3.37	12.29*	2.86	34.37
71	0.57	-1.13	-0.13	2.03*	4.83*		1.87		4.6*		15.3	-2.07	1.17	8.1*		
72																
73	0.5	-2.4	0.5	-6.83*	-2.47*	0.73	-0.5		2.4*	4.63	-2.07	3.87	4.83	3.13*		
74	-3.1	-3.47	0.97	1.27*	-3.27*	-0.1	2.73	14.5	0.9*	1.93	-8.77	-0.53	0.47	2.97*	2.73	
75	0.27	0	2.07	0.8*	0.43	-1.5	1.2	-7.63	0.87*	4.0	-0.8	1.0	4.57	7.47		25.67
76	0	0.4	-3.1	1.33	2.1	-1.77	-0.3	1.87	0.8*	-2.63	0.9	0.38	0.77	1.63	2.7	9.27
77	-2.77	1.43	2.83	2.67	-8.6	1.37	3.9	6.87	-7.23	1.03	-0.87	-1.47	-4.97	-6.0		-7.1
78																
79	-0.23	0.17	0.23	0*	-1.3	0.7	-2.6	-4.17	0.77	2.63	4.67	0.37	1.03	0.83*	0.2	-2.53
80	0.1	0.47	-0.13	2.03*	2.2	0.8	-2.23	-1.4	2.7*	-0.17	-0.53	0.7	-4.23	-3.53	0.3	2.33
81	0.33	-1.13	0.6	1.3	-1.2	1.33	0.63	9.0	1.2	12.73	3.1	-2.13	-4.2	-8.6		-6.4
120	-0.66	-0.33	0.3	1.33	-0.8	0.57	-3.97	-0.9	3.6	0.2	3.07	-1.87	-5.83	0.17	-10.4	-4.03
121	-0.03	1.47	0.1	1.1*	3.7	-0.03	2.6	0.37	5.93*	9.43	7.8	2.1	1.83	-3.57	3.07	-19.8
160	0.4	-0.2	0.05	3.55*	2.5	0.75	1.45	2.15	3.55*	-1.0	-4.65	0.9	-1.05	1.3	0.7	2.1
161	0.8	1.05	0.25	-2.45	0.55	0.45	0.15	-0.75	1.15	2.25	0.15	0.55	0.5		-0.85	0.9
162																
163	0.4	-0.9	0.57	0*	-1.4	-0.97	-2.33	-2.67	-4.2	-0.43	-0.1	-0.63	-1.67	-8.7	-2.73	-8.13

Figure 5.2 Stability of cumulative average percentage deviations for calcium in WIIS





The data show good agreement among many countries for most analytes. Agreement appears excellent for electrolytes in particular, as shown by the average deviations (irrespective of sign) in Table 5.15. Method differences may have contributed to the residual differences for some analytes. For example, countries using predominantly chemical (indicated by asterisks in Table 5.14) or enzymic procedures for urea, glucose, urate and cholesterol assay could be expected to differ in bias. Little difference was shown for urea (+0.4% and -0.3%) and glucose (+0.6% and -0.5%), but urate (+2.1% and -1.5%) and cholesterol (+6.0% and -2.6%) showed the anticipated marked effects.

In general, considering only countries 20-28 and 79-163 gives the closest concordance. These represent Western Europe, North America, the Middle East, Japan, South Africa and Australasia, and their comparability thus appears better than that among schemes in Asia (70-78) and Eastern Europe (50-55). Within each of these groups, however, there are exceptions. These may be generalised (eg countries 50 and 81) or with respect to individual analytes only (eg bilirubin and albumin in country 23; cholesterol and magnesium in country 163), but inspection failed to suggest any reason for these.

Because the laboratories participating in the WIIS may not be fully representative of their national EQAS, the scheme organisers were each asked in August 1985 to provide data from their own scheme. The data requested were the average percentage deviations (for the 3-5 laboratories representing their scheme in the WIIS) over the same period (September 1984 to August 1985) relative to their national scheme. These would be relative to

Table 5.15 Overall average cumulative average percentage deviation from WIIS consensus in Trials 101-112, with corresponding data and average percentage deviation of national EQAS from WIIS consensus (Table 5.18) only for countries providing information on national performance

	All countries		Countries giving information		
	n	WIIS	n	WIIS	NEQAS v WIIS
Sodium	29	0.56	11	0.4	0.9
Potassium	29	0.76	11	0.8	1.4
Chloride	29	0.91	11	0.9	1.4
Urea	29	1.90	13	2.2	3.7
Glucose	29	2.15	14	2.0	3.3
Calcium	28	1.13	13	0.7	1.4
Phosphate	29	1.93	10	1.8	2.4
Iron	26	4.74	7	1.7	4.8
Urate	29	3.02	14	3.3	4.1
Creatinine	28	3.22	12	2.4	3.7
Bilirubin	29	3.20	13	2.5	4.8
Total protein	29	1.47	14	1.6	2.3
Albumin	29	3.11	10	1.9	3.3
Cholesterol	28	4.67	13	3.7	5.3
Lithium	23	2.64	6	1.4	3.3
Magnesium	26	8.17	8	3.0	5.1

designated values from reference laboratories, or to overall or method-related consensus values. These figures are given in Table 5.16. Unfortunately, responses were received initially from only very few organisers, and further requests were therefore sent. These emphasised that confidentiality agreements would not be broken by reporting average data for several laboratories, and clarified the situation where some of the WIIS participants were also reference laboratories for the national scheme. The response was still disappointing, with to date only 14 returns from the 29 schemes for which WIIS data were available.

These data on the bias relationship between WIIS participants and their national scheme can then be combined with the data from the WIIS to yield an assessment of the bias between each national EQAS and the international (WIIS) consensus:

National scheme relative to WIIS =

Average % deviation from WIIS - Average % deviation from NEQAS

This is exemplified for the UKEQAS in Table 5.17. Here for analytes such as sodium and potassium the two deviations cancel, giving better agreement between UKEQAS and WIIS consensus values, whereas for others such as chloride the resultant deviation becomes worse. The positive biases for urea and urate may be due to the then predominant use in the UK of chemical rather than enzymic procedures, with the reverse effect contributing to the apparent negative bias for cholesterol.

The resultant values for all schemes providing data on national scheme performance are given in Table 5.18. Unfortunately, the apparent divergence of many NEQASs from the WIIS consensus was greater than that of their group of 'representative' laboratories

Table 5.16 Percentage deviations of WIIS participants from their national EQAS

Country	Na	K	Cl	Urea	Gluc	Ca	Phos	Fe	Urate	Creat	Bili	TP	Alb	Chol	Li	Mg
20	0.21	0.39	0.19	-0.73	0.51	-0.66	-0.48	-1.37	0.11	-0.61	1.63	-0.53	1.01	-1.6	2.18	-0.26
21																
22																
23																
24	-2.5	-2.5	-0.02	-0.02	-0.73	2.05	3.6	6.2	-11.6	2.0	-0.3	3.8	-4.5	-9.8	-0.75	2.5
25	0.56	0.05	-0.37	0.46	-1.8	1.09	-0.33	2.83	-2.72	0.88	0.11	-0.43	-4.16	0.78		3.17
26	1.13	1.4	-1.22	-7.2	-0.5	-1.35	-5.85	2.28	-1.5	4.9	-0.3	1.1		0.7	-1.05	6.5
27																
28																
50	-1.39	-2.32	-2.12	-5.74	-6.7	4.26	-4.35	9.2	10.5	8.26	-8.31	-5.83	3.27	6.46	7.78	-10.7
51																
52	-0.72	-0.42	-0.62	1.67	0.45	-0.1	1.22		-2.04	1.42	0.89	0.5		-4.23		
53																
54																
55																
70																
71				4.43	3.61				1.1		8.03	-1.76	2.73	-5.25		
72																
73	0.78	0.2	1.57	-2.51	-5.25	0.58	2.12		3.2	5.64	-6.54	-0.7	3.01	7.39		
74																
75																
76	-2.3	-1.6	0.6	-1.9	12.7	-1.5	0.9		4.6	0.5	-0.8	0.1	6.0	4.1		
77				7.6	5.1	0.95			-0.73	0.25	2.9	-0.53		9.1		
78																
79	0.8	-1.0	1.3	2.35	-1.25	0	-1.5	2.25	1.45	-1.4	7.2	1.15	2.9	-0.25	-3.85	2.3
80																
81																
120	-0.25	-1.46	0.15	2.66	-2.44	2.13	-4.47		-0.49	2.05	-6.38	0.45	4.42	0.64		-2.42
121																
160																
161	-0.8	2.15	-4.75	7.33	-2.61	1.3		4.06	0.04	5.52	-15.57	4.33	2.8	-2.52	-4.2	-0.29
162																
163					-0.3	-1.6			3.6			7.3		-0.3		

Table 5.17 Comparison of performance in the UKEQAS for General Clinical Chemistry and the WIIS. *the 5 UK participants in the WIIS; see section 5.6.3 for explanation

	Percentage deviation		
	UK* v WIIS	**UK* v UKEQAS**	**UKEQAS v WIIS**
Sodium	+0.2	+0.2	0.0
Potassium	+0.4	+0.4	0.0
Chloride	-0.2	+0.2	-0.4
Urea	+0.9	-0.7	+1.7
Glucose	+2.2	+0.5	+1.7
Calcium	+0.1	-0.7	+0.8
Phosphate	-1.9	-0.5	-1.4
Iron	+2.0	-1.4	+3.4
Urate	+4.2	+0.1	+4.1
Creatinine	+1.2	-0.6	+1.8
Bilirubin	+1.1	+1.6	-0.6
Total protein	-0.2	-0.5	+0.4
Albumin	-1.2	+1.0	-2.2
Cholesterol	-2.8	-1.6	-1.2
Lithium	+2.3	+2.2	+0.2
Magnesium	-1.7	-0.3	-1.4

Table 5.18 Percentage deviation of national EQAS from WIIS consensus. Derived from cumulative average percentage deviations in WIIS (Table 5.14) and information from national EQASs (Table 5.16); see section 5.6.3 for explanation

Country	Na	K	Cl	Urea	Gluc	Ca	Phos	Fe	Urate	Creat	Bili	TP	Alb	Chol	Li	Mg
20	-0.01	-0.02	-0.39	1.66	1.71	0.75	-1.4	3.37	4.06	1.81	-0.56	0.36	-2.24	-1.23	0.15	-1.41
21																
22																
23																
24	2.13	1.65	-2.85	-5.98	-0.04	-2.0	-2.6	-5.8	4.83	-5.75	-2.23	2.07	0.8	6.8	0.72	-3.2
25	0.17	1.45	0.24	-0.23	2.53	-1.59	-0.7	0.7	-1.95	-4.81	-5.28	0.56	5.69	-2.51		5.03
26	-0.4	-0.6	0.62	5.97	2.43	2.32	4.82	-2.28	-3.83	-7.1	0.97	-2.7		-3.97	-1.35	-8.33
27																
28																
50	1.46	2.92	2.52	2.61	3.97	-4.33	6.92	-10.1	-13.0	-10.43	8.61	3.9	-0.77	1.44	-10.28	14.83
51																
52	0.42	0.32	-0.91	-3.07	1.68	-2.27	2.18		-0.86	-0.53	0.18	1.27		8.26		
53																
54																
55																
70																
71				-2.4	1.22				3.5		7.27	0.31	-1.56	13.35		
72																
73	-0.28	-2.6	-1.07	-4.32	2.78	0.15	-2.62		-0.8	-1.01	4.47	4.57	1.82	-4.26		
74				-4.93	-13.7	0.42			-6.5	0.78	-3.77	-0.94		-15.1		
75																
76	2.3	2.0	-3.7	3.23	-10.6	-0.27	-1.2		-3.8	-3.13	1.7	0.28	-5.23	-2.47		
77																
78																
79	-1.03	1.17	-1.07	-2.35	-0.05	0.7	-1.1	-6.42	-0.68	4.03	-2.53	-0.78	-1.87	1.08	4.05	-4.83
80																
81																
120	-0.41	1.13	0.15	-1.33	1.64	-1.56	0.5		4.09	-1.85	9.45	-2.32	-10.25	-0.47		-1.61
121																
160																
161	1.6	-1.1	5.0	-9.78	3.11	-0.85		-4.81	1.11	-3.27	15.72	-3.78	-2.3		3.35	1.19
162																
163					-1.1	0.63			-7.8			-7.93		-8.4		

participating in the WIIS, demonstrated in the average deviations in Table 5.15. One effect which may have contributed to this would be a misunderstanding on the part of some NEQAS organisers; thus they may have provided data with incorrect sign (ie the bias of the scheme relative to the group of laboratories participating in the WIIS), or data for their own laboratory alone.

An alternative contributory effect could have been that the laboratories selected were not in fact representive of their national scheme. This might reflect a failure in the initial choice (organisers may have selected the most technologically advanced and outward-looking laboratories, whose performance could reflect this) or, more importantly, an unwanted 'success' of the scheme if participants used it to approach more closely the international consensus at the expense of agreement in their NEQAS.

5.6.4 Appraisal of the scheme

The study showed that such a scheme design is practicable. NEQAS participants could be selected by their national organiser, and returned results reliably for the specimens distributed. Some organisers expressed doubts regarding the validity of the design, and several were only willing to cooperate for a limited period.

Is the use of the WIIS consensus value as reference point valid? Firstly, this was intended only as a stable yardstick to which countries' performance could be related, rather than being claimed to be accurate and reflect the true value. Experience (see section 5.2 above) had indicated consensus values to be highly reproducible within schemes, borne out by differences averaging <1% for repeated distributions (101/109 and 102/108) of

the same material, so the stability condition should have been satisfied. Secondly the WIIS incorporated NEQASs with very different approaches to determining target values (see section 3.4.3), including the use of consensus values and of reference laboratory values; many scheme organisers would have satisfied themselves that their approach yielded accurate and reproducible targets. Such heterogeneity would be unlikely to lead to a biased consensus value from the WIIS. Apart from cholesterol, which showed method-related effects, the close agreement demonstrated in Table 5.19 suggests that it should also be close to the true value. These data may be compared with average scores in the UKEQAS given in Table 4.3, confirming the initial impression that CVs were only one third higher than in the UKEQAS.

Inter-country agreement appeared from the WIIS data (Table 5.14) to be quite good, particularly for the most mature assays and the more developed countries. The effects of analytical method seemed to be confined to urate and cholesterol assay, and it would be advisable to take the method used into account for at least these analytes in any more detailed further study. The more definitive procedure of also making allowance for biases of the WIIS laboratories relative to other participants in their NEQAS (Tables 5.16 and 5.17) confirmed UKEQAS consensus values to be close to the international consensus.

The effects shown in Tables 5.15 and 5.17 indicate, however, two failings in the scheme. One appears to reflect the reluctance of NEQAS organisers to divulge performance data, which might conceivably be overcome with more intensive contact. The other is more serious (unless it simply represents poor initial selection of laboratories) in that participants may have misunderstood, or

Table 5.19 Average VISs in the WIIS for Trials 101-112, with ratio to those in UKEQAS for General Clinical Chemistry (Table 4.3)

	Average VIS	Ratio to UKEQAS
Sodium	86	1.40
Potassium	78	1.41
Chloride	95	1.23
Urea	84	1.52
Glucose	56	1.27
Calcium	78	1.22
Phosphate	69	1.27
Iron	59	1.29
Urate	84	1.31
Creatinine	72	1.59
Bilirubin	48	1.29
Total protein	88	1.33
Albumin	76	1.44
Cholesterol	95	1.98
Lithium	47	1.37
Magnesium	82	1.50

ignored, the WIIS's objective and used it successfully as a conventional EQAS.

This latter effect could be beneficial to patient care, since results on patients served by these laboratories would be in better agreement with the international consensus and thus more accurate, and such a change in bias could stimulate a more general movement in the country towards greater accuracy. On the other hand there would be worse interlaboratory agreement within the NEQAS, which is less desirable. These short-term effects seem, however, to be the cost of progress, and the long-term outcome would outweigh this temporary perturbation.

The objective of the WIIS was assessment of the situation rather than correction of any between-country discrepancies, so this should be classed as a partial failure rather than a success of the scheme. Nevertheless it appears to indicate the remarkable power of EQA in influencing laboratories' practice, even where this was not intended.

5.7 Summary

Consensus values are highly reproducible on repeated distribution of the same material through an EQAS. Consensus values obtained in different schemes show close agreement, whether the schemes are dependent or independent.

Comparisons using liquid human sera show close agreement of consensus values with definitive methods for several analytes in the UKEQAS for General Clinical Chemistry. Study of values assigned using reference methods to a single manufacturer's materials confirm that consensus values in the UKEQAS for General Clinical Chemistry have not been subject to any 'drift' away from

accuracy.

Feasibility studies in the WIIS show that a small group of laboratories which perform well in their own national EQAS can be used to reflect the accuracy base of their scheme in international comparison. In most cases good agreement was found, though as expected greater variability was found for developing than for the most developed countries. UKEQAS consensus values were confirmed to be in good agreement with the international consensus.

Attempts to adjust this comparison according to these laboratories' bias in their NEQAS led to apparent divergences in some cases, however, due perhaps to difficulties of interpretation. Alternatively results from these laboratories may be more accurate than is general within their country, due in part to an (unwanted) influence of participation in the WIIS on their performance.

Chapter 6:

THE EFFECTS OF ANALYTE LEVEL ON INTERLABORATORY AGREEMENT

6.1 Introduction

For a comprehensive assessment of performance, EQASs must distribute specimens with analyte concentrations covering the range encountered clinically. The scheme design must therefore take into account any effect of analyte level on interlaboratory agreement, or misleading information may be generated.

If interlaboratory agreement varies substantially with analyte level it may not be valid to combine performance estimates obtained at different levels, or such combination should be confined to concentration bands. A curvilinear relationship between CV and analyte level is normally seen in the 'precision profile' from IQC data (eg Jeffcoate, 1981; Ekins, 1983) and would be expected in EQA data. This has, however, never been fully demonstrated for the organic and inorganic constituents of serum, though Roehle and Voigt (1986) have recently published preliminary data for three analytes.

Such information has policy implications for organisers of schemes such as the UKEQAS, eg in the selection of analyte levels for survey and scoring. The question is indeed of particular importance for the UKEQAS for General Clinical Chemistry and its sub-schemes, since VI scoring (see Appendix II and Chapter 4) assumes that the interlaboratory CV is effectively independent of concentration over the range surveyed.

6.2 Studies in the UKEQAS for General Clinical Chemistry

6.2.1 Study design

This study was undertaken as the first stage of an investigation of the effects of species of origin and of manufacturing technique on interlaboratory agreement, as discussed in Chapter 14. Prior inspection of data from the scheme had shown little fluctuation from distribution to distribution in the spread of results provided that materials distributed were of satisfactory quality. This reproducibility, due perhaps to the relatively large number of participants (about 400 results for each distribution), is discussed in Chapter 5.

A two-year period, covering 40 distributions, should have yielded a sufficiently large database for assessment without excessive complications from improvement in participants' performance. The study period was chosen such that the method classification used was unchanged throughout the period, though the CCVs for lithium and magnesium were not established until part way through.

A similar two-year period of data (but with a slightly different method classification) was also available to validate the conclusions. This was essential because previous work (eg Wilding et al, 1979) had demonstrated the importance of testing findings on an independent set of data.

The study design included examination of several indicators of performance (Table 6.1) in two years' data from the scheme. To characterise any relationships, graphical presentation against the truncated mean (Appendix I.1; the most reliable estimate of analyte level) of all performance indicators was chosen.

Appendix III.2.1 gives full details of the study. As described

Table 6.1 Indicators of performance used in study of relationships between analyte level and interlaboratory agreement

Indicators of overall performance:

- overall CV
- recalculated CV
- average Variance Index Score (VIS)

Indicators of discrepant performance:

- number of results excluded in the truncation
- number of VISs greater than 200
- number of VISs greater than 300
- number of VISs of 400
- percentage of results excluded in the truncation
- percentage of VISs greater than 200
- percentage of VISs greater than 300
- percentage of VISs of 400

Indicators of method performance:

- recalculated CV
- average VIS
- difference of recalculated method mean from recalculated mean
- percentage difference of recalculated method mean from recalculated mean

therein, many of the indicators appeared unhelpful, as overall CVs on untrimmed data are more likely to be influenced by gross errors ('blunders') on the part of individual participants than by any underlying relationship with analyte level. Examination of the recalculated CV and the average VIS therefore constituted the main part of the study. Data classified according to method were examined for a more limited range of analytes, to confirm that the conclusions drawn also held for individual method groups.

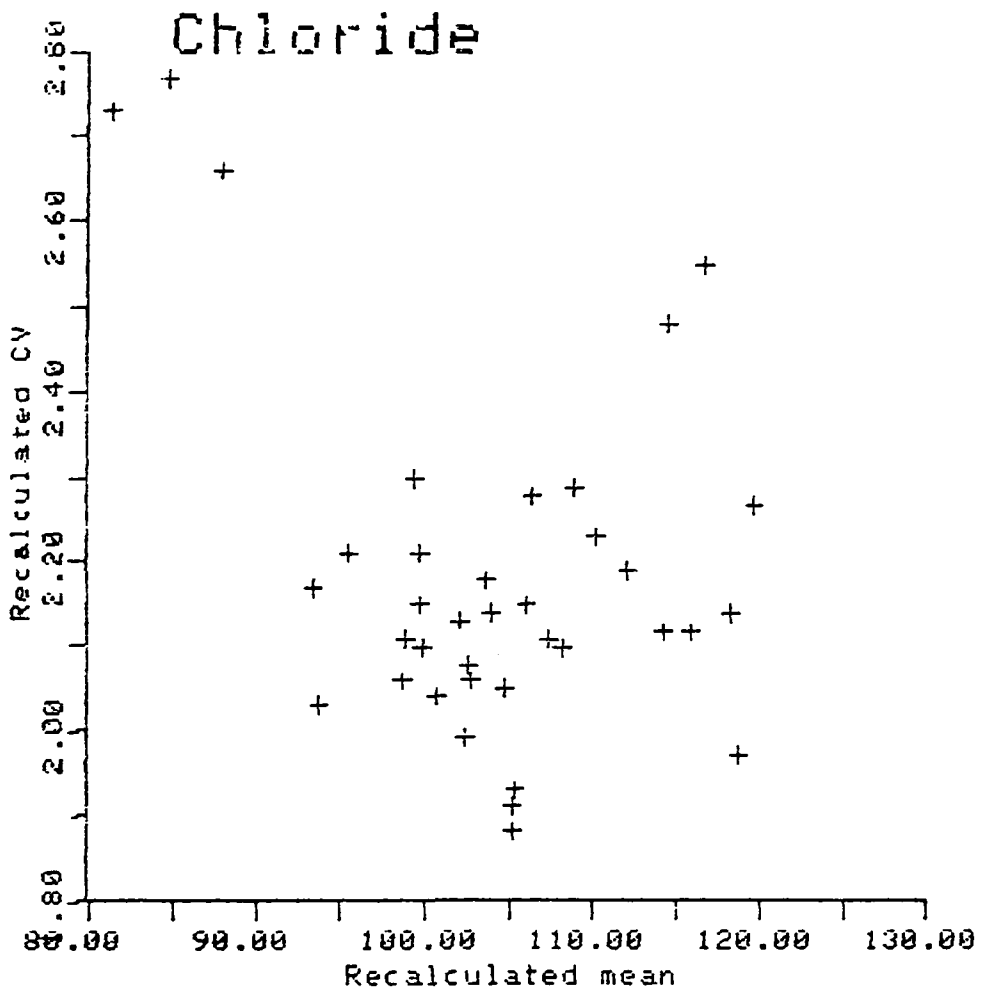
6.2.2 Overall data

In all cases some variability in interlaboratory agreement, as reflected by recalculated CV or average VIS, was found. For some analytes, eg chloride and iron, no consistent relationship of the recalculated CV or average VIS with analyte level was discernible (Figure 6.1). For others such as urea and bilirubin, however, definite level-dependence lay beneath an apparently random variation in the indicators.

In these cases interlaboratory agreement deteriorated rapidly as levels approached and decreased through the reference interval (Figures 4.1 and 6.2). The patterns for glucose, creatinine and cholesterol were similar. Apart from cholesterol at levels below 2.5 mmol/L, average VISs did not usually exceed 100, with the best average VISs attained being around 40. During this period the average VIS for all analytes was about 65.

Graphs from the validation period yielded conclusions virtually identical to those drawn from the initial evaluation of the first database studied, illustrated in Figure 6.3 for glucose. In some cases there were slight differences in the magnitude and variability of the relationship, due probably to a somewhat different selection of materials being distributed in the two

Figure 6.1 Relationship with the recalculated mean of recalculated CV for chloride and of average CV for iron in study period



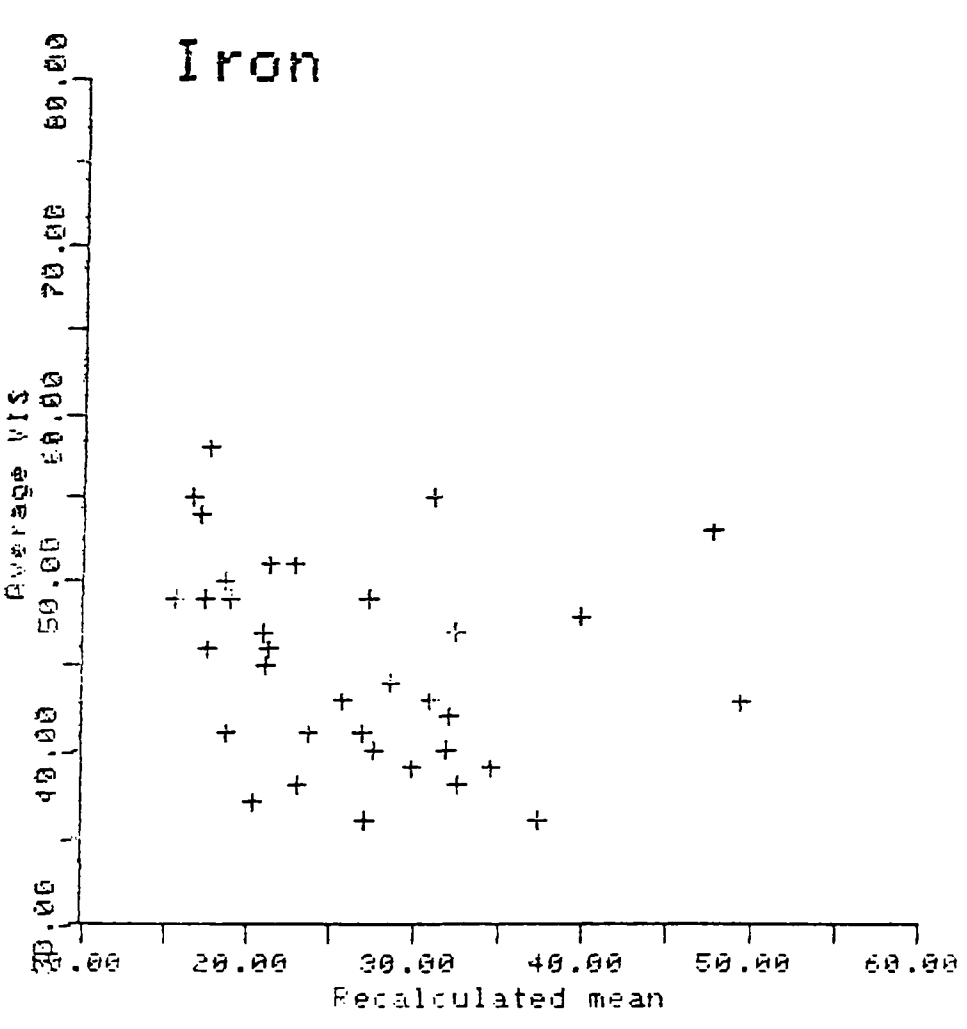


Figure 6.2 Relationship with recalculated CV and average VIS for bilirubin in study period

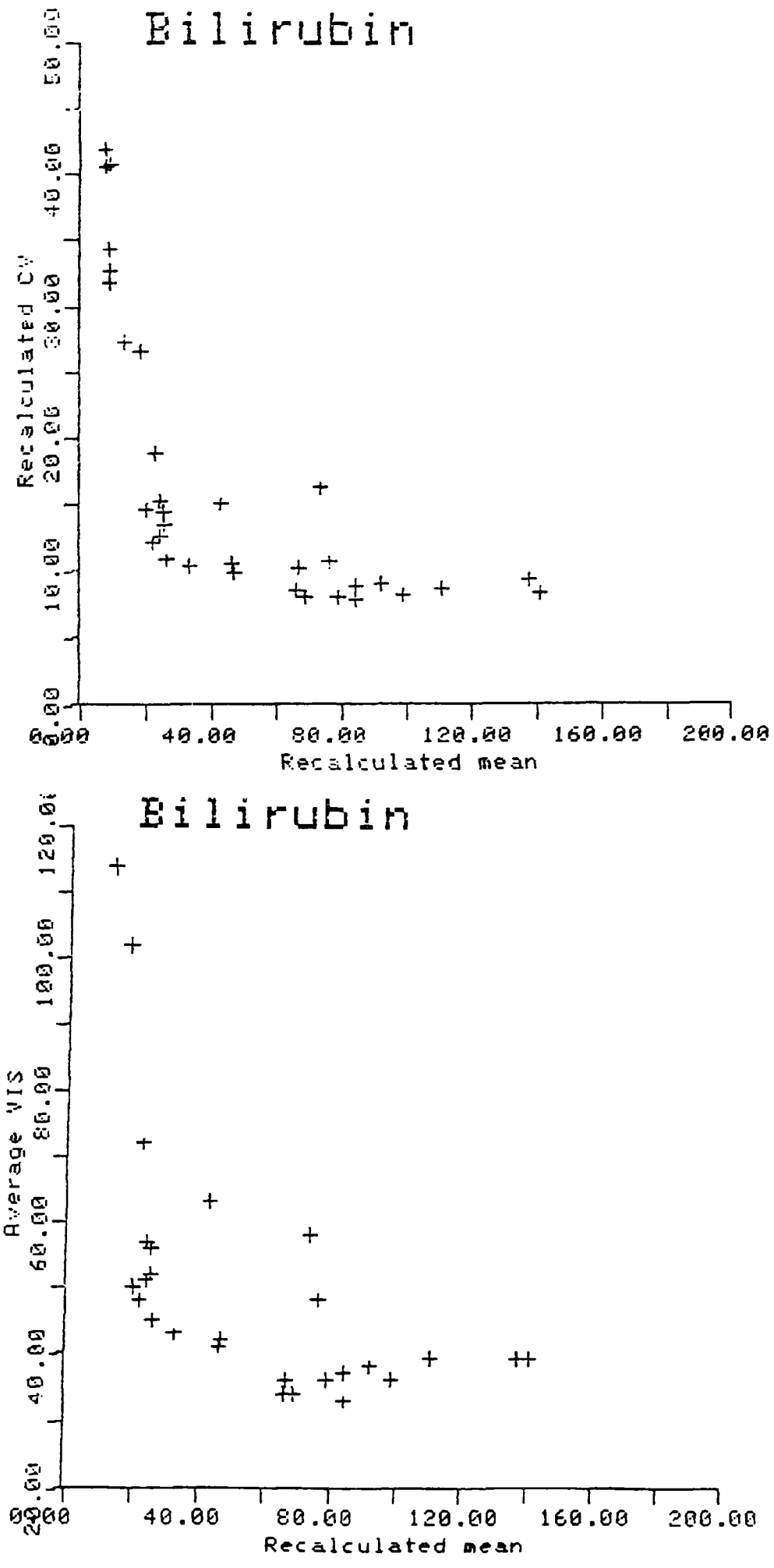
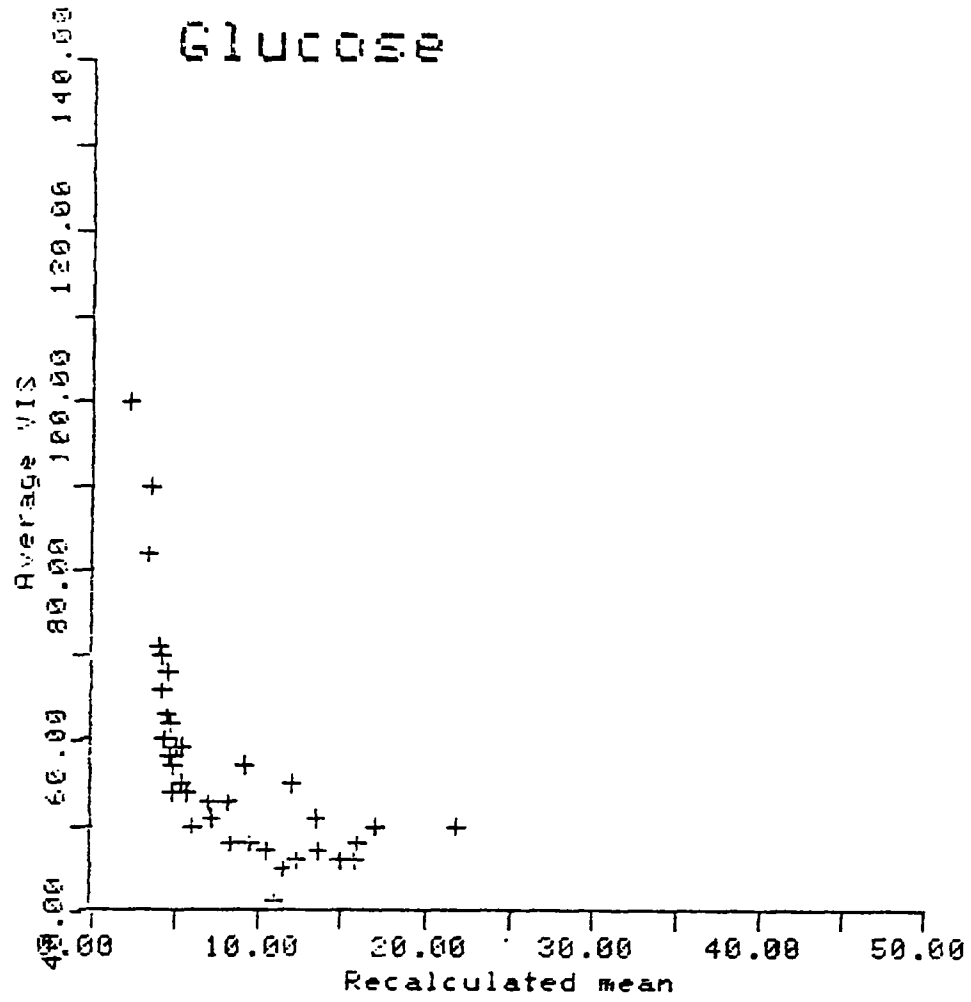
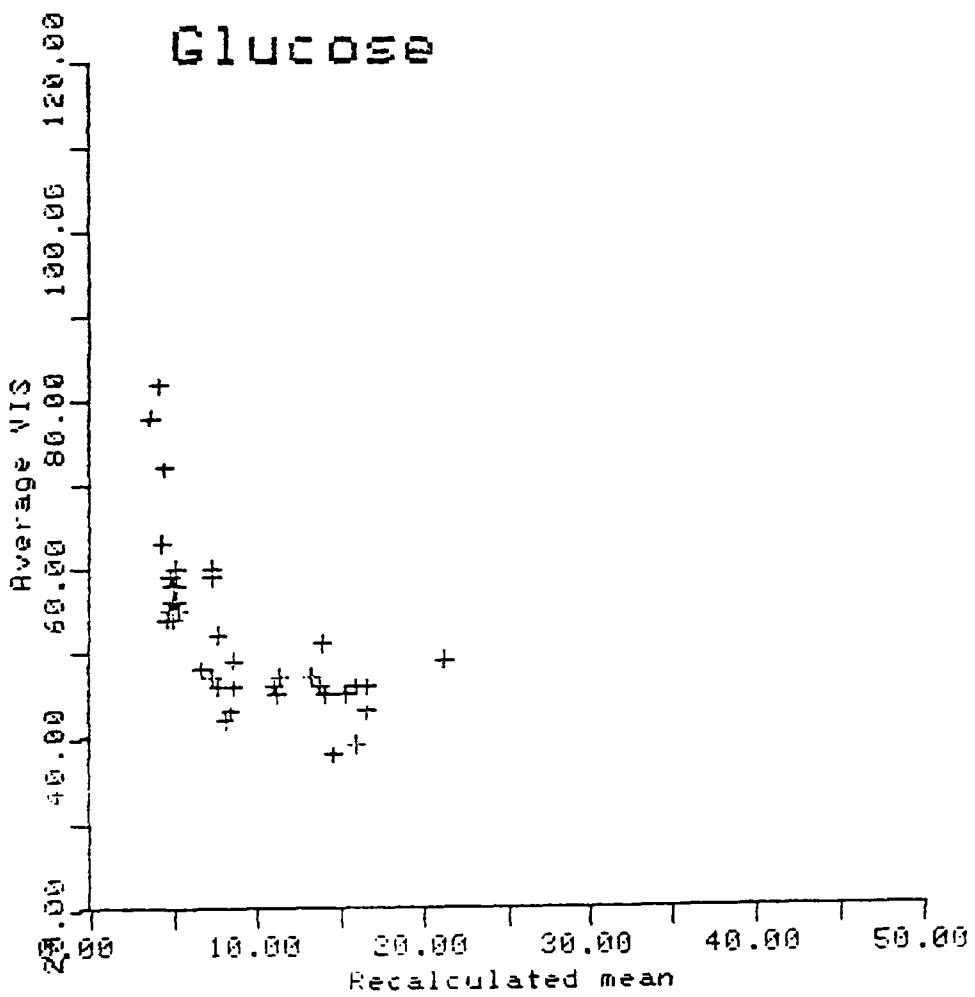


Figure 6.3 Relationship with average VIS for glucose in study and validation periods

Study:



Validation:



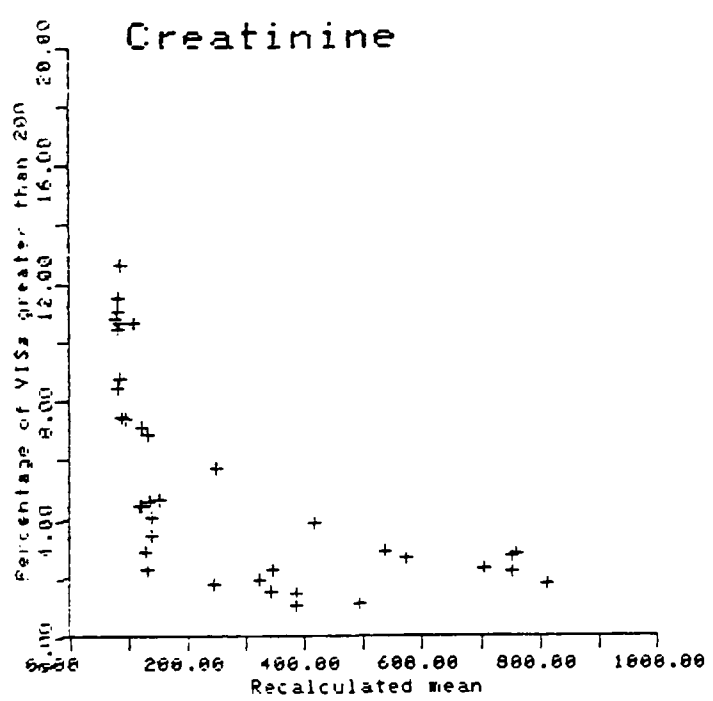
periods. For lithium and magnesium, however, relationships emerged which had not been discernible initially, as these analytes were introduced into the scheme during the study period.

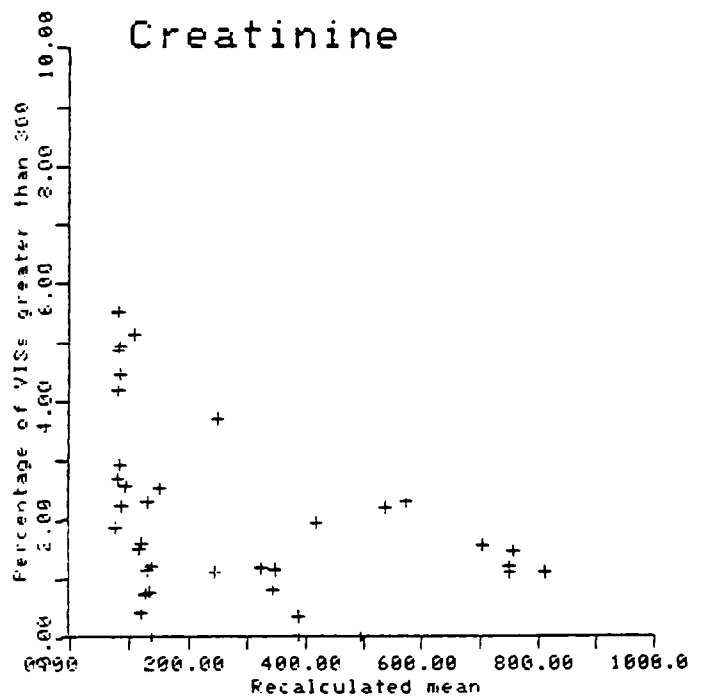
Examination of the indicators of discrepant performance (Table 6.1) yielded relationships between the percentage of VISs >200 and analyte level similar to those for recalculated CV and average VIS, though with lesser clarity. The variability increased and the clarity of the relationship decreased further as the percentages of VISs >300 and of VISs of 400 were examined, exemplified for creatinine in Figure 6.4. No relationship between the percentage of results excluded in the truncation and analyte level was apparent for any analyte.

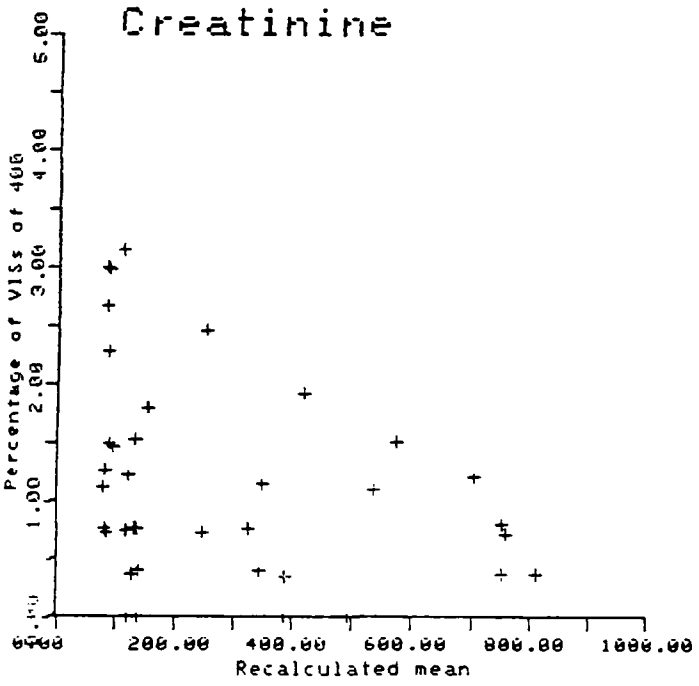
Since the level-dependences observed are reflected in average VISs, and supported by the VIS indicators of discrepant performance, they cannot be due to the effects of intermethod differences or of little-used methods (no VISs are calculated if <15 results contribute to a method mean). Furthermore they were almost all confirmed by the data from the validation period and may therefore be considered real phenomena rather than artefacts.

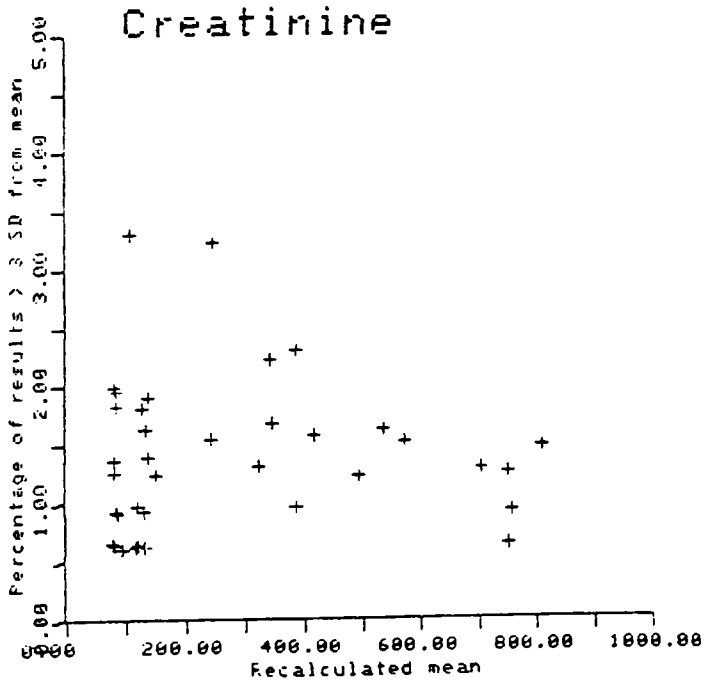
The extent of the level-dependence seen for, eg, urea, creatinine, bilirubin and cholesterol also leads to difficulties in interpreting VISs (or indeed any system of assessment based on percentage deviations from the designated value) if the range of levels in the specimens distributed is not controlled. Interlaboratory agreement (ie laboratory performance) for some analytes is demonstrably worse at normal levels. This confirms the subjective impression that average performance appears to

Figure 6.4 Relationship with percentages of VISs >200, >300 and 400 and of numbers of results excluded for creatinine









deteriorate if several materials with low or normal levels are distributed within a short time, and the UKEQAS policy of distributing a balance of materials at differing analyte levels (see Chapter 3) is thus supported.

The dependence seen was the 'classical' U-shaped precision profile relationship described from IQC data for analytes such as potassium and urea (Figure 4.1). The range of concentrations studied did not, however, extend for all analytes above the level of optimal agreement (eg Figures 6.2 and 6.3). These observations were of importance for UKEQAS operation, and prompted recommendations to the Steering Committee on External Quality Asssment for General Clinical Chemistry (SCEQCC) that some of the limits on the mean (outside which VISs are not calculated) were no longer appropriate. The lower limits for bilirubin and cholesterol were accordingly raised to 17 from 9 umol/L and to 2.5 from 1.3 mmol/L respectively; the upper limit for glucose was also raised, to 30 from 22.2 mmol/L.

Exceptions to the general relationships were found for several materials. In some cases behaviour was exceptional for all analytes, in others for one or two analytes only. Almost all of these had been detected at the time of distribution, so that no VISs were calculated, and the availability of this well-studied database increases the reliability of assessment for future distributions.

6.2.3 Method-related data

As expected, the method-related data confirmed the existence of differences in specificity for some analytes, eg between chemical and enzymic procedures for glucose, urate and cholesterol, and differences in performance for others. Differences in the pattern

of relationship with analyte level were also revealed. These appeared to be associated primarily with the degree of interlaboratory agreement, less clear-cut relationships being observed for those methods with higher CVs.

For example, examination of the recalculated CVs for creatinine (Figure 6.5) reveals that, though similar patterns exist for the AAII/SMA and Other (ie kinetic Jaffe) groups the variability within the kinetic group is almost twice that within the continuous flow group. Differing patterns for different methods were observed for some analytes. For instance with the two ostensibly similar glucose analyser groups (Figure 6.6) there was a smaller though less consistent bias for Beckman instruments.

The effects of automation in yielding improved interlaboratory agreement and average VISs were also shown. For example the AAII/SMA group for urea performed better than the Manual urease group, which also gave a correspondingly less clear relationship with analyte level (Figure 6.7).

6.3 Studies in other schemes

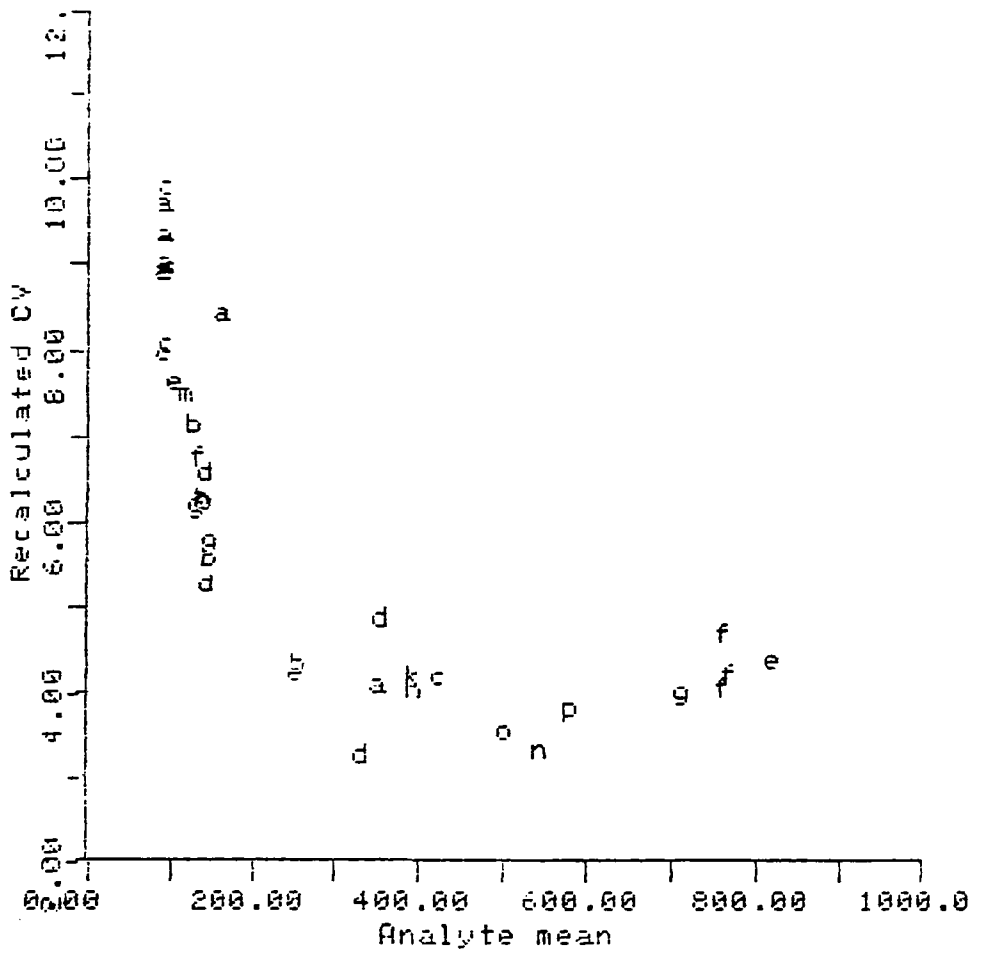
Similar studies may be undertaken for the analytes in other schemes. Many of these assays are less mature, with correspondingly greater interlaboratory variance.

6.3.1 UKEQAS for Salicylate and Paracetamol

Examination of performance for these analytes revealed definite relationships, similar to those described above for cholesterol and bilirubin. Figure 6.8 shows the relationship for salicylate, and Figure 8.3 the much more marked relationship for paracetamol; the latter is discussed in detail in section 8.4.1.

Figure 6.5 Relationship with recalculated CV for creatinine by AAII/SMA and by Other

AAII/SMA:



Other:

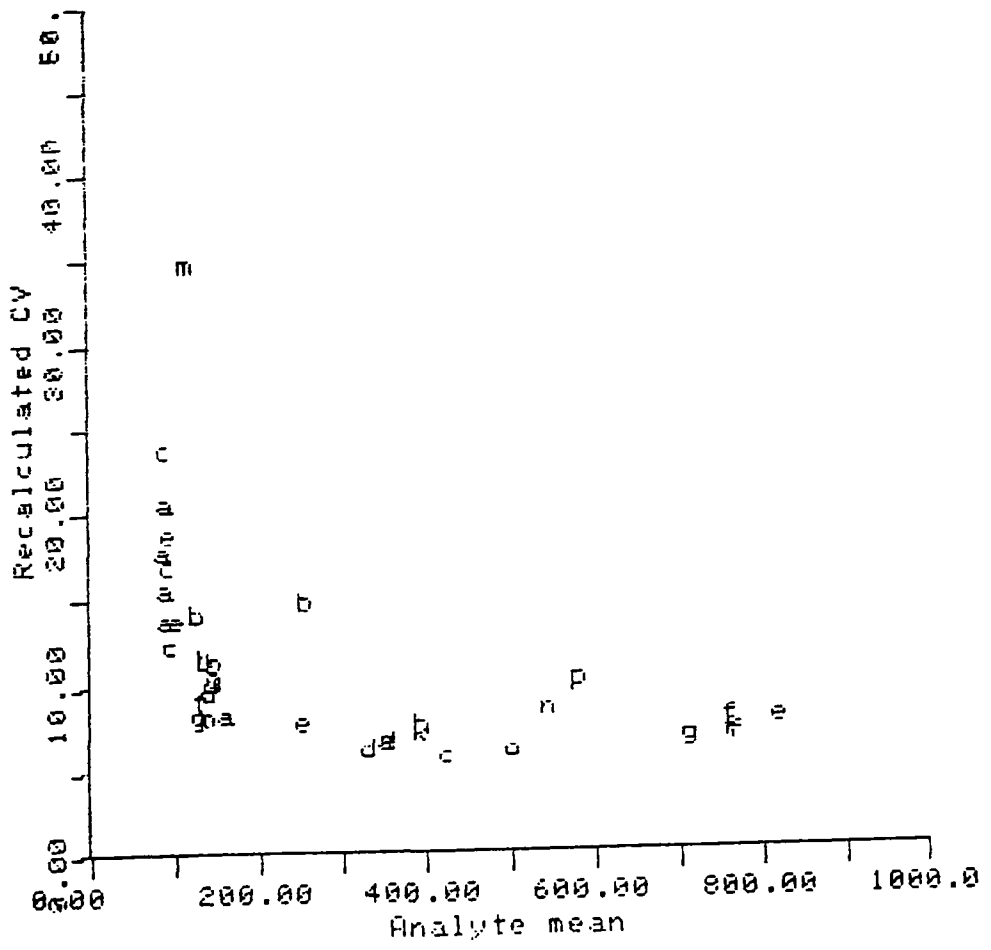
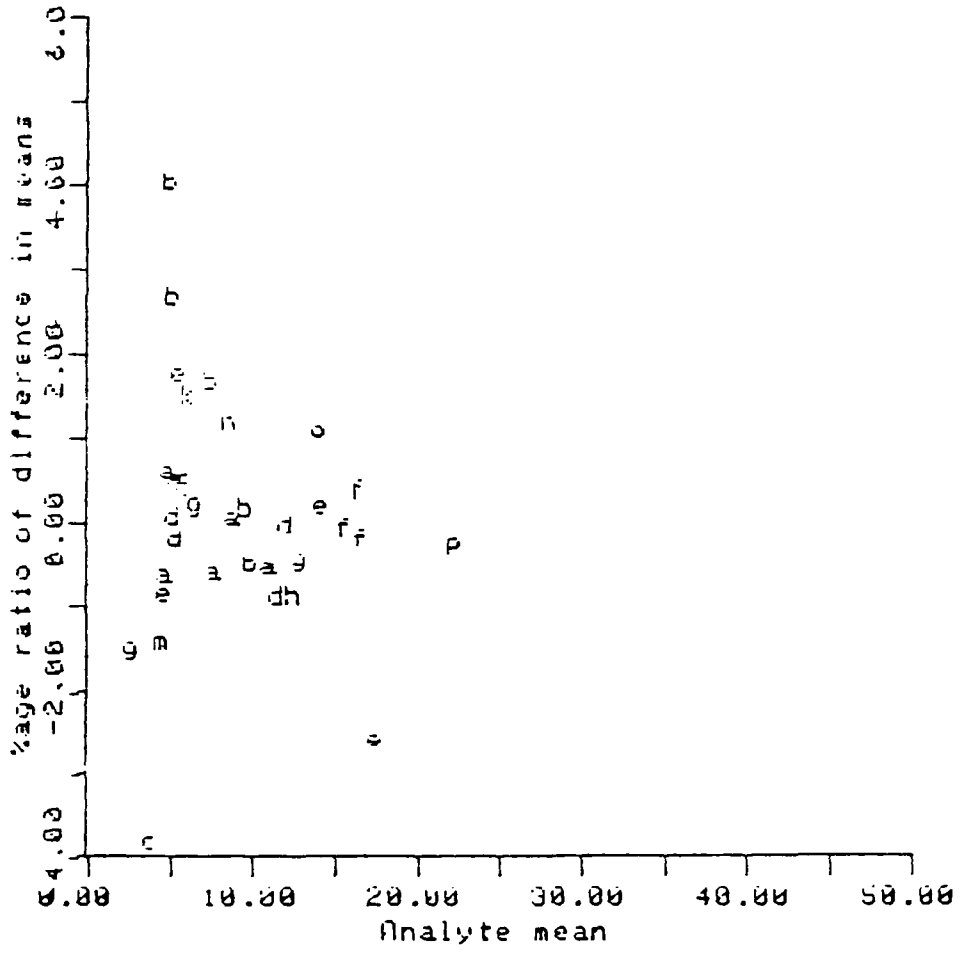


Figure 6.6 Relationship with percentage difference from recalculated mean for glucose by Beckman Glucose Analyzer and by YSI Glucose Analyzer

Beckman:



YSI:

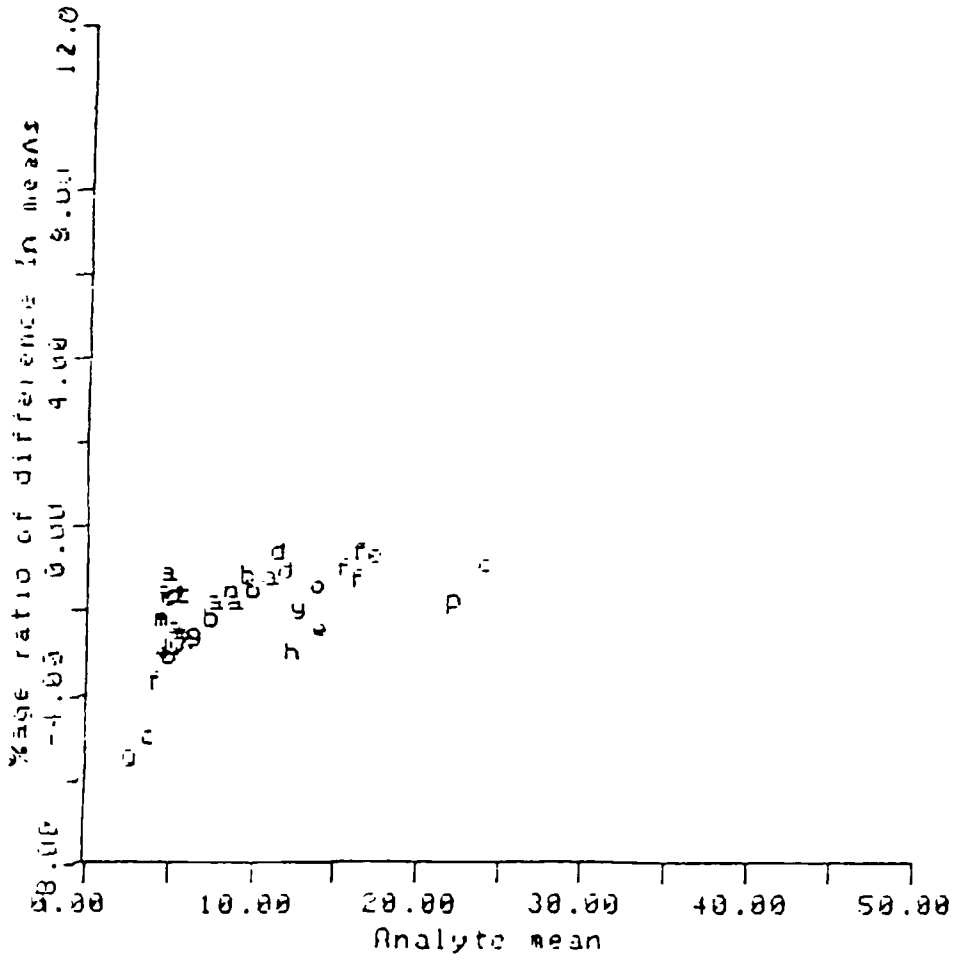
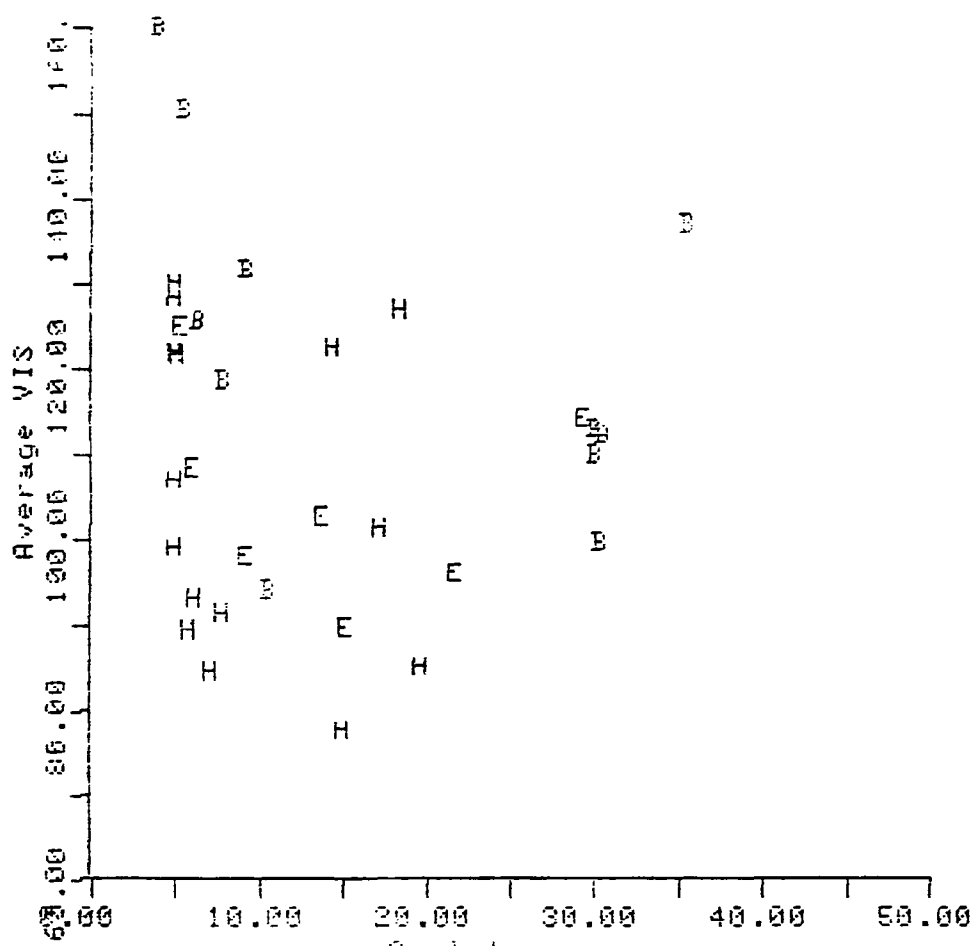


Figure 6.7 Relationship with average VIS for urea by Manual urease and by AAII/SMA

Manual urease:



AAII/SMA:

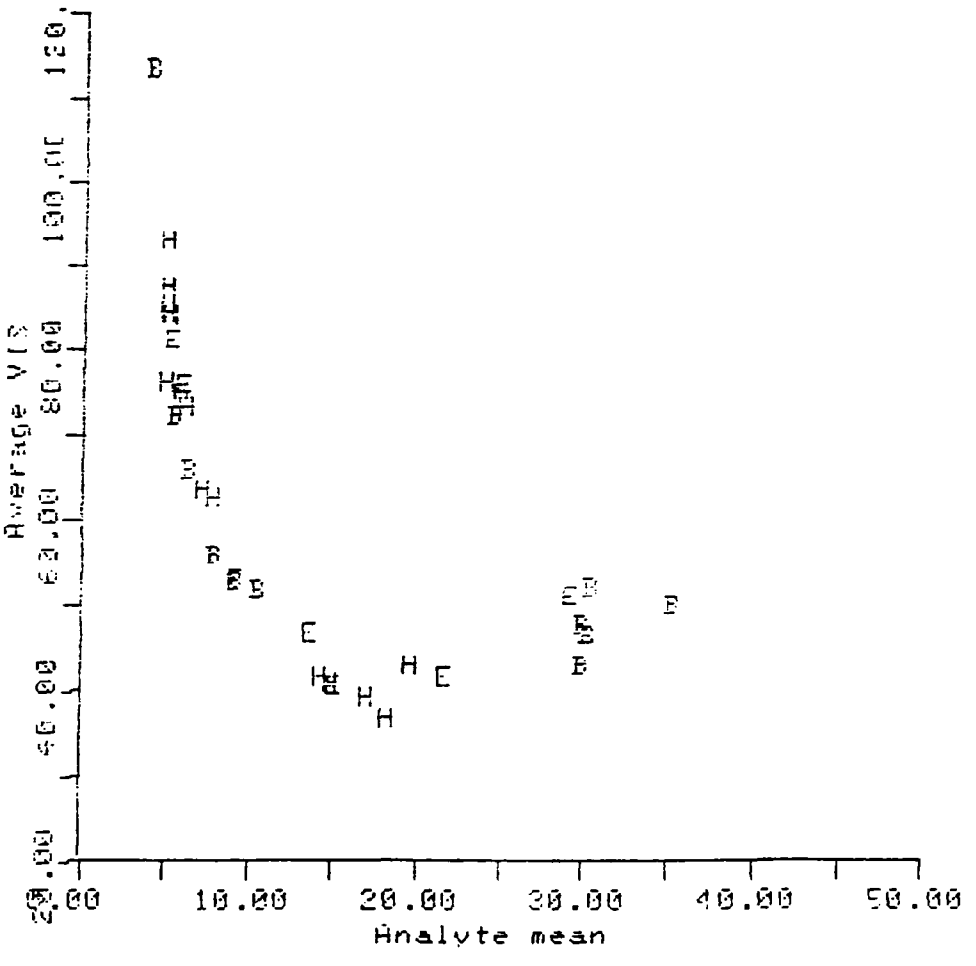
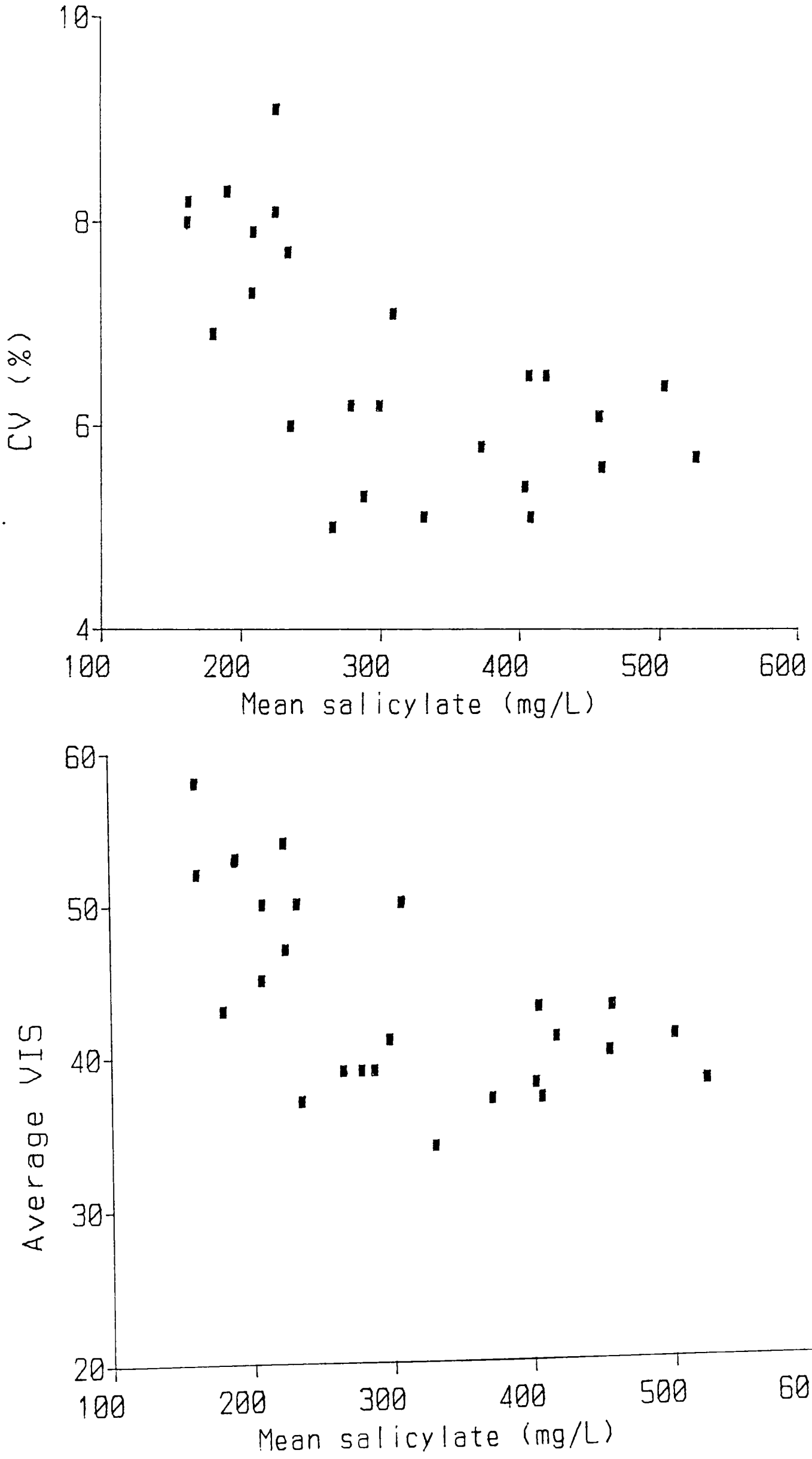


Figure 6.8 Relationship between interlaboratory agreement (CV; average VIS) and mean salicylate concentration in UKEQAS for Salicylate & Paracetamol, 1985-1986



These data underline the need for EQAS specimens to cover a range of concentrations so that trends in performance can be truly reflected. Fortunately, the degree of dependence for paracetamol is decreasing, as a greater proportion of participants adopt the enzymic procedures with their lesser concentration-dependence (Figures 8.2 and 8.3).

6.3.2 UKEQAS for Lead in Blood

Figure 6.9 demonstrates the relationship between interlaboratory agreement and lead concentration lying beneath the scatter in the early years (1973-1974) of this scheme's operation. With improvements in participants' performance during the following years, the level-dependence became less apparent: thus by 1978 much of the effect had gone and the data for 1983 show no real indication of any dependence (Bullock et al, 1986c).

Does this pattern, here demonstrated in two schemes, reflect a true loss of concentration-dependence with improved agreement? This may be so, but it would appear more likely that improvements in participants' procedures are widening the 'trough' region of the relationship, ie that the region of optimal method performance has been increased.

6.3.3 UKEQAS Enzyme Surveys

For other assays, however, neither a striking relationship nor independence of concentration is seen. Figure 6.10 illustrates this for two enzyme activity assays, AST and CK. The points are widely scattered and no relationship is clearly discernible, though there is a suggestion of worse agreement at low CK activities.

Why is this? One possible explanation is that the interlaboratory

Figure 6.9 Relationship between interlaboratory agreement (CV) and mean in UKEQAS for Lead in Blood in 1973-1974 (■), 1978 (o) and 1983 (●)

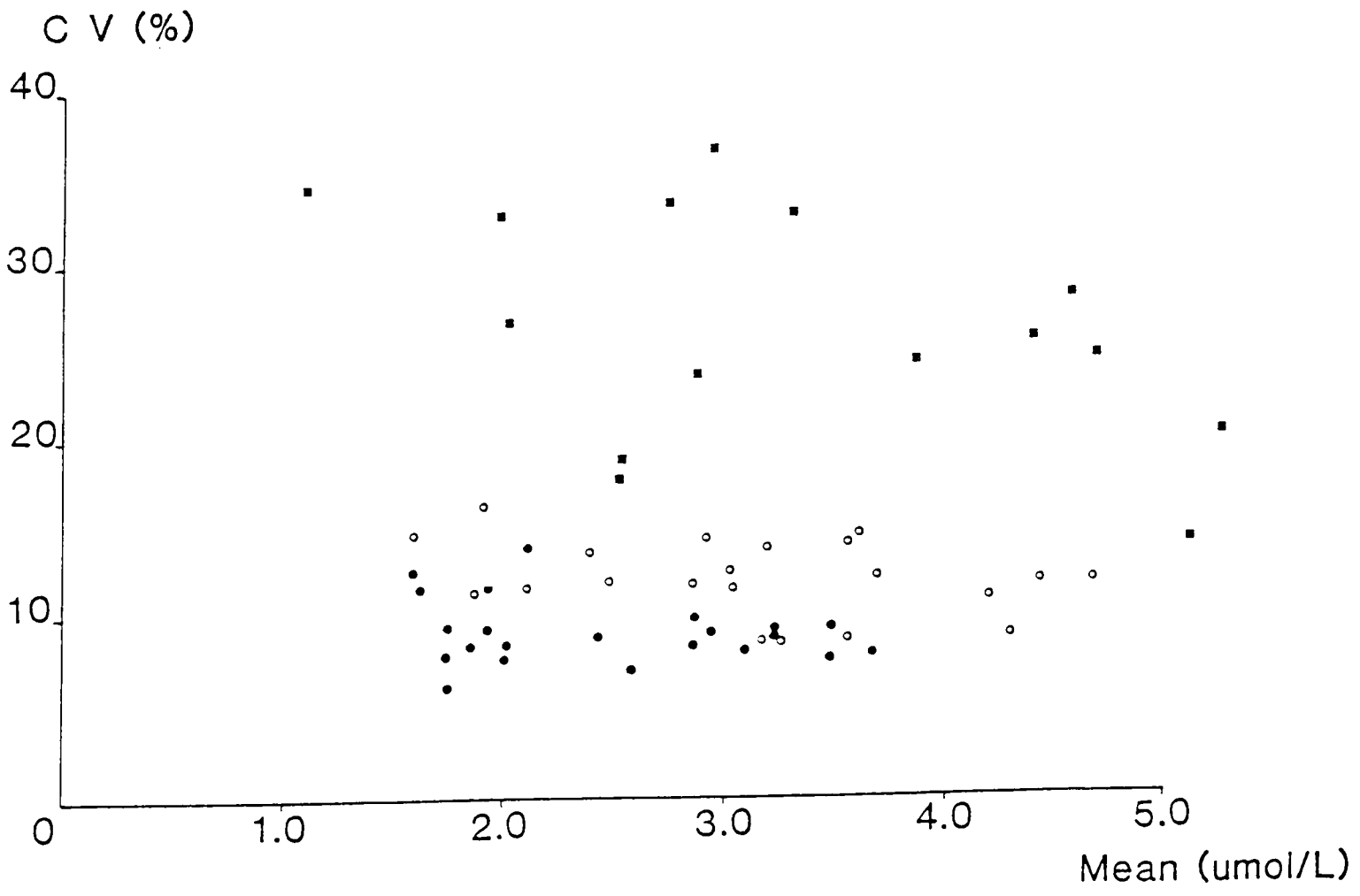
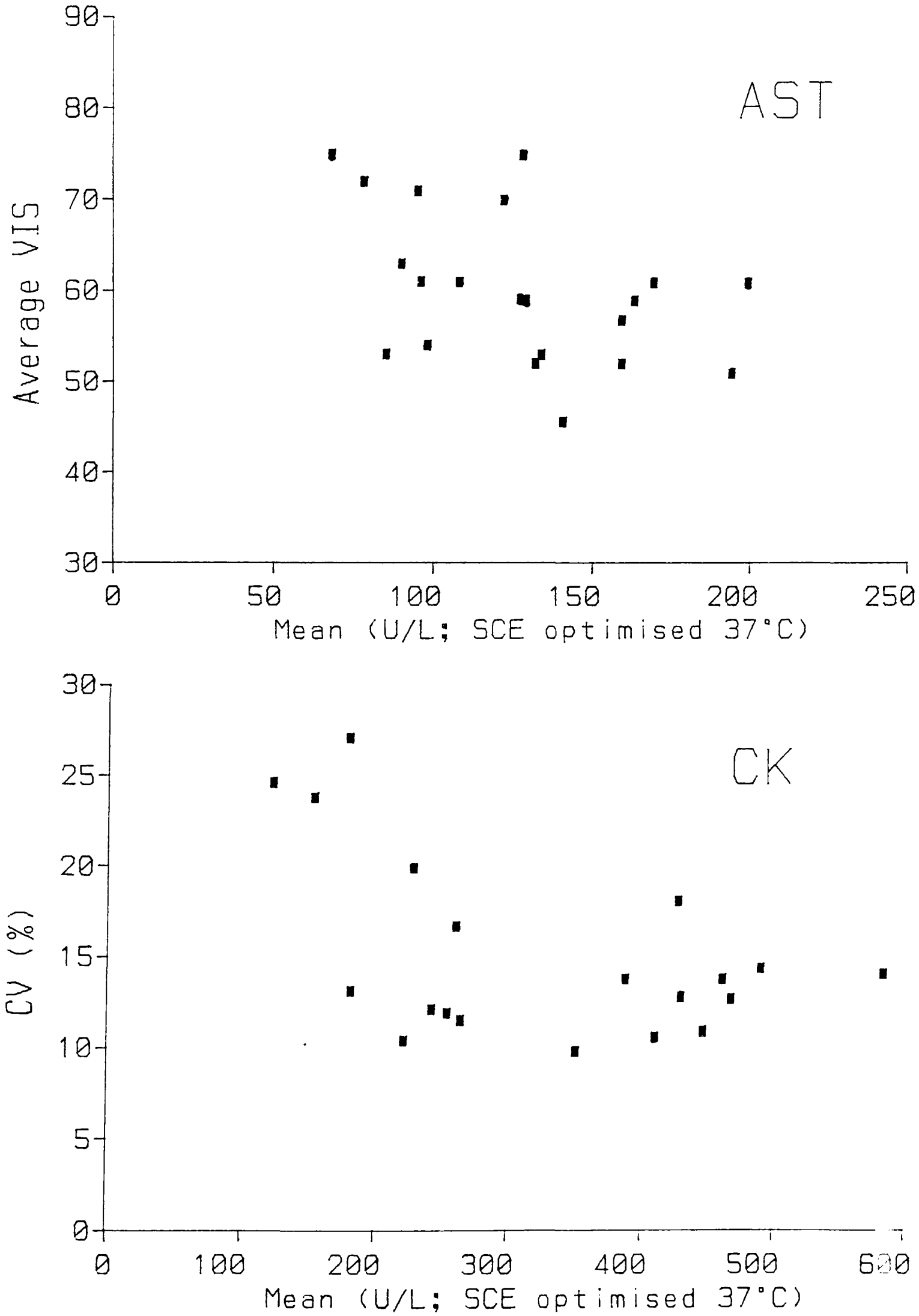


Figure 6.10 Relationship between interlaboratory agreement and mean activity for AST and CK in UKEQAS Enzyme Surveys and UKEQAS for General Clinical Chemistry, 1985-1986. Average VIS for all participants for AST; CV for SCE optimised NAC-activated procedures at 37°C for CK



variance is too great, even within 'reliable' method groups (Bullock et al, 1986b) for an underlying relationship to be resolved at present, though future improvements in agreement may reveal one. Alternatively, there may simply be insufficient data to delineate such a relationship. A further possibility is that there are greater differences in the behaviour of the various materials distributed (as discussed in Chapter 13), and these differences are masking any relationship which exists.

It might be possible to investigate some of these aspects through examination of data from the 'calibration' studies (Chapter 13), which yield greater agreement, and from distributions using similar materials. Such examination, however, does not appear to clarify the situation: there are even fewer data points, and in the latter case investigation is hampered by the tendency of the groups of materials to have similar enzyme activities.

6.4 Summary

For many analytes EQAS data show a relationship between analytical performance and analyte concentration. In these cases interlaboratory agreement is best at intermediate concentrations, though in some the range of concentrations surveyed does not extend high enough to demonstrate deteriorating agreement.

The effects of such relationships on performance assessment must be minimised, eg by confining scoring to a band of concentrations (see Chapter 4).

The interrelationships between concentration-dependence and interlaboratory agreement are complex. Relationships are more readily discernible for methods with better performance, so improvement over time may lead to improved delineation of such

relationships. In other cases, however, the degree of dependence may be reduced as assays become more mature and performance improves.

Chapter 7:

EXTERNAL QUALITY ASSESSMENT OF NON-QUANTITATIVE TESTS: AMINOACID INVESTIGATIONS

7.1 Introduction

Though the initial EQA surveys were conducted in the field of syphilis serology testing (Cumming et al, 1935), EQA has developed most readily and widely for the quantitative assays in clinical chemistry and haematology, as reviewed by WHO (1981), Whitehead and Woodford (1981) and ECCLS (Leblanc et al, 1985a; Lewis et al, 1986). Some progress has been made in microbiology (Leblanc et al, 1985b and 1985c) and immunology, but little beyond very localised systems, often almost exclusively for continuing education rather than quality asessment, in histopathology (Whitehead and Woodford, 1981).

Truly quantitative assays obviously present fewer problems for the design and operation of an EQAS. Thus, as discussed in Chapter 3, appropriate identical specimens can be produced in sufficient numbers for national survey, and the results can readily be evaluated by simple statistical techniques against a target derived by consensus or reference laboratory analyses. Scoring systems can also be devised, as demonstrated in Chapters 4 and 9, to facilitate further assessment of the state of the art and of individual laboratory performance.

Quality assurance is required for semi-quantitative and qualitative as well as for quantitative investigations. For IQC this often takes the form of use of positive and/or negative controls in laboratory assays, though such QA procedures are

rarely applied to urinary dipstick tests and even blood glucose test strips read visually or using reflectance meters (see section 2.2). Such procedures, however, are essentially 'pass/fail' systems and do not readily lend themselves to any form of performance monitoring over time.

What then are the problems associated with EQA of such assays? There appear to be four main groups:

- provision of appropriate specimens
- realistic survey design
- criteria for performance evaluation
- design of a scoring system

7.1.1 Specimen provision

As in any EQA activity, the specimens distributed should have properties resembling closely those of clinical specimens. This may be more difficult for qualitative investigations involving the identification of a pattern than the relatively simple quantitation of an organic or inorganic component in serum. For example, inherited metabolic diseases cause secondary changes in metabolism which may be reflected in the composition of blood and urine. Thus though cystinuria is characterised by a high urinary excretion of cystine, addition of cystine to normal urine will not yield a specimen which mimics urine from a cystinuric patient since the other aminoacids sharing a tubular reabsorption mechanism with cystine (ornithine, arginine and lysine) will not also be present in increased amounts. In such instances, therefore, material from authentic clinical cases is highly desirable wherever this is at all feasible.

Similar considerations apply to 'specialised' specimen presentations. For example, in screening for PKU and neonatal

hypothyroidism it would be inappropriate to provide laboratories with serum since routine specimens are received as dried blood spots on filter paper or liquid whole blood in capillaries.

7.1.2 Survey design

A particular difficulty with many of these investigations is treating EQA specimens in exactly the same way as clinical specimens. This is relatively easy for quantitative assays, which are done routinely, often on an automated system, and reported similarly. Non-quantitative tests are generally done manually, and may require considerable human input in terms of interpretation. An EQA specimen will also lack the personal involvement and clinical urgency associated with the investigation of real cases, though the potential also exists for excessive care to be taken.

The circumstances of laboratories may also differ. Thus for example in urinary aminoacid investigation some participants may only be trying to determine whether the specimen is abnormal and should be referred to a specialist laboratory for further investigation, whereas the latter must endeavour to make a definitive diagnosis of the underlying disorder. The scheme design must accommodate this diversity, and in particular must not penalise laboratories for failing to make an identification which they would not be expected to make in their clinical role.

7.1.3 Performance criteria

Several attempts have been made to assess the quality of urinary dipstick tests, within and among hospitals (eg Simpson and Thompson, 1978; Shephard et al, 1982), which have revealed wide divergence in the results reported. The latter group adopted a

tentative criterion for satisfactory performance in semi-quantitative tests of negative where the analyte was absent and of the correct value plus or minus one positive colour block at finite concentrations (Shephard et al, 1982; Fraser, 1983). For qualitative assays specimens should be identified correctly as positive or negative, but here there may be difficulties in determining the cut-off concentration and in assessing the significance of false positive and false negative results at levels close to this cut-off.

7.1.4 Scoring systems

This problem has so far only been thoroughly examined in microbiology, since EQA of this type of assay in clinical chemistry has been confined to occasional surveys. Here a system based on the "degree of correctness" of the participant's response has been used; this was evolved initially for bacterial isolation and identification surveys, but has since been modified for other microbiological applications (Leblanc et al, 1985c). In essence the system reduces to the following classification of responses:

Fully correct	+2
Partly correct	+1
Wrong, or no return	0
Badly wrong	-1

This system has been criticised in that it combines non-return with identification performance (though some laboratories fail to return only for 'difficult' specimens, justifying such combination in this scheme) and that gross errors can be compensated for by correct identification of other specimens. The latter may also be true of other scoring systems, and suggests

that particular note should be taken of any scores of -1 in addition to an appraisal of average scores. Nevertheless, the system has provided a means of identifying participants with the best and worst performance, and can highlight the incidence of gross errors.

7.1.5 EQA of non-quantitative screening assays

What then is the most appropriate form of survey design for this type of assay? Investigations of aminoacid disorders provide an appropriate model within clinical chemistry. Here false negative results are detrimental to the individual subject concerned and must be avoided if the screening procedure is to be effective. False positives are also far from desirable since they have a cost in terms of increased work and anxiety even if further investigation confirms the subject as negative. Both factors must be borne in mind in devising an appropriate scheme.

The development of the UKEQAS for PKU Screening provides an example of many of the issues mentioned above, complemented by points arising from the initial UKEQAS surveys of urinary aminoacid investigations.

7.2 Development of the UKEQAS for Phenylketonuria (PKU) Screening

Following the demonstration through survey (section 2.4.5) of analytical problems in these screening assays, a scheme was established in the UK in 1980. Its design and subsequent development were determined through consideration of the above-mentioned elements.

7.2.1 Specimen provision

Blood from patients with PKU would be ideal, but the disease is

rare and patients should be on dietary treatment until adulthood to reduce their blood phenylalanine (PheA) concentration. Use of such clinical specimens is therefore infeasible, and spiked blood must be used. The screening tests rely on assessment of PheA itself rather than the pattern of aminoacids in blood, so spiking with PheA alone would be satisfactory. Normal adult blood levels of around 60 umol/L give a useful baseline, providing 'negative' specimens which can be spiked with PheA to yield any concentration up to the 1500-2000 umol/L seen at presentation of 'classical' PKU.

The need to spike the blood and allow equilibration necessitates use of an anticoagulant, whereas specimens for screening are collected directly onto cards or into capillaries. The blood will be somewhat more dilute, especially if citrate/phosphate/ dextrose/adenine (CPDA) rather than heparin is used. Unfortunately, however, only CPDA-anticoagulated blood could be obtained from the Blood Transfusion Service, and this material was therefore used; there was no indication that this would be unsuitable for any of the methods used. The blood used would, by special arrangement, be available the day following collection (after testing for HB_SAg and, more recently, antibody to HIV) to avoid any problems in fluorimetric procedures due to a rise in blank absorbance with aging of the blood.

Various cards were in use in different Regions, including differing filter papers. Since spreading properties are paper-dependent, it was important to provide specimens on the same type of card normally received by each laboratory. Those participants not using HMR 101/6 cards (obtained centrally from DHSS) were therefore asked to provide a supply of their own

cards. Laboratories receiving liquid blood specimens were similarly asked to provide tubes or capillaries.

The EQAS specimens thus resembled the specimens received for screening in all but the inclusion of CPDA anticoagulant. This might affect spreading properties, but many laboratories used materials based on outdated blood bank blood for calibration and IQC.

7.2.2 Scheme design

Each survey would comprise several specimens, covering a range of PheA concentrations. This should provide a more complete assessment performance at each survey rather than relying on cumulation of data over time; multiple specimen distributions also include a more realistic contribution from transposition and other specimen identification errors, which are of major importance in screening procedures. The specimen preparation procedure was relatively complex, and blood collection required special arrangements, so more comprehensive but less frequent surveys gave the most effective use of these resources. In later stages of the scheme two-monthly distributions were maintained, but distributions were less frequent in the earlier phases.

Since laboratories must process specimens rapidly to enable diagnosis and treatment of affected cases as early as possible, a fairly short period, usually around two weeks, was allowed for results return. Laboratories failing to return two consecutive surveys were approached to ascertain the reasons, and in one such case a postal processing delay which also affected clinical specimens was identified and rectified. As usual, reports were prepared for distribution as soon as feasible after the closing date.

The screening methods used were varied, including microbiological inhibition (Guthrie test), fluorimetric and chromatographic procedures. The latter were mostly qualitative procedures only, the others being at best semi-quantitative. Since numerical results were not reported routinely, and the primary outcome of screening relates to the action taken, the emphasis in the scheme was on clinical action. Thus participants were asked to categorise each specimen according to a range of potential actions (Table 7.1) derived from an initial questionnaire to these laboratories. PheA concentrations, as a single figure or range, were also requested where the participant's method would provide these, to assist in differentiation between analytical and interpretation errors if an incorrect action was recommended.

7.2.3 Performance assessment and scoring systems

With this type of investigation consensus values are not reliable, and an externally-derived target must be used. Weighed-in values could not be used since the endogenous PheA content of the blood was unknown, though estimates were available as an independent check. The true PheA concentrations were therefore estimated by aminoacid analyser in two laboratories which did not provide a PKU screening service. The mean PheA was included in the report, provided the two results agreed with each other and the intended content.

In the initial surveys this information was provided to participants in tabular form (Figure 7.1), allowing comparison of their own return with the 'reference laboratory' PheA and other participants' returns. Categorisation of returns in terms of false positives and missed cases was possible after two

Table 7.1 Coded alternative laboratory actions which would normally be taken following the first analysis of a specimen by the routine screening procedure, used in the UKEQAS for PKU Screening. A combination of two or more codes may be returned

Code	Action
A	No action beyond issuing a negative report
B	Repeat analysis by the same screening procedure using the same card
C	Confirm analysis by an alternative method using the same card, in own or other laboratory
D	Request further specimen 2-3 weeks later
E	Request repeat specimen for analysis by the same screening procedure
F	Request repeat specimen for confirmation by an alternative method
G	Summon infant to hospital urgently, or ensure that medical staff institute similar immediate investigation

Figure 7.1 Example of report format in UKEQAS for PKU Screening: Summary of coded actions and PheA concentrations returned by participants. Action codes as in Table 7.1

EXTERNAL QUALITY ASSESSMENT OF PKU SCREENING
SURVEY 20 - APRIL 1986

Lab	Phenylalanine (umol/l) Specimen						Action(s) (Code)					
	1+	3+	6	2	4	5	1+	3+	6	2	4	5
Reference	194	194	229	253	496	669						
15							A	CA	CA	CA	CF	FG
29							A	BD	BD	BD	BG	BG
32	250	220	200	250	450	840	B	B	A	B	BCD	GC
44							A	A	A	AB	GG	CG
51	220	230	235	320	490	660	BC	BC	BC	BCD	BCEF	BCG
60	178	210	214	214	427	531	A	A	BA	BA	BE	BE
61							A	A	A	CD	CD	F
72							A	A	BG	A	BG	BG
92	220	220	235	250	515	760	A	A	BD	BD	BE	BFG
94	60-120	60-120	60-120	120	450	720	A	A	A	A	BE	CG
98							GCEF	GCEF	GCEF	GCEF	GC	GC
204	300	280	311	316	487	673	BCE	BCE	BCE	BCE	BCE	BCE
279	120-240	120-240	120-240	240-360	480-600	720-1200	A	A	A	BE	CFG	CFG
289							A	A	A	A	BE	BE
315	250	250	360	350	500	710	BE	BE	BE	BE	BEG	BEG
437							BE	A	A	BE	EFG	EFG
492	<240	<240	<240	240-360	240-360	360-480	A	A	A	BE	BE	BE
599	<240	240-700	<240	240-700	700	>700	A	F	A	F	F	FG
906	210	220	270	310	530	710	D	D	D	D	E	G
922	180	195	210	280	500	600	A	A	A	CF	CF	CF
926	<120	<120	<120	<120	240-360	360-480	A	A	A	A	BF	FG

surveys, and the apparent incidence (7% and 8% respectively, with 14% of participants failing to summon the infant urgently for specimens with a PheA level >1000 umol/L) gave cause for concern.

A more discriminating means of performance assessment therefore appeared necessary, since introduction of such scoring systems had led to improvement in quantitative assays (Chapters 4 and 9). A system based on that for the microbiology UKEQAS (see section 7.1.4 above) was then devised. This was based on the identification of certain actions as inappropriate for the PheA concentration (stratified into bands), with weighting according to the severity of the error, as shown in Table 7.2. Participants received scores for each specimen distributed, with cumulation over four distributions.

This system was introduced at Survey 4, and a number of complaints were received. These related mostly to the weightings, which for example would give a score for classifying as worthy of further investigation two specimens with PheA levels approaching the upper limit of the reference interval (eg at around 200 umol/L) equal to that for missing a case with a level about 1500 umol/L. Participants using a cut-off of 200 or 260 rather than 240 umol/L were also penalised unduly, and participants requested that the system be reconsidered.

A meeting with a representative selection of screening laboratory directors was therefore arranged to discuss the scheme. The differences in practice among screening centres, for reasons of geography and differences in the local paediatric services, were emphasised, but agreement was reached on the major objectives of screening:

Table 7.2 First scoring system in UKEQAS for PKU Screening, used in Survey 4, November 1981. Action codes as in Table 7.1; score = -3 for non-return of results in a survey (6 specimens). Criterion for unacceptable performance: a total score of -3 or worse in two or more of 4 consecutive surveys

Phenylalanine (umol/L)	Appropriate	Inappropriate	Highly inappropriate
<240	A (B, C)	D - F	G
240 - 360	B - F		A, G
361 - 480	E, F	B - D, G	A
481 - 800	E, F (G)	D	A - C
>800	G with F		A - F without G
Score for each specimen	0	-1	-2

- identification of PKU cases with grossly elevated PheA levels, for urgent confirmation and treatment
- detection of PheA elevations as possible PKU cases, for further investigation
- identification of babies with normal PheA levels as normal, with no requirement for further action

A simplified scoring system was then devised, based upon these three categories. Participants would be asked to classify each specimen as one of:

- 'normal' (N), requiring no follow-up
- 'intermediate' (I), requiring follow-up
- 'high' (H), requiring urgent follow-up

The boundaries between these categories would be at 240 and 700 umol/L, as determined by the DV from the reference laboratories. Participants classifying the specimen correctly would score zero, whereas any misclassification would score the difference between the DV and the furthest boundary crossed. For example, classification of a specimen with a DV of 220 umol/L as N would score zero, as I would score 20, and as H would score 480.

This system should yield scores related more closely to the severity of the error, so the examples cited above of two false positives and one missed case would give scores of 2 x 40 = 80 and 1260 respectively. Because of imprecision around 240 umol/L some 'noise' would be expected and participants were unlikely to obtain zero scores, but laboratories making major errors should be readily identifiable. Example sections of participants' reports tabulating the scores for the survey and the cumulated scores over 4 surveys are shown in Figure 7.2. This system appeared to be accepted much better by participants, and there

Figure 7.2 Example of report format in UKEQAS for PKU Screening.
A: Summary of classifications by participants, with corresponding scores
B: Summary of total scores in and average scores for most recent surveys

EXTERNAL QUALITY ASSESSMENT OF PKU SCREENING
SURVEY 20 - APRIL 1986

Lab	Classification						Score					
	Specimen						Specimen					
	1+	3+	6	2	4	5	1+	3+	6	2	4	5
Reference (PheA, umol/l)	N 194	N 194	N 229	I 253	I 496	I 669						
15	N	N	N	I	I	H						+31
29	N	I	I	I	H	H		+46	+11		+204	+31
32	I	I	N	I	I	H	+46	+46				+31
44	N	N	N	N	I	I				-13		
51	N	N	N	I	I	I						
60	N	N	N	N	I	I				-13		
61	N	N	N	N	N	I				-13	-256	
72	N	N	I	N	I	I			+11	-13		
92	N	N	I	I	I	H			+11			+31
94	N	N	N	N	I	H				-13		+31
98	I	I	I	I	H	H	+46	+46	+11		+204	+31

UK EQAS FOR PKU SCREENING - PERFORMANCE SCORES

Lab	Total score for Survey				Average score
	17	18	19	20	
15	549	0	114	31	174
29	90	0	114	292	124
32	246	0	114	123	121
44	90	0	114	13	54
51	0	86	0	0	22
60	0	86	0	13	25
61	79	0	17	269	91
72	0	0	114	24	35
92	0	-	0	42	14
94	0	0	70	44	29
98	0	86	114	338	135

were no further complaints. Figure 7.2 also shows the effect on laboratory 29's scores of over-reaction, particularly for Survey 20.

Examination of average scores for all participants proved less useful than in other schemes (section 4.5), however, due to volatility. This is attributable to their greater dependence upon the PheA concentrations surveyed (ie proximity to the classification boundaries) and the relatively small number (around 30) of participants. The system was nevertheless useful in practice and in assessing the effects on performance of factors such as method and workload, as described in sections 8.4 and 11.2.2, provided scores were averaged over a sufficiently long period.

7.3 Establishment of UKEQAS Surveys of Urinary Aminoacid Investigations

Following demonstration in Regional schemes of poor quality in this investigation (Green A, Holton JB, Worthy E, 1985, personal communications), the UKEQAS Steering Committee was persuaded that a national assessment of performance was necessary. A programme of two exploratory surveys was therefore agreed for 1986, and the need to to continue survey was accepted following review of the results.

7.3.1 Specimen provision

The investigation comprises the production and interpretation of an aminoacid pattern, produced by separation in one or two dimensions by chromatographic and/or electrophoretic means; one- and two-dimensional thin layer chromatography (TLC) are used most commonly. Since most metabolic disorders, whether inherited or

acquired, lead to widespread alterations in the pattern rather than to discrete changes in the concentration of one or two aminoacids or metabolites it appeared essential to use specimens from authentic clinical cases. This should provide additional interest and motivation for participants, and permit refutation of any suggestions that the specimens were unsatisfactory for a laboratory's particular procedures.

Such materials are difficult to obtain, however, requiring considerable effort and cooperation from the scientific advisors for these surveys. Urine could most readily be obtained from cases with disorders which are relatively benign; with disorders which are treated urine would only be useful when obtained at time when control was poor, and such cases are often associated with lack of understanding and cooperation by the patient's family.

Specimens would almost certainly not be obtainable from disorders which present acutely and in very young children. To cover the full range of disorders likely to be encountered clinically, and in particular those which must be identified as a matter of great urgency, later surveys would have to use 'synthetic' specimens. Because some of the characteristic patterns are complex such specimens would first need to be tested through one of the Regional schemes.

Following assessment by questionnaire of laboratories' requirements, a 5 mL specimen volume was chosen. Though some participants claimed to need more, it was felt that this volume was more characteristic of that provided in many clinical cases; the total volume available was also limited, and slightly less was distributed on some occasions. No preservative was included

in the first survey, but following reports from several participants of bacterial contamination merthiolate (100 mg/L) was added to all subsequent specimens.

7.3.2 Scheme design

Two specimens would be included in each survey, to give the maximum information yield that could be supported by the supply of urine from authentic cases. Errors in specimen identification might thus also be detected.

It is unrealistic to expect participants to carry out investigations such as these in the absence of clinical information, but decisions as to what is appropriate are difficult: many clinical details would be non-informative or suggest strongly the diagnosis. Where possible the details accompanying the initial specimen from the individual case concerned would be used, though directly leading comments (such as "brittle curly hair" from a case of argininosuccinic aciduria) would be omitted.

Though the clinical details might not indicate a need for urgent investigation (requiring results to be available on the same day), reports should be generated on all clinical specimens within 5 working days. A deadline for return of results of 16 days from specimen despatch was therefore determined. Despite this, around 20% of participants failed to return results, and many had a turnround longer than 5 days (Table 7.3).

The format in which results were requested also presented difficulties. Some Regional schemes allowed free text returns, which were then evaluated in detail by the organiser, but this was not viable with 150 rather than 30 participants. Some

Table 7.3 Turnround times for UKEQAS Surveys 2 and 3 of Urinary Aminoacid Investigation, October 1986 and February 1987. Interval between specimen receipt and results despatch (working days); 19 (18 for Survey 3) laboratories provided insufficient information

Turnround	Survey	
(working days)	2	3
<1	3	2
2	13	6
3	6	10
4	10	5
5	11	7
6	6	7
7	12	9
8	19	16
9	4	10
10	8	12
11	6	10
12	3	5
13	-	1
14	1	1
15	1	1
>15	3	4

structuring of response was therefore essential, and participants were asked to classify each specimen as normal, as showing generalised aminoaciduria or as showing a specific increase in one or more aminoacids (or groups of aminoacids where the technique could not resolve these). There was also provision for results of auxiliary tests, such as pH, creatinine content and 'spot test' results, as shown in Figure 7.3. These could provide useful data on the variance of these investigations and their influence on the overall response given, eg participants who were apparently misled into identifying argininosuccinic acid as cystine following a false positive cyanide/nitroprusside test.

The report format was dictated by this design. A table of participants' responses (Figure 7.4) was supplemented by summaries of the spot test results (Figure 7.5) and comments from the scientific advisors. In addition to the overall standard of responses with regard to the aminoacid patterns, these comments addressed problems associated with the apparently poor turnround, the advisability of loading chromatograms relative to creatinine content and the susceptibility of spot tests to error.

7.3.3 Performance assessment

Each participant thus received information on the correct answer (in terms of the patient's diagnosis) and the variability of responses. How then should they assess their performance? Here the variations in laboratory circumstances must be considered. Some merely act as a 'filter' in deciding which specimens are abnormal and need referring to a specialised laboratory for further investigation, whereas the latter must make a definitive diagnosis.

Figure 7.3 Format of results document in UKEQAS Surveys of Urinary Aminoacid Investigation

UK EQAS SURVEYS OF URINARY AMINOACID INVESTIGATION

Laboratory: | | | | |

SURVEY 2, October 1986 SPECIMEN 2A (red label)

1. Origin of specimen

Age: 3 years **Sex**: Male **Source**: Ward

Clinical details: Developmental delay

2. Dates

Specimen receipt: : :86 Chromatographic analysis: : :86 Results despatch: : :86

3. Qualitative response

Please indicate below whether you found the specimen normal, or whether there was a generalised aminoaciduria or whether specific aminoacids were present in increased concentration:

Normal?	Generalised?	Specific increases?*

* please use conventional 3-letter abbreviations (eg Met for methionione); indicate clusters of aminoacids which migrate together in your procedure as follows, eg Lys/Orn/Arg

4. Spot test results

Please indicate below the results of any spot tests you carried out on this specimen.

pH | | | . | | Creatinine | | | . | | mmol/L

	Negative	Trace	Positive	Strong positive
Albustix/Protein				
Ferric chloride/Phenistix				
Reducing substances				
Acetest/Ketostix				
2,4 DNP				
Cyanide/nitroprusside				

5. Comments

Please complete and return as soon as possible, and by Friday 31 October 1986 to:

UK EQAS for General Clinical Chemistry, Clinical Chemistry Department, Queen Elizabeth Hospital, BIRMINGHAM B15 2TH

Figure 7.4 Format of report in UKEQAS Surveys of Urinary Aminoacid Investigation: Specimen description and summary of responses by participants

UK EQAS SURVEYS OF URINARY AMINOACID INVESTIGATION

SURVEY 2, OCTOBER 1986 SPECIMEN 2A (red label)

1. Origin of specimen

Age: 3 years Sex: Male Source: Ward

Clinical details: Developmental delay

The specimen was obtained from a patient with citrullinaemia. The chromatographic pattern showed an increased amount of citrulline which was confirmed by quantitative analysis. The specimen had a pH of 6.0 and a creatinine of 11.4 mmol/L; it was negative by ferric chloride, Clinitest, Acetest and 2,4 DNP, and showed a trace by cyanide/nitroprusside.

2. Reports from individual participants

Lab	Method	Normal	Generalised	Specific increase(s)
1	1A	Y		
2	3A			Cys
5	1B/5C			Cit
7	1A			?Cit ?Lys/Arg
10	1B			Arg/Cit/Gln
11	2A	Y		
13	7C	Y		
15	1B/2B			Gln His/Cys
17	1B/2B	Y		
20	1A/5A			
25	2A	Y		Gln
26	1A	Y		
31	5C			Cit
33	1B	Y		
38	2A/5A			Cit
39	1A			
47	1A/2A			Hom
48	3A	Y		
49	7A			Cit
51	1B/2B			Cit
52	1A	Y		
54	1A			Cys His/Lys/Orn/Arg ?Cit/?HomoCys
57	1B	Y		
58	1A	Y		
59	1A/1B			Cit/Arg
61	6B/8B			Glu/Gln
62	2A/2B			Cit
64	1B/2B			Ser/Cit Ala Asp/Gly/Glu
66	5C			Cit
67	1A/2A	Y		
68	3A/8B	Y		
72	1A/7A			His/Cys/Arg

Figure 7.5 Format of report in UKEQAS Surveys of Urinary Aminoacid Investigation: Summary of spot test results

4 Specimen 2A

3. Summary of pH, creatinine and spot test results

i. pH

<5.75	3	X
6.0	69	XXXXXXXXXXXXXXXXXXXXX
6.5	27	XXXXXXXXX
7.0	8	XXX

ii. Creatinine (mmol/L)

< 9.5	2	X
10.0	7	XX
11.0	41	XXXXXXXXXXXXX
12.0	40	XXXXXXXXXXXX
13.0	7	XX
>13.5	4	X

iii. Spot tests	Negative	Trace	Positive	Strong positive
Albustix/Protein	41	28	1	
Ferric chloride/Phenistix	83	3		
Reducing substances	96			
Acetest/Ketostix	84	3		
2,4 DNP	67			
Cyanide/nitroprusside	48	23	11	

If a laboratory undertakes such investigations, however, it should be aware of the conditions which may be encountered (eg Holton, 1982) and their characteristic patterns, so diagnoses are neither missed nor delayed and the specialist laboratory receives helpful information with the referred specimen. Errors can have tragic consequences for patient care, particularly where acutely ill neonates are concerned.

Provision of a competent service thus requires expertise in interpretation, usually acquired by experience or through training in a specialised centre. Though aminoacid chromatography is often perceived as an occasional activity which provides interest but is of little consequence, if a laboratory receives requests only infrequently patient care may be best served by immediate referral. Indeed several laboratories suspended their service following the institution of the UKEQAS surveys, some because they recognised their staffing was insufficient to provide a satisfactory service and one due to serious errors in their responses to survey specimens.

For these reasons it is appropriate to use a single standard for judging performance, that of a competent general laboratory which should be capable not only of recognising promptly that an abnormality is present but also of making a tentative identification of the aminoacid(s) involved before referring for further investigation or confirmation. Specialist laboratories should obviously be capable of an accurate identification.

A simple statement that a specimen is "not normal" should not be regarded as satisfactory, and laboratories should be expected to

know which aminoacid or group of aminoacids is present in abnormal amounts. Instances where a specimen from a patient with argininosuccinic aciduria was identified as "a case of gross homocystinuria", supported by quantitation of the homocystine and metabolites present, are plainly also unsatisfactory.

Laboratories should also be capable of recognising normal urines as not requiring further investigation, to avoid unnecessary anxiety for parents and clinicians and unnecessary use of laboratory resources. One of the first 6 specimens was normal, and disturbingly several participants failed to classify it thus: most indicated minor abnormalities only, but others reported more serious errors such as the presence of tyrosine, citrulline, or valine and methionine. In this survey three participants returned the same response (which was correct for one of the specimens) for both specimens, indicating failures in their specimen identification procedures, which are as important for these investigations as for quantitative assays.

7.3.4 Scoring systems

The judgements discussed above are still to a large degree subjective, and it might be expected that a scoring system could simplify such interpretation, as in other cases (see Chapter 9). Unfortunately, however, where individual laboratories vary in their capability and the detail provided in their responses devising a scoring system is fraught with difficulties; here the situation is much less well characterised than even that for PKU screening, discussed in section 7.2.3 above. Any system to assist interpretation by individual participants thus appears impractical.

Table 7.4 Summary of tentative categorisation of participants' responses in UKEQAS Surveys 1-3 of Urinary Aminoacid Investigation. See section 7.3.4 for explanation of categories

Lab	Method	Specimen					
		1 A	1 B	2 A	2 B	3 A	3 B
1	1A	+	+	-	+	+	-
2	3A	+	+	-	+	-	-
5	1B/5C	+	+	+	+	+	+
7	1A	X	X	+	-	+	-
10	1B	+	+	+	+	+	-
11	2A	+	-	-	+	-	-
13	7C	X	X	-	+	X	X
15	1B/2B	+	+	-	+	+	-
17	1B/2B	+	+	-	+	-	
20	1A/5A	+	+	X	X	X	X
25	2A	X	X	-	+	+	
26	1A	X	X	-	+	+	-
31	5C	+	-	+	+	+	-
33	1B	-	-	-	+	+	+
38	2A/5A	+	+	+	+	+	+
39	1A	X	X	X	X	+	-
47	1A/2A	/	/	-	+	+	+
48	3A	+	+	-	+	+	+
49	7A	+	+	+	+	+	-

It may nevertheless be feasible to provide a categorisation which would be helpful for another application, that of identifying laboratories apparently experiencing problems and which may benefit from a more detailed appraisal and the provision of advice. Table 7.4 represents a preliminary attempt at such a classification, using the following symbols:

Correct	+
Unclassifiable response	
Wrong	-
No return	X
Specimen not provided	/

As can be seen, patterns begin to emerge from such a classification after only three surveys (6 specimens). For example, laboratories 5 and 38 appear to perform reliably, whereas 2, 11 and 33 made frequent incorrect responses and 13, 20 and 39 failed to make regular returns.

It is premature to use such a system in reports to participants, since many of the classifications cannot be completely objective, but it promises to serve well in indicating to the scheme's advisors which laboratories' returns should receive particular attention. This will complete the service provided by these surveys, which initially provided only an assessment of the state of the art but have progressed to assessment of individual laboratories' performance and, in a limited way, of method performance.

7.4 Summary

The requirements for EQA of semi-quantitative and qualitative investigations differ from those for quantitative assays, discussed in Chapter 3. These differences centre on:

- specimen provision
- scheme design
- performance assessment
- scoring systems

Provision of appropriate specimens can cause difficulties. Presentation must be in a form as close as possible to that of routine specimens and, where feasible, specimens from authentic clinical cases should be used.

Scheme design must also approach the routine clinical situation and avoid appearing to be an 'artificial' exercise.

Performance appraisal can be complicated by differences in laboratory circumstances and practices. The fundamental objective of carrying out the investigation, often a clinical action, must be identified and satisfying this treated as the minimum criterion for satisfactory performance. Some participants should be aiming higher than this, but patients are more at risk from a laboratory which fails in the basic objective.

Scoring systems can be devised for such investigations, but to ensure general applicability these must be simple and concentrate on the essentials. More detailed, but more subjective, appraisals may be valuable for use by scheme organisers in identifying participants in apparent need of assistance.

Chapter 8:

EXTERNAL QUALITY ASSESSMENT IN THE CHOICE OF ANALYTICAL PROCEDURES

8.1 Introduction

A clinical chemist who has to select an analytical procedure must consider many factors. The overall objective is a procedure which yields results of the greatest clinical usefulness in the diagnosis and management of patients. Such a procedure must ideally be accurate, precise, rapid and robust, yet be practical within the constraints of the laboratory circumstances. Such a combination of performance characteristics may well be unattainable in practice, given the lack of maturity of many analytical procedures, and a compromise may need to be sought. To arrive at such a compromise may, however, require more care than fulfilling the ideal.

The potential sources of information on the performance characteristics of candidate procedures include:

- information from reagent and instrument manufacturers
- the scientific literature
- evaluations carried out previously or conducted as part of the selection procedure
- information from colleagues using the procedure now or in the past
- EQAS results

All are useful to some extent, but criticism may be directed at each.

8.1.1 Reagent and instrument manufacturers

Suppliers of instruments and reagents are not in general philanthropic, their primary aim being to increase sales of their products. Thus the information they provide, though it may be accurate in that it could be defended in law, is suspect from the point of view of selectivity.

For example, where precision data are quoted their derivation (from optimal or routine conditions) is often unclear, and the concentrations at which they are obtained may have been selected to give the most favourable impression. In comparisons with the performance of alternative procedures data from inferior methods may be given. In method comparison studies, the method against which the supplier's procedure has been assessed may itself not have been fully validated, eg an enzymic urea assay claimed to give improved specificity might be demonstrated as having excellent correlation with a routine method based on diacetylmonoxime. The International Federation of Clinical Chemistry (IFCC) has provided guidelines for the presentation of claimed performance characteristics (Okuda, 1984; Rubin et al, 1984) which have improved the situation slightly, but most manufacturers do not adhere to these.

8.1.2 Scientific literature

This should be more objective, since reports will have been refereed before publication. Authors intent on demonstrating the superiority of their procedure, however, may be equally selective in their choice of data for publication and will frequently choose the best performance from that obtained over a period. They may also not record the problems experienced, nor any deficiencies in the method. Even where objective data are

reported, the performance may reflect in part the 'hands of the devoted' effect and not be reproducible in other laboratories. Reference and selected methods (eg those of the American Association for Clinical Chemistry) do undergo a rigorous refereeing procedure and transferability studies, but these are small in number and may not be widely applicable.

8.1.3 Evaluations

Properly constructed evaluations (eg Broughton et al, 1974; Büttner et al, 1979b; Logan et al, 1983) should yield reliable information about the performance characteristics of a procedure. Thus precision and accuracy, and any additional factors such as potential interferences in the analysis, can be addressed. The component (eg dispensing and photometric) subsystems, safety, reliability and servicing of instruments may also be investigated thoroughly. The accuracy of many clinical chemistry assays is difficult to assess in such evaluations if true recovery studies cannot be undertaken; accuracy is therefore frequently assessed as bias relative to some other procedure or relative to values assigned to QC materials through EQAS distribution or by manufacturers.

Thorough evaluations, however, are time-consuming and therefore not carried out on all procedures; their availability through the conventional scientific literature is also limited, and many prospective users may be unaware of their existence. Secondly, they are conducted over a relatively short period and since the procedure is not in true routine use even the 'routine conditions variance' (Whitehead, 1977) may not reflect accurately actual performance. These limitations in time and effort will apply particularly to evaluations undertaken by the individual

laboratory, especially if several procedures are to be assessed. Furthermore, manufacturers often alter products after an evaluation and thus undermine the applicability of the conclusions.

8.1.4 Information from colleagues

Colleagues using the procedure should be able to give a good assessment of its true routine performance. This would include such aspects as precision, ease of control and acceptance by laboratory staff. The procedure's performance in EQASs will also be an important element of the information.

The comments of laboratories which were faced with a similar choice, and those of laboratories which have since abandoned the method in question, can be particularly valuable. Identification of such users may be difficult, however, and instrument or reagent suppliers may suggest satisfied rather than dissatisfied customers if experiences differ.

8.2 External quality assessment data in method selection

The results obtained in EQASs yield much information about the relative performance of analytical procedures, and should be used in method selection for a wide range of reasons.

Firstly and most importantly, EQAS data reflect in large part the routine performance of methods. Thus each participant laboratory is reporting results obtained from their routine procedure, ie the procedure which is used for release of results for patient care, which is subject to an internal QC program, and which is being carried out with none of the special precautions to which evaluation assays may be subject. Though some laboratories may apply special treatment to EQAS specimens, as exemplified in the

studies of Rumley and Roberts (1982) and of Rowan et al (1984), such treatments have only a minor effect on the EQAS performance of an individual laboratory. The effects will be diluted when a group of laboratories is considered, and it is implausible that the application of such treatment would be concentrated within particular method groups, so this factor should not affect conclusions about the relative performance of methods. As a consequence of this, however, EQAS performance may give a slightly favourable picture of the routine performance obtained, though the effects of 'blunders' (eg transcription or transposition errors in the participating laboratory) due to the additional step in reporting EQAS results may cancel this out. EQAS data will only reflect performance of the assay when it is in control, since results from analytical batches shown to be out of control by the participant's IQC programme will not be reported, and EQAS performance will thus appear better. EQA cannot normally give any assessment of this aspect of method performance, ie the proportion of batches rejected, unless such information is specifically requested through the scheme.

The data on method performance from an EQAS are averaged over all laboratories employing the procedure. In practice, performance is not determined exclusively by the analytical method used, and some laboratories will attain superior results because of their closer attention to other aspects of quality assurance. These disparities will be reflected in EQAS performance and in most cases there will be a considerable range of achievement within each method grouping. Again it is unlikely that the better laboratories would disproportionately make a similar choice of method (unless there were indeed methods with compelling reasons in their favour), so this should also not invalidate the

usefulness of EQAS data in method selection. Such diversity may in fact be an excellent indicator of method quality, in that a greater than average spread of attainment for a method probably indicates susceptibility to problems. For example, an analytical principle may be unsuited to particular equipment, it may be difficult to control and its performance therefore be related to IQC programme efficiency, or there may be variability in the quality of reagents from alternative suppliers. Each of these factors may make the method less robust and indicate a need for caution in selecting such a method.

The main information provided by EQAS data is on the degree of interlaboratory agreement. Variance in a scheme arises from within-laboratory imprecision (both short- and long-term) and between-laboratory differences in bias. Such differences in bias may be real, provided that the behaviour of the material distributed is identical to that for clinical specimens. In some cases, however, interactions between analytical methods and the specimen distributed may prevent judgement against other than a value assigned for each method group and no overall assessment of accuracy may be made; this problem is discussed in detail in Chapters 12 and 13.

The information obtainable thus depends upon scheme design. The most informative design would be one in which authentic clinical specimens were distributed repeatedly, to assess within-laboratory precision (or, more strictly, reproducibility of results over a long period), with studies using analyte-free matrix to assess baseline security and addition of pure analyte to assess recovery (eg Hunter and McKenzie, 1979). This is the ideal, however, and can rarely be attained in practice, except

within relatively small schemes and where analyte-free matrix can be obtained. More limited scheme designs, where only agreement within a method group can be assessed reliably, should still yield sufficient information to enable comparison of methods on this basis alone. Such approaches can also be used to assess the effects of reagent kit (Hayes et al, 1985), instrument (Westwood et al, 1986) or modifying analytical procedures (Groom, 1985c).

Overall, therefore, EQAS results provide an important source of information on method performance and facilitate reliable choices of analytical procedures. In addition, EQAS performance is valuable in monitoring whether the correct choice of procedure has been made. The selection of a method for serum calcium assay, an assay with great clinical importance, provides a useful illustration of the use of EQAS data.

8.3 The selection of a method for serum calcium assay

Consider the situation of a clinical chemist in 1980 selecting a method for serum calcium assay (a decision most important for patient care, since the assay is critical for diagnosis and treatment and in general the state of the art falls short of analytical goals derived from biological variation). Information would have been available from the sources discussed above, but EQAS results would have formed the most readily accessible and comprehensive database for initial assessment.

8.3.1 Assessment of EQAS data

A wide variety of method principles for serum calcium assay were available at that time. These included atomic absorption photometry, titrimetry and dye-binding procedures employing a number of chromophores. Table 8.1 shows the information provided

Table 8.1 Average interlaboratory agreement for serum calcium assay in UKEQAS for General Clinical Chemistry, January-June 1980 (n = 10 distributions)

	n	Mean (mmol/L)	CV (%)
Overall	335	2.67	3.6
EDTA titration	19	2.62	4.0
Atomic absorption	50	2.65	3.8
AutoAnalyzer I	37	2.69	3.2
AAII or SMA system	123	2.67	2.9
Vickers M300/D300	20	2.69	2.4
Manual methylthymol blue	21	2.64	4.7

by averaging the data from the UKEQAS for General Clinical Chemistry reported to participants for the first 6 months of 1980. The classification used in 1980 was based on the method principle, except in the case of automated procedures which were divided on the basis of instrument type.

It is readily apparent from these data that the automated procedures (AAI, AAII/SMA and Vickers groups) show the best interlaboratory agreement, ie these methods appear overall to be the most reliable. Atomic absorption and EDTA titration are somewhat less reliable, though the titrimetric procedure is less widely used, and the methylthymol blue (MTB) method group is apparently the least satisfactory. In no case is there evidence of a large bias relative to other procedures, though the automated procedures do tend to give results higher than those by atomic absorption, the principle on which reference methods (eg Cali et al, 1973) have been based.

Within each group there are variations due to individual laboratory factors, to robustness of the procedure, and to the reliability of the instrumentation used. Thus the EDTA titration and MTB groups are essentially manual procedures, the atomic absorption group uses a single type of instrument, whereas the AutoAnalyzer and Vickers groups employ automated systems of varying degree of sophistication. In the situation of increasing workload an automated procedure would be advisable and the obvious choice would be an AutoAnalyzer II or SMA; AutoAnalyzer I and Vickers systems were by this time no longer in production.

The choice of a continuous flow system may, however, not have been in complete accord with the work pattern of the laboratory. An increasing dissatisfaction with the concept of 'profiling' and

movement towards discretionary testing had made discrete analysers a new and more popular choice for many laboratories, their greater versatility with regard to carrying out a range of different assays at different times on the same instrument being particularly attractive. In this case the choice would be one of method principle rather than of procedure. Improving instrumentation for atomic absorption assays and the possibility of also carrying out other cation assays on the same instrument could, however, have made this theoretically more accurate procedure the choice for some laboratories.

Which method principle should then be selected? Discrete analysers are limited to dye-binding assays, and here the choice essentially lies between o-cresolphthalein complexone (CPC) and MTB, though methods based on, for example, alizarin have also been proposed. The UKEQAS data favour CPC, as the predominant dye in continuous flow procedures, rather than MTB, which shows substantially worse interlaboratory agreement. This difference may at first sight be due solely to the degree of automation, but discrete analyser procedures using MTB were in fact also included within the group described as "Manual MTB".

This illustrates one problem of method assessment using EQAS data, as the method classification used may be designed primarily to avoid the consequences of matrix effects and not take account of all differences between methods which are of interest. Thus from the 1980 data there is no way of distinguishing between the performance of MTB assays performed manually, by means of a centrifugal analyser, or by means of any other type of discrete analyser.

Therefore the EQAS data indicate that an automated procedure based on CPC should be best, with laboratory circumstances dictating whether continuous flow or a discrete analyser system should be used. The performance of discrete systems could not be forecast directly from the EQAS data since such methods lay outside the 1980 classification, and CPC on a continuous flow system would consequently be the 'safest' choice.

8.3.2 Appraisal of the conclusions

How accurate were these conclusions? This can also be determined through examination of EQAS data. Thus Table 8.2 shows the UKEQAS results from the corresponding period of 1986.

Now, with a larger number of participants than in 1980, CPC procedures as a whole appear in fact not to have the most reliable performance. The best interlaboratory agreement is now shared with continuous flow CPC by the automatic titrator and atomic absorption groups. There have been continuing advances in atomic absorption instrumentation over the last few years, and current equipment is more reliable than that available in 1980. The automatic titrator performance also demonstrates that EQA data cannot forecast future developments, and thus cannot yield information on novel instrumentation until it is in routine use; evaluation data should therefore be used in conjunction to alleviate this problem.

Table 8.2 raises a number of other issues about the use of EQAS data in studies of method performance. Firstly, as discussed in Chapter 4, the use of an appropriate scoring system will give an additional and more readily interpreted way of assessing cumulative data. Thus the MRVISs are akin to average CVs in yelding an overall assessment of interlaboratory comparability

Table 8.2 Average interlaboratory agreement for serum calcium assay in UKEQAS for General Clinical Chemistry, January-June 1986 (n = 12 distributions), with average running scores at June 1986

	n	Mean (mmol/L)	CV (%)	MRVIS	SDBIS
Overall	412	2.82	3.2	61	65
Atomic absorption	26	2.78	2.7	56	57
Manual/discrete CPC	188	2.82	3.3	68	76
Continuous flow CPC	140	2.83	2.3	48	49
Methylthymol blue	25	2.77	3.9	81	84
Automatic titrator	27	2.75	2.4	51	54

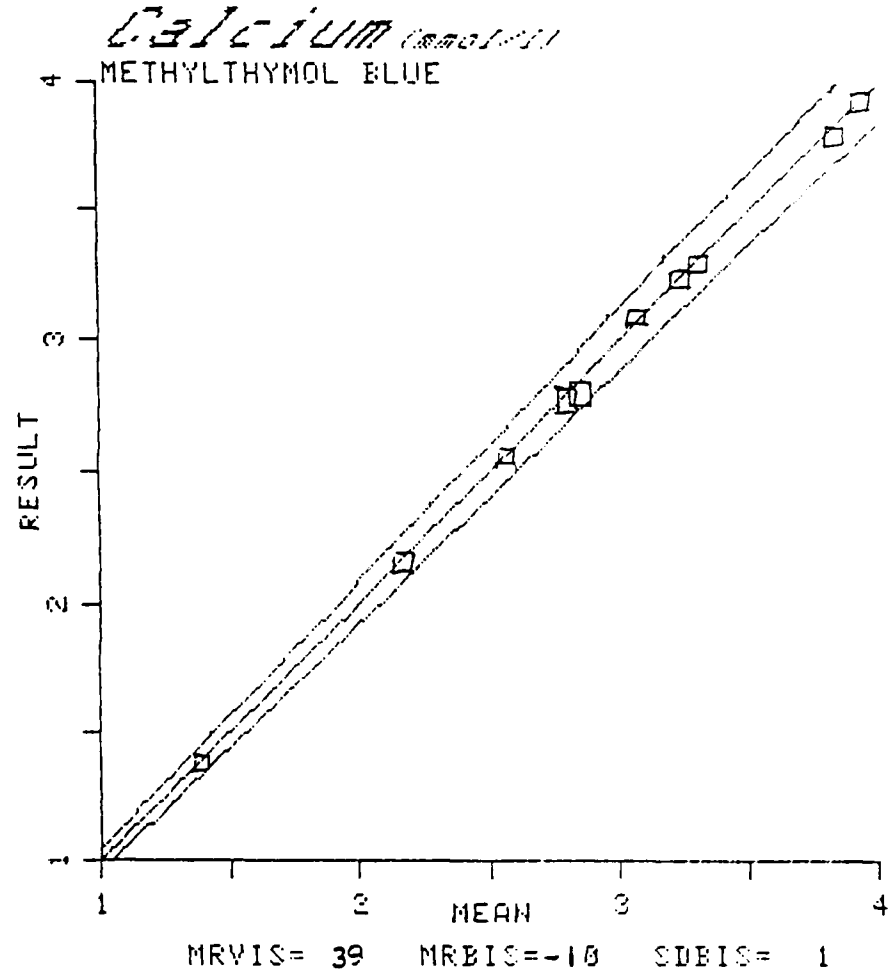
and general method reliability. This will pertain to reliability within the method group if a method-related figure is used as designated value. The SDBIS then yields information on each laboratory's consistency of bias (or, in many circumstances, within-laboratory repeatability). The MRBIS yields information on bias, provided that method is not taken into account in derivation of designated values though bias information can be obtained for subclassifications within a method group, eg for different manufacturers' instruments within an indirect ISE group for sodium and potassium. The average MRVIS and SDBIS figures confirm the overall impressions of relative performance discussed above, though they also suggest the possibility that laboratories in the manual/discrete analyser CPC group show slightly more inconsistency of bias.

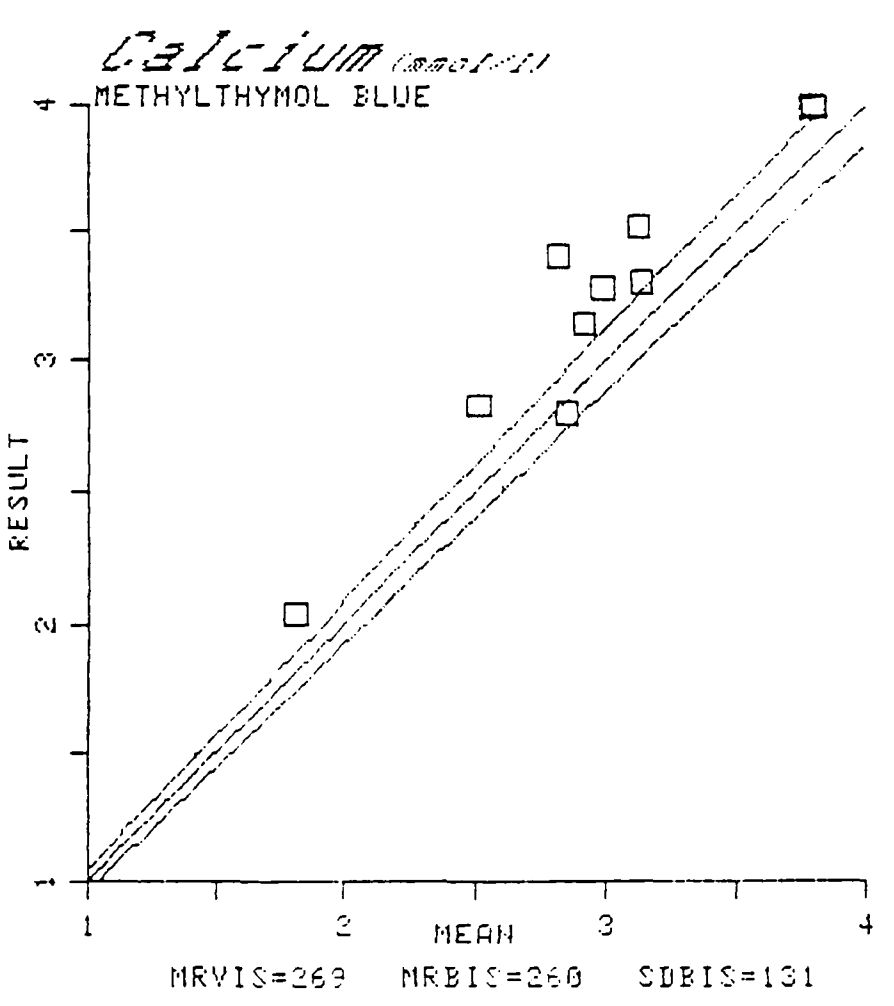
The inhomogeneity of some method groups is apparent from Table 8.3, which gives running scores for the MTB calcium method group from 1983. There is considerable diversity in performance, with some laboratories showing satisfactory and others grossly unsatisfactory performance. Examination of the scores for individual laboratories showed the primary problem to be one of inconsistency of bias; very few participants had a large and consistent bias. These data were circulated to participants in 1984, with the recommendation that users of this method should re-assess their performance and consider a change of method if this was unsatisfactory. The MTB method has since decreased in use but it has not been eliminated completely. The same diversities in performance persist, as demonstrated in the graphs of laboratory result against method mean for two participant laboratories using this procedure shown in Figure 8.1.

Table 8.3 Average and range of running scores for serum calcium assay by methylthymol blue method group in UKEQAS for General Clinical Chemistry, 1983 and 1984

	MRVIS	MRBIS	SDBIS
November 1983 (distribution 257; n = 33)			
Average	74	-5	80
Range	26 - 198	-128 - +148	25 - 235
June 1984 (distribution 267; n = 36)			
Average	80	-3	86
Range	35 - 198	-142 - +129	23 - 196

Figure 8.1 Graphs of laboratory result against method mean for calcium by methylthymol blue in UKEQAS for General Clinical Chemistry, 1986, for participants with (A) good performance, and (B) poor performance





The EDTA titration method is no longer used by UK laboratories, which may be related to the demonstration by EQASs of its overall inferiority. Before its disappearance in 1983, however, the remaining group of very few laboratories using EDTA titrimetric procedures performed reliably. Thus the average CV for 2-3 laboratories in late 1982 was 2.0%, which is good in comparison with the majority of average CVs in Tables 8.1 and 8.2. This confirms previous experience with methods which seem to require considerable effort to maintain: some participants can attain excellent performance with them, but the majority of laboratories attempting to apply them experience difficulties. The manual dithizone procedure for blood lead (Table 8.4) provides a further example of this situation. Overall, the dithizone method shows poor performance (as judged by MRVIS) relative to other methods, but of the five laboratories in the group one showed good performance. Studies through the scheme of reproducibility and recovery of added lead (Bullock et al, 1986c) confirmed these conclusions (Table 8.4).

The effect, which may be likened to the 'hands of the devoted' effect in new methods, suggests considerable caution in interpreting EQAS data obtained from a group of a few laboratories only. Besides the purely statistical considerations, the performance of these laboratories is thus much less likely to be representative of the method's routine performance than is performance information derived from a larger group of participants.

8.4 Assessment of factors other than interlaboratory agreement

As discussed above, some EQAS designs can also yield information on aspects of performance such as accuracy and turnround time,

Table 8.4 Average MRVIS and recovery and repeatability studies in UKEQAS for Lead in Blood, 1980, with ranges for dithizone method group

	n	MRVIS	Recovery of added lead (%)	Within-lab CV (%)
Overall	90	72	98.1	8.6
Delves cup	33	62	93.8	8.1
Electrothermal atomisation	33	65	102.6	7.5
Other atomic absorption	14	63	103.2	8.1
Dithizone (range)	5	90 (50-163)	79.6 (48-123)	20.3 (12.1-35.9)

which may be of particular importance in some assays.

8.4.1 Accuracy

Paracetamol and salicylate are added to drug-free equine serum to yield specimens for the UKEQAS for Salicylate and Paracetamol, and recovery may therefore be assessed.

Figure 8.2 shows the relationships between percentage recovery and added paracetamol for the enzymic and Glynn & Kendal method groups. The enzymic procedures show a much more quantitative recovery of paracetamol, with less concentration-dependence. There is also a clear relationship between interlaboratory agreement and analyte level (Figure 8.3) so simple examination of average CVs may not be sufficient to determine the relative merits of the various methods. Figure 8.3 therefore compares the relationships for the enzymic and the Glynn & Kendal method groups, from which the superiority of the enzymic procedures is readily apparent (Bullock, 1987).

These impressions are confirmed by the average MRVISs (December 1986) of 41 for enzymic and 73 for Glynn & Kendal users; the average SDBISs, of 48 and 73 respectively, also indicate superior within-laboratory consistency for the enzymic methods. Demonstration of these facts by the scheme has been associated with a large swing towards the enzymic methods, used by 57% of participants by December 1986 but only 33% two years earlier. The evidence for any link between these two facts is purely circumstantial, and the reagent manufacturers have also promoted their products, but it is reasonable to assume that EQAS results have contributed towards this change in practice. Such a contribution may of course be mediated by the early introduction

Figure 8.2 Relationship between percentage recovery and paracetamol added for enzymic and Glynn & Kendal method groups in UKEQAS for Salicylate and Paracetamol, 1985-1986

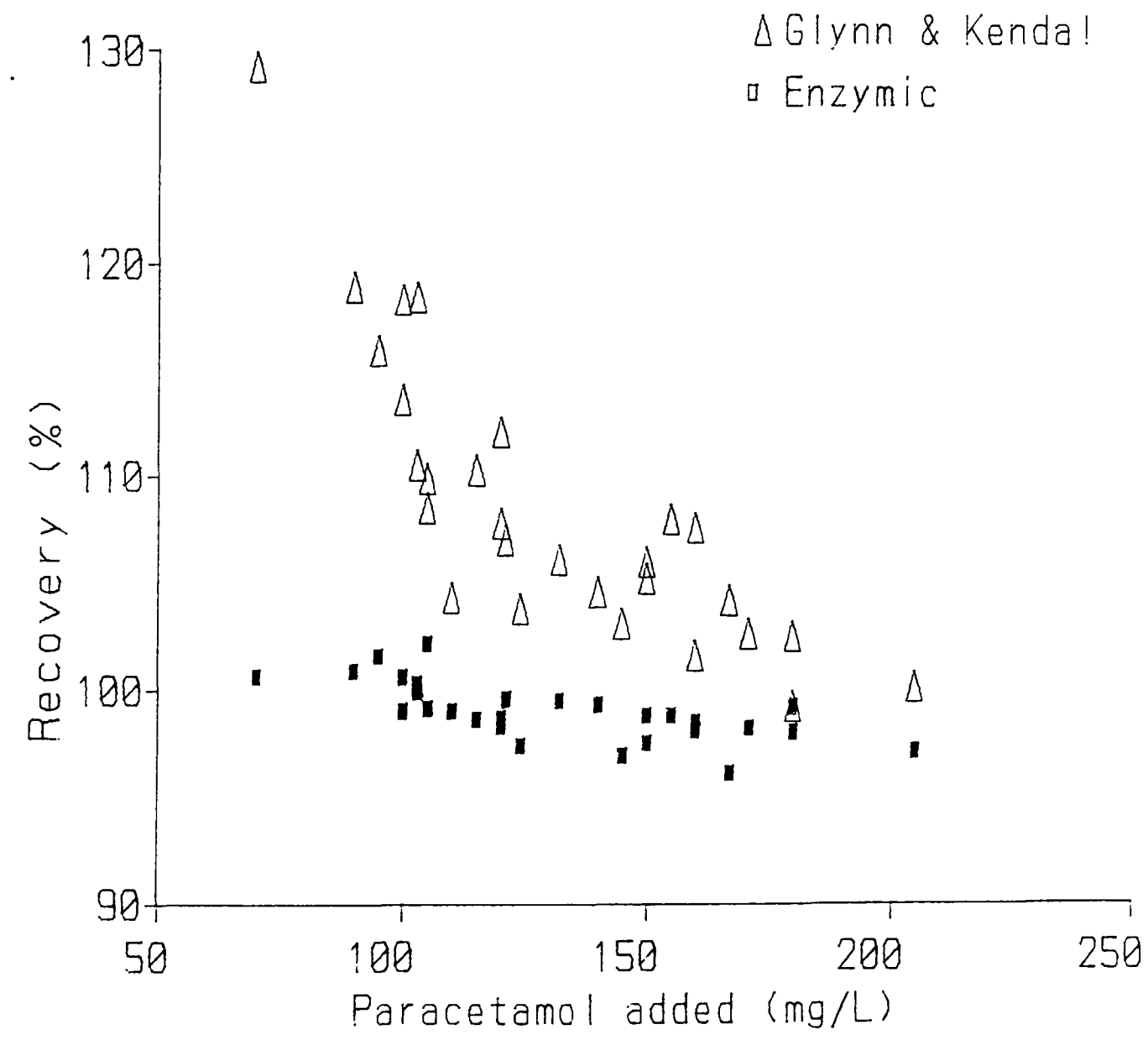
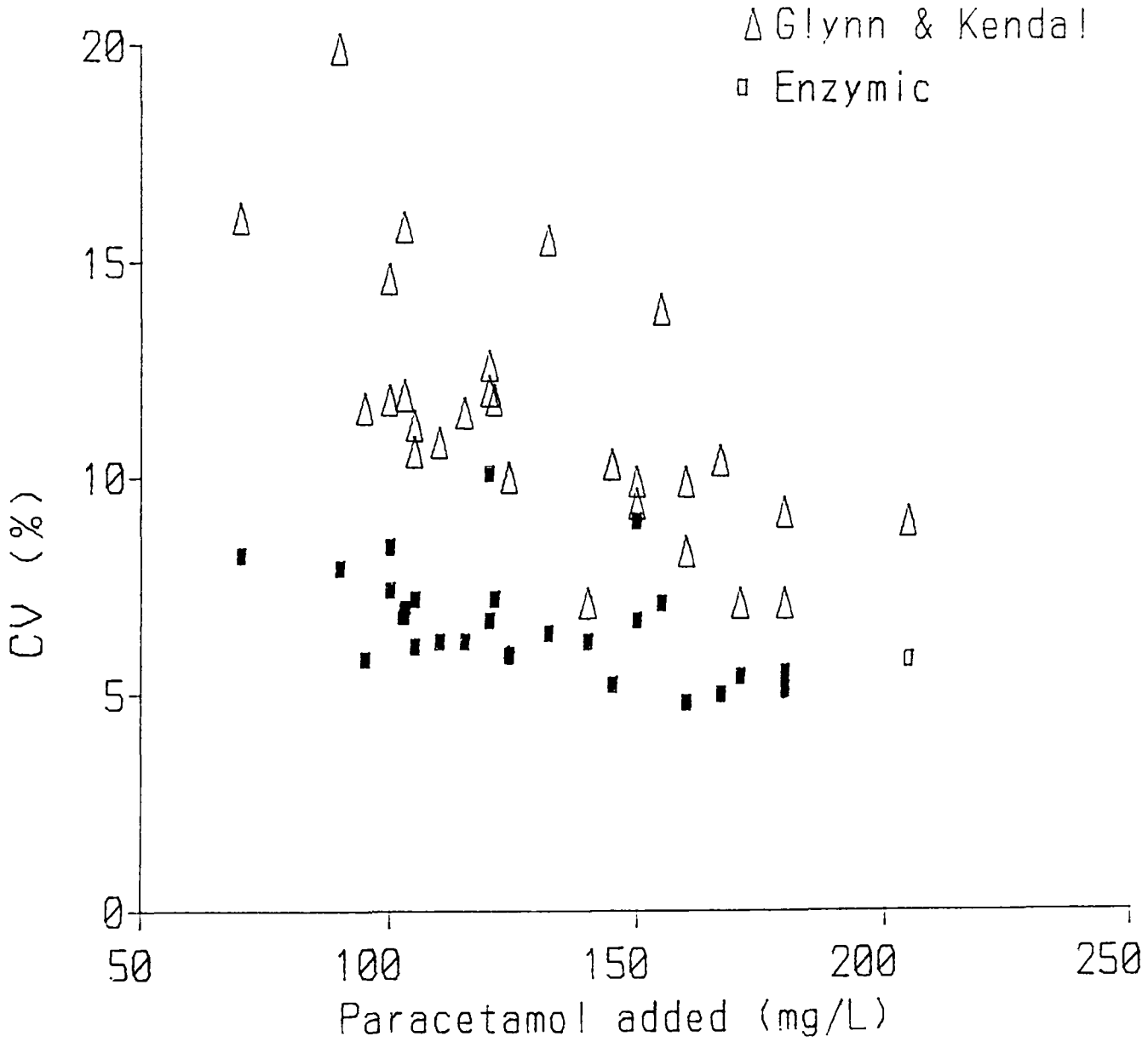


Figure 8.3 Relationship between interlaboratory agreement (CV) and paracetamol added for enzymic and Glynn & Kendal method groups in UKEQAS for Salicylate and Paracetamol, 1985-1986



of scoring of individual laboratories' performance (see Chapter 9) as well as by the presentation of overall method-related data.

The design of the UKEQAS for PKU Screening also allows assessment of a number of aspects of method performance. Accuracy of overall interpretation can be quantitated here through the scoring system for participants' classification of each specimen as having a 'normal', an 'intermediate' or a 'high' phenylalanine level (Chapter 7; Appendix I.2.7). Since errors of classification are infrequent, and the scheme has few participants overall (and consequently even fewer in some method groups), performance must be cumulated over a long period to gain a reliable assessment. Figure 8.4 shows the average score per survey (6 specimens) over Surveys 6-19, with participants classified according to their method; these are broad groupings, with many variations introduced by individual laboratories. From this it is apparent that the performance of the four method groups is very similar, perhaps suggesting that each laboratory has now controlled most aspects of their own method. Indeed, the two laboratories having high average scores obtained the contributory scores early in their participation.

8.4.2 Turnround time

In addition to analytical performance, the UKEQAS for PKU Screening also endeavours to assess turnround time through requesting information on the dates of specimen receipt, specimen analysis and results despatch. Turnround is important in this assay because the response to treatment is better the earlier treatment is started (Williamson et al, 1981; Smith, 1985).

In some cases delays before receipt of specimens have been found to be due to delays within the recipient hospital's postal

Figure 8.4 Relationship between analytical method and average score for Surveys 6-19 in UKEQAS for PKU Screening

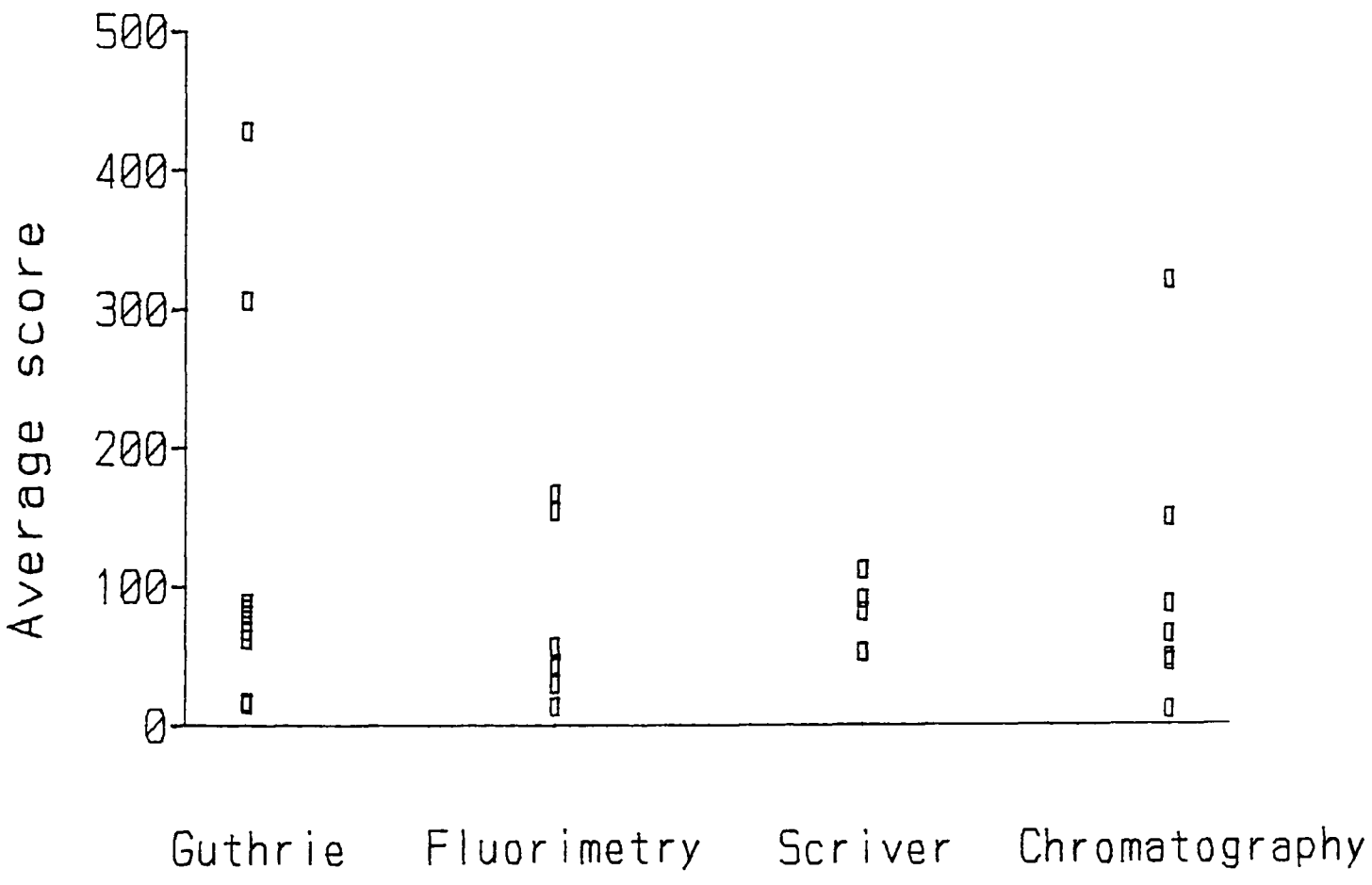


Figure 8.5 Relationship between analytical method and average turnround time (delay between specimen receipt and analysis; working days) for Surveys 17-20 in UKEQAS for PKU Screening

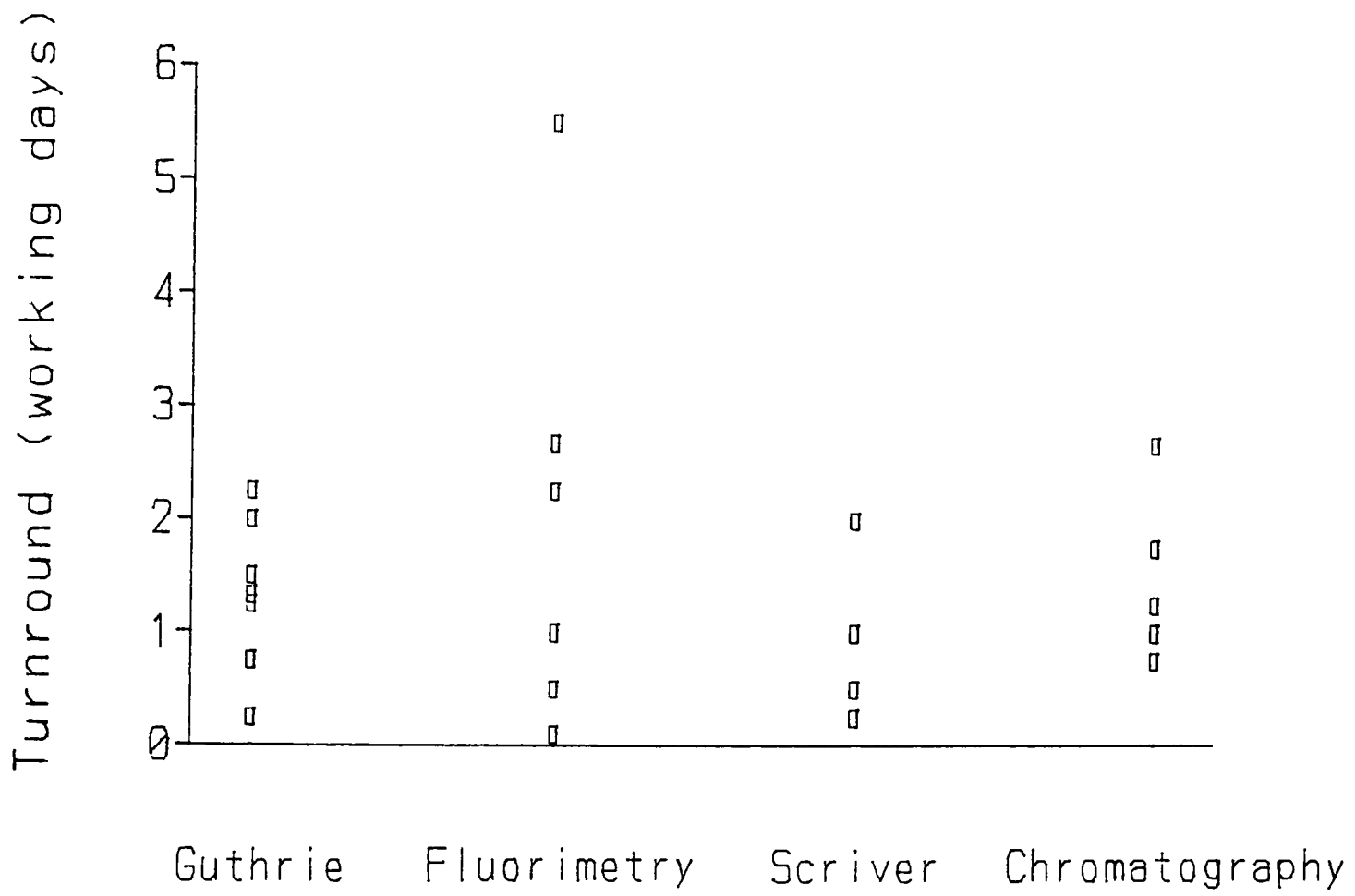
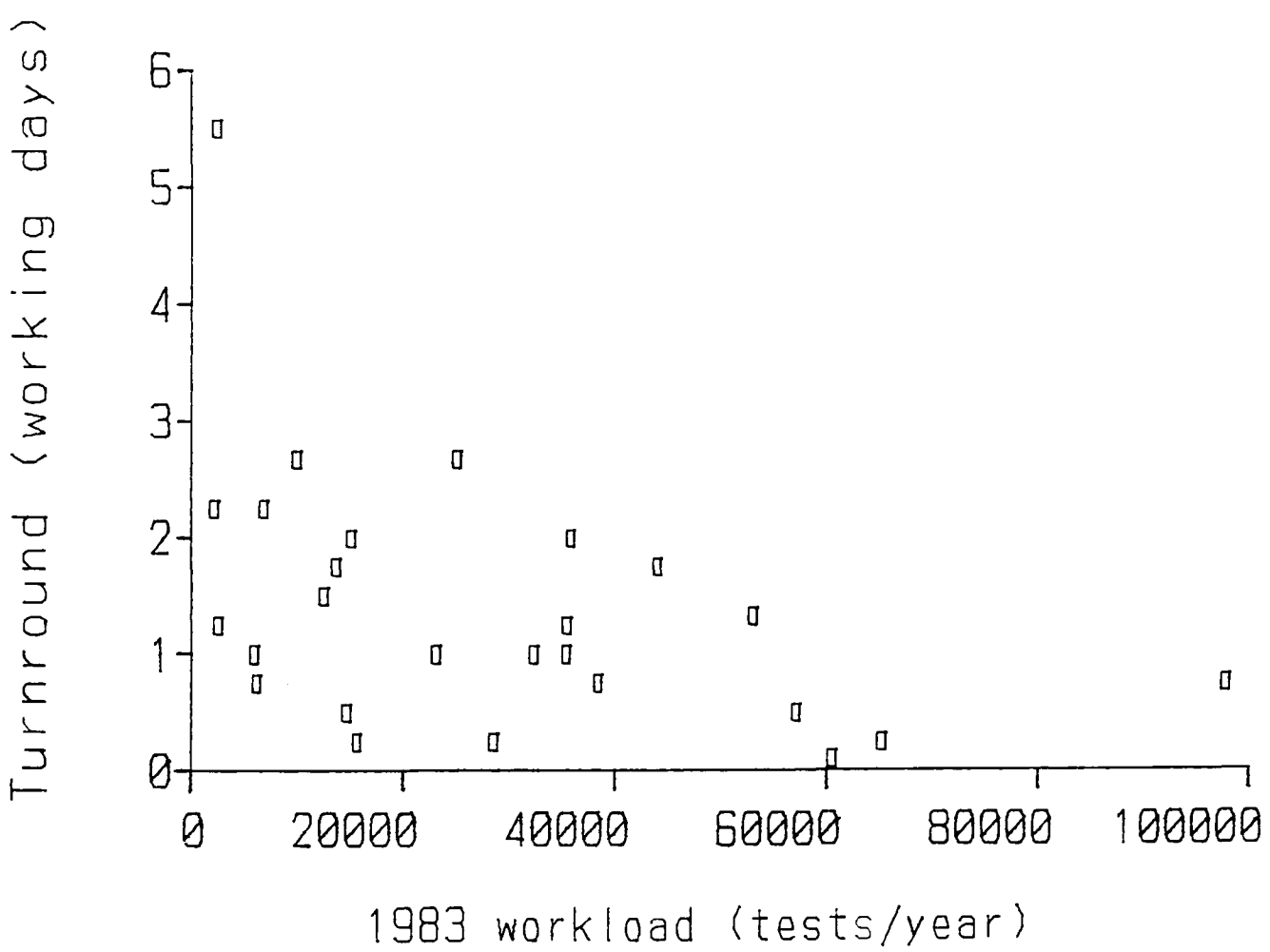


Figure 8.6 Relationship between 1983 workload and average turnround time (delay between specimen receipt and analysis; working days) for Surveys 17-20 in UKEQAS for PKU Screening



arrangements, and investigations have led to elimination of this problem for clinical screening specimens as well as for EQA specimens. The most useful interval susceptible to study is that between specimen receipt and analysis. The relationship between method and average interval (in working days) between specimen receipt and analysis over Surveys 17-20 is shown in Figure 8.5. Here the Guthrie and Scriver groups seem to have a slightly faster turnround of specimens, with greater variability within the other groups. As Figure 8.6 demonstrates, however, this may be related in part to the relationship with annual workload for PKU screening.

Total turnround (specimen receipt to results despatch) has not been studied in detail since the nature of the scheme prevents participants from dealing with the results in their routine manner: specimens are grouped on a single request/results document whereas routine specimens are dealt with independently and turnround time is likely to be shorter for specimens having raised phenylalanine levels.

8.5 Summary

EQAS data provide an excellent medium for assessing the routine performance of analytical procedures in current use, though the estimate may be slightly more favourable than would be obtained in practice. Use of a scoring system can be a more convenient and sensitive means of performance assessment than average interlaboratory CVs. This also facilitates identification of heterogeneity of performance, indicating susceptibility to problems, within method groups.

Scheme designs may also enable aspects of performance other than

interlaboratory agreement, such as accuracy and turnround time, to be assessed. For example, schemes incorporating known additions of pure analyte to analyte-free matrix provide a reliable assessment of method accuracy. The assessment of relative biases in schemes subject to matrix effects is, however, much less reliable.

Overall, the ability to assess method performance is limited by the degree of discrimination available in the classification used in the scheme, and new methods are not immediately susceptible to study.

Chapter 9:

SCORING AS A STIMULUS TO IMPROVED LABORATORY PERFORMANCE

9.1 Introduction - the assessment of performance

Data reduction through use of a scoring system is invaluable in rendering the performance information from EQA easier to interpret in terms of the assessment of interlaboratory agreement over geography (ie among schemes) and time (ie within a scheme), as described in Chapter 4. The original intention behind the introduction of scoring systems, however, was to make EQAS information more comprehensible to participants (see section 3.4.2).

Comparison of results with targets in the form of overall or method-related means or histograms is difficult. If a number of analytes (10 to 20 in many schemes covering general clinical chemistry) are involved the problem becomes more complex, even for a single specimen distribution only. The assessment 'by eye' of the data accumulated over several distributions is yet more difficult, and anything other than a gross change in performance is effectively impossible to detect.

A scoring system makes such assessment much simpler through the cumulation of information in more readily comprehensible form. This cumulation may refer to a single analyte only, with use of results covering a period, to permit appraisal of performance relative to other participants at that time or to the individual laboratory's past performance; for the latter, the score must be independent of other participants' performance, as discussed in section 4.7.2.

Appropriate scoring systems can also permit combination of data from more than one analyte, and so provide an overall assessment of the laboratory's overall performance. Again appraisal can be against others or against previous performance. Such measures of 'performance at a glance' are extremely powerful distillations of EQAS information and are thus a useful managerial tool, providing the limitations discussed below are recognised.

The most primitive systems give information in qualitative form, eg the 'pass/fail' criteria applied by licensing schemes (Bundesärztekammer, 1971; Stamm, 1975), and thus give only the crudest reflection of performance. More sophisticated systems such as VI scoring yield quantitative information as a numeric score, retaining the potential for easy interpretation and being susceptible to graphical presentation (see Chapter 10). These have been of great assistance in enabling laboratories to recognise the existence of suboptimal performance and in stimulating them to improve towards this goal.

Most systems are based on an estimate of total error, but further refinement can also provide guidance on the type of contributory errors. Schemes usually rely on precision information gained separately through the laboratory's IQC programme to assist in interpretation (though some do assess imprecision through repeated distribution of the same specimens), and provide estimates of bias and its consistency. A bias may be inconsistent due to poor within- or between-assay precision, but the potential presence of other contributory factors such as nonlinearity or other concentration-dependent bias, short- or long-term accuracy changes and specimen/method interactions usually precludes such a simplistic interpretation.

Application of a good scoring system can then provide a robust and reliable assessment of overall performance, indicating improvement or deterioration and identifying any need for improvement. It should then also assist in the resolution of any problems indicated through provision of more detailed information for the individual analytes concerned, since different approaches are required for problems arising from eg erroneous calibration and nonlinearity. To avoid unnecessary sifting through irrelevant information (and thus defeat the object of using a scoring system) the scoring system and report format should enable the participant to proceed in a stepwise manner through a hierarchy of increasingly detailed performance data, deciding at each stage whether it is necessary to proceed further.

Though detailed information must be provided by scoring systems, this sophistication may make it increasingly difficult to relate performance scores to patient care, especially where they are determined by consideration of the state of the art rather than clinical requirements. In such contexts there is also perhaps a need for a 'pass/fail' system in addition, to indicate in a very simple manner whether performance is good enough to satisfy medical decision criteria.

9.2 Hierarchical interpretation of scoring

The most helpful scoring system will fail in its objective if the scores are not presented in such a way as to simplify the interpretation. Each participant has only limited resources (in term of time, effort and ability) to devote to this interpretation, and experience suggests that those in most need

of acting on EQA data devote or choose to devote the least.

A clinical chemist receiving a report needs to make decisions on a series of questions, which are usually self-terminating when a negative answer is given:

- do I have a major overall problem?
- which analytes are contributing most to this?
- are these problems significant?
- what is the source of the errors in each case?

A well-designed combination of scoring system and report format can assist considerably in this process, and thus contribute to patient care not only through stimulation of improvement where this is indicated but also through removing the need for unnecessary investigation.

The application of VI scoring (defined in Appendix II) in the UKEQAS for General Clinical Chemistry will be used as an example of such application. Consider then the report of which the primary and an example secondary pages are shown in Figures 9.1 and 9.2 respectively.

9.2.1 The OMRVIS and assessment of overall performance

The first element to be considered is the laboratory's OMRVIS, which represents the average of the 40 most recent VISs. The value of 66 here will have been derived from roughly the last 3 fortnightly distributions (6-8 weeks), since the laboratory is receiving scores for about 14 analytes.

The most direct comparison is with the mean OMRVIS for all participants (60), which shows the laboratory to be slightly worse than average. A more appropriate comparison is with the mean OMRVIS for participants in the same size group, ie with the

Figure 9.1 Primary pages of report for participant laboratory in UKEQAS for General Clinical Chemistry

U.K. EXTERNAL QUALITY ASSESSMENT SCHEME

FOR GENERAL CLINICAL CHEMISTRY

Clinical Chemistry Department, Queen Elizabeth Hospital, Birmingham B15 2TH, U.K. (Tel. 021-472 1311 Ext. 3172)

MATERIAL DISTRIBUTED: ARMTROL NORMAL, LOT 901 (BOVINE SERUM) LAB. No. [redacted] DATE: 27-JUL-87

DISTRIBUTION NUMBER: 334

477 LABS PARTICIPATED IN THIS DISTRIBUTION

	NA	K	UREA	GLUC	CALC	FE	URATE	CREAT	BILI	T.PROT	CHOL	MG	AST	ALT	LD	CK
RECALCULATED RESULTS EXCLUDING THOSE OUTSIDE ± 3 SD:																
No. of RESULTS	459	459	460	449	419	221	350	421	394	399	336	269	335	249	183	285
No. EXCLUDED	4	4	1	7	7	5	7	4	8	5	-	4	4	5	1	2
MEAN VALUE	139.3	4.06	6.0	5.5	2.39	24.8	0.255	103.6	17.5	71.7	3.74	0.776	36.1	30.8	639.6	118.6
STD. DEVIATION	1.5	0.09	0.3	0.3	0.06	2.1	0.023	10.3	2.5	2.2	0.20	0.062	7.1	5.6	155.7	28.4
COEFF OF VAR.	1.1	2.11	5.9	4.8	2.54	8.5	9.008	10.0	14.2	3.1	5.30	7.933	18.7	18.1	25.9	23.9
METHOD MEAN	139.37	4.05	6.06	5.51	2.442	24.26	0.253	108.8	16.88	71.88	3.74	0.739		29.4		124.2
YOUR RESULT	139.0	4.20	5.70	5.00	2.35	24.0	0.250	120.0	17.0	73.0		0.74		42.0		106.0
YOUR BIS	-16	131	-104	-120	-94	-7	-15		4	40		-62		248		-79
YOUR MRVIS	50	71	106	46	55	33	62		12	57		40				72

Overall Mean Running Variance Index Score 66

AVERAGE OMRVIS. ALL LABS 60

AVERAGE OMRVIS. SIZE GROUP I 64

AVERAGE OMRVIS. SIZE GROUP II 57

AVERAGE OMRVIS. SIZE GROUP III 56

SIZE GROUP 3 (124 LABS)

```
  1- 10 '
 11- 20 '
 21- 30 '
 31- 40 'XXXXXXXXXXX
 41- 50 'XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
 51- 60 'XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
--> 61- 70 'XXXXXXXXXXXXXXXXXXXX
 71- 80 'XXXXXXXXX
 81- 90 'XXXXXX
 91-100 'XX
101-110 'X
111-120 '
121-130 '
   >130 '
```


U.K. EXTERNAL QUALITY ASSESSMENT SCHEME

FOR GENERAL CLINICAL CHEMISTRY

Clinical Chemistry Department, Queen Elizabeth Hospital, Birmingham B15 2TH, U.K. (Tel. 021-472 1311 Ext. 3172)

MATERIAL DISTRIBUTED: ARMTROL NORMAL, LOT 901 (BOVINE SERUM) LAB. No. [redacted] DATE: 27-JUL-87

DISTRIBUTION NUMBER: 334

477 LABS PARTICIPATED IN THIS DISTRIBUTION

	ALP	AMS
RECALCULATED RESULTS EXCLUDING THOSE OUTSIDE ± 3 SD:		
No. of RESULTS	358	329
No. EXCLUDED	2	1
MEAN VALUE	112.4	255.7
STD. DEVIATION	18.6	82.5
COEFF. OF VAR.	16.5	32.3
METHOD MEAN	100.0	363.2
YOUR RESULT	97.0	294.0
YOUR BIS	-19	-166
YOUR MRVIS	40	164

Overall Mean Running Variance Index Score 66

AVERAGE OMRVIS. ALL LABS 60

AVERAGE OMRVIS. SIZE GROUP I 64

AVERAGE OMRVIS. SIZE GROUP II 57

AVERAGE OMRVIS. SIZE GROUP III 56

SIZE GROUP 3 (124 LABS)

```
  1- 10 '
 11- 20 '
 21- 30 '
 31- 40 'XXXXXXXXXXX
 41- 50 'XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
 51- 60 'XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
--> 61- 70 'XXXXXXXXXXXXXXXXXXXX
 71- 80 'XXXXXXXXX
 81- 90 'XXXXXX
 91-100 'XX
101-110 'X
111-120 '
121-130 '
   >130 '
```

©Copyright The data in UKEQAS reports are confidential to the NHS, and participants should contact the scheme Organiser before quoting data from the scheme.

Figure 9.2 Example secondary pages of report for participant laboratory in UKEQAS for General Clinical Chemistry

UKEQAS FOR GENERAL CLINICAL CHEMISTRY LABORATORY [redacted] DISTRIBUTION 334 27-JUL-97

RESULTS FOR ASI (U/L)
(EXCLUDING VALUES OUTSIDE +/- 3.0 S.D.)

	NO.	MEAN	S.D.	C.V.
OVERALL	335	38.1	7.1	18.7
SCE OPTIMISED (TRIS), 37'C	142	38.0	4.6	12.1
OTHER "OPTIMISED", 37'C	112	37.2	3.6	9.8
CONTINUOUS FLOW UV, 37'C	36	43.2	5.6	13.1
OTHER "OPTIMISED", 30C	11	31.4	19.6	62.3
PYRIDOXAL PHOS/OPTIMISED, 37'C	13	58.5	10.8	18.5
OTHER METHOD OR TEMPERATURE	17	29.7	11.2	37.8

YOUR RESULT:

YOUR BIAS INDEX SCORE:

YOUR MRVIS : YOUR MRBIS : YOUR SDBIS :

RESULTS FOR ALT (U/L)
(EXCLUDING VALUES OUTSIDE +/- 3.0 S.D.)

	NO.	MEAN	S.D.	C.V.
OVERALL	249	30.8	5.6	18.1
-->SCE OPTIMISED (TRIS), 37'C	111	29.4	4.1	14.0
OTHER "OPTIMISED", 37'C	45	32.5	4.1	12.5
CONTINUOUS FLOW UV, 37'C	21	38.7	8.0	20.6
OTHER "OPTIMISED", 30'C	8	24.4	6.8	27.8
PYRIDOXAL PHOS/OPTIMISED, 37'C	4	34.5	4.8	13.9
OTHER METHOD OR TEMPERATURE	15	25.3	8.9	35.2

YOUR RESULT: 42.0

YOUR BIAS INDEX SCORE: 248

YOUR MRVIS : YOUR MRBIS : YOUR SDBIS :

RESULTS FOR LD (U/L)
(EXCLUDING VALUES OUTSIDE +/- 3.0 S.D.)

	NO.	MEAN	S.D.	C.V.
OVERALL	183	639.6	165.7	25.9
SCE OPTIMISED (P TO L), 37'C	62	735.9	50.6	6.9
OTHER "OPTIMISED" P TO L, 37'C	77	699.2	64.8	9.3
L TO P, 37'C	24	297.7	22.4	7.5
OTHER METHOD OR TEMPERATURE	14	504.2	208.2	41.3

YOUR RESULT:

YOUR BIAS INDEX SCORE:

YOUR MRVIS : YOUR MRBIS : YOUR SDBIS :

RESULTS FOR CK (U/L)
(EXCLUDING VALUES OUTSIDE +/- 3.0 S.D.)

	NO.	MEAN	S.D.	C.V.
OVERALL	285	118.6	28.4	23.9
-->SCE OPTIMISED (NAC), 37'C	223	124.2	22.9	18.4
ACB/DGKC OPTIMISED, 30'C	14	76.0	16.0	21.0
GLUTATHIONE ACTIVATED, 37'C	14	115.9	20.8	18.0
OTHER METHOD OR TEMPERATURE	27	103.7	44.1	42.5

YOUR RESULT: 106.0

YOUR BIAS INDEX SCORE: -79

YOUR MRVIS : 72 YOUR MRBIS : 48 YOUR SDBIS : 77

UKEQAS FOR GENERAL CLINICAL CHEMISTRY LABORATORY [redacted] DISTRIBUTION 334 27-JUL-97

RESULTS FOR ALP (U/L)
(EXCLUDING VALUES OUTSIDE +/- 3.0 S.D.)

	NO.	MEAN	S.D.	C.V.
OVERALL	358	112.4	18.6	16.5
SCE/DGKC OPTIMISED (DEA), 37'C	145	124.1	12.0	9.6
CONTINUOUS FLOW 4NPP/AMP, 37'C	86	116.0	11.1	9.6
-->MANUAL/DISCRETE 4NPP/AMP, 37'C	62	100.0	13.9	13.9
OTHER METHOD OR TEMPERATURE	55	93.2	18.7	20.1

YOUR RESULT: 97.0

YOUR BIAS INDEX SCORE: -19

YOUR MRVIS : 40 YOUR MRBIS : 23 YOUR SDBIS : 52

RESULTS FOR AMYLASE (U/L)
(EXCLUDING VALUES OUTSIDE +/- 3.0 S.D.)

	NO.	MEAN	S.D.	C.V.
OVERALL	328	255.7	82.5	32.3
PHADEBAS TABLET TEST, 37'C	188	246.6	24.5	9.9
-->BCL COLORIMETRIC, 37'C	70	363.2	43.3	11.9
OTHER METHOD OR TEMPERATURE	62	171.0	98.5	57.6

YOUR RESULT: 294.0

YOUR BIAS INDEX SCORE: -166

YOUR MRVIS : 164 YOUR MRBIS : -154 YOUR SDBIS : 43

average performance of laboratories broadly similar in terms of workload. The laboratory is in group III, representing an annual workload of >500,000 tests/year, and the comparison figure is 56. Thus this laboratory's performance is somewhat worse than the average for larger laboratories, and may indicate a cause for concern.

Further evidence is provided by the frequency distribution (histogram) of the OMRVISs for participants in this size group. This shows not only that the laboratory's OMRVIS of 66 is worse than the average but also that an appreciable proportion of laboratories in a similar situation can attain an OMRVIS <50, and about 10% a score below 40. The best performance attained should be the goal of each participant, and even those with better than average performance should ideally also strive for this (subject to the reservations discussed in section 2.2).

Thus the evidence available suggests that the laboratory's performance is less satisfactory than that of others, indicating a need for more detailed investigation.

9.2.2 Assessment of performance for individual analytes

What is then the source of this apparent deficiency? This information is essential before proceeding further. It is also useful in assessing priorities for allocating improvement effort, since effort will be most effectively applied where performance is initially worst.

Examination of the BISs for this distribution shows particularly poor scores for potassium, urea, glucose, ALT and amylase, so are these the analytes requiring attention? Possibly not, since the quality of individual results and hence the resulting BISs is

influenced by the analyte concentrations (see Chapter 6) in the specimen distributed and by random variation.

To gain the benefit of cumulation over time (and hence over a range of analyte concentrations) the MRVIS for each analyte must be considered. This is the mean of the 10 most recent VISs for the analyte, usually derived from about 6 months' distributions. In this case the MRVISs for most analytes appear satisfactory, apart from the values of 164 for amylase, 106 for urea, 72 for CK and 71 for potassium. These indicate clearly that the laboratory's higher than average OMRVIS stems primarily from the unsatisfactory performance for these 4 analytes, and attention must be given to improving performance for these before the overall situation can be improved.

Is the performance for these analytes in fact unsatisfactory, or is it merely poor relative to that of other participants but quite adequate for the clinical application of the assay? Certainly the CCVs, which 'scale' the VISs and MRVISs for each analyte, are based solely on the state of the art (in 1972; their continuing validity is discussed in section 4.8) and might therefore prove misleading; the issues are discussed in section 9.6 below. Removal of this scaling, however, reveals these MRVISs to represent average deviations from the DV of 18.9% for amylase, 13.3% for CK, 6.0% for urea and 2.1% for potassium. Intuitively such errors still appear significant. Though they satisfy the analytical goals proposed for within-laboratory CVs (Tables 2.1 and 9.1), the appropriateness of the goals for urea and for enzymes is questioned in section 9.6.

Table 9.1 Comparison, in terms of VIS, of performance attained in UKEQAS for General Clinical Chemistry with analytical goals (Fraser CG, personal communication)

	CCV	Average VIS	Analytical goal
Sodium	1.6	61	25
Potassium	2.9	55	90
Chloride	2.2	77	36
Urea	5.7	55	119
Glucose	7.7	44	29
Calcium	4.0	64	20
Phosphate	7.8	54	50
Iron	15.0	46	89
Urate	7.7	64	56
Creatinine	8.9	45	26
Bilirubin	19.2	37	66
Total protein	3.9	66	36
Albumin	7.5	53	23
Cholesterol	7.6	48	36
Lithium	11.0	34	-
Magnesium	10.0	55	11
Osmolality	2.9	46	-
AST	12.5	59	62
ALT	17.3	52	114
LD	13.2	58	27
CK	18.5	62	194
ALP	15.5	57	17
Amylase	11.5	80	-

9.2.3 Appraisal of contributory factors

To obtain more detailed performance data the secondary pages (Figure 9.2) must be considered; if no potential problems had been revealed by the assessment above then consideration could have been confined to the primary page of the report.

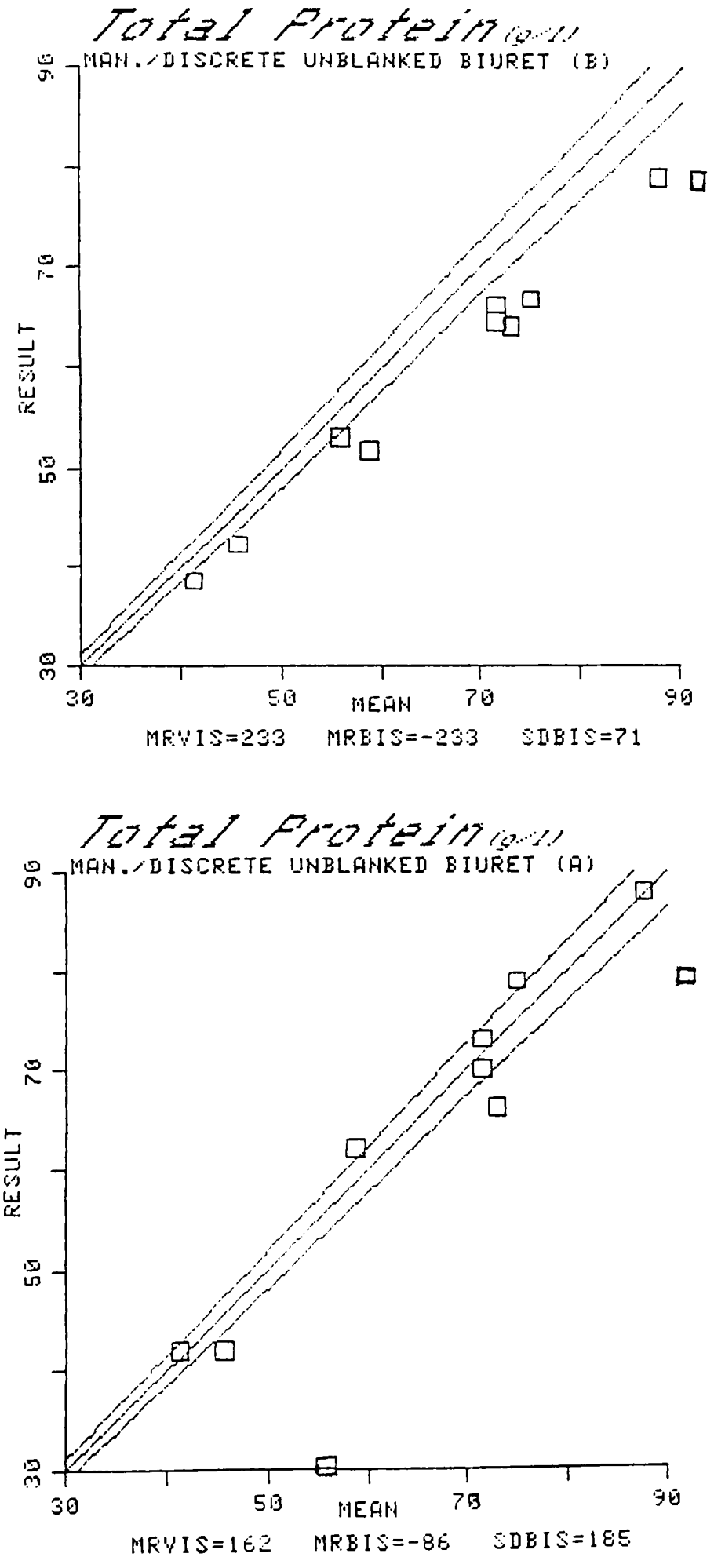
These pages give additional guidance in the form of data for each analyte, including the MRBIS and SDBIS (Appendix II.2; Bullock and Wilde, 1985). These provide information on the components of variance contributing to the MRVIS, the MRBIS representing a cumulated estimate of bias relative to the DV (the method mean in this scheme) and the SDBIS the consistency of this bias.

Thus a large (positive or negative) MRBIS and small SDBIS indicate a consistent proportional bias, a small MRBIS and large SDBIS an inconsistent bias, and large MRBIS and SDBIS a mixed picture. For ideal perfomance both MRBIS and SDBIS (and hence the MRVIS) should be small. These situations are exemplified in Figure 9.3, showing contributory results. Imprecision provides the source of the inconsistent bias for urinary oestrogen assay (Bullock and Wilde, 1985).

Apart from the simple case of consistent proportional bias, due in most cases to use of an inappropriate calibrant value, the differential diagnosis requires appraisal of the actual results contributing to the poor scores. Thus the scoring system reaches the limit of its usefulness, having indicated the existence of a problem and some of its likely causes, and there is no substitute in these circumstances for examination of individual results.

Here a graphical presentation as in Figure 9.3, discussed in more

Figure 9.3 Demonstration of the effects of bias and imprecision (inconsistent bias) upon the relationship between laboratory result (y axis) and designated value (x axis)

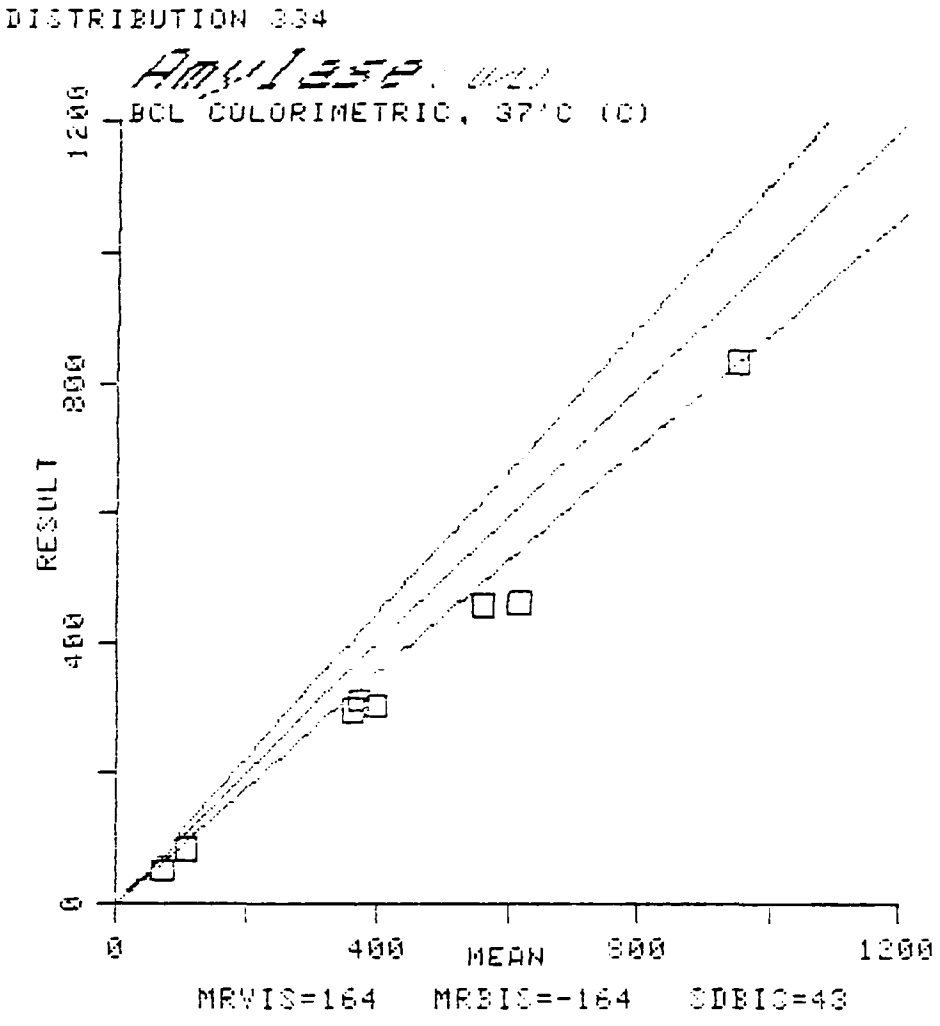


detail in section 10.5.2, proves invaluable in identifying the underlying analytical problem. Figures 9.4 and 9.5 show the relation of the results for the 4 problem analytes to the corresponding DVs. These demonstrate the consistent proportional negative bias for amylase, apparent compensated bias (probably arising from a combination of inappropriate calibration and blanking procedures; Lever et al, 1981) for urea, inconsistent bias (probably reflecting an imprecise assay) for CK, and somewhat inconsistent positive bias for potassium. Examination of the individual results revealed no consistent trend with time (section 10.5.2).

9.2.4 Method assessment

The secondary pages also provide information on method-related performance. This is of great benefit in assessing whether and how a laboratory should change method, as discussed in Chapter 8. The data are presented as mean, SD and CV for the specimen distributed, which suffer from the problems outlined in section 9.2.2 for individual VISs. Cumulative information is, however, available from the scheme Organisers in the form of average scores for each method group and for more detailed subclassifications, exemplified for sodium in Table 9.2. Here the average MRVISs provide guidance on the relative overall performance of each group, and the average SDBISs on the robustness of the method. The average MRBISs indicate any consistent biases between the method groups (eg Bullock et al, 1986c), though if the method mean is used as DV (as in the UKEQAS for General Clinical Chemistry) these indications will only be valid for the subclassifications within each method group. These aspects are discussed in more detail in sections 8.3 and 12.3.

Figure 9.4 Relationship between laboratory result (y axis) and designated value (method mean; x axis) for amylase and urea in UKEQAS for General Clinical Chemistry. * denotes result outside graph limits; lines are x=y and at BIS=±100



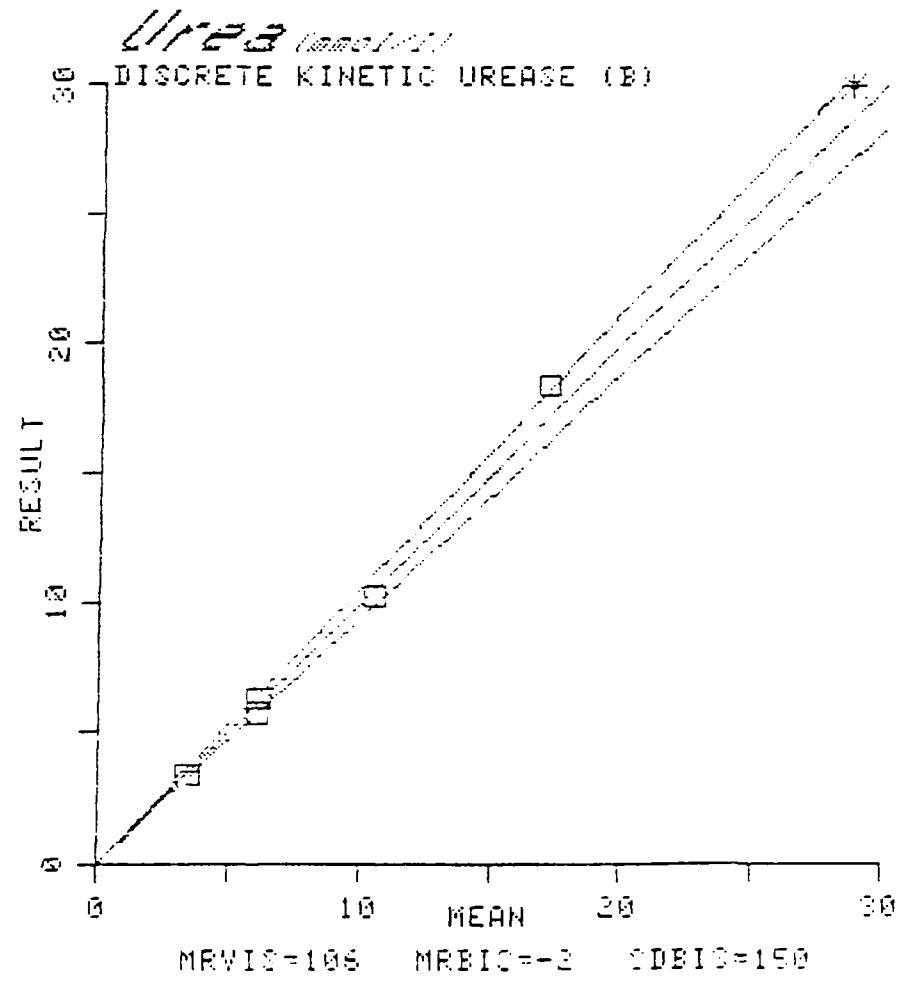


Figure 9.5 Relationship between laboratory result (y axis) and designated value (method mean; x axis) for CK and potassium in UKEQAS for General Clinical Chemistry

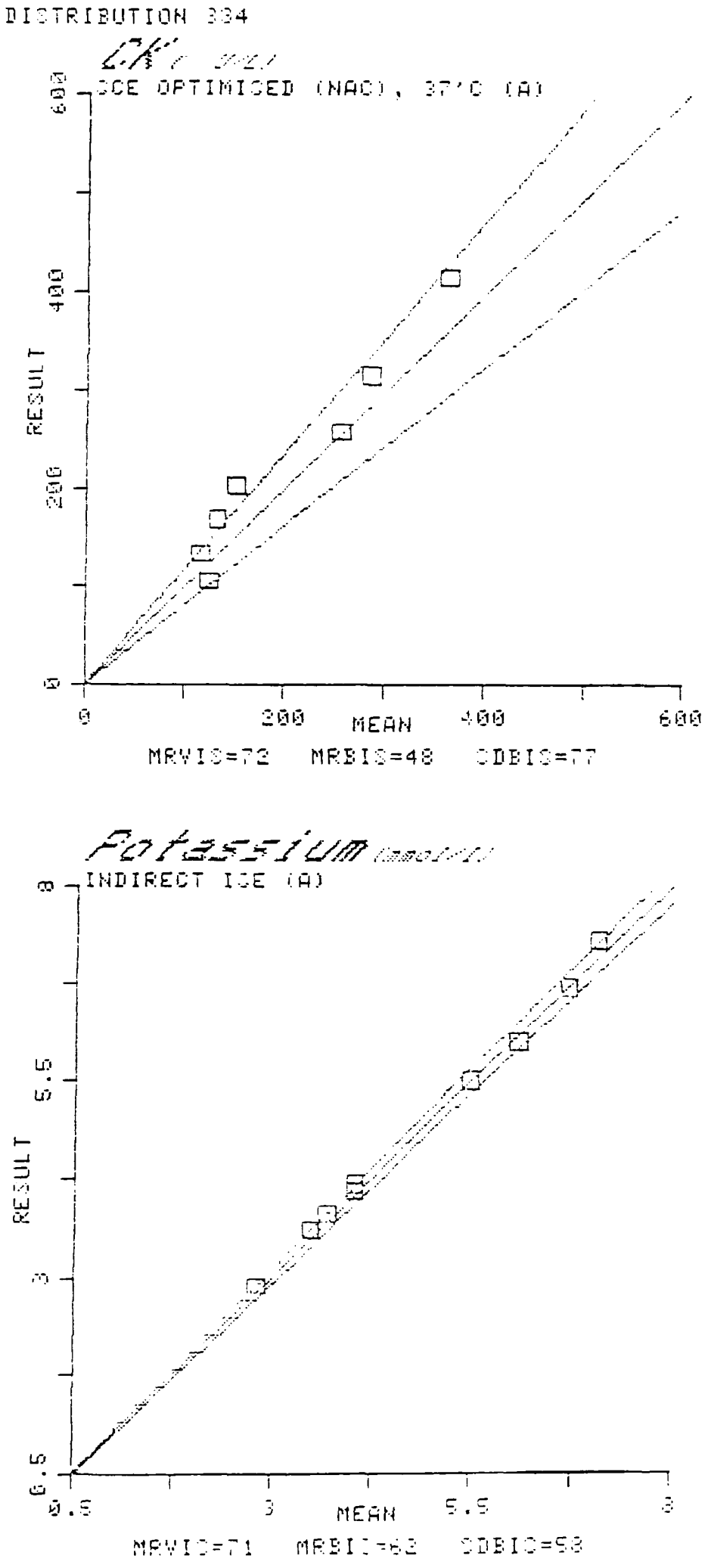


Table 9.2 Average running scores for method groups and subgroups for sodium assay in UKEQAS for General Clinical Chemistry, 1987. The Miscellaneous group is scored against the overall mean; no scores are calculated for the Other group, nor for groups with <15 results contributing to the method mean; *direct ISE is direct potentiometery on undiluted sample

	n	MRVIS	MRBIS	SDBIS
Overall	**487**	**63.0**	**-**	**71.9**
Continuous flow flame photometry				
Overall	102	57.2	-	65.6
Technicon flame photometer	90	55.7	-1.2	63.9
Other instrument	12	68.3	+20.0	78.8
Flame photometer with integral dilutor				
Overall	132	61.8	-	71.9
IL	66	63.7	+1.2	75.4
Corning	51	58.8	+8.4	68.1
Other instrument	14	62.8	-28.2	67.4
Indirect ion-selective electrode				
Overall	227	65.0	-	73.2
Technicon	76	65.4	-2.2	75.0
Beckman	100	58.5	+7.5	65.2
Other instrument	49	77.9	-8.0	87.0
Miscellaneous				
Flame photometer without dilutor	10	84.3	-18.2	91.2
Other method				
Overall	38			
Direct ISE* - IL	4			
Direct ISE* - Corning	21			
Direct ISE* - Nova	1			
Direct ISE* - other instrument	12			

9.3 Assessment of performance relative to other laboratories

One of the first thoughts of many EQAS participants on receiving a report is how their performance compares with that of their colleagues in other laboratories. This desire may be motivated by scientific curiosity or by the human competitive instinct.

A frequency distribution (histogram) of results provides such a means of comparison, but has limitations. The main disadvantage is that the comparison is based solely on analysis of a single specimen whereas scoring systems incorporate cumulation of data from a number of specimens over a period and a range of concentrations. Assuming that this is not an objection, however, how useful is a histogram for such comparison?

Certainly it shows how close the laboratory's result is to the target (mean, median, mode or externally-assigned designated value) relative to other participants' results. This may be useful on a qualitative basis, but the judgement is thus dependent upon the state of the art since histograms are usually presented in a 'scaled' form to fit consistently within the report. Thus the range commonly approximates to ± 2 SD limits from the mean and this is often interpreted as defining 'acceptable performance', with many participants assessing acceptability merely by whether their result falls within the histogram limits. Such interpretation is obviously superficial, reducing to an arbitrary 'pass/fail' criterion. Furthermore results lying close to the histogram limits have been known to be rationalised as showing that "there are 6 laboratories further away than we are".

More importantly these judgements are based on the general standard of performance of participants in this scheme, which may

in turn be unsatisfactory. For this reason the same laboratory's result might appear unacceptable in one EQAS but relatively satisfactory in another with worse interlaboratory agreement.

A 'hybrid' presentation incorporating scoring parameters may therefore be more useful. Such a presentation is illustrated in Figure 9.6, taken from the International EQAS (Appendix I.3.1). This uses an 'orthodox' histogram format for the central portion, taken as the range of concentrations which would give a VIS of 100 or less, with results further from the mean being grouped into VIS bands and denoted by symbols also suggesting their less satisfactory character.

For reliable comparison with other participants' performance, however, a scoring system is essential, which may take one of a number of forms, as reviewed in section 4.2. Assessment in terms of SD difference (SDD or 'Z score'; eg Wellcome Diagnostics, 1984) appears simple and reliable, and has the added attraction of scaling for differences in attainable performance between analytes, so it is ideal for comparing simply with other participants in the scheme. It suffers, however, from the same deficiency as the histogram presentation, ie dependence upon the performance of other participants and consequently does not offer an objective performance assessment.

Consider for example a laboratory which obtains a result of 146 mmol/l for sodium, for which the DV is 140 mmol/L. In a scheme such as the UKEQAS for General Clinical Chemistry a typical SD would be 1.7 mmol/L, yielding an SDD of 3.5, but interlaboratory agreement in the International EQAS is worse and the typical SD of 3.2 mmol/L yields an SDD of only 1.9. Thus SDDs do not provide an objective appraisal of a laboratory's performance, though they

Figure 9.6 'Hybrid' histogram format used in International EQAS. Central portion represents the concentration range yielding a VIS <100, with results lying outside being grouped into bands according to VIS

```
INTERNATIONAL EXTERNAL QUALITY ASSESSMENT SCHEME

                                HISTOGRAM DISTRIBUTION FOR URATE
                                I*****I****I*****I****I*****I****I*****I
  9    VIS       >200           *.........                              *
  3    VIS   151-200            *XXX                                    *
  5    VIS   101-150            *XXXXX                                  *
                                *                                       *
  2     4.62-   4.73            *OO                                     *
  8     4.73-   4.84            *OOOOOOOO                               *
  3     4.84-   4.95            *OOO                                    *
 11     4.95-   5.06            *OOOOOOOOOOO                            *
  8     5.06-   5.17            *OOOOOOOO                               *
  6     5.17-   5.28            *OOOOOO                                 *
  2     5.28-   5.39            *OO                                     *
                                *                                       *
  5    VIS   101-150            *XXXXX                                  *
  2    VIS   151-200            *XX                                     *
  7    VIS       >200           *.......                                *
                                I*****I****I*****I****I*****I****I*****I
    NUMBER OF RESULTS                5   10   15   20   25   30   35
                                NUMBER OF RESULTS IN HISTOGRAM  71

                                HISTOGRAM DISTRIBUTION FOR BILIRUBIN
                                I*****I****I*****I****I*****I****I*****I
  0    VIS       >200           *                                       *
  7    VIS   151-200            *XXXXXXX                                *
  6    VIS   101-150            *XXXXXX                                 *
                                *                                       *
  8     1.86-   1.98            *OOOOOOOO                               *
 13     1.98-   2.11            *OOOOOOOOOOOOO                          *
 10     2.11-   2.24            *OOOOOOOOOO                             *
  9     2.24-   2.36            *OOOOOOOOO                              *
  8     2.36-   2.49            *OOOOOOOO                               *
  4     2.49-   2.62            *OOOO                                   *
  5     2.62-   2.74            *OOOOO                                  *
                                *                                       *
  2    VIS   101-150            *XX                                     *
  0    VIS   151-200            *                                       *
  2    VIS       >200           *..                                     *
                                I*****I****I*****I****I*****I****I*****I
    NUMBER OF RESULTS                5   10   15   20   25   30   35
                                NUMBER OF RESULTS IN HISTOGRAM  74
```

do yield a reliable assessment relative to that of the other scheme participants and cumulation in terms of average SDDs is possible both over analytes and over time.

Variance Index scoring, however, overcomes this problem through use of a fixed Chosen Coefficient of Variation (CCV). This provides a means of scaling according to analyte to give scores in a 'common currency' (as are SDDs), but the score is also independent of the other participants' performance. Cumulation over both time (giving analyte MRVISs) and analytes and time (giving an OMRVIS) is also desirable. Average scores and their distribution then provide a ready comparison with other laboratories' performance, as described in section 9.2.1 above.

9.3.1 Competition in EQA

The competitive instinct was cited above as one factor motivating attempts to improve performance, and scoring facilitates competition in EQA, particularly when it includes a 'league table' presentation of scores. Anything which improves comparability of results should be welcomed as benefiting patient care, but care must be exercised in exploiting this urge. Firstly it may encourage an attitude of improvement for improvement's sake, irrespective of clinical requirements, so that attaining good performance in EQASs becomes an end in itself.

More disturbingly, this may lead to a dissociation between the procedures used for EQA specimens and clinical specimens. Thus if assay replication and other favourable treatments (eg Rumley and Roberts, 1982; Rowan et al, 1984) are used EQAS performance

will no longer be an objective reflection of that normally attained in the laboratory. This is fundamentally dishonest, though such 'cheating' only deceives the participant laboratory into a false impression of the reliability of their assays.

For these reasons most scheme organisers try to avoid an excessive competitive element while still encouraging a healthy striving to emulate the performance of the best laboratories. For example, presentation of OMRVISs in a league table format in the UKEQAS for General Clinical Chemistry was discontinued in the late 1970s because some participants were reportedly more concerned with their position than with the OMRVIS itself.

9.4 Assessment of improvements in performance

Similar arguments apply to this application of EQAS data, which is critical for EQA to be successful in stimulating and monitoring improvement. Participants need a robust and reliable reflection of their current performance to appraise their progress.

As described in section 9.1, simple comparison with DVs or histogram presentations is most unlikely to be sufficiently sensitive and discriminating to be helpful. 'Pass/fail' assessment systems are also far too crude a reflection of performance.

SDDs could be used, but are not reliable in this context. Their dependence upon the observed SD makes them susceptible to influence from changes in the state of the art, which may be greater than changes for individual laboratories. In most schemes there is (or should be) a continuing trend towards better interlaboratory agreement, leading to decreasing SDs and

consequently higher SDDs for a participant with unchanging performance. Any appraisal of individual laboratories' performance would thus have to take these changes into account, again a complex procedure.

The VI scoring system, however, overcomes this problem by scaling using a fixed denominator, the CCV. The influence of other participants' performance is then confined to their contribution to the DV if consensus values are used, and interlaboratory variability appears to be reflected only in the reproducibility of consensus values and not their accuracy (see section 5.2).

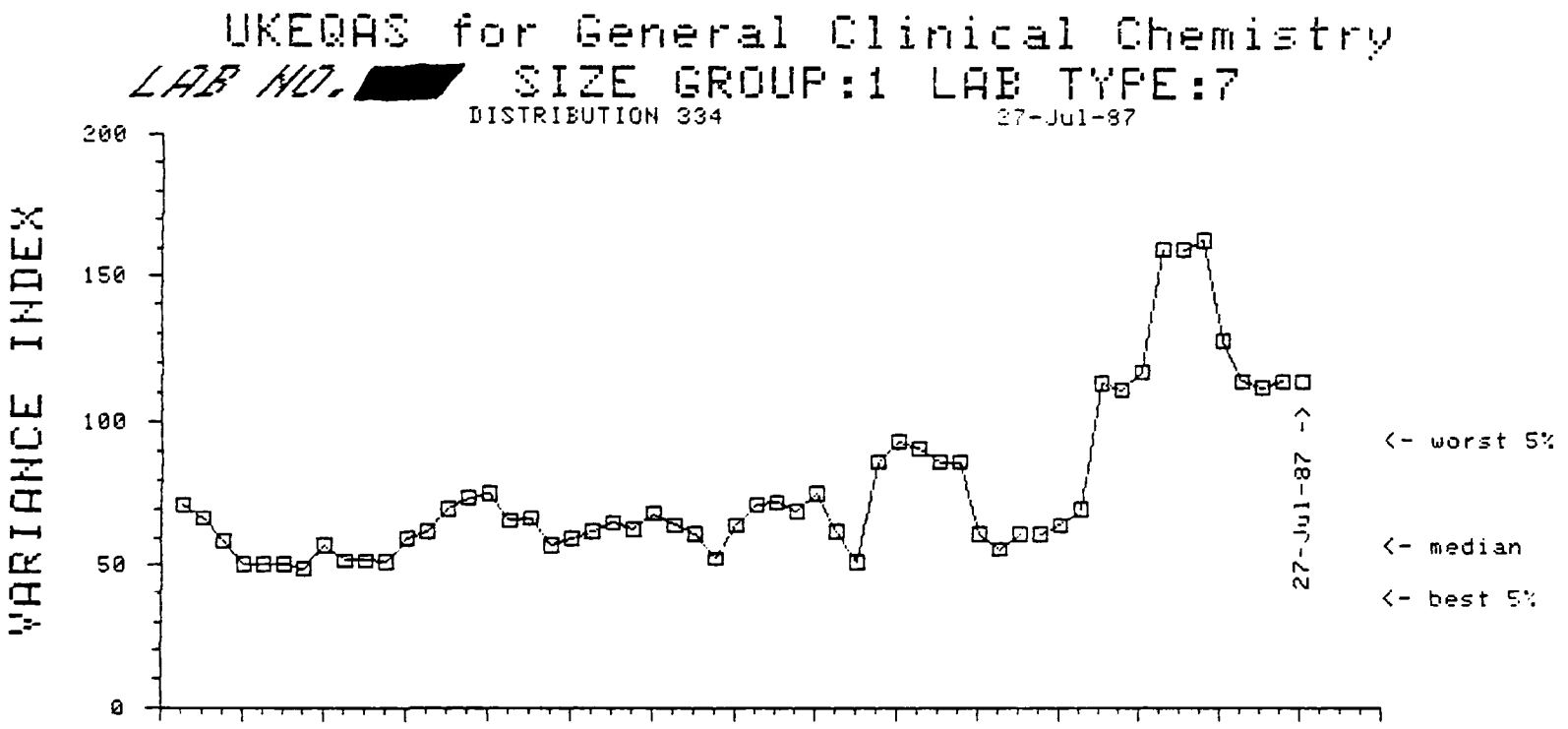
9.4.1 VI scoring in assessing changes

Examples demonstrate this application of VI scoring. Here deteriorating overall performance in the UKEQAS for General Clinical Chemistry is clearly shown by plotting OMRVIS against time, the graph covering a period of about two and a half years as discussed in section 10.4.1 below (Figure 9.7). For a single analyte, such as blood lead assay, similar treatment of the MRVIS reveals trends in performance (Figure 9.8).

Is this pattern seen in all such schemes? Figure 9.9A shows the changes in the average OMRVIS for all participants in the Middle East EQAS. There was for some time no appreciable improvement in the average score, and only a 10% improvement over 6 years despite the inclusion of scoring from the start of the scheme. This differs considerably from the experience with UKEQASs, as described above. The situation for individual laboratories, however, may be much more encouraging (Figure 9.9B).

Why is this? Possible reasons relate to changes in participant composition during the operation of the scheme. Two main factors

Figure 9.7 Graph of OMRVIS against time for a participant with deteriorating performance in UKEQAS for General Clinical Chemistry

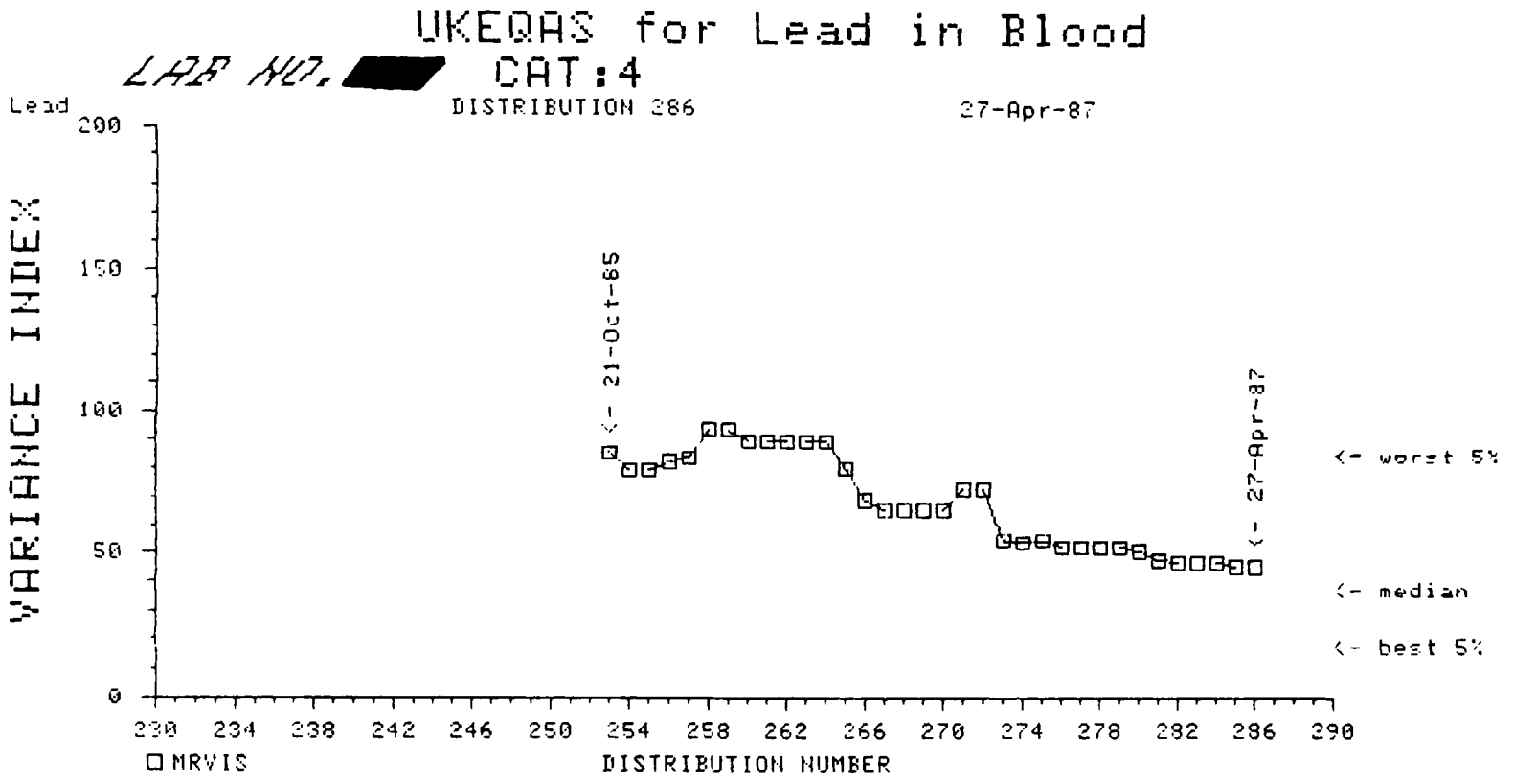


	NA	K	UREA	GLUC	CALC	FE	URATE	CREAT	BILI	T.PROT
MRVIS	122	103	139	79	176				65	143
MRBIS	10	11	-99	-39	-29				-13	-34
SDBIS	186	166	173	93	218				83	194

	CHOL	MG	AST	ALT	LD	CK	ALP	AMS	CL	PHOS
MRVIS	116					92				93
MRBIS	-41					71				32
SDBIS	160					79				125

	ALB	LI	OSMOL
MRVIS	33		
MRBIS	8		
SDBIS	38		

Figure 9.8 Graphs of OMRVIS against time for participants with (A) improving and (B) deteriorating performance in UKEQAS for Lead in Blood



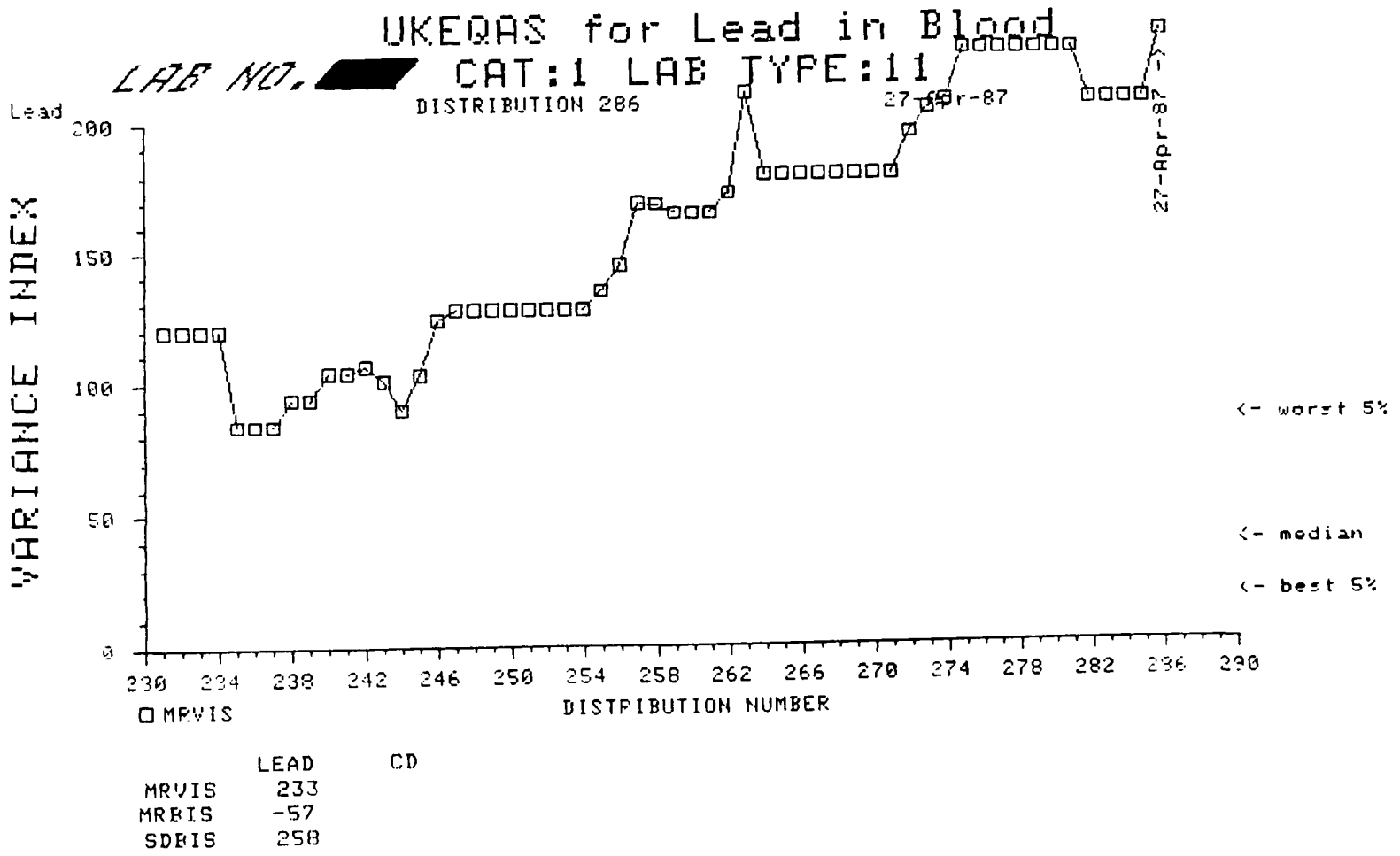
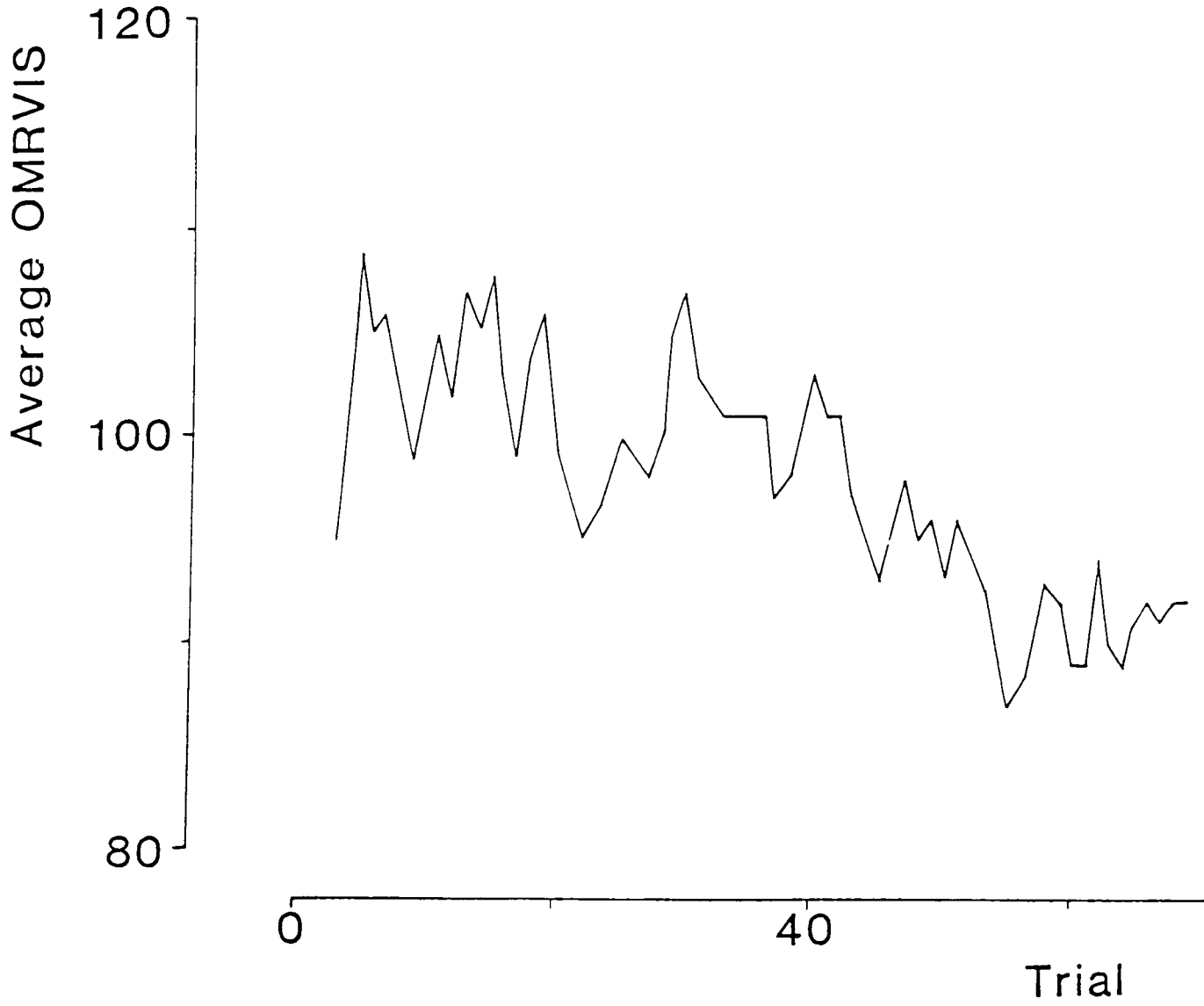
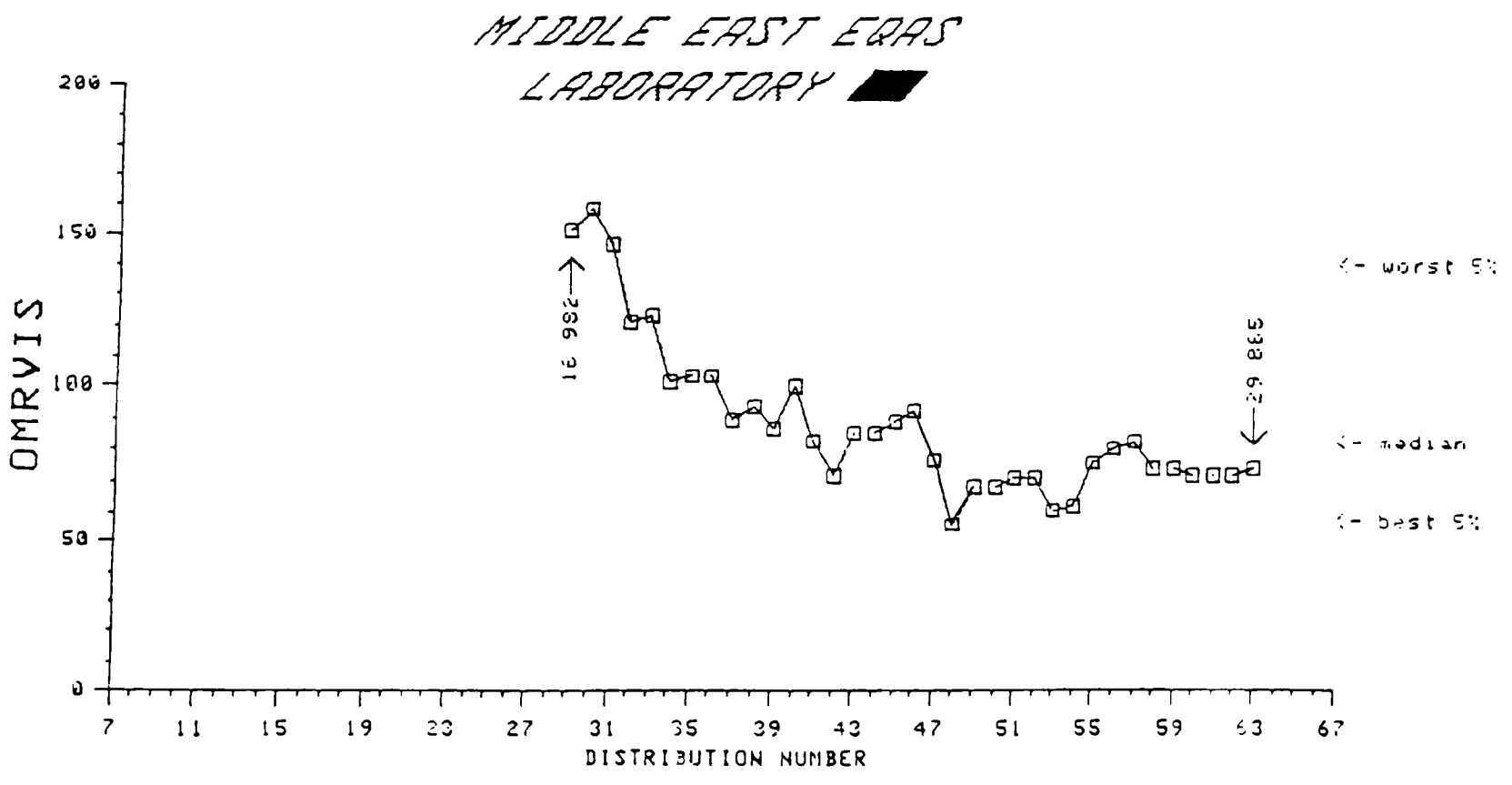


Figure 9.9 Graphs against time of (A) average OMRVIS for all participants and (B) OMRVIS for an individual participant in Middle East EQAS





are likely to affect the situation. Firstly, experience suggests that laboratories joining an EQAS go through a 'learning phase'. During this period they discover from the reports that their performance is not optimal, they investigate the problems revealed, and when they resolve these their performance begins to improve. Figure 9.9B shows an example of this.

Some participants, however, unfortunately do not respond to EQA reports in this way, and following a period of apparent poor performance they seem to lose their confidence in the scheme and cease to return results. In the MEEQAS such behaviour led to removal of the laboratory, to direct resources towards those laboratories willing to learn from their EQA performance and participate regularly and fully. It is also not feasible, for economic and geographic reasons, to offer advice and assistance to participants experiencing problems in any more than a cursory way through correspondence. Again the situation differs from that within the UK, where National Quality Assurance Advisory Panels (Browning, 1984; Walker, 1985) have been established for this purpose.

Overall, therefore, a continued influx of new participants coupled with the gradual removal of those, usually with high OMRVISs reflecting performance problems, which did not return results regularly can lead to an average OMRVIS which is relatively stable. As more laboratories were recruited and confidence in the scheme increased, however, there was a growing tendency towards improvement in performance as assessed by the average OMRVIS (Figure 9.9A).

The additional potential of the VI scoring system can also be exploited by plotting the MRBIS and SDBIS as well. Thus Figure

9.10 shows the effect on performance indices for paracetamol of changing from an unsatisfactory chemical procedure to one of the more specific and reliable enzymic procedures (Bullock, 1987; see section 8.4.1). After this change the gradual trend to less bias (smaller MRBIS) and better overall performance (lower MRVIS) over 10 distributions (the running scores being cumulated from 10 BISs or VISs) can be seen. During this period the SDBIS increased, since there are two groups of differing BISs contributing, before falling again to reflect the true performance of the enzymic procedure.

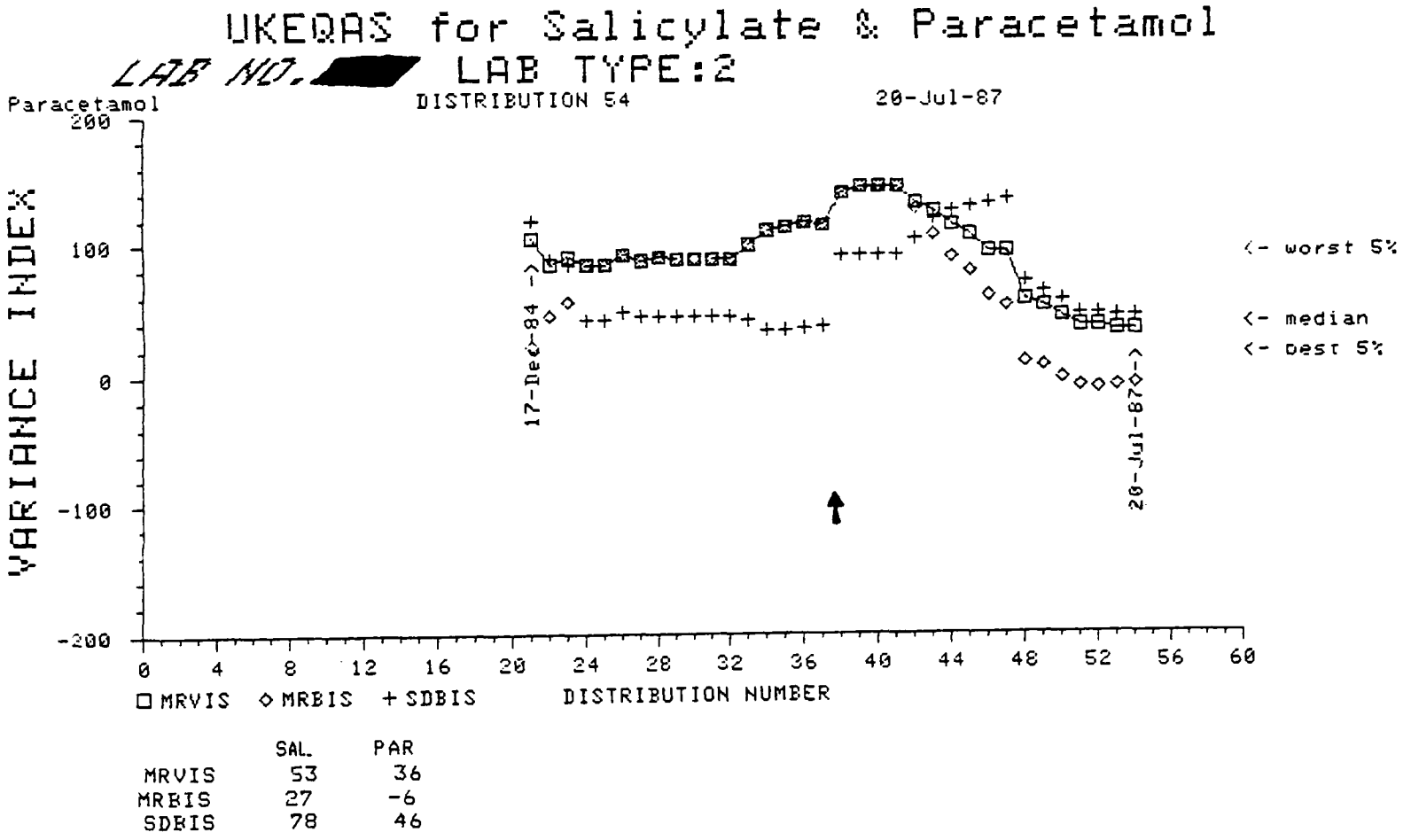
9.4.2 Cumulation period

This 'lag' prompts consideration of another point relevant to scoring system design, namely the period over which data should be cumulated.

Ideally the scoring parameter used should be robust and stable, yet respond immediately to reflect fully any change in performance. Obviously, however, these are incompatible and a compromise must be reached through balancing the conflicting requirements. The balance chosen depends upon circumstances and priorities, so if interlaboratory agreement is good a shorter period may be allowable since individual scores are likely to be more homogeneous.

For example, OMRVIS is calculated in the UKEQAS for General Clinical Chemistry from 40 contributing VISs. During the 1970s and early 1980s many laboratories received scores for about 10 analytes in each two-weekly distribution, giving an OMRVIS covering a period of about two months which was thus fairly sensitive to changes in performance but reasonably stable for

Figure 9.10 Graph of running scores for paracetamol against time for a participant in the UKEQAS for Salicylate & Paracetamol, 1984-1987, showing effect of changing from chemical to enzymic assay procedure



	SAL.	PAR
MRVIS	53	36
MRBIS	27	-6
SDBIS	78	46

most participants. With the increase in 1986 to survey of 18 analytes per distribution the OMRVIS perhaps covers too short a period and a change to use of 50-60 scores may now be indicated. In contrast, the monthly distribution frequency and return of results for fewer analytes by many participants in the International EQAS indicated 30 scores as a more appropriate basis.

For individual analytes a longer period must be tolerated, or instability could result. Here a 6-month period appears realistic, and thus 10 scores were selected for MRVIS, MRBIS and SDBIS calculation. Though some have suggested a longer cumulation (eg 40 scores; Wilson D et al, personal communication) deriving a running score from two years' performance will fail to reflect adequately changes in performance. Is 6 months too long, however? UKEQASs for hormone assay, though including more (24-30) specimen assessments, still use a 6-month window for performance assessment in terms of BIAS and VAR (Bacon et al, 1983) and this compromise appears to be generally acceptable to participants.

9.5 The detection of unsatisfactory performance

How can one determine which participants (if any) do not perform to a satisfactory standard? Again simple comparison with means is a complex procedure. Identification of laboratories producing 'outliers' or results at the extremes of the distribution might be helpful, but neither lends itself to ready examination of results from more than a single distribution. Thus this appears another potential application for a scoring system.

Though it does permit cumulation, use of SDDs will always identify some laboratories as deficient, eg the worst 5% of participants, whereas their performance may be adequate for

clinical application. More importantly, the criterion is dependent upon the performance of other participants, as described in section 9.3 above, so a laboratory is more likely to escape such classification if the general standard is poor.

Here a simple 'pass/fail' system seems ideal. If a sufficiently high proportion of a laboratory's results are classified as unacceptable, then the laboratory itself may be deemed unsatisfactory. There is flexibility in selection of the proportion required to fulfil this criterion, but it is inflexible with respect to the original definition of acceptability: if this is adjusted then considerable reprocessing of past data will be needed to provide continuity of surveillance.

Systems similar to VI scoring are more flexible in application. Firstly the basic scores calculated are objective, in that they are independent of other participants' performance. Secondly the thesholds applied for appraisal can be varied: they are not 'tied' to any definition of the acceptability of individual results. Finally scores may be cumulated either across the range of assays offered by each laboratory (as an OMRVIS) or for individual analytes (as MRVISs), and in the latter case the MRBIS and SDBIS can also be appraised.

9.5.1 Application to UKEQASs

Such flexibility offers procedures suited to a wide range of applications. Within the UKEQAS for General Clinical Chemistry, participants' OMRVISs were used to select participants whose poor performance relative to other laboratories appeared to merit further examination. Initial appraisal was then through graphs of

OMRVIS against time, to assess whether this selection represents an isolated instance or a continuing situation. Consideration of this information, together with graphical presentations of results for individual analytes (see sections 9.2.3 above and 10.5.2), facilitated decisions on the need to approach the participant with an offer of advice and assistance.

These procedures were initially carried out by the scheme's Steering Committee, and later assumed by the National Quality Assurance Advisory Panel for Chemical Pathology (Browning, 1984; Walker, 1985) after the inception of these Panels in 1976. Similar protocols were subsequently applied using MRVISs for individual analytes within the UKEQASs for Lead in Blood, Urinary Pregnancy Oestrogens and Salicylate and Paracetamol. Within the urinary oestrogens scheme the SDBIS was used as an additional screeening index, since within-laboratory consistency of performance is the prime determinant of acceptability in the clinical application of serial monitoring within individual patients (Oakey, 1980; Bullock and Wilde, 1985).

With the continuing expansion of the general clinical chemistry scheme, which now includes 25 analytes, the risk of unacceptable performance for one or more analytes being concealed by satisfactory performance for the others (leading to a satisfactory OMRVIS) is increased. This indicates a need for closer surveillance of individual analytes in addition to surveillance of overall performance, and this has been addressed through the MRVISs.

Thus at each distribution graphs of laboratory result against DV (section 10.5.2) are produced for all participants with an MRVIS exceeding a threshold for each analyte. These are assessed in

conjunction with the participant's MRVIS, MRBIS, SDBIS and method, and helpful interpretive comments are added by the scheme Organisers. The purpose is twofold: to ensure that the laboratory is aware of the apparent problem, and to assist in its resolution. The activity is conducted by the organisers rather than the Panel since such problems are mostly isolated analytical difficulties and do not appear to result from the major disorders of laboratory management which require Panel intervention (Browning, 1984).

Initially all thresholds were set at an MRVIS of 150, so that performance similar relative to the overall state of the art was highlighted for each analyte. This appeared undesirable, however, and adjustments were made to emphasise those analytes which are more critical for patient care. Thus for example the thresholds for potassium and calcium were lowered, whereas those for chloride and urate were raised. These changes were, however, relatively minor (giving a range of 130-160) and arbitrary in that they were not derived from medical requirements expressed as analytical goals.

9.6 The basis of assessment - state of the art or clinical requirements?

The relative merits of assessing laboratory performance against attainable standards or medical needs has always generated controversy. In the early years of EQA assessment against the state of the art was the only feasible procedure, tempered by later consideration of the effects of any errors on clinical care. Indeed, there had been no realistic estimate of clinical requirements, apart from Tonks' criterion of errors not to exceed one quarter of the 'normal range', as the reference interval was

then called (Tonks, 1963) and various statements derived from subjective or even arbitrary views of individual clinicians (eg Barnett, 1968), as discussed by Fraser (1983).

Against this background many scoring systems were established using attainable performance as a baseline. For example, scoring in terms of SDDs or 'Z scores' had been widely practised in many surveys as a way of compensating for differences in interlaboratory agreement at varying analyte levels (Ley and Ezer, 1974; Wellcome Diagnostics, 1984). This also enabled expression in terms of a 'common currency' for all analytes, with similar performance relative to other laboratories giving scores of similar magnitude. In GFR the concept was amended to define limits of acceptability in terms of the SD obtained by reference laboratories, though the effect was similar (Stamm, 1975).

SDDs, however, remained dependent upon the general standard of performance, as shown in sections 9.3 and 9.4 above, and an advance was required. This came with the adoption of the best interlaboratory trimmed CVs attained in the UK in 1972 as the Chosen Coefficients of Variation (CCVs) in the VI system (Whitehead et al, 1975). The objective was to scale against analyte level and analyte performance in deriving an index of overall performance which could be used as an indicator of changes over time.

Other schemes chose systems related in some way to perceived clinical needs. Thus the Netherlands scheme (Jansen et al, 1977) assessed deviations from the DV in terms of a points scale, with for example the same percentage deviation giving a worse score for calcium than for urate. The principal disadvantage of this

type of system is that it precludes combination of the scores obtained for the analytes surveyed unless all participants offer the same range of analytes. Otherwise a laboratory could artefactually 'improve' their assessment by refraining from reporting calcium results, which would score poorly.

In recent years more objective analytical goals for imprecision and total laboratory error have been established for many commonly-determined analytes, in relation to biological variation (Subcommittee on Analytical Goals, 1979; Fraser, 1983). These have been adopted as criteria of acceptability in some EQASs, such as that in Australia (Bowyer et al, 1981), giving performance standards which should be realistic estimates of clinical requirements. Performance criteria of similar derivation have also been proposed for use in GFR (Stamm, 1982), though they have not yet been endorsed in full.

These introductions have largely been in schemes using 'pass/fail' criteria, for licensing or educational purposes. Since analytical goals are not at present met for most analytes (see Tables 2.1 and 9.1), these goals have not been implemented in their entirety or almost all laboratories would fail; this entails a further assumption of the level of performance which is acceptable. The quantitative performance information supplied through such schemes is rather limited in most cases, so the main emphasis in performance assessment remains on scoring systems based on the state of art. The problem of combining scores for different analytes remains, irrespective of the advances in derivation of the medical needs.

With the availability of analytical goals, state of the art systems can now be related to clinical requirements. Thus for

example the goal for calcium of 0.8% (Fraser CG, personal communication) can be combined with the CCV of 4.0% to yield an acceptability criterion of 20 VIS (Table 9.1). Comparison of this figure with the average MRVIS of 67 for calcium at December 1986 in the UKEQAS for General Clinical Chemistry emphasises the problems entailed by the use of medical needs in EQA.

Table 9.1 demonstrates that among the non-enzyme analytes average performance satisfies such goals only for potassium, urea, iron and bilirubin. For urea and bilirubin, however, clinical situations such as the detection of changes at elevated concentrations may make more stringent demands than the biological variation in normal subjects upon which these goals are primarily based. Similar arguments apply also to enzymes such as ALT and CK, for which the derived 'goals' of 19.7% and 35.8% CV (Fraser CG, personal communication) appear intuitively too permissive.

Given this continuing controversy, which basis is better? The choice depends on the purpose of the assessment, with state of the art assessment being more convenient and versatile in most EQASs, and clinical requirements providing a more objective appraisal of the extent to which the results of laboratory investigations can fulfil their potential in patient care. On balance, the best approach seems to be the use of a system such as VI for 'routine' application, with analytical goals being used to assist rational choice where particular decisions (such as the relative need for improvement for two analytes) have to be made.

9.7 Summary

A satisfactory scoring system facilitates participants'

interpretation of information derived from EQA. Ideally the system should allow stepwise interpretation of increasingly detailed information:

- indication of whether overall performance is satisfactory
- indication of which analytes contribute most to overall variance
- indication of source of problem for each analyte

The scoring system should allow participants to compare their performance with that of other laboratories at the same time, and with their own past performance.

The scoring system should also assist in the detection of participants experiencing performance difficulties, so that assistance may be offered to them.

Variance Index (VI) scoring appears to fulfil these requirements for a scoring system, though it has some disadvantages. These relate to its being based upon the state of the art, and to potential problems caused by changes in this among analytes.

Scoring systems may be based on the state of the art or clinical requirements, though choice of the latter precludes combination of scores for several analytes. The best compromise is to use the former, and incorporate the latter to assist in appraisal of the clinical relevance of the scores.

Chapter 10:

THE USE OF GRAPHICAL PRESENTATIONS OF EXTERNAL QUALITY ASSESSMENT DATA

10.1 Introduction

In any situation involving the appraisal of information graphical presentations can make the data much more readily comprehensible. EQA data are no exception to this general principle, though such presentations have not been used extensively by many schemes.

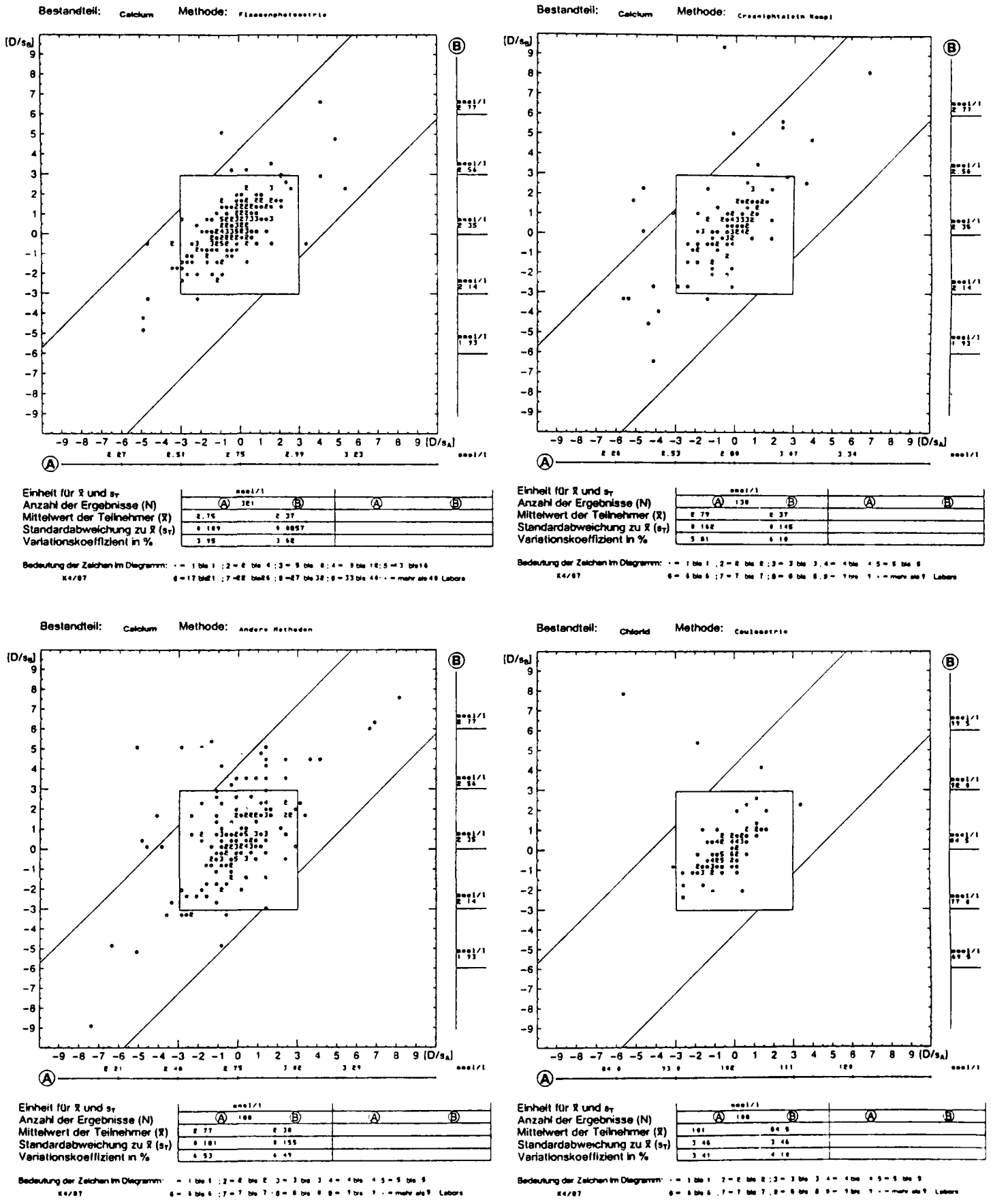
Graphical presentations of data can assist at each stage of the interpretive process in EQA, and examples of these will be presented in turn. Finally the reasons for their relatively sparse use will be examined.

10.2 Youden plots

The first true graphic presentations used in EQA were Youden plots, in which each participant's result for one specimen is related to their result for another specimen assayed at the same time (Skendzel and Youden, 1969). This was a natural extension of the application of such graphs in IQC procedures. It yields a crude impression of the contributions of bias and imprecision to overall variance, as discussed in section 3.2.3.

The presentation is therefore of limited value, and may be best employed in licensing or other schemes using an externally-derived DV to summarise participants' results without emphasising the consensus value. The licensing scheme in GFR provides an example of such use (Figure 10.1). Unfortunately, to provide uniformity of interpretation (since the acceptability limits are ± 3 SD) the axes are rescaled in multiples of the SDs obtained by

Figure 10.1 Youden plot of results in participants' reports from DGKC Ringversuch 4/87 in GFR



the reference laboratories, which makes location of the individual participant's result pair difficult.

10.3 Frequency distributions

Frequency distributions, in the form of histograms, scattergrams and bar charts, have been the main graphical representation used in EQA from the earliest surveys (Belk and Sunderman, 1947; Wootton and King, 1953) to demonstrate the dispersion of participants' results.

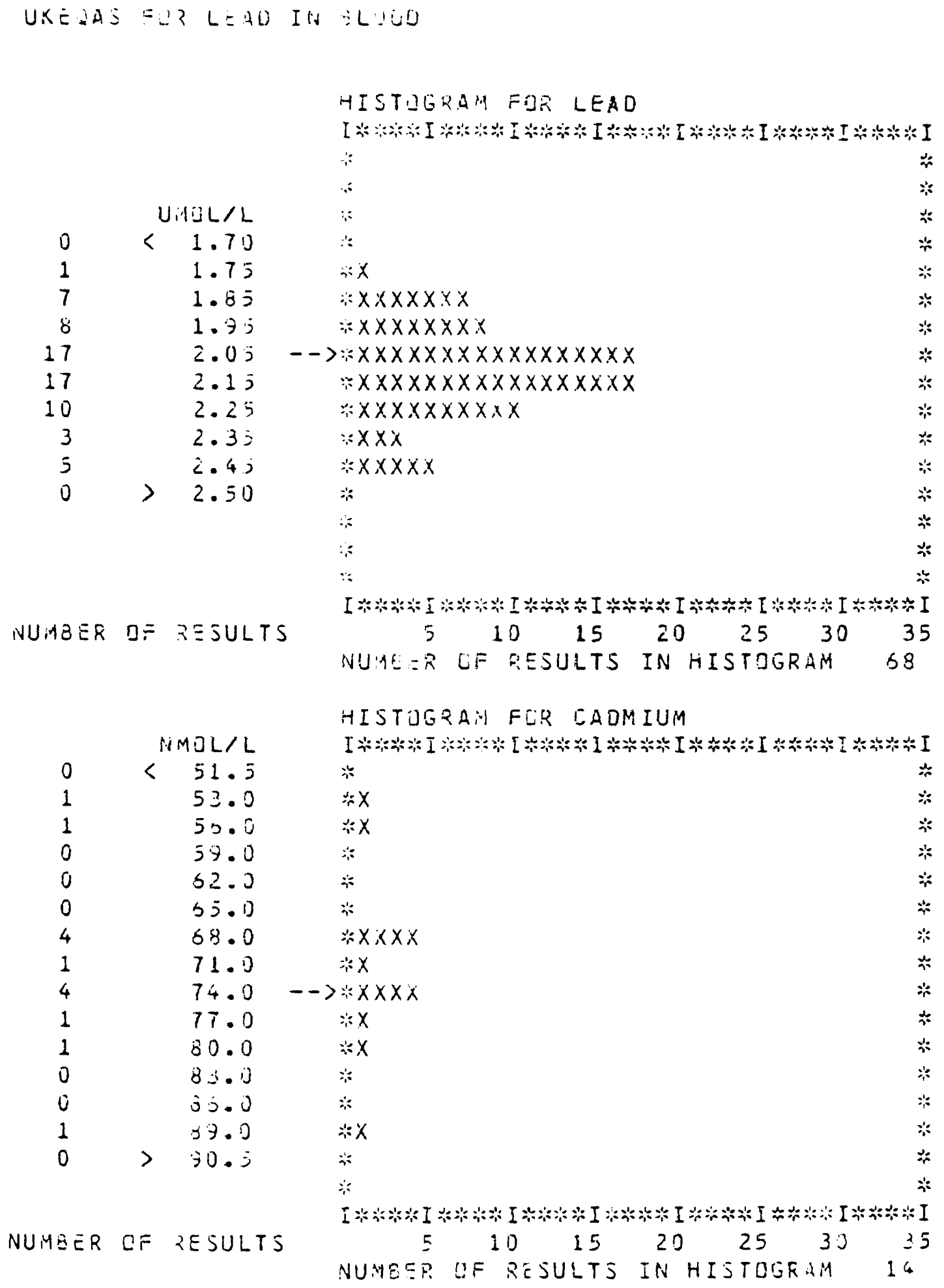
10.3.1 The distribution of results

Such histograms continue to enjoy wide use in EQA reports as a convenient means of displaying all results received. For use in participants' reports they are usually scaled (often at the DV $\pm$2 SD) to fit into a defined space within the report, and other refinements include the indication of where the individual laboratory's result lies. This type of presentation is exemplified in Figure 10.2 for lead and cadmium in the UKEQAS for Lead in Blood.

Despite their ready production and wide usage, however, these histograms convey only limited information, as discussed in section 9.3. Their principal contribution is to reveal the nature of the distribution of results, whether Gaussian or other, as a reassurance for scheme organisers and participants. The interpretation of individual performance on this basis is frequently ineffective (see section 9.3), though modification of their format to reinforce criteria of acceptability, as in Figure 9.6, may extend their utility.

It therefore appears appropriate to suppress their inclusion in participants' reports when a scheme reaches a certain maturity.

Figure 10.2 Histogram presentation of results for lead and cadmium in UKEQAS for Lead in Blood



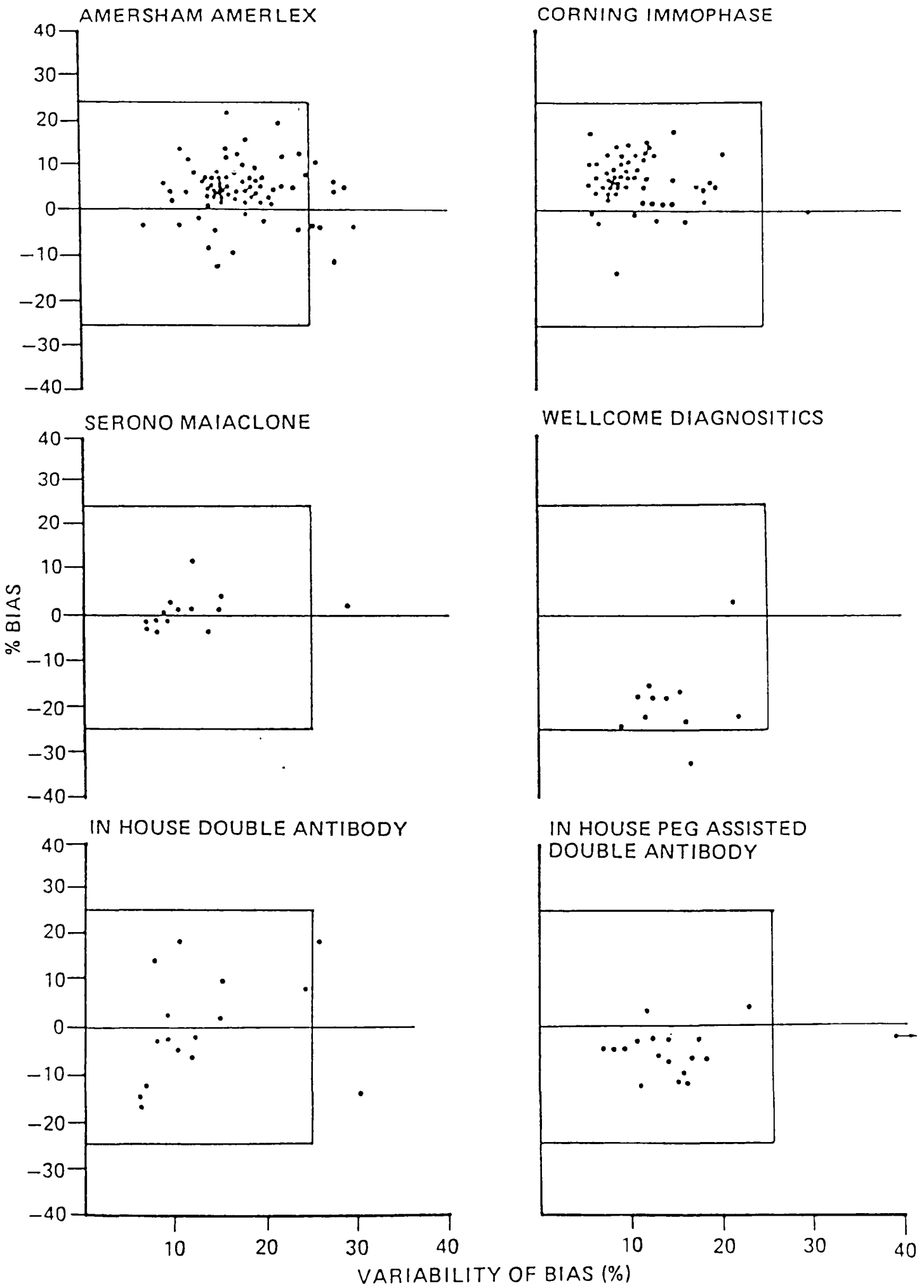
This stage is dependent upon balancing the useful content against the resources consumed in their production. The principal factors are the assay maturity, reflected in the interlaboratory agreement, and the number of participants. Thus the information content is minimal where a large number of laboratories produce gaussianly-distributed results with very close agreement; sodium assay in many schemes is a prime example. The UKEQAS for General Clinical Chemistry therefore ceased during 1984 to provide histograms of results in participants' reports. Where there are few participants, however, the distribution of results is more variable (especially with less reliable assays) and there would also be less saving from their omission.

10.3.2 The distribution of scores

The distribution of scores can be more informative. Such histograms (eg that in Figure 9.1) show the best performance attained by participants and thus provide an additional stimulus to improvement by other laboratories, as explained in section 9.2.1.

In the more detailed examination of performance for an individual analyte, a graphical presentation of bias against consistency of bias conveys useful information on the performance of individual participants and of analytical procedures. Figure 10.3 shows an example of a 'football pitch' presentation of BIAS and VAR data for TSH in the UKEQAS for Thyroid-related Hormones. The limits of acceptable performance form the 'goal area', with the origin representing the ideal performance of a consistent zero bias from the ALTM. In such a summary presentation it is easy for a laboratory to assess its performance in terms of bias and consistency, relative to other laboratories and to criteria of

Figure 10.3 Presentation of participants' BIAS against VAR for TSH in UKEQAS for Thyroid-related Hormones. 'Goal area' represents limits of acceptable performance (±25% for BIAS, 25% for VAR)



acceptability.

This also provides a means for a combined assessment of these two aspects of performance, which are usually interpreted separately, as an indication of overall performance. Thus a laboratory may be content with its BIAS and VAR of +17% and 17%, since they are well within the criterion of <25%, and believe its performance superior to that of a participant with a VAR of 23%. If, however, equal weight is attached to both indices (which is in turn a matter to be decided for each analyte) then the other laboratory will be closer to the ideal if it has a bias of less than 7%. Indeed, the rectangular 'goal area' might be better replaced by a semi-oval objective of weighted total error; a similar argument would replace the square zone of acceptability in Figure 10.1 with a circle (in SDs: in reality an ellipse; Roehle et al, 1986).

The separate presentations in Figure 10.3 permit method performance to be evaluated more readily than can be done from the average scores alone. This can be purely by eye, as for example the pattern of negative bias for the Wellcome group in Figure 10.3, or using pattern recognition techniques (Jansen et al, 1981; Jansen, 1983). The use of such information is discussed in Chapter 8.

10.4 Assessment of performance over time

The graphical presentation of performance, usually in the form of scores, over a period can be most valuable in the assessment of any changes which have occurred. Such considerations apply both to any cumulation over analytes as an index of overall performance and to scores for a single analyte.

10.4.1 Overall performance

Section 9.2.1 describes the procedure for appraisal of laboratories with apparently poor overall performance in the UKEQAS for General Clinical Chemistry, using plots of OMRVIS against distribution. An example of such a laboratory is given in Figure 9.7, and Figures 10.4A and 10.4B show further examples of consistently good and improving performance. This presentation summarises performance over about two and a half years, including contribution from around 7500 individual results or scores for this laboratory, in a readily comprehended format. Similar presentations have been or could be devised for other schemes and scoring systems.

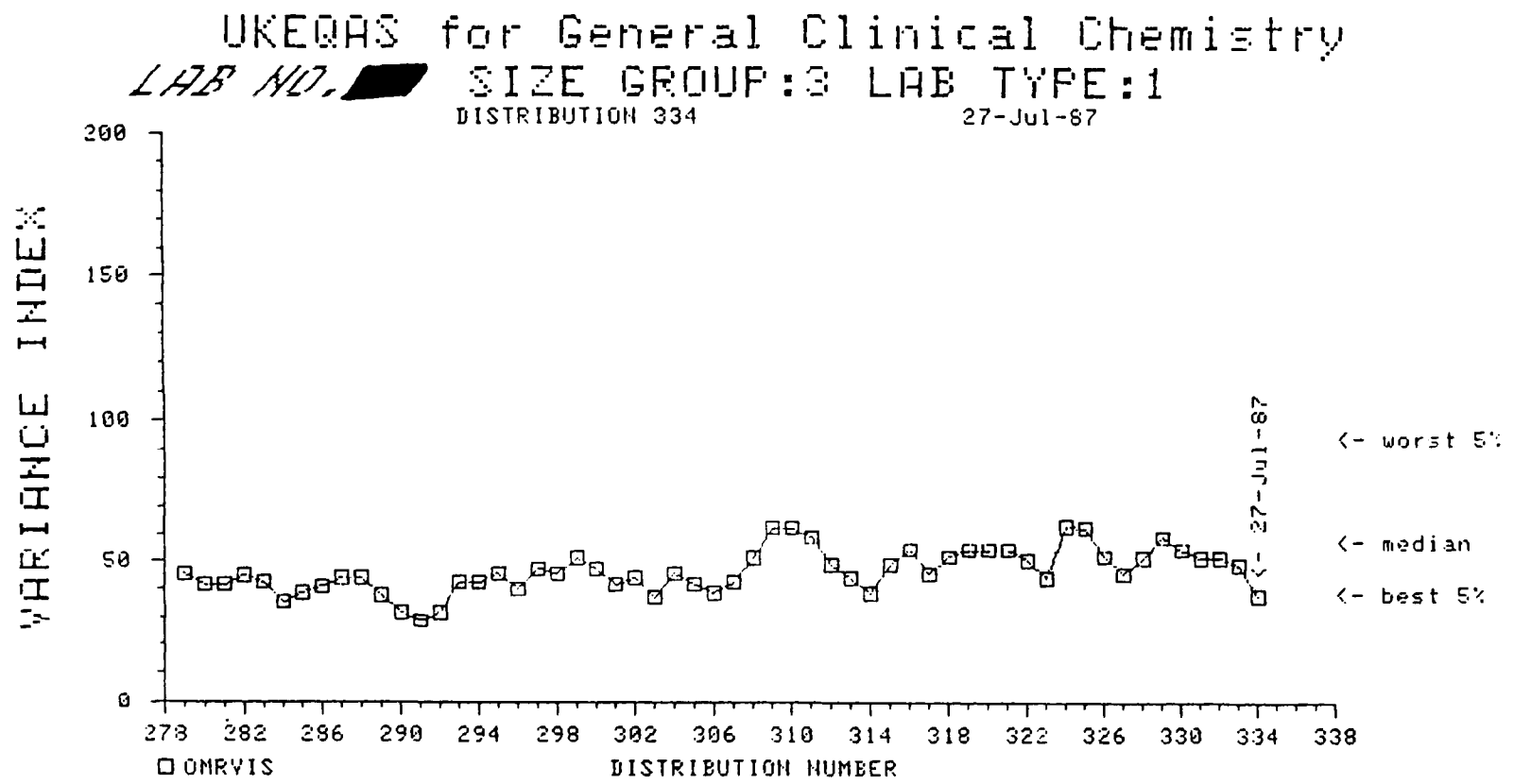
By use of conventions, other information can also be included. For example non-return of results for a distribution is denoted by the lack of a line connecting this score to that for the preceding distribution. The OMRVIS remains the same because it is not updated; Figure 9.8 contains many such examples.

10.4.2 Performance for individual analytes

Scores for individual analytes within multi-analyte EQASs or in single-analyte schemes can be treated in exactly the same manner. Thus Figures 10.5 and 10.6 show plots of MRVIS against distribution for calcium in the UKEQASs for General Clinical Chemistry and for blood lead assay. Interpretation is exactly as for the overall situation discussed above, except that only 60 results are covered by the period of the graph.

The same format can be used for more detailed performance indices, such as MRBIS and SDBIS. Figure 10.7 gives an example of these in conjunction with the MRVIS. Here the laboratory has a negative bias, consistent apart from specimen 270 (BIS = -379).

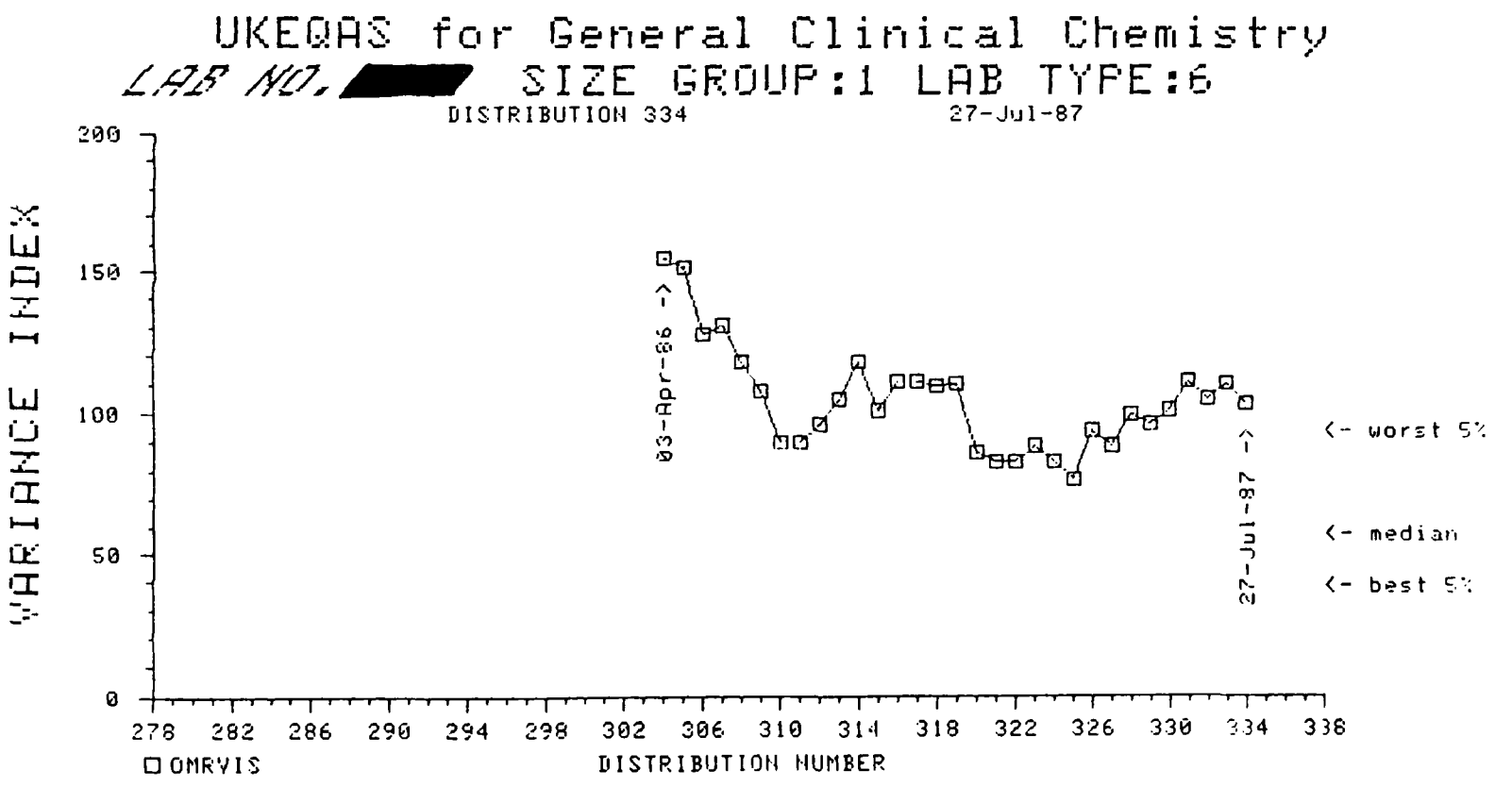
Figure 10.4 Changes in OMRVIS for participants with (A) good and (B) improving performance in UKEQAS for General Clinical Chemistry



	NA	K	UREA	GLUC	CALC	FE	URATE	CREAT	BILI	T.PROT
MRVIS	38	56	46	23	42	18	38	37	30	50
MRBIS	-9	2	21	-12	16	4	-22	-7	13	47
SDBIS	52	69	50	27	51	22	47	51	45	35

	CHOL	MG	AST	ALT	LD	CK	ALP	AMS	CL	PHOS
MRVIS	68	48	59	42	56	54	56	151	42	23
MRBIS	-58	-48	59	6	30	12	-50	139	-5	-5
SDBIS	69	27	38	65	55	67	60	142	54	28

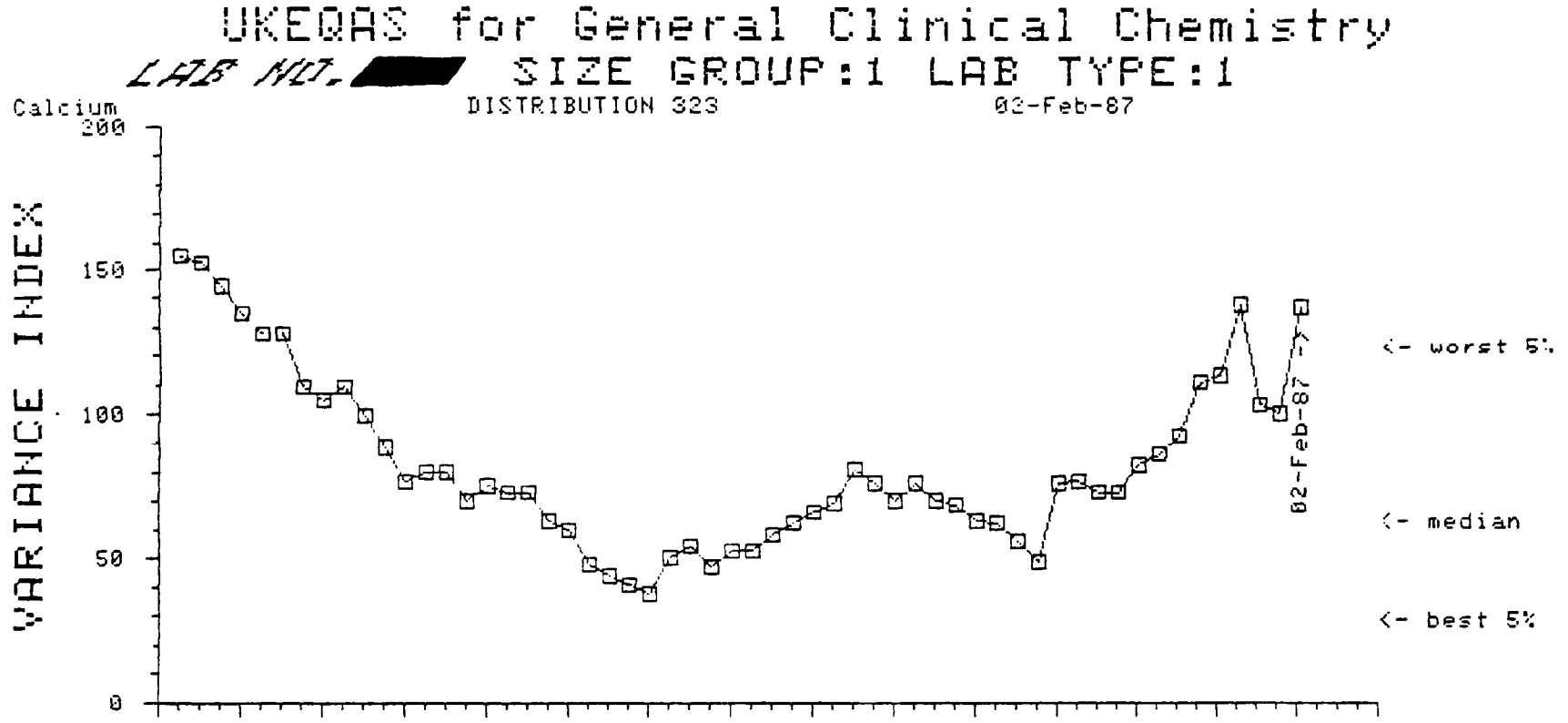
	ALB	LI	OSMOL
MRVIS	33	34	44
MRBIS	-17	-33	-4
SDBIS	35	21	54



	NA	K	UREA	GLUC	CALC	FE	URATE	CREAT	BILI	T.PROT
MRVIS	146	157	68	67	107				40	
MRBIS	-87	-15	-36	67	82				-2	
SDBIS	148	175	86	112	148				47	

	CHOL	MG	AST	ALT	LD	CK	ALP	AMS	CL	PHOS
MRVIS	112									106
MRBIS	-112									106
SDBIS	46									78

Figure 10.5 Changes in MRVIS for calcium for a participant in UKEQAS for General Clinical Chemistry

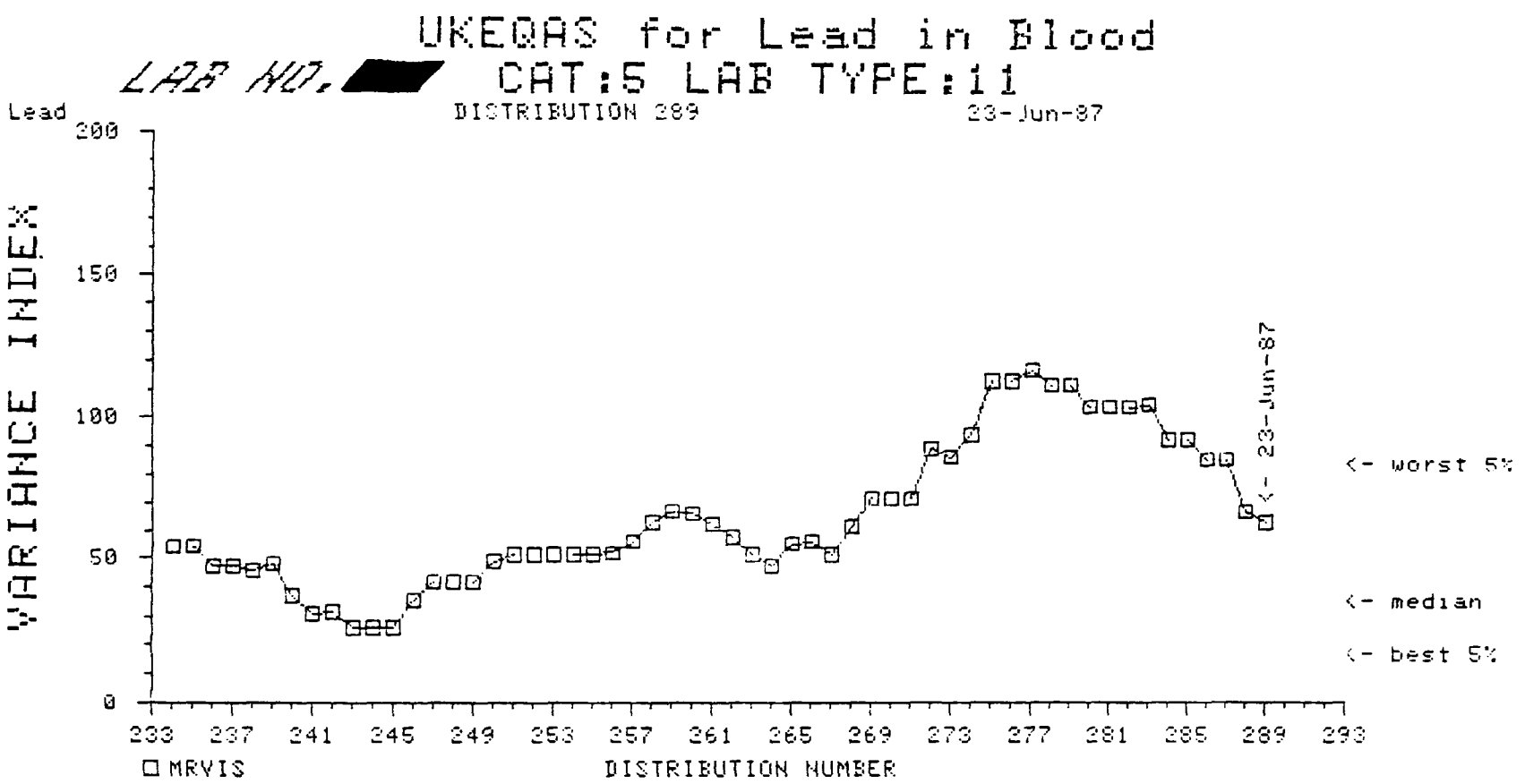


	NA	K	CL	UREA	GLUC	CALC	PHOS	URATE	CREAT	BILI
MRVIS	40	53		46	55	137	51	74	103	33
MRBIS	13	21		41	31	-60	-21	29	103	12
SDBIS	46	67		64	63	165	84	91	38	47

	T.PROT	ALB	MG	AST	ALT	LD	ALP	AMS	FE	CHOL
MRVIS	46	44		35	75		75		20	
MRBIS	-15	-8		8	-55		-61		3	
SDBIS	59	60		44	54		72		27	

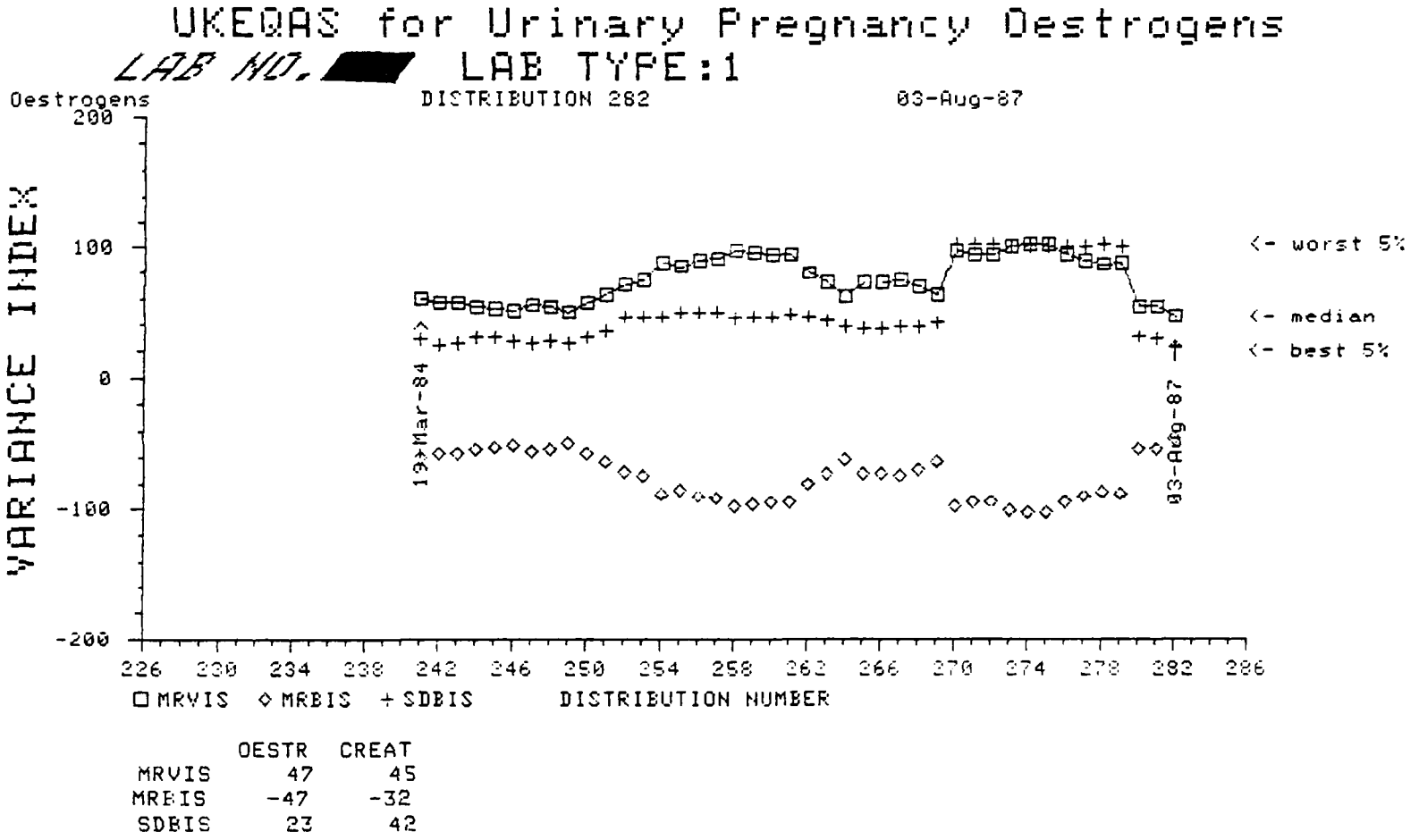
	LI	OSMOL	CK
MRVIS		57	
MRBIS		57	
SDBIS		29	

Figure 10.6 Changes in MRVIS for a participant in UKEQAS for Lead in Blood



	LEAD	CD
MRVIS	63	99
MRBIS	36	46
SDBIS	80	111

Figure 10.7 Changes in MRVIS, MRBIS and SDBIS for a participant in UKEQAS for Urinary Pregnancy Oestrogens



	OESTR	CREAT
MRVIS	47	45
MRBIS	-47	-32
SDBIS	23	42

The interpretation of Figure 9.10, showing changes in performance for paracetamol assay, is similarly discussed in section 9.4.

10.5 Assessment of results for an individual analyte

Useful though graphical presentations are in the assessment of overall and analyte-related performance, it is in the appraisal of the individual results obtained for an analyte that they are of most advantage in EQA data interpretation. The procedures used depend upon the scheme design.

10.5.1 Linearly-related specimens

The general procedures have been discussed in section 3.2.3 above. In summary, the results are related to the proportions in each specimen to yield a plot of result against analyte concentration, as demonstrated in Figure 10.8. Here the data reveal patterns of bias, imprecision and nonlinearity. Regression analyses can provide statistical evidence to suggest and confirm such patterns, but they are much more obvious and striking when presented in graphical form.

Though the specimens involved are not linearly-related, the 'intensive EQA' scheme design described in section 3.2.3 also uses this type of presentation in addition to VIS parameters. Interpretive comments are added after consideration by the scheme organisers, since three points provide only an approximate guide to the laboratory's true performance and experience is needed to extract maximal information. Interpretations are clear in many cases, however, such as the patterns of bias and imprecision revealed in Figure 10.9.

10.5.2 Cumulation procedures

An alternative procedure to obtain similar information on the

Figure 10.8 Example plots of laboratory result against analyte concentration for a set of linearly-related specimens. Patterns of bias, imprecision and nonlinearity shown

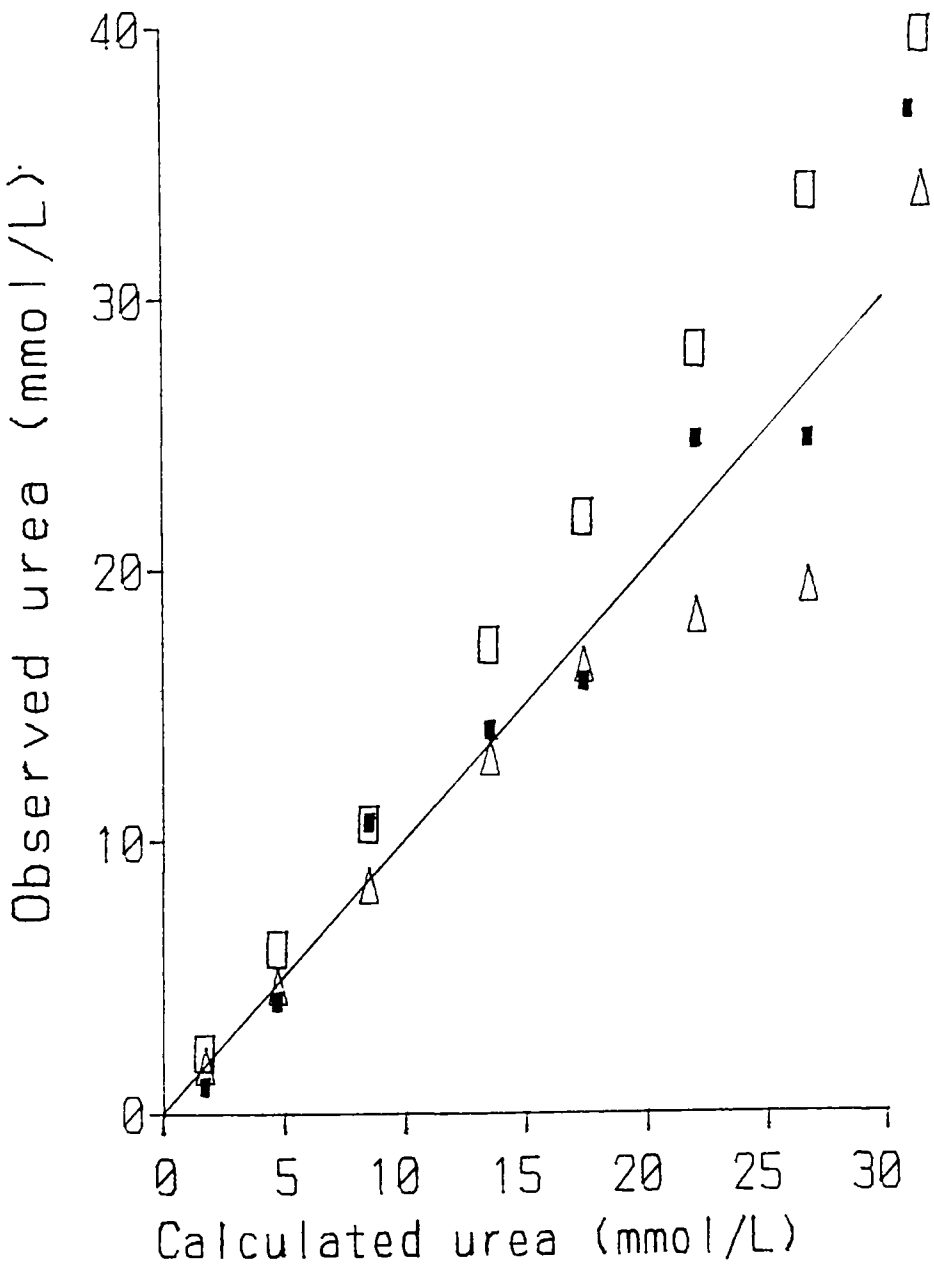
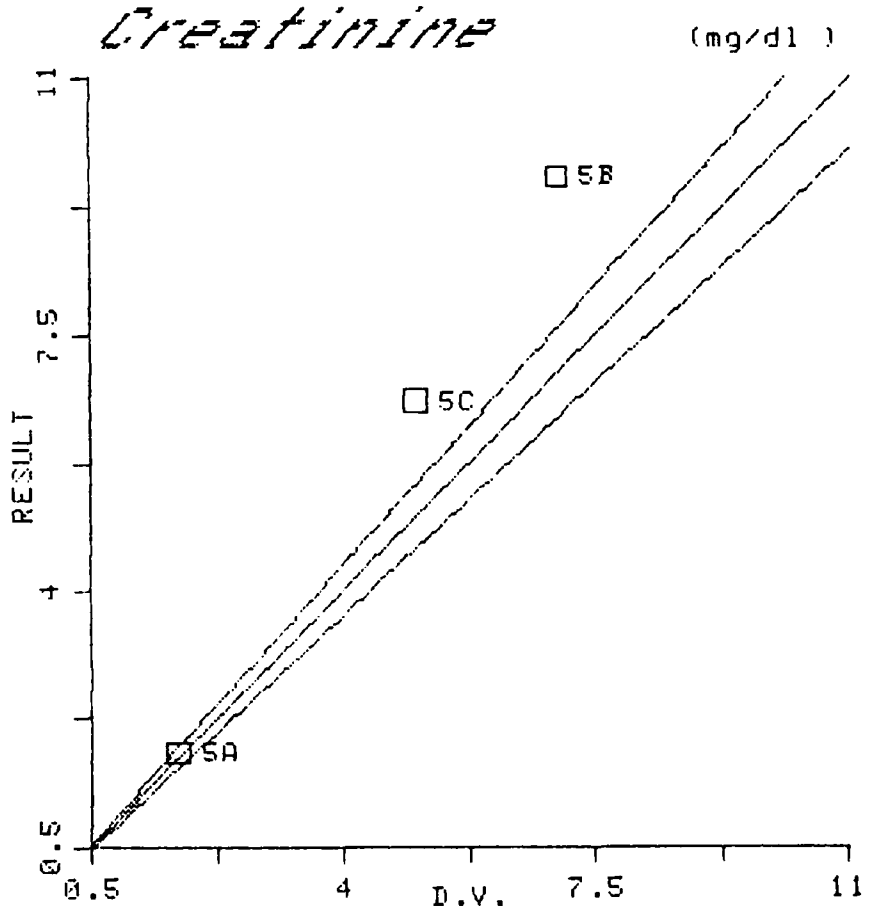
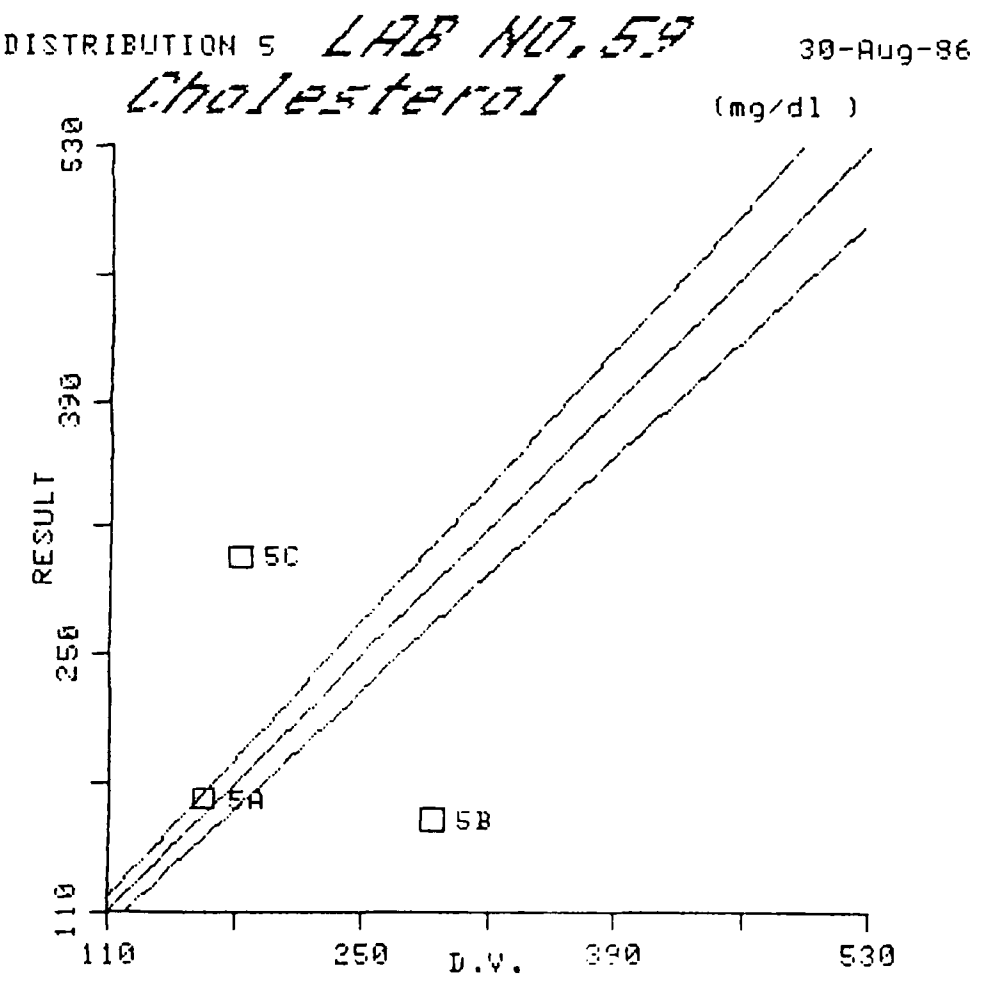


Figure 10.9 Graphs of laboratory result against designated value for participants in an 'intensive' EQA scheme, showing (A) bias and (B) imprecision





relationship between a laboratory's results, as outlined in section 3.2.3, is to cumulate data from a number of distributions.

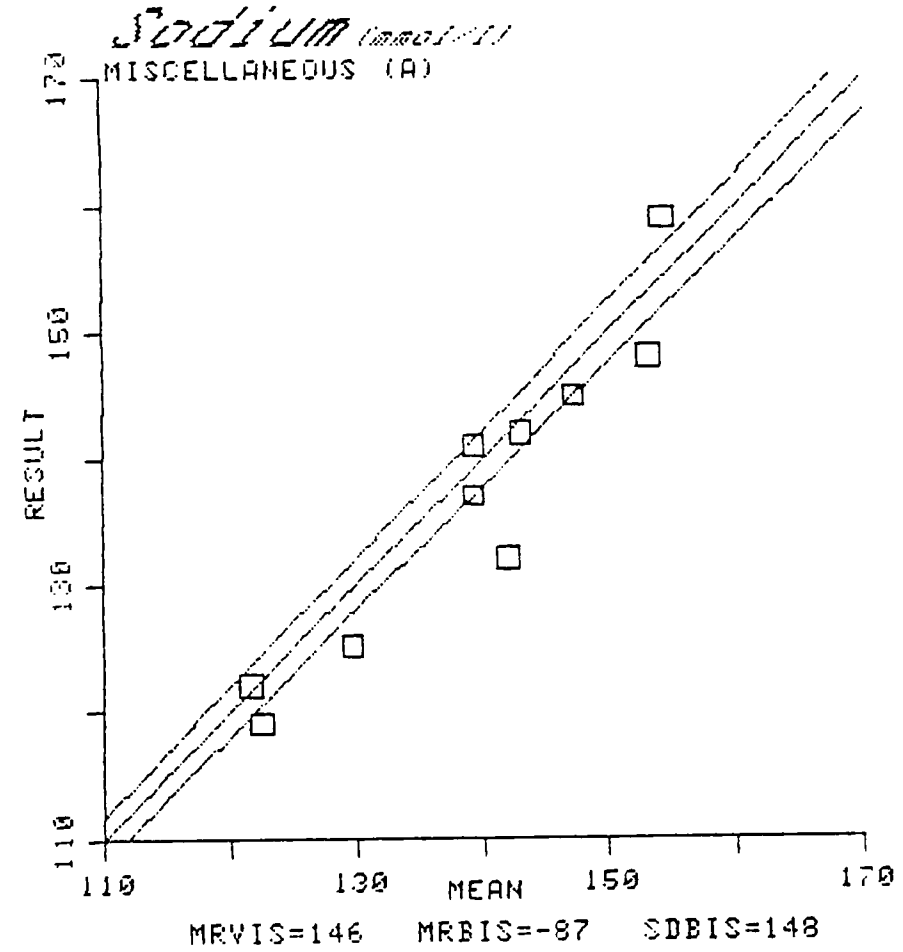
Appraisal of scores such as the MRBIS and SDBIS or BIAS and VAR gives a reasonable impression of the errors contributing to a laboratory's overall performance, but unless there is excellent performance or simply a consistent proportional bias (eg for amylase in Figure 9.4) examination of this type of presentation is essential to gain a reliable assessment. This is necessary in particular to prevent equating an index such as SDBIS or VAR with within-laboratory imprecision, since these in fact reflect consistency of bias and an inconsistent bias may be due to one or more of a number of factors.

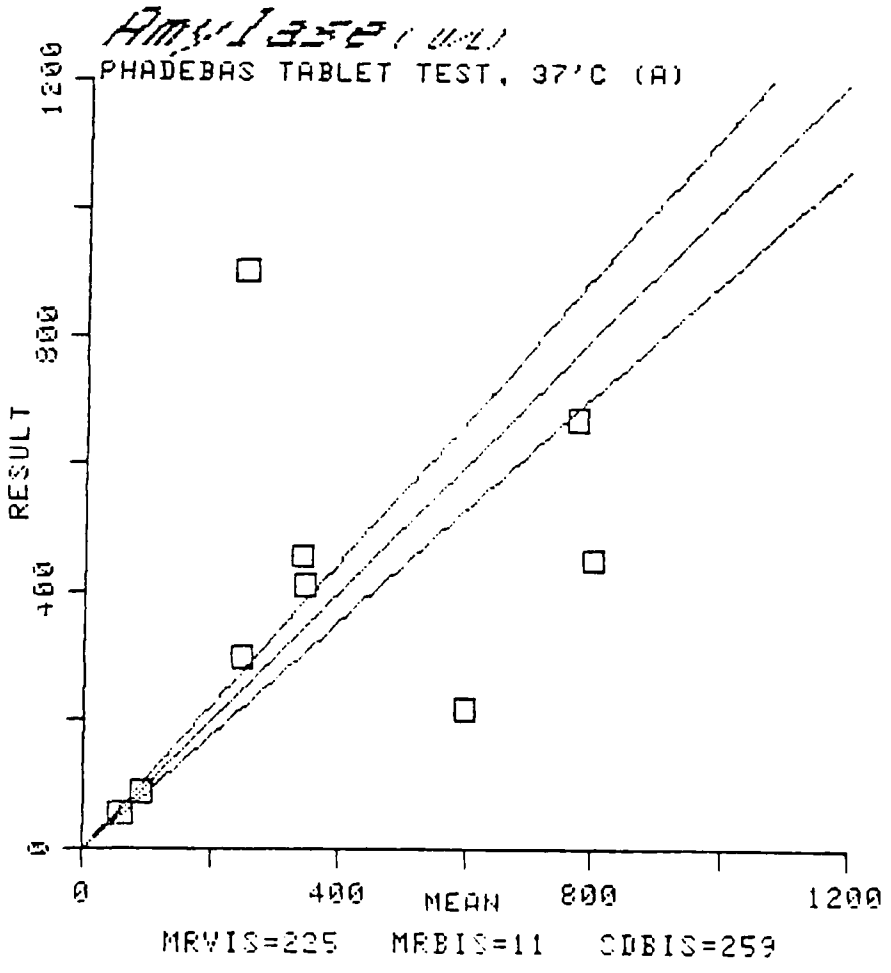
These presentations, exemplified in Figure 10.10 for two analytes in the UKEQAS for General Clinical Chemistry, thus reveal the type of errors contributing to the laboratory's overall variance. As outlined in section 9.2.3 above, a number of patterns are commonly seen:

- good performance
- consistent proportional bias
- compensating proportional and constant biases
- nonlinearity
- imprecision
- short-term changes in bias

To distinguish between imprecision and short-term bias changes and to assess whether performance is improving it is essential to have information on the temporal relation between the points. Such plots, however, do not contain this unless the points are

Figure 10.10 Graphical displays of results against designated values for sodium and amylase for participants in UKEQAS for General Clinical Chemistry. See Figure 10.11 for tabular displays





identified by their distribution number (eg Figure 10.9). Simultaneous comparison with a chronological tabulation of the data, as shown in Figure 10.11, is a convenient means of gaining this extra dimension; the contribution of any changes in method can also be thus assessed. Here the variability indeed appears to reflect imprecision for amylase, but a change from positive to negative bias in the case of sodium.

The benefit of the graphical presentation can be gauged from comparison with Figure 10.12, a tabular display of similar data for TSH from the UKEQAS for Thyroid-related Hormones. This gives the result and ratio to the DV for each result in a 6-month period, arrayed in chronological (left to right) and concentration (top to base) order. This presentation has advantage in schemes with multi-specimen distributions and if each specimen is distributed repeatedly, especially if changes in bias from distribution to distribution are common. In most cases it is more difficult to distinguish concentration-dependence, however, and the format is much less informative in schemes with single-specimen distributions and few repeated distributions.

How many distributions should be included in these plots? The number used should be sufficient to delineate the relationship, yet not cover such a long period as to allow changes in performance to obscure it. This obviously requires compromise, with choice extending from one (as in the intensive scheme design mentioned above) to over 20, though 10 results (covering 5-10 months with fortnightly or monthly single-specimen distributions) appears suitable for many applications. Even so, the temporal relationship must always be examined to identify any changes in bias which have occurred during the period covered.

Figure 10.11 Tabular displays of results and designated values for sodium and amylase for participants in UKEQAS for General Clinical Chemistry. See Figure 10.10 for graphical displays

S O D I U M

DATE	RESULT	MEAN	M E T H O D
2-MAR-87	159.0	154.2	MISCELLANEOUS (A)
16-MAR-87	141.0	139.2	MISCELLANEOUS (A)
30-MAR-87	125.0	129.7	MISCELLANEOUS (A)
13-APR-87	142.0	143.0	MISCELLANEOUS (A)
11-MAY-87	122.0	121.6	MISCELLANEOUS (A)
1-JUN-87	145.0	147.2	MISCELLANEOUS (A)
15-JUN-87	148.0	153.2	MISCELLANEOUS (A)
29-JUN-87	119.0	122.4	MISCELLANEOUS (A)
13-JUL-87	132.0	142.0	MISCELLANEOUS (A)
27-JUL-87	137.0	139.3	MISCELLANEOUS (A)

A M Y L A S E

DATE	RESULT	MEAN	M E T H O D
2-MAR-87	218.0	597.4	PHADEBAS TABLET TEST, 37'C (A)
16-MAR-87	903.0	249.1	PHADEBAS TABLET TEST, 37'C (A)
30-MAR-87	57.0	57.7	PHADEBAS TABLET TEST, 37'C (A)
13-APR-87	408.0	345.0	PHADEBAS TABLET TEST, 37'C (A)
11-MAY-87	89.0	91.5	PHADEBAS TABLET TEST, 37'C (A)
1-JUN-87	455.0	339.0	PHADEBAS TABLET TEST, 37'C (A)
15-JUN-87	670.0	774.9	PHADEBAS TABLET TEST, 37'C (A)
29-JUN-87	450.0	803.2	PHADEBAS TABLET TEST, 37'C (A)
13-JUL-87			
27-JUL-87	298.0	246.6	PHADEBAS TABLET TEST, 37'C (A)

Figure 10.12 Tabular display of results and ratio to designated value for TSH for a participant in UKEQAS for Thyroid-related Hormones

U K.
EXTERNAL QUALITY ASSESSMENT SCHEME

THYROID STIMULATING HORMONE
mU/L

ANALYSIS OF BIAS

LAB [redacted] METHOD SS

~~DISTRIBUTION~~

	67	68	69	70	71	72
POOL 208						1.1 0.88
POOL 209						1.2 0.95
POOL 205					1.3 0.86	
POOL 200					0.5 0.33	1.5 0.97
POOL 59			2.0 0.87			
POOL 60		1.8 0.72			2.1 0.85	
POOL 50	2.0 0.77	1.8 0.69	2.2 0.85		1.5 0.59	
POOL 67				1.3 0.50		
POOL 51	2.0 0.69					
POOL 52	2.3 0.68					
POOL 53		2.5 0.64				
POOL 54		3.3 0.72	4.2 0.91			
POOL 55	3.6 0.75					
POOL 68		2.9 0.56		2.0 0.39		
POOL 56	3.3 0.65					
POOL 207						4.8 0.85
POOL 206					1.0 0.18	2.6 0.45
POOL 57			5.8 0.91			
POOL 58			6.8 0.93			
POOL 69				4.0 0.51		
POOL 70				2.1 0.21		

10.6 Use of graphical presentations by EQASs

Since graphical presentations are of great benefit in improving the convenience and reliability with which conclusions may be drawn from EQA data, why have they not been used much more widely? The main reasons relate to their applicability and the resources required.

10.6.1 Application

In most schemes covering reasonably 'mature' assays the majority of participants have acceptable performance. Though general improvement remains a goal of the scheme, across-the-board provision of detailed graphical information is probably unnecessary, since most participants probably do not require it, and potentially counter-productive, since reports (as discussed in section 3.4.1) should be concise and routine provision of overly detailed information can lead to it being ignored.

Resources should therefore be concentrated on the areas where they will be of most use in improving the reliability of patient care, ie in addressing first the problems of those participants with the worst performance. Such presentations convey the greatest amount of information where a problem exists, as discussed in section 9.2. Where performance is good they provide little more than reassurance, which could be derived more economically from a well-designed scoring system: eg if a laboratory has an MRVIS of 30 there is probably little scope for improvement and a plot of results against DV would reveal no more than do their MRBIS and SDBIS of -23 and 29 respectively.

Provision of graphical presentations only for participants with apparent problems also has other benefits. The laboratories are

much more likely to take note of them, since they do not form part of the usual report from the scheme and should therefore stimulate awareness that a problem exists and hopefully activity towards its resolution. The laboratory may also gain the psychological benefit of feeling the someone is aware of their existence and problems and is taking an interest in them.

Though provision of the more detailed graphical presentations should be concentrated on those most in need, the periodic distribution of graphs summarising performance scores against time appears beneficial. This can be combined with a review by the scheme organisers and relevant comments added where there appear to be problems, and may also stimulate participants to maintain such graphs on a prospective basis.

10.6.2 Resource implications

In essence provision of true graphical presentations is very expensive in resource terms. This has been so from the earliest days, when they had to be prepared by hand, and the advent of computers has not so far altered this.

Histograms, however, can readily be produced on any printer, and this may explain their continued popularity in EQAS reports. Participants and organisers alike feel that graphical presentations are useful, and provision of histograms in some way satisfies participants' appetites and salves organisers' consciences. As outlined in sections 9.3 and 10.3.1 above, however, these formats are far from ideal and their suppression should be considered since they may be conveying little useful information.

'True' graphics have required investment in computer hardware

dedicated to their production. Initially this would have been a plotter, expensive and very slow though capable of producing very high quality output, and the associated software for data extraction and display; graphical output was therefore restricted to the most essential cases. A later introduction was the graphics terminal with associated printer, still expensive but faster (especially when combining text with graphics); though the quality available was inferior output was sufficiently detailed for most purposes. More recently development of increasingly sophisticated software has permitted use of much cheaper dot matrix printers to produce output of quality similar to that of graphics terminals even from microcomputers. Such production, however, remains relatively slow and it is still difficult to combine graphical presentations within the body of a printed report, especially since large schemes require use of a line printer to produce the large volumes of high quality text output which are needed.

Greater utilisation of graphical presentations may become feasible in the future, as the development of computer peripherals continues at an ever-increasing rate. Modern software offers flexibilty in the combination of text and graphics within the same screen framework, and laser or electrostatic printers the possibility of generating very high quality output for report production. Figure 10.13 shows an example of such a report format, from the Netherlands NEQAS. This combines tabular statistical and bar chart presentations of data from the current distribution, and also a summary of the laboratory's results from past distributions in a composite Youden plot.

Such technology is still new, however, and not yet fully

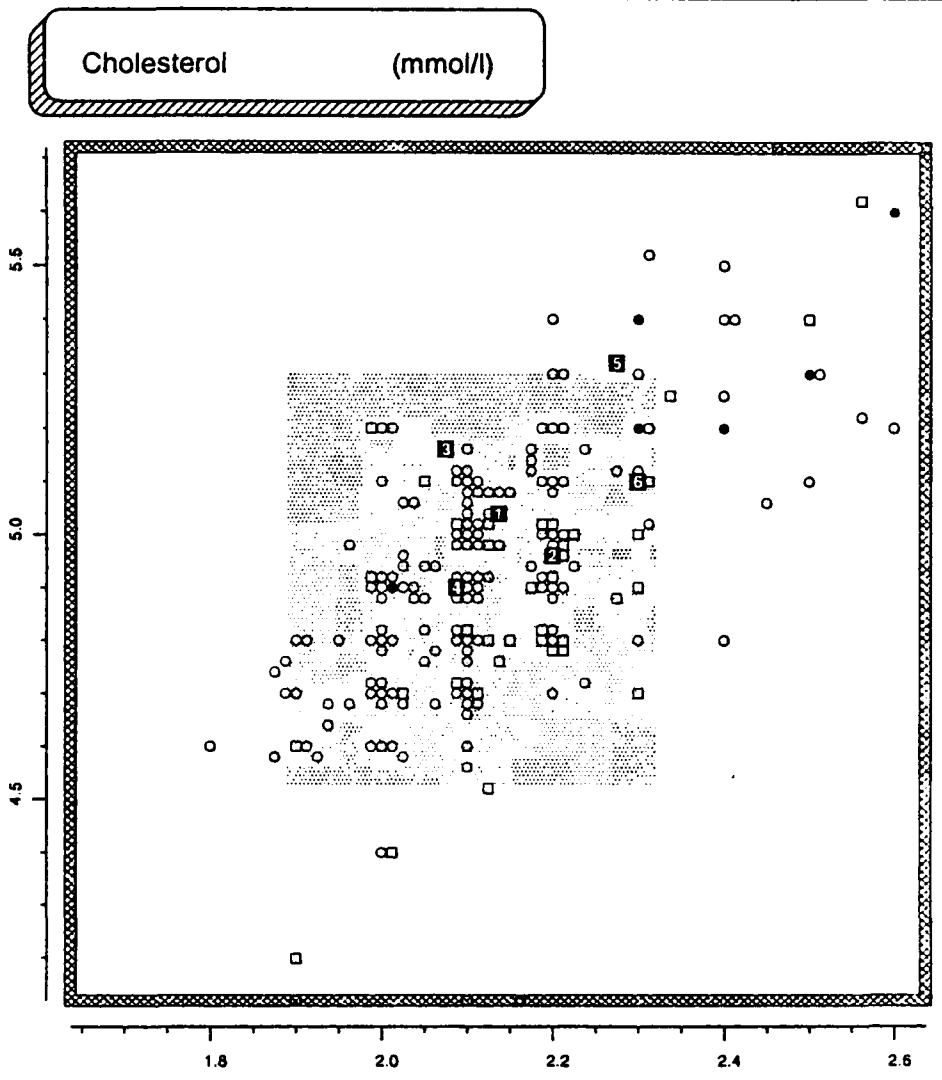
Figure 10.13 Eaxample of mixed text and graphical presentation of data in participant's report from Netherlands national EQAS

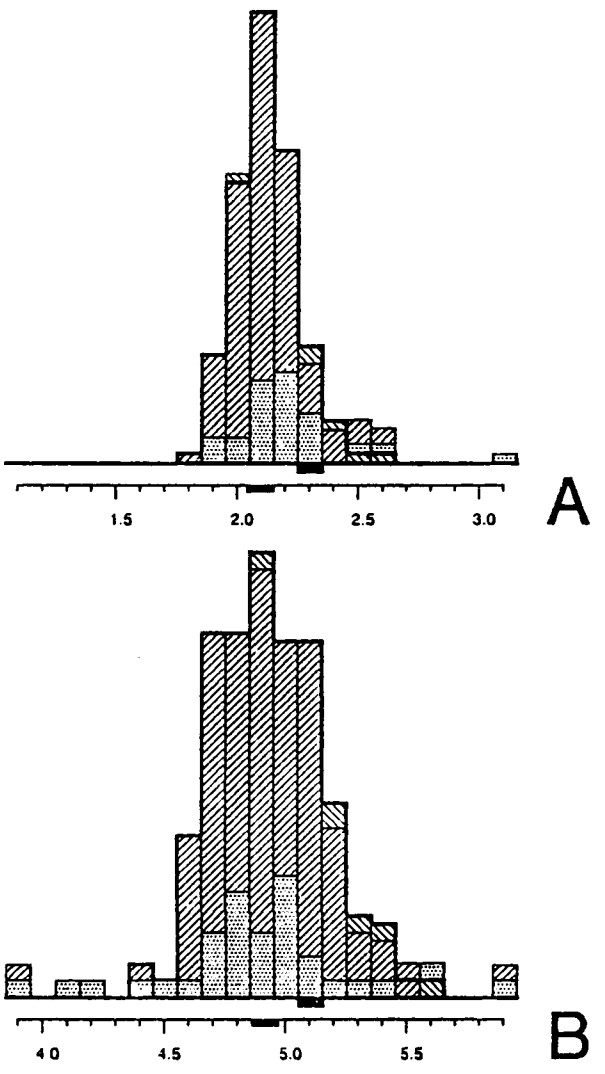
S K Z L

Externe enquête 80
datum enquête: 7 april 1986
print datum: 15 april 1987
monsters: Nyegaard Seronorm/Pathonorm

De heer dr. S.K.Z.L. Bewaking 4
Ziekenhuis "Onder Controle"
Kwaliteitsstraat 1
1234 AB Netjesdorp pagina 17

Cholesterol (mmol/l)





Methode	A gem.	A s.d.	A v.c.	A ntot	A uitb	B gem.	B s.d.	B v.c.	B ntot	B uitb
● Liebermann + modificaties, manueel	2.42	.13	5.4%	6	1	5.27	.23	4.4%	6	0
Consensus waarde	2.42	.13	5.4%	6	1	5.27	.23	4.4%	6	0
□ Enzymatisch, manueel	2.15	.12	5.4%	41	3	4.91	.19	3.8%	41	5
O Enzymatisch, automatisch, CFS	2.18	.11	5.1%	16	0	5.04	.19	3.7%	16	0
O Enzymatisch, automatisch, discreet	2.08	.10	4.7%	130	10	4.91	.20	4.0%	130	3
O Enzymatisch, automatisch, kinetisch	2.02			2	0	4.68			2	0
Consensus waarde	**2.10**	**.11**	**5.1%**	**189**	**13**	**4.92**	**.20**	**4.0%**	**189**	**8**
6 Uw resultaten	**2.30**					**5.10**				

reliable. It also remains relatively slow, with very high capital and operating costs to attain the throughput required. Though its use is feasible in a scheme with 150 participants and two-monthly distributions, it is not yet practicable for schemes such as the UKEQAS for General Clinical Chemistry with its 550 paricipants and two-weekly distributions. Such systems nevertheless offer great promise for future application; use of colour graphics may similarly become feasible as more reliable and economic dot matrix and ink-jet printers are developed.

10.7 Summary

Graphical presentations assist greatly in facilitating the interpretation of EQA data, for participants and scheme organisers.

Simple presentations such as histograms of results have been widely used because of their ready production, though their information content is limited. Modifications may increase their usefulness, but if there is good interlaboratory agreement their omission from reports should be considered.

Graphical displays of performance scores against time incorporate much information and are very valuable in the assessment of participants' progress. Laboratories should be provided with such graphs at intervals and encouraged to maintain them between these times.

In the investigation of factors contributing to overall error for an individual analyte, plotting the participant's results against designated values is most valuable. Such graphs may be derived from distribution of multiple specimens, or from data cumulated over 5-20 distributions. Where information is cumulated

the presentation must include, or be supported by, information on the time relationship of the results.

Presentations other than simple histograms have been used only infrequently by most schemes, partly to confine such effort to cases where special effort is needed to stimulate improvement. A major limitation, however, has been the cost and time involved in such production. Newer technology now offers the prospect of cheaper and faster systems to provide high quality output, combining graphical and text elements.

Chapter 11:

STUDIES OF FACTORS AFFECTING PERFORMANCE

11.1 Introduction

External quality assessment provides an excellent means of assessing the analytical performance of individual laboratories and stimulating improvements where these are needed, and also of assessing the overall state of the art and any changes, as discussed above. For participants to improve, however, guidance on factors which may affect performance is helpful so they can take any necessary action.

The analytical procedure (instrument, reagents and method) used is one of the major factors involved, and EQA data have been shown in Chapter 8 to be valuable in assessing the relative reliability of the procedures used by scheme participants. Can information derived from EQA surveys and schemes also assist in the appraisal of other contributory factors? Experience confirms this expectation, and the application of EQAS data in the Nuffield survey will first be reviewed to illustrate general principles.

11.1.1 The Nuffield survey of factors affecting analytical performance in clinical chemistry laboratories

Though earlier appraisal of UKEQAS data (Whitehead et al, 1973) had revealed that participants with greater workloads performed better, this survey (Maclagan et al, 1980), sponsored by the Nuffield Provincial Hospitals Trust, was the first major systematic attempt to identify factors contributing to laboratory performance. The working party, drawn from all clinical chemistry

professions, obtained information by questionnaire from and visited in 1976-1977 an apparently representative sample of 68 UKEQAS participants. The indicator of performance used was the OMRVIS, shown in those 40 laboratories participating in both schemes to correlate significantly with repeatability-related indices in the Wellcome Group QC Programme. The lesser correlation with the Wellcome bias index might be expected, since VISs are calculated relative to method means in the UKEQAS.

Using a variety of statistical procedures, the study identified several factors which were related to analytical performance. The elements considered included both objective data, such as laboratory costs and workload, and subjective impressions of staff quality and morale. Factors aiding differentiation of laboratories with good and poor performance were also studied. Factors with the strongest relation to performance included:

- large laboratory size (whether assessed by workload, staffing, area, equipment cost, etc), though the relationship was no longer significant when workload was expressed relative to other size parameters (eg as requests/staff)
- cost of calibration serum/request, and use of serum-based calibrants
- management team scores (rather than scores from individual professional groups)
- independence of clinical chemistry laboratory, with a full-time head of department

Causal relationships could not be inferred, however, and many factors were significantly inter-related. Though the working party noted exceptions to these relationships, consideration of

the various factors appeared to enable identification of a general pattern likely to be associated with good performance. Elements such as management and organisation are also vital, since no single factor or combination could predict accurately the performance of all laboratories. Importantly, the study excluded consideration of the choice of analytical method, discussed in Chapter 8 above, though automated equipment (predominantly SMA systems) appeared to yield better performance, as noted earlier (Whitehead et al, 1973).

The Nuffield survey thus delineated a number of often inter-related factors associated with performance for general clinical chemistry analyses, and stressed the importance of laboratory management and organisation within the general meaning of quality assurance. Are these conclusions still valid, and do they apply equally to all clinical chemistry investigations?

11.2 The influence of laboratory workload on performance

Overall laboratory workload influences and is influenced by many other aspects of laboratory size, as reported above. Thus for instance departments with high clinical chemistry workloads are likely to use automated rather than manual procedures, to have more and better-qualified staff, and to be independent of other pathology disciplines. All these would intuitively be expected to be conducive to better performance for at least those assays forming the bulk of this work.

When workload becomes excessive in relation to the resources, however, performance is likely to suffer. Thus a significant relationship between performance ranking in the Wellcome scheme and technologist workload was found in Canada by Whitlow and Campbell (1983).

11.2.1 Effects on overall performance

Data from the UKEQAS for General Clinical Chemistry, exemplified by the observations of Whitehead et al (1973) and Maclagan et al (1980), have continued to bear out the main expectation. This effect remains substantial, with in 1987 average OMRVISs of 64 and 56 for laboratories with workloads of <100,000 and >500,000 tests/year (Figure 9.1).

The relationship is consistent, as demonstrated in Figure 11.1 where the differences persist through increases in average OMRVIS. These increases were due to a period in which distribution of a succession of 'normal' materials, and to the incorporation of VISs for enzyme activity assays into OMRVIS calculation. Table 11.1 demonstrates that similar relationships also hold within sub-groups of UKEQAS participants, eg those in Eire.

11.2.2 Effects for single analytes

The balance of contributions changes when individual determinations are considered. Though the overall factors should still have an effect, other more specific aspects also become important. These relate mostly to familiarity with the procedure used, which is likely to be greater with the assay of more frequent and larger batches. IQC procedures are also facilitated and more efficient in such circumstances.

Table 11.2 demonstrates this association for blood lead assay (Bullock et al, 1986c). Plotting performance (as MRVIS) against workload did not show a clearcut relationship, but classification according to the criterion of acceptability for occupational monitoring under the Lead at Work Regulations (Health and Safety

Figure 11.1 Average OMRVIS for groups of participants in UKEQAS for General Clinical Chemistry, 1984-1986. Size group I <100,000, II 100,000-500,000, and III >500,000 tests/year; arrows show a period of predominantly 'normal' materials and the inclusion of VISs for enzymes in OMRVIS calculation

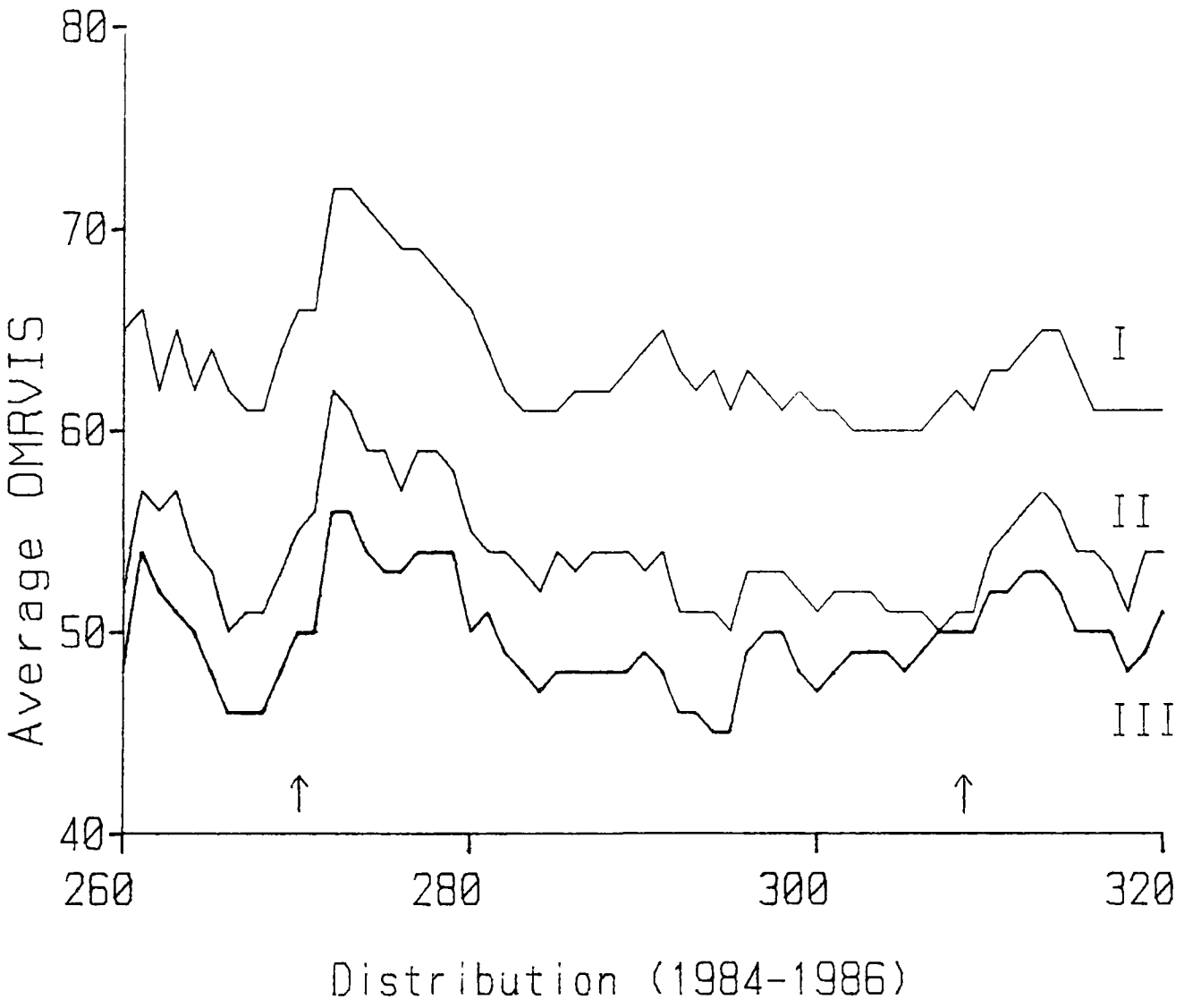


Table 11.1 Performance (average OMRVIS) for groups of participants in UKEQAS for General Clinical Chemistry, December 1986. See Figure 11.1 for size group definitions

	n	Average OMRVIS
All participants	531	57
- size group I	195	62
- size group II	208	55
- size group III	123	53
UK NHS	405	55
- size group I	99	60
- size group II	192	55
- size group III	114	53
UK private sector	59	63
- size group I	56	65
Eire	25	62
- size group I	11	64
- size group II	11	62
- size group III	3	52
Armed forces - UK	8	63
- overseas	6	81
Pharmaceutical industry	11	63
Overseas	5	70
Veterinary	5	71
University (clinical research)	3	53
Equipment/reagent manufacturer	3	43
Other laboratory	2	59

Table 11.2 Performance (1981) of participants in the UKEQAS for Lead in Blood classified according to annual blood lead assay workload (1980) and interval between batches. p <0.001 in both cases

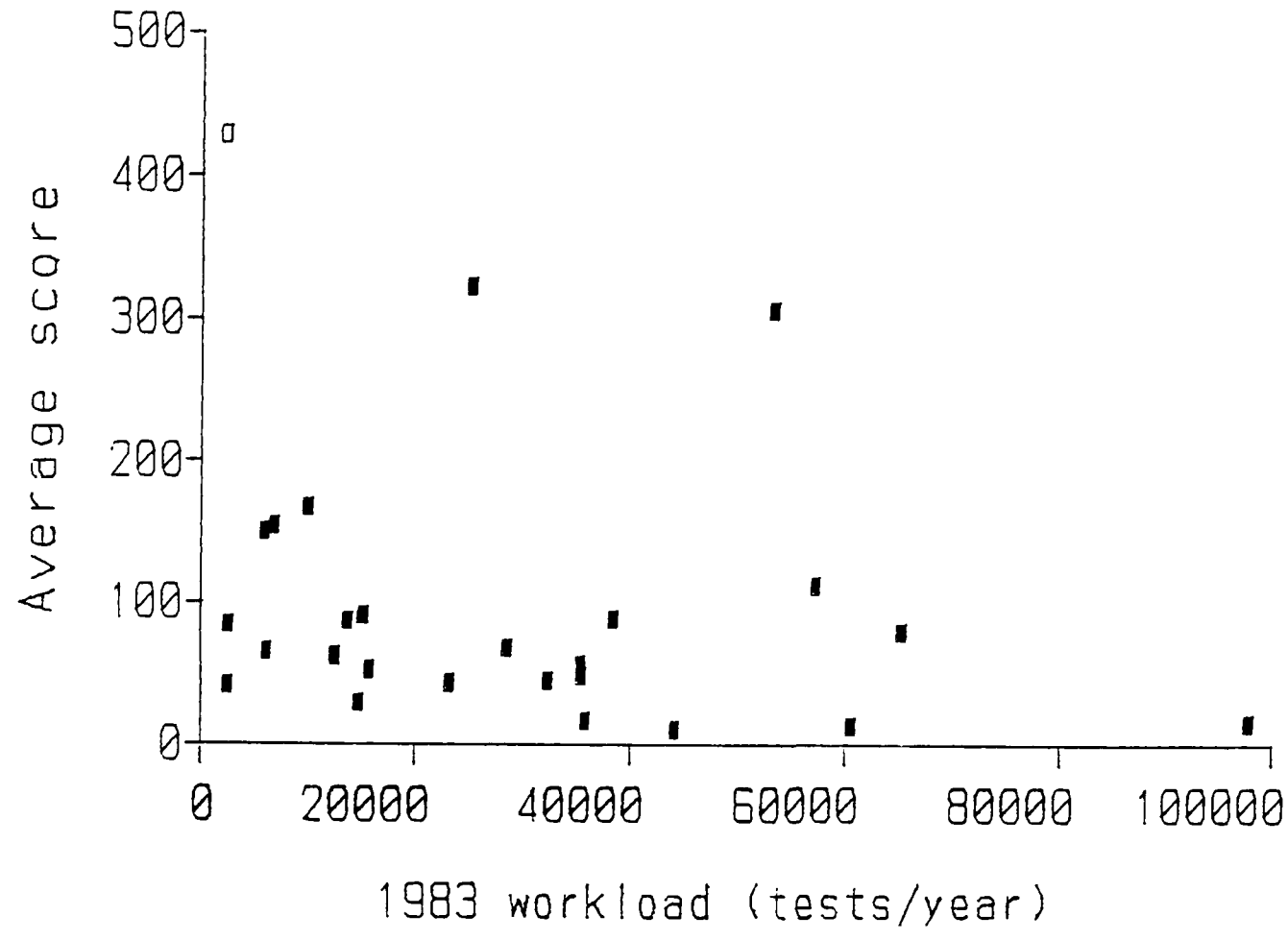
	Number of laboratories	
	MRVIS 0-80	MRVIS >80
Annual workload:		
0 - 1000 tests/year	25	16
>1000 tests/year	39	1
Average batch interval:		
1 day - 1 week	54	7
>1 week	10	10

Commission, 1980) reveals significant associations for both workload parameters. The assay is specialised and technically demanding, so correlation with the workload for the individual analyte would be expected, whether carried out in a specialised occupational or environmental monitoring laboratory or in a section within a much larger clinical laboratory.

Neonatal PKU screening provides a similar situation, and Figure 11.2 shows a graphical presentation of performance (Appendix III.3.1) against annual workload. Performance can be expressed in terms both of average scores and of turnround time (see section 8.4). Here there is also no clearcut relationship, but a definite association with both aspects of performance is apparent.

For urinary pregnancy oestrogen assay the IQC problems have probably become increasingly important as the clinical use of this determination has declined in recent years. Many laboratories have ceased to offer a service as physical monitoring of foetal well-being has superseded biochemical means, and most other participants have a decreasing workload. Batch frequency has also tended to decrease, despite the dependence of clinical usefulness upon rapid availability of results (Wilde and Oakey, 1975). The situation in 1985 is summarised in Figure 11.3. Again no clearcut distinctions can be made, but definite trends are apparent. Recent deterioration in overall performance, following previous major improvements (Figure 3.1; Bullock and Wilde, 1985), confirms that an assay done in smaller numbers and (in many laboratories) less frequently is likely to show suboptimal performance.

Figure 11.2 Relationship to annual workload of (A) average score for Surveys 15-19 and (B) turnround time in UKEQAS for PKU Screening



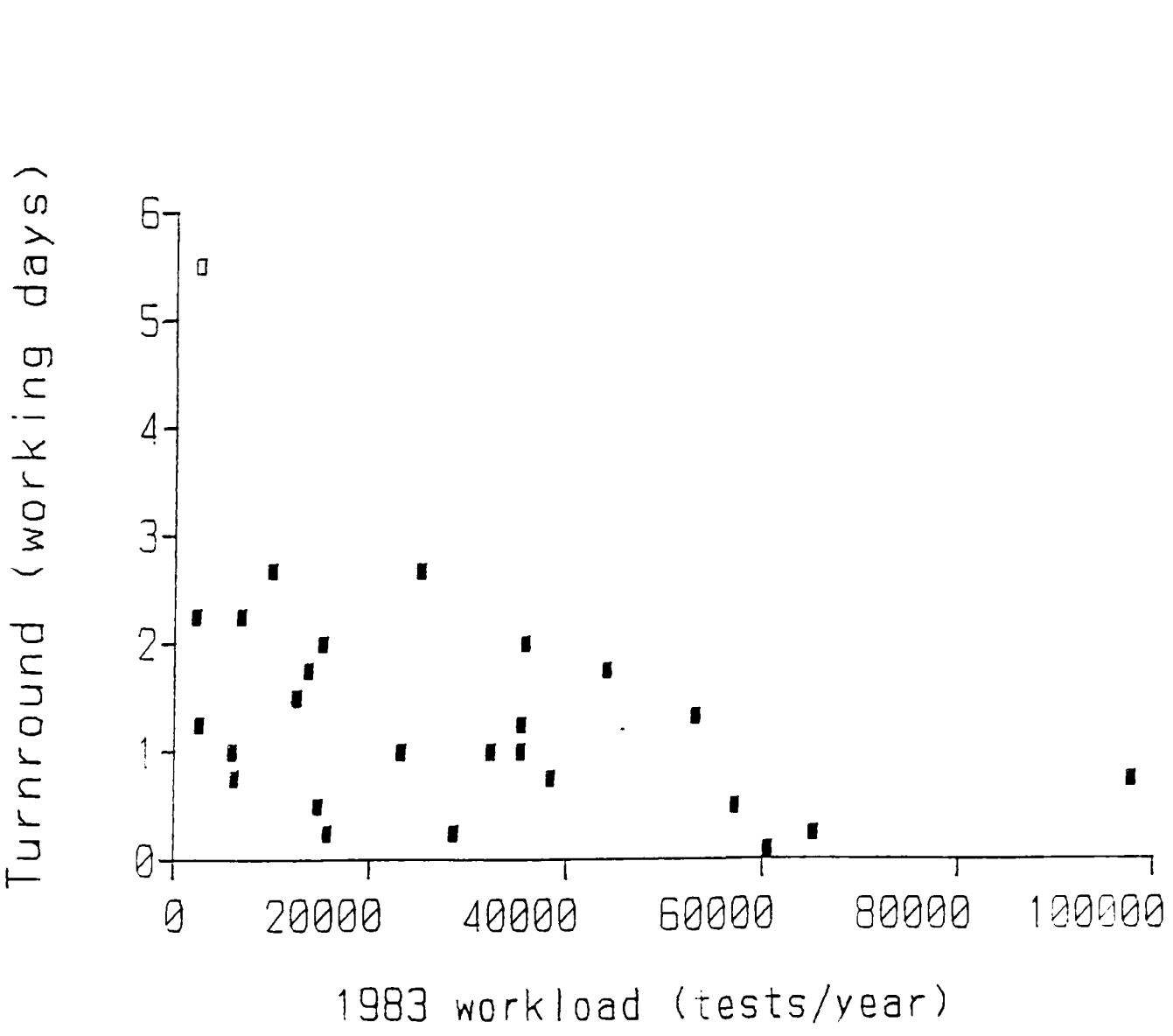
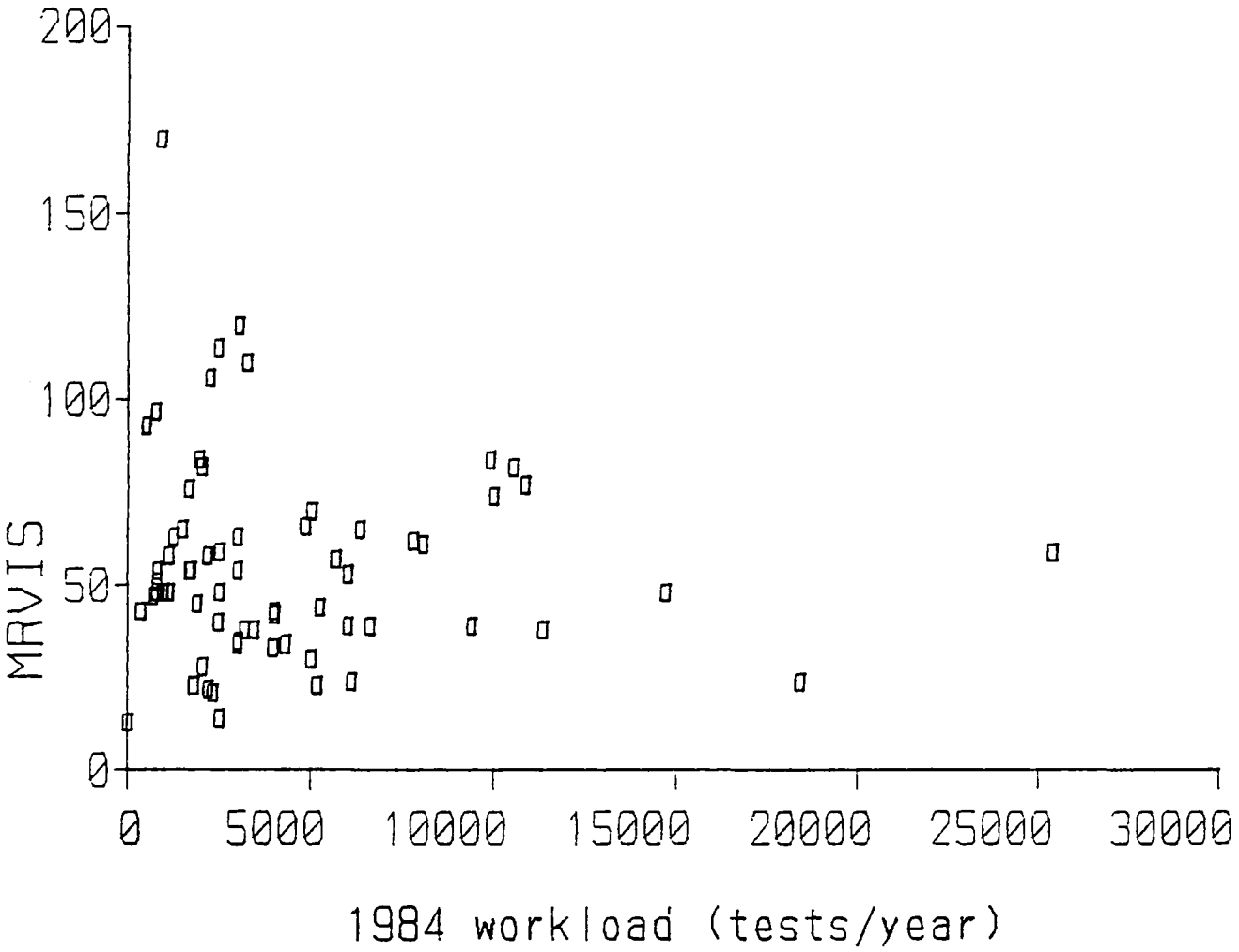


Figure 11.3 Relationship to (A) 1984 annual workload and (B) 1985 batch freqency of performance (MRVIS at June 1985) in UKEQAS for Urinary Pregnancy Oestrogens



11.3 The influence of laboratory type on performance

The type of laboratory may also have a bearing upon performance. In some cases this might be due to differing assay requirements, and in others to variation in the facilities available.

Thus examination of performance in the UKEQAS for Lead in Blood (Table 4.1; Bullock et al, 1986c) reveals differences in average MRVIS. The laboratories carrying out the assay for environmental monitoring purposes perform better, perhaps reflecting their need to provide precise and accurate results at lower lead concentrations. Differences according to the type of establishment were much less consistent, suggesting that the assay purpose is the primary factor determining performance for an individual determination.

For overall performance, however, the only main distinction that might be made would be between assays carried out for diagnosis or monitoring of therapy and those for screening. Since very few UK laboratories deal primarily with the latter, this is not susceptible to reliable study. The type of laboratory providing the analytical service may also be a determinant of performance, and this can be studied more readily through UKEQAS data.

Table 11.1 thus gives average performance data for the various groups of laboratory participating in the UKEQAS for General Clinical Chemistry. The overwhelming majority are UK National Health Service (NHS) laboratories, with several small more specialised groups such as those in the pharmaceutical industry conducting assays in drug toxicity testing. The other groups of appreciable size whose performance may then be compared with that of the UK NHS laboratories are the UK private sector and laboratories in Eire.

On overall appraisal the performance in the private sector appears considerably worse than that of NHS laboratories. One potential confounding factor is laboratory size, discussed in section 11.2.1 above. Most of the private sector laboratories are in size group I (<100,000 tests/year), and comparison should therefore be with this group. This, however, fails to account fully for the difference seen. Other factors are undoubtedly involved, therefore, and it must be remembered that many of these laboratories are small multidisciplinary units without full-time supervision by a medical or non-medical clinical chemist. The association of such factors with poor performance was highlighted by the Nuffield survey (section 11.1.1 above; Maclagan et al, 1980) and later by the Advisory Panel, who described a "small laboratory syndrome" deriving primarily from professional isolation which appeared to be linked with poorer performance standards (Browning, 1984). The performance of the three larger private sector laboratories (average OMRVIS 29) compares favourably with that of the larger NHS laboratories. An additional influence is the relatively recent (1984-1986) establishment of some private sector laboratories, and their consequent position on the 'learning curve' of EQAS participation as discussed in section 9.4.1.

The performance of the Irish laboratories also appears worse to that in the UK NHS. Again many of these are smaller laboratories and the comparison should therefore be with the corresponding size groups in the UK, though this still suggests inferior performance. These participants are mainly the larger laboratories in Eire, so the factors described above apply to a lesser degree. Nevertheless, geographical isolation may also

contribute to their problems. Examination of data from previous years suggests that improvement is occurring gradually. For example their average OMRVIS in 1982 was 77 (UKEQAS overall average 62), with 4 of 25 laboratories having a OMRVIS >100, whereas now none have such a score.

11.4 Summary

EQA data provide a convenient means for identifying factors which may be associated with good or poor performance. The Nuffield survey provides a good example of a systematic study of such factors. Problems of such studies include the inability to establish a causal relationship and the strong inter-relation among many of the factors. This may lead to identification of a general pattern rather than of specific factors.

The main factor identified is large laboratory size, conveniently reflected as annual workload, provided resources are commensurate with the workload undertaken. UKEQAS data confirm this association of laboratory size with better overall performance. For individual determinations there is also an association with batch frequency. The clinical application of the assay may be important, eg in blood lead assay for environmental monitoring.

A second important factor is the laboratory's organisation and general attitude to maintaining professional standards. The effects of professional and geographical isolation in inducing a "small laboratory syndrome" associated with poor performance are seen through comparison of UK private sector and Irish laboratories with those in the UK NHS. Such comparisons must also take into account the association of laboratory size with performance.

Chapter 12:

THE SUITABILITY OF QUALITY CONTROL MATERIALS FOR EXTERNAL QUALITY ASSESSMENT

12.1 Introduction

Any quality assurance activity can only be as reliable as the reference or QC material used. Effective EQA thus demands that the specimens distributed are stable and have properties which reflect faithfully the behaviour of clinical specimens; defined as "fidelity" by Fasce et al (1973).

As described in section 3.3, clinical chemistry is concerned primarily with determinations on serum or plasma specimens from (human) patients and subjects; serum will be used as the main example, though similar considerations apply to other biological fluid or tissue specimens. To be most effective, therefore, QA procedures should ideally use serum from these patients or subjects as the QCM. Such materials, however, present problems of instability, infectivity and restricted availability. Commercial products, especially those stabilised by freeze-drying, are convenient and relatively free from these problems and are therefore used widely. The properties and limitations of such specimens must be considered, and in relation to their intended use since a QCM may be suitable for one analytical situation but not another (Stamm, 1979; Büttner et al, 1980a).

The main categories of QCM use, also discussed in Chapters 13 and 14 are:

- precision control in IQC
- bias control in IQC

- bias control in EQA
- calibration of assays
- method comparison

12.1.1 Precision and bias control in IQC

Where QCMs are used in IQC, their properties must be as close as possible to those of clinical specimens to ensure that faults in the analytical method leading to erroneous results for specimens from patients are detected.

Ideally, therefore, materials used for IQC should be as responsive to changes in assay conditions as are clinical specimens, since in this way all 'out of control' situations should be detected while no satisfactory batches would need to be repeated. These requirements apply irrespective of whether the material is used for precision control only (with a value defined within the user laboratory), or has an assigned value for the method used and is also used in bias control.

12.1.2 Bias control in EQA

In EQA, the QCM must be stable, and the precision for the QCM should be identical to that for clinical specimens. If the designated value against which a participant's result is assessed is derived using an identical analytical procedure, eg for CK an assessment of laboratories using the 'European' NAC-activated procedure at 37^{o}C against the consensus value for that method, there is no further requirement on QCM properties (though an animal-based QCM would not be ideal for assessing, for example, albumin assay).

In other cases, however, the designated value is for all methods, or for a group of methods (eg methods for AST without addition of

pyridoxal phosphate). Such applications are exemplified by the 'new concept' proposed in GFR (Stamm, 1982). Here the QCM's response to differing analytical procedures within the grouping used should be the same as the response of clinical specimens. This requirement is for '**commutability**'.

This property was defined originally by Fasce et al (1973) as "the ability of an enzyme material to show interassay activity changes comparable to those of the same enzyme in human serum", but the clearer and more general definition "the consistency of the relationship between results obtained by different analytical methods for control specimens and patients' specimens" (Bullock et al, 1980b; Broughton et al, 1981) is preferable.

Commutability is of vital importance in EQA if information on laboratory and method performance is to be reliable (Bretaudiere et al, 1974 and 1981a; Rej et al, 1984). The QCMs distributed must therefore be commutable or the scheme design be such as to render negligible the effects of any lack of commutability: the use of method means as designated values is one such solution. If not, participants will not have confidence in the information generated by the EQAS and the scheme will fail in its objectives.

12.1.3 Calibration of assays

For a calibration material used for a single method, which is assumed to be in control, the only requirements are for a known activity (ie reliable assigned value) by that method and for this activity to be stable. If a calibration material is to be used to convert results obtained by different methods, however, there is also a requirement for it to be commutable with fresh human sera for the methods concerned, as discussed in Chapter 13 and for enzyme activity assays by Moss et al (1985).

12.1.4 Method comparison

Again for method comparison the primary requirement is for commutability, with long term stability being of lesser significance. Methods for any analyte should, however, be evaluated primarily with clinical specimens rather than with QCMs.

Additionally, comparisons or cross-calibration between methods are valid only if specimens from patients are also commutable between the methods: ie there is no between-specimen variation in the relationship between results obtained using the two methods. Any link between methods must therefore be made by assaying clinical specimens by both methods, and then checking that use of the reference material as a calibrant leads to full recovery of the activities (ie obtaining the same numerical results) for the specimens (Moss et al, 1985).

High density lipoprotein (HDL) cholesterol assay provides a good example of the problems which may be encountered in the application of QCMs in IQC and EQA, due to differences in properties. The use of lyophilised materials with direct-reading ISE instruments for sodium and potassium assay provides further examples of such difficulties. The requirements for enzyme reference materials and calibration materials are considered in more detail in Chapter 13.

12.2 High density lipoprotein (HDL) cholesterol assay

Following the demonstration of an inverse correlation with the incidence of ischaemic heart disease in epidemiological studies such as the Framingham Study (eg Castelli et al, 1977, Gordon et al, 1977) there has been considerable interest in recent years in

the measurement of HDL cholesterol in serum. Though the interpretation of results in individual subjects remains controversial, many laboratories have introduced this assay (eg Ballantyne, 1984).

Methods for HDL isolation have been known for many years (eg Cohn et al, 1946; Burstein et al, 1970). Based on this work, many methods and variants thereof have been proposed for selective precipitation of other lipoproteins followed by estimation of cholesterol in the supernatant; such methods are widely used and have been compared (eg Warnick et al, 1979). Ultracentrifugal separation of HDL, though regarded as a reference procedure, is unsuitable for routine use, and though electrophoretic separation has been proposed it has been shown to be insufficiently precise (Goldberg, 1978; Bullock et al, 1980a).

Since interlaboratory agreement must be excellent to permit the use of common criteria for interpretation, and this had been shown to be poor in the UK (Bullock DG, Carter TJN, Whitehead TP, personal communication) and overseas (Boerma, 1979; Hainline et al, 1980; Warnick et al, 1980), a move towards towards standardised methods, previously shown to be effective in lipid assays (Lippel et al, 1978), appeared desirable. To this end, provisional recommendations for the two precipitation procedures - those based on phosphotungstate/Mg^{2+} (PhT) and heparin/Mn^{2+} (Hep) - used most widely were published for comment (Whitehead et al, 1979).

For this or other initiatives to be successful, however, suitable materials must be available for calibration, IQC and EQA. The choice of a QCM for HDL cholesterol assay is particularly

difficult: lipoproteins are complex entities of variable and species-dependent composition and have limited stability, and the precipitation procedures rely on their physical rather than chemical properties. Fresh human serum is therefore the ideal, but this is neither possible for IQC nor practicable for EQA and it would therefore be of great advantage to supplement the inadequate procedures (Bullock et al, 1980b) with lyophilised materials.

Some lyophilised QCMs, however, were known to lack fidelity and, in particular, shown to be non-commutable with fresh sera even for simple organic analytes and for enzyme activity assays (eg Bretaudiere et al, 1974; Saidi, 1979; van Helden et al, 1979). Combination of factors such as non-human origin of the base serum, lack or excess of physiological interferents, addition of impure or non-physiological compounds and extracts with the disruptive effects of freeze-drying on lipoproteins suggests that the situation for HDL cholesterol assay could be considerably worse.

In this situation empirical study of a broad selection of materials is essential, and such an investigation was therefore undertaken. Several of the materials appearing to have properties suitable for use in EQA were then distributed to a small group of laboratories in a limited survey, to validate this conclusion.

12.2.1 Investigation of commercial QC sera

The study (Appendix III.5; Bullock et al, 1980b) comprised assessment of the mean HDL cholesterol concentration and within-batch imprecision for each QCM, by both provisionally recommended precipitation procedures (Whitehead et al, 1979). These methods had previously been shown to yield good agreement for clinical

specimens of fresh serum. Pooled fresh serum from patients and 25 commercial and commissioned sera (Table III.3), chosen to cover a range of manufacturing procedures, species of origin, additives, constituent concentrations and presentation (liquid or lyophilised), were studied.

Table 12.1 shows the great diversity in results obtained, with some sera (eg serum S) exhibiting very poor precision. With clinical specimens giving SDs of 0.015-0.02 mmol/L, a criterion for acceptability for use in IQC of an SD 0.03 mmol/L or less and a CV of 4% or less was proposed (Bullock et al, 1980b). Most sera (15 for the PhT and 14 for the Hep procedure) satisfied this criterion, but in some cases (eg sera K, Q, T and Y) there were striking differences in behaviour. For some of these (K and Q) the methods also differed in accuracy.

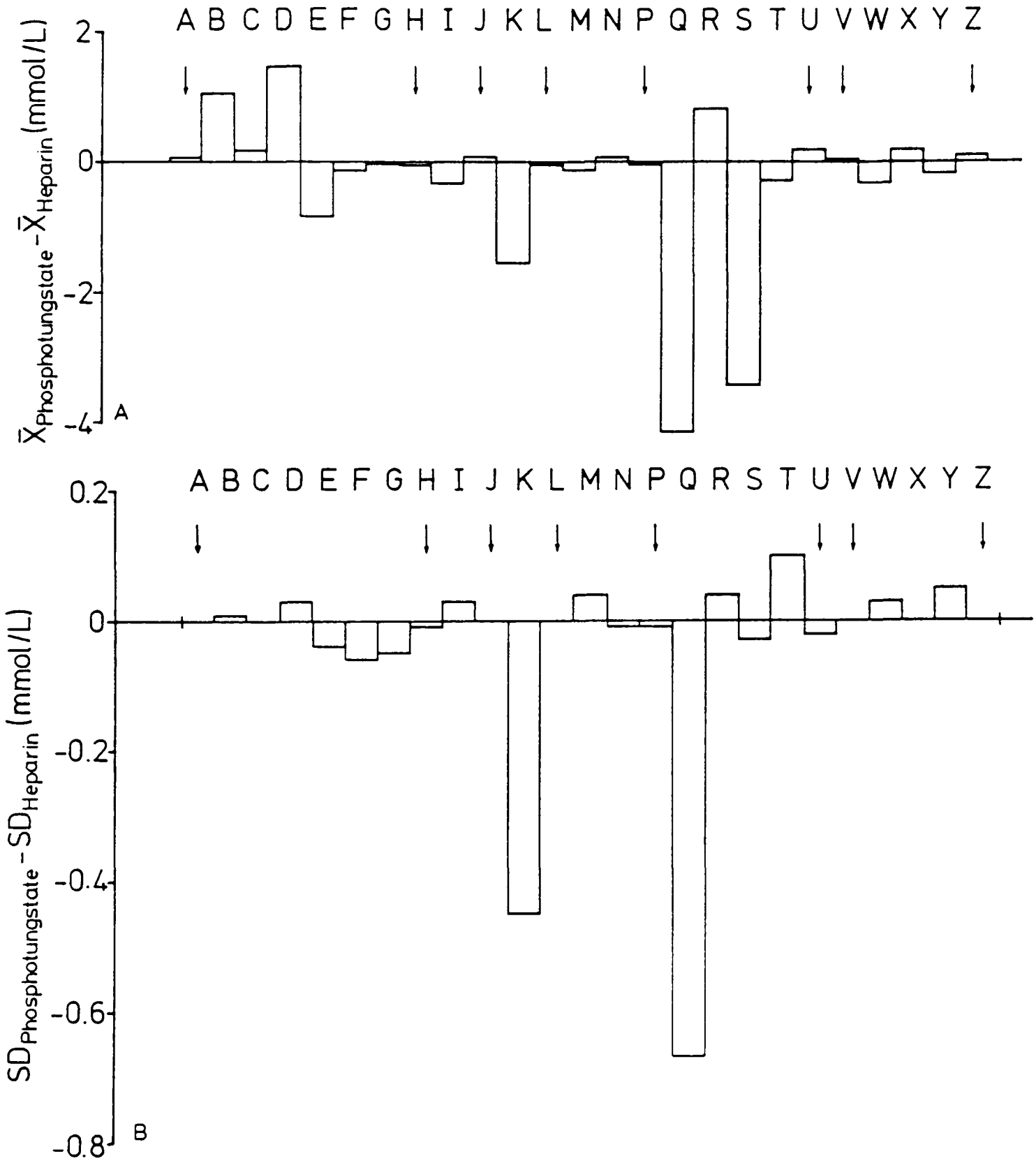
When the results were expressed in terms of the difference in mean (Figure 12.1A) or SD (Figure 12.1B) for the two procedures further differences were revealed. Differences in accuracy were not confined to the Hep procedure, and appeared to be related to incomplete precipitation of other lipoproteins. Major differences, indicating potential unsuitability for at least one method, would preclude use in EQA. A criterion of a maximum difference of 20% between the mean concentrations, which should (unlike those for sera E and F) not exceed the range usually encountered, was therefore taken.

Application of these criteria left eight sera, indicated by arrows in Figure 12.1, which appeared to be suitable for EQA. None of these was of bovine origin, but all the equine sera studied and only a quarter of the human-based materials satisfied

Table 12.1 Mean and within-batch precision (SD) of HDL cholesterol assay by PhT and Hep procedures on pooled patients' sera and 25 QC sera. In mmol/L; n = 10 (n = 20 for pooled sera) unless specified (* n = 9; $^{+}$ n = 5)

Serum	Phosphotungstate/Mg^{2+}		Heparin/Mn^{2+}	
	Mean	SD	Mean	SD
Pools	1.40	0.017	1.39	0.015
	1.50	0.02	1.43	0.015
A	0.87*	0.02	0.80*	0.02
B	2.14	0.03	1.09	0.02
C	0.77	0.03	0.61	0.03
D	2.32	0.09	0.83	0.06
E	3.17	0.02	4.00	0.06
F	4.03	0.04	4.17	0.10
G	2.09	0.03	2.11	0.08
H	1.23	0.01	1.28	0.02
I	1.47	0.08	1.80*	0.05
J	0.92	0.02	0.84	0.02
K	1.42	0.03	2.99	0.48
L	0.88	0.01	0.94	0.01
M	1.43	0.06	1.58	0.02
N	2.55	0.03	2.50	0.04
P	1.38	0.02	1.40	0.03
Q	1.61	0.02	5.78	0.69
R	1.56	0.06	0.77^{+}	0.02
S	2.04	0.29	5.48^{+}	0.32
T	2.14	0.13	2.43	0.03
U	1.22	0.01	1.06	0.03
V	1.27	0.02	1.24	0.02
W	3.29	0.05	3.62	0.02
X	0.78	0.03	0.64	0.03
Y	1.71	0.06	1.90	0.01
Z	0.79	0.02	0.70	0.02

Figure 12.1 Differences between the PhT and Hep precipitation procedures for the 25 sera studied, in (A) mean results and (B) within-batch SD. Arrows denote sera meeting the criteria for use in EQA (CV <4%, SD <0.03 mmol/L, and difference between means <20%)



the criteria, as did the single liquid (equine) serum. Only one of the 6 human-based sera intended for control of lipid assays appeared suitable.

To confirm that the between-day precision was reflected by the within-batch data obtained in Table 12.1, 6 sera with a variety of properties were then assayed on 10 successive working days. The CVs (Table 12.2) were, as expected, greater than the within-batch imprecision, especially by the Hep procedure. Correspondence appeared reasonable in most cases, and the within-batch data could then be taken to represent performance for this assay.

The study thus demonstrated that some of the materials were suitable for IQC of one or both of the analytical procedures investigated. A more restricted group appeared acceptable for use in EQA as well, with comparable concentrations by both procedures which lay within the physiological range, though only fresh serum serum specimens should be used to assess intermethod differences. Surprisingly, suitability was not confined to human-based products, with equine sera performing well and 'special' lipid controls badly.

12.2.2 Use of lyophilised sera in an EQA survey

Having determined which of the materials in the single-laboratory study appeared to be suitable for such use, 6 were distributed in an EQA survey. This included a relatively small number (14) of laboratories collaborating in the method standardisation initiative (Whitehead et al, 1979), and which had previously received specimens of fresh liquid human serum (John, 1983; see Appendix I.5.1).

Table 12.2 Comparison of within-batch and between-day mean and precision (CV) of HDL cholesterol assay for 6 sera. n = 10 unless specified (* n = 9); means in mmol/L

	Phosphotungstate/Mg^{2+}		Heparin/Mn^{2+}	
	Within	**Between**	**Within**	**Between**
Serum D				
Mean	2.32	2.46	0.83	0.74
CV (%)	3.7	4.7	7.5	7.1
Serum G				
Mean	2.09	1.98	2.11	2.13*
CV (%)	1.4	4.1	3.7	4.6
Serum N				
Mean	2.55	2.39	2.50	2.36
CV (%)	1.0	3.4	1.7	2.2
Serum P				
Mean	1.38	1.36	1.40	1.37
CV (%)	1.8	3.3	1.9	6.3
Serum U				
Mean	1.22	1.23	1.06	1.06
CV (%)	0.8	3.3	3.3	6.0
Serum W				
Mean	3.29	3.18	3.62	3.46
CV (%)	1.5	5.5	0.7	1.9

Table 12.3 shows the interlaboratory agreement obtained for HDL and total cholesterol in Survey 3, using the commercial materials. This is compared in Table 12.4 with data obtained in surveys using fresh sera. In addition to groups using the PhT and Hep procedures, two laboratories employed polyethylene glycol (PEG) as precipitant.

The wide spreads of CVs are attributable primarily to the small numbers of participants in each group and to the effects of analyte concentration upon interlaboratory agreement (shown for the Hep group in Figure 12.2A, and discussed in general terms in Chapter 6). Nevertheless the same conclusions could be drawn from the lyophilised specimens as from the liquid sera, namely the positive bias of Hep relative to PhT procedures and the similar between-laboratory variability for these two methods (Table 12.4). The apparently worse performance for total cholesterol was attributable to the rather lower analyte concentrations in this group of predominantly animal-based materials (Figure 12.2B).

12.3 Sodium and potassium assay using direct-reading ion-selective electrodes (ISEs)

Instruments employing potentiometry on undiluted sample have been used increasingly both within and outwith laboratories for sodium and potassium assay (Buckley et al, 1984). Factors contributing to this growth include their greater convenience for emergency situations, particularly the capability for whole blood samples thus obviating the need for specimen centrifugation and separation, and commercial pressure.

In contrast to flame photometry and ISE systems incorporating sample dilution prior to measurement (indirect ISEs), such procedures should in theory estimate the ionic molal activity,

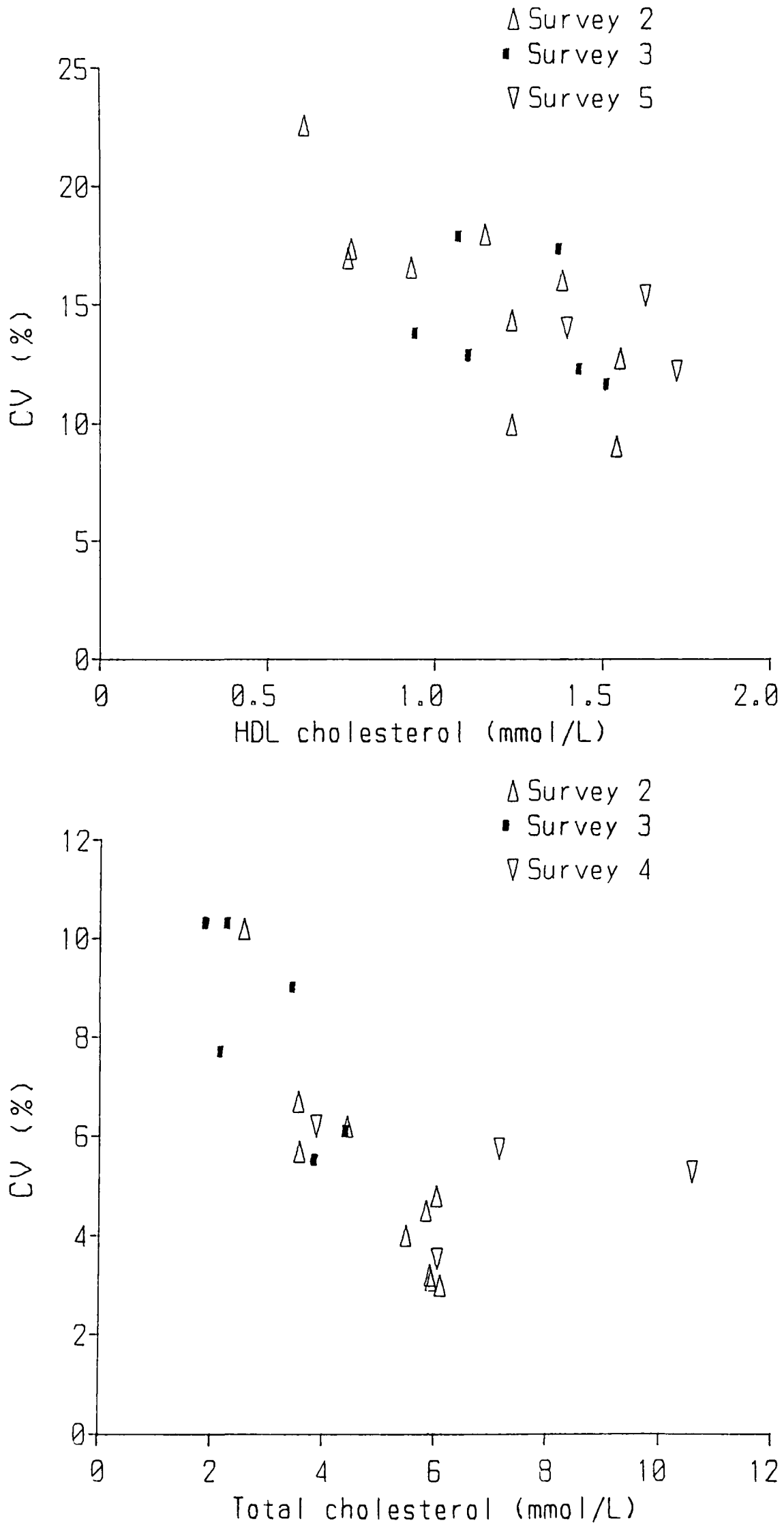
Table 12.3 Intralaboratory precision, mean results and between-laboratory agreement for HDL and total cholesterol in Survey 3, 1980

	Serum					
	H	**J**	**L**	**P**	**U**	**Z**
HDL cholesterol (mmol/L)						
Intralaboratory CV (%)	0.7	2.5	3.8	1.9	3.1	2.5
Overall (n=14 laboratories)						
Mean	1.32	0.93	1.00	1.35	1.27	0.91
CV (%)	13.3	20.3	15.2	17.1	16.1	10.5
PhT (n=7)						
Mean	1.25	0.88	0.92	1.34	1.26	0.90
CV (%)	13.7	16.3	15.7	15.2	14.4	9.4
HepT (n=5)						
Mean	1.43	1.07	1.10	1.51	1.37	0.94
CV (%)	12.3	17.9	12.9	11.7	17.4	13.8
PEG (n=2)						
Mean	1.27	0.79	1.02	1.04	1.09	0.89
Total cholesterol (mmol/L)						
Intralaboratory CV (%)	1.8	2.4	1.9	3.9	3.1	3.4
Enzymic (n=12)						
Mean	2.14	3.44	4.40	2.27	1.87	3.84
CV (%)	7.7	9.0	6.1	10.3	10.3	5.5

Table 12.4 Summary of intralaboratory precision, mean results and between-laboratory agreement for HDL and total cholesterol in Surveys 2 (1979), 3 (1980), and 4 or 5 (1981). 10 specimens in Survey 2, 6 in Survey 3 (Table 12.3), 4 in Survey 4 and 3 in Survey 5

	Survey 2	Survey 3	Survey 4 or 5
HDL cholesterol (mmol/L)			
Intralaboratory CV (%)	3.81	2.42	0.81
Overall (n=12-14 laboratories)			
Mean	1.07	1.13	1.51
(Range)	(0.55 - 1.52)	(0.91 - 1.35)	(1.37 - 1.66)
CV (%)	12.7	15.4	12.6
(Range)	(7.1 - 21.4)	(10.5 - 20.3)	(11.8 - 13.3)
PhT (n=6-7)			
Mean	1.05	1.09	1.43
CV (%)	10.2	14.1	16.8
(Range)	(5.0 - 19.0)	(9.4 - 16.3)	(15.4 - 17.7)
Hep (n=5)			
Mean	1.11	1.24	1.58
CV (%)	15.4	14.3	13.9
(Range)	(9.1 - 22.6)	(11.7 - 17.9)	(12.3 - 115.5)
PEG (n=1-2)			
Mean	1.03	1.02	1.58
Total cholesterol (mmol/L)			
Intralaboratory CV (%)	2.11	2.75	1.71
Enzymic (n=12-13)			
Mean	4.95	2.99	6.93
(Range)	(2.57 - 6.04)	(1.87 - 4.40)	(3.87 - 10.61)
CV (%)	5.1	8.2	5.2
(Range)	(3.0 - 10.2)	(5.5 - 10.3)	(3.5 - 6.2)

Figure 12.2 Relationship with (A) HDL and (B) total cholesterol concentration of between-laboratory agreement (CV) for liquid and lyophilised sera. HDL cholesterol determined by Hep procedure



claimed to be the most useful indicator of physiological activity. The results and reference intervals should then differ appreciably from those for plasma sodium concentration. The systems are, however, complex and most manufacturers have endeavoured to provide results comparable with concentrations, through manipulation of the electrodes, liquid junctions and calibration material composition and assigned values (Broughton and Maas, 1984; Buckley et al, 1984).

These factors lead to considerable difficulties for measurements on clinical specimens (eg Broughton et al, 1985; Smith et al, 1986), in view of the complex interrelationships with specimen composition, especially protein and lipid content. It has been recommended that such instruments should yield results in agreement with those by flame photometry for specimens with normal protein and lipid concentrations, ie the same reference intervals should be applicable (Broughton and Maas, 1984).

The problems associated with provision of suitable materials for calibration, IQC and EQA may be correspondingly large. Calibration cannot be dissociated from instrument design, and is therefore essentially a matter for manufacturers rather than users. For precision control, the requirement is for precision comparable with that for clinical specimens (see section 12.1.1 above). Requirements for accuracy control in IQC and for EQA are, however, more stringent, requiring either commutability with clinical specimens or an appropriate grouping of methods and instruments to avoid any adverse effects of a lack of commutability.

12.3.1 UKEQAS for General Clinical Chemistry

Throughout the development of the scheme new method groups have been introduced as new analytical procedures, using novel principles or instruments, came into use. Thus separate groupings were provided for direct and indirect ISE procedures, which were increasing in use during the early 1980s. When more than 15 results were received VISs were calculated; as indirect ISEs are used more widely in laboratories, particularly in the Technicon SMAC and Beckman Astra instruments, this stage was reached earlier for this group. No problems were apparent from the EQA data and such use has continued to increase, with 43% of sodium results in December 1986 being obtained by indirect ISE methods.

The direct ISE group acquired VISs later, and observation that the scores seemed to be higher than those for other methods prompted a more detailed analysis in December 1984. Table 12.5 compares the average scores obtained by the direct ISE group with those for all participants, with further analysis of the performance of the sub-groups using IL, Corning and other instruments.

The worse than average performance is confirmed by the higher average MRVISs, both for the group as a whole and for the sub-groups. Differences in bias among the various manufacturers' instruments are revealed by the non-zero average MRBISs for the sub-groups. In addition to these between-manufacturer differences, the variations in MRBIS within each sub-group (eg for Corning instruments the ranges were -70 to +25 for sodium and -107 to +47 for potassium) indicate differences between individual laboratories. These may result from differences between instrument models or from differences between

Table 12.5 Comparison of performance (average running scores) for sodium and potassium assay by direct ISE instruments in UKEQAS for General Clinical Chemistry, 1984 |MRBIS| is the average MRBIS disregarding its sign

	Overall	Direct ISE			
		All	IL	Corning	Others
n	429	26	5	15	6
Sodium					
MRVIS	64	77	98	71	69
MRBIS	-	-	-23	-28	+39
|MRBIS|	-	39	29	42	45
SDBIS	71	90	118	83	74
Potassium					
MRVIS	61	89	107	87	76
MRBIS	-	-	+10	-41	+26
|MRBIS|	-	42	18	52	46
SDBIS	69	99	127	94	81

laboratories in the settings used. These factors combine to give an average CV during 1984 of 1.67% for sodium by direct ISE, considerably greater than the average of 1.30% for all participants.

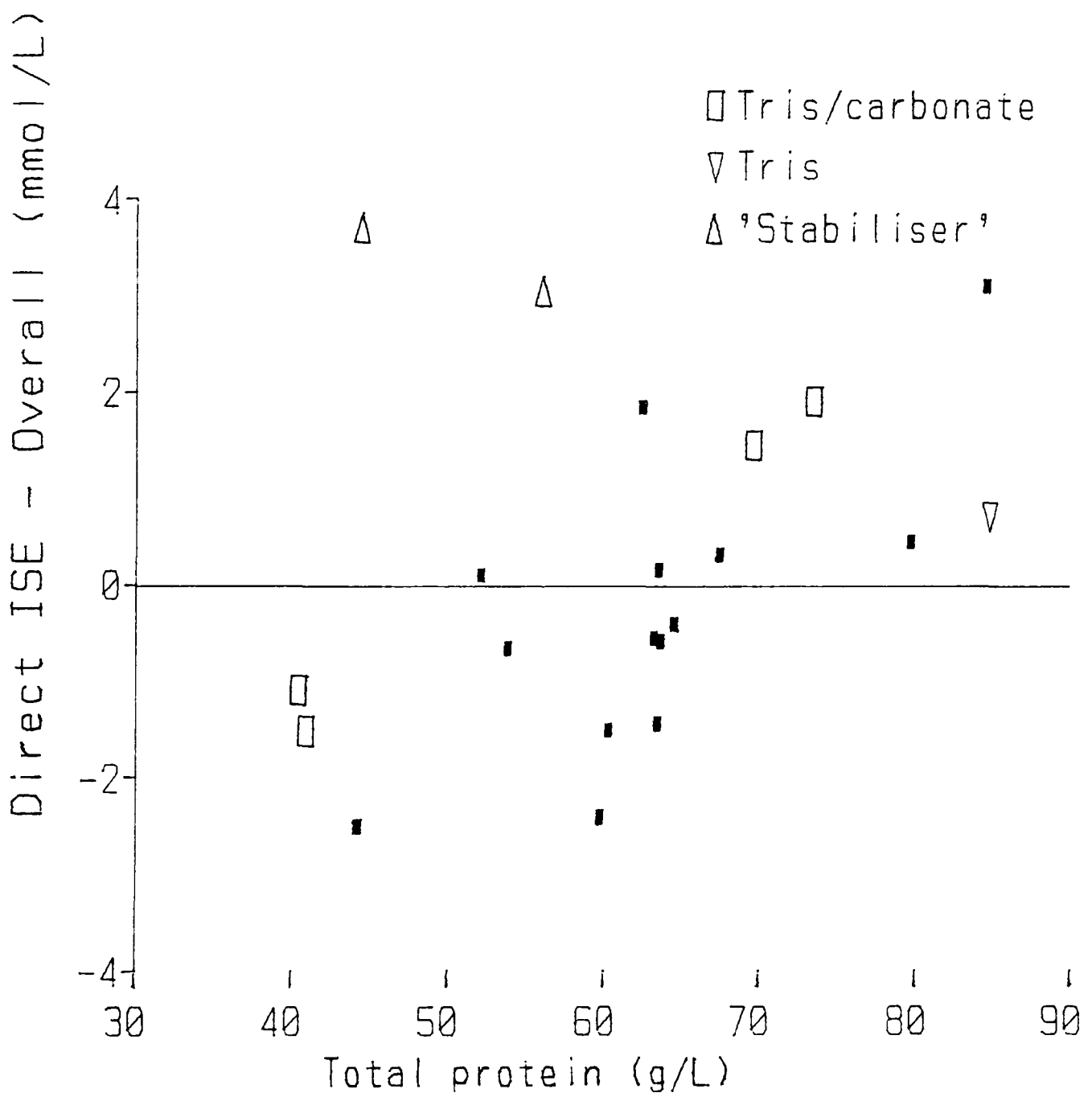
More importantly the high SDBISs, with only three laboratories achieving scores lower than the average for all participants, suggest that inconsistent bias could stem from interactions between these methods and the materials distributed. Could this then be due to non-commutability?

Studies (eg Broughton et al, 1985; Smith et al, 1986) have shown a consistent relationship for clinical specimens between total protein concentration and 'ISE - flame difference' (the amount by which the direct ISE result exceeds the result by flame photometry) for several commercial instruments. The exact relationship differed but in each case a positive correlation was observed. Did the UKEQAS data also show this pattern?

Figure 12.3 provides this comparison for the 21 materials distributed during 1984, from which no consistent pattern emerges, in apparent confirmation of non-commutability. Examination of the materials involved (Table I.2) reveals several with divergent manufacturing process. In particular one contained Tris buffer, 4 were to be reconstituted with a Tris/carbonate diluent rather than distilled water and two QCMs contained an (unspecified) 'stabiliser'. These are identified separately in Figure 12.3, and indeed behave differently from the remaining materials.

The scatter for the remainder, however, still appeared substantially greater than that for clinical specimens though no

Figure 12.3 Relationship with total protein of difference for sodium between direct ISE and overall mean in UKEQAS for General Clinical Chemistry, 1984. Materials with Tris/carbonate diluent, containing Tris buffer and containing 'stabiliser' are identified



further specimens with unusual additives or manufacturing treatments were apparent. The overall conclusion was thus that the materials distributed in the UKEQAS did not truly reflect the performance obtained with clinical specimens and corrective action was therefore required.

The basic problem is one of diversity within the grouping, leading to a mean which is not representative of each instrument. Even classification separately by manufacturer would not provide homgeneous groups capable of yielding valid method means, the primary way in which the effects of non-commutability can be circumvented. The alternative procedure of not scoring for distributions where the material shows atypical properties was also infeasible, since firstly there appeared to be no ready means to determine unsuitability in advance and secondly the UKEQAS computer programs do not provide the facility to exclude individual method groups (in addition to the 'Other' group) from scoring.

The action taken in January 1985 was to reclassify these procedures within the 'Other' group, so they no longer received VISs for any distribution. This also took account of the relatively small proportion of participants (still only 8% at December 1986) using direct ISE instruments. Users were advised to monitor their performance in the scheme, and to contact the instrument manufacturer if their bias appeared to be consistent.

Subsequent re-examination of data from the scheme confirmed the continuation of these problems, as would be expected without drastic changes in the type of material distributed. Indeed the very design of these instruments, with undiluted serum, plasma or whole blood being presented to the potentiometric system, should

suggest the likelihood of major matrix effects. The situation thus remains unresolved, though the possible future growth in the use of direct ISE procedures suggests that a procedure similar to that described in section 12.2.1 above for HDL cholesterol be applied to determine the suitability for direct ISE instruments of materials to be distributed in the scheme.

12.4 Summary

The primary requirements for materials to be used in internal quality control are stability and precision similar to that for clinical specimens. Materials for external quality assessment, however, must be commutable with clinical specimens, unless designated values are obtained by the same method.

A protocol for assessment of suitability was tested for HDL cholesterol assay. This included comparison of the within-batch CVs and relationship between the mean values for two common analytical procedures. Criteria based on the findings for fresh sera from patients were proposed, which were satisfied by 8 of the 25 QC materials studied. Such a design could form the basis of a general procedure for assessing the suitability of QCMs for use in EQA.

The suitability of 6 of these was further tested by use in a survey of HDL cholesterol assay in 14 laboratories. Taking account of the effects of analyte concentration, the results were similar to those obtained previously and subsequently by these participants on fresh sera, supporting the validity of this protocol.

Examination of UKEQAS data revealed inhomogeneity both among and within instrument groups for sodium and potassium using direct

potentiometry on undiluted specimen (direct ISE). This divergence was further manifested in different behaviour than for clinical specimens. The inhomogeneity reflects the susceptibility of these procedures to matrix effects, resulting in a lack of commutability.

Such effects prevented scoring of the performance for direct ISEs, since the small number of participants using them precluded the allocation of individual method groups. Application of a protocol similar to that used for HDL cholesterol assay might enable identification of commutable materials and thus permit some reliable assessment of their performance.

Chapter 13:

THE EFFECTS OF CALIBRATION ON INTERLABORATORY AGREEMENT

13.1 Introduction

The results obtained for an assay in clinical chemistry are dependent not only upon the specimen analysed and the method used but also upon the calibration procedure. Where pure compounds in aqueous solution (primary calibrants; Büttner et al, 1979a) are available and applicable their use is recommended. This procedure derives from good practice in analytical chemistry, and the inclusion of calibrants (then termed "standards"; "calibrant" is preferred, to avoid confusion with standards of performance) with each analytical batch has long been recommended to improve performance by compensating for variations in conditions (eg Henry and Segalove, 1952).

Such primary calibrants are not, however, available for all analytes (eg many protein species) nor are they applicable in all analytical procedures. If the procedure is susceptible to matrix effects (see Chapter 12 above) a secondary calibrant (eg a serum-based material; Büttner et al, 1979a) is essential to avoid errors due to lack of commutability between calibrant and clinical specimens. Such secondary calibrants must then be calibrated (ie values be assigned to them) against the relevant primary calibrant or other reference point, eg an International Reference Preparation.

Some potential problems in the assignment of values to QC and calibration materials have been discussed in Chapters 3 and 12 above. Discordant results arising from use of individual

suppliers' calibrants have indeed been reported, eg from EQA surveys of immunoglobulin assay (Ritchie and Rippey, 1982; Chambers et al, 1984). In these cases the manufacturer's protocol for calibrating against the WHO International Reference Preparation was presumably in error, and recalibration was undertaken when this deficiency was exposed by the EQA data. Similar studies have shown that use of a single method and calibrant yields better between-laboratory agreement (Rowe et al, 1970 and 1972).

Further problems arise when the analyte is not a defined chemical or protein species. For example, "pregnancy oestrogens" in urine vary in relative composition from subject to subject and from day to day in individual subjects. The relative non-specificity of many of the methods used further compounds the difficulties, and though the primary clinical interest is in day-to-day changes in oestrogen excretion within individuals (Wilde and Oakey, 1975) there are obvious advantages of between-laboratory agreement in terms of common criteria for interpretation. Similar considerations apply in any assay for a heterogenous mixture of species, such as total protein in urine.

Enzyme activity assays pose a particular problem, in that the numerical result is entirely dependent upon the analytical conditions used. Here attention has focused primarily on the reaction temperature in view of its large effect on activity, but substrate, cofactor and buffer identity and concentrations are probably more important. The trend towards the use of agreed or recommended methods, such as those proposed by the IFCC Expert Panel on Enzymes (Bowers et al, 1979) and national societies in for example GFR (German Society for Clinical Chemistry, 1970 and

1972), Scandinavia (Committee on Enzymes, 1974) and the UK (Association of Clinical Biochemists, 1980), has reduced but not eliminated this problem. Several different method recommendations are current, however, and with many clinical chemists unwilling to change their methods (Fleck and Colley, 1982) and the increased use of instrument-dependent methods it seems unlikely that the goal of a single method only for each enzyme will ever be reached. Lack of comparability of enzyme results will thus probably remain a problem in the foreseeable future, if the present practice of calculating results on the basis of the known absorption coefficient of a reaction product (Peake et al, 1984) continues to prevail. Such considerations have led to questioning of whether the standardised method approach alone can bring about interlaboratory agreement, and the potential benefits of using calibration materials for enzyme activity assays have been explored by a number of authors in recent years (eg Jansen and Jansen, 1983; Bowers and McComb, 1984; McComb and Bowers, 1985; Moss et al, 1985; Bullock et al, 1986b).

In such circumstances use of a common calibration material by all laboratories might be expected to improve interlaboratory agreement. Indeed an ability of laboratories more reliably to compare the analyte content of two specimens than to determine their absolute concentration was apparent from the earliest EQA surveys (Wootton, 1956). EQA offers a valuable means to study this important hypothesis, through the distribution of two or more specimens for analysis together in the same analytical batch with recalculation of results to mimic the use of one as a common calibrant. The data thus obtained fairly reflect the likely impact of such calibration, being derived from routine conditions and from the many method variants used in participant

laboratories.

13.2 Specific proteins in serum

In an attempt to overcome the problems mentioned above, a working calibrant for these assays (SPS-01) has been developed for use in the UK and calibrated against WHO preparations (Milford Ward et al, 1984). The UKEQAS for Specific Proteins (Appendix I.2.5; Chambers et al, 1984 and 1987) was used to assess the effects of using this as a common calibration material, both in validating its proposed utility and in monitoring the effects of its availability.

13.2.1 SPS-01 calibration study

The response of many immunochemical methods for protein assay is non-linear, so participants were asked to use this material as calibrant for the survey. Thus results using routine calibration procedures were thus not available for the same specimens, and data from the preceding and following surveys were therefore used for comparison purposes (Appendix III.4.4; Chambers et al, 1987).

The improvements in interlaboratory agreement are demonstrated in Tables 13.1 and 13.2 (which show the average CVs for distributions before, with and after SPS-01), and also in Figure 13.1 for IgG, IgM and C3. The CVs with SPS-01 were lower and their scatter less than before; patterns for IgA and A1-AT were similar to IgG, and that for C4 to C3.

It is clear from this study that overall between-laboratory variation is significantly reduced when a common reference preparation is used for assay calibration. The most marked changes were for C3 (Figure 13.1) and A1-AT, with an approximate halving of the CV from 23-24% to 12-13% (Table 13.2). The

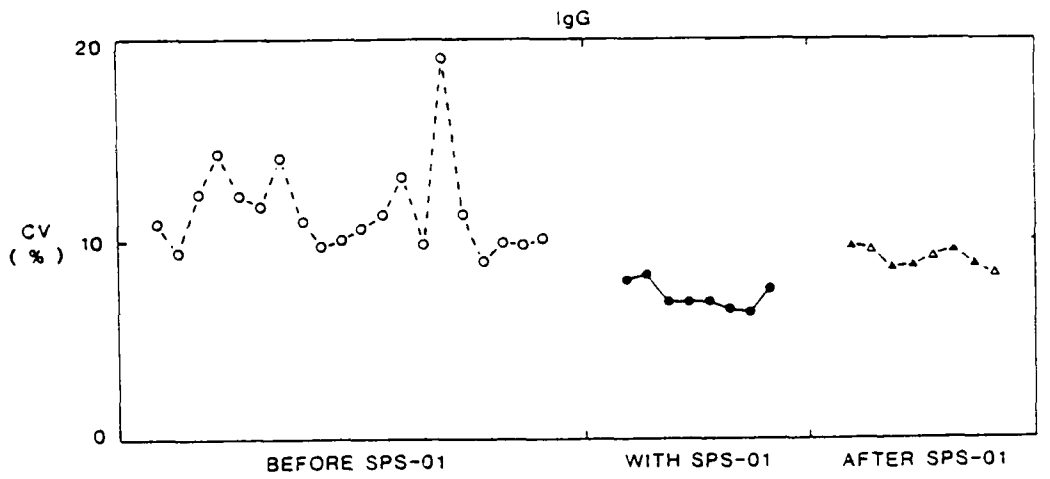
Table 13.1 Mean between-laboratory CV for immunoglobulins before, with and after SPS-01. p denotes the probability of the CV with or after SPS-01 differing significantly from before SPS-01, calculated by the Mann-Witney U test. NS = not significant, i.e. p>0.05

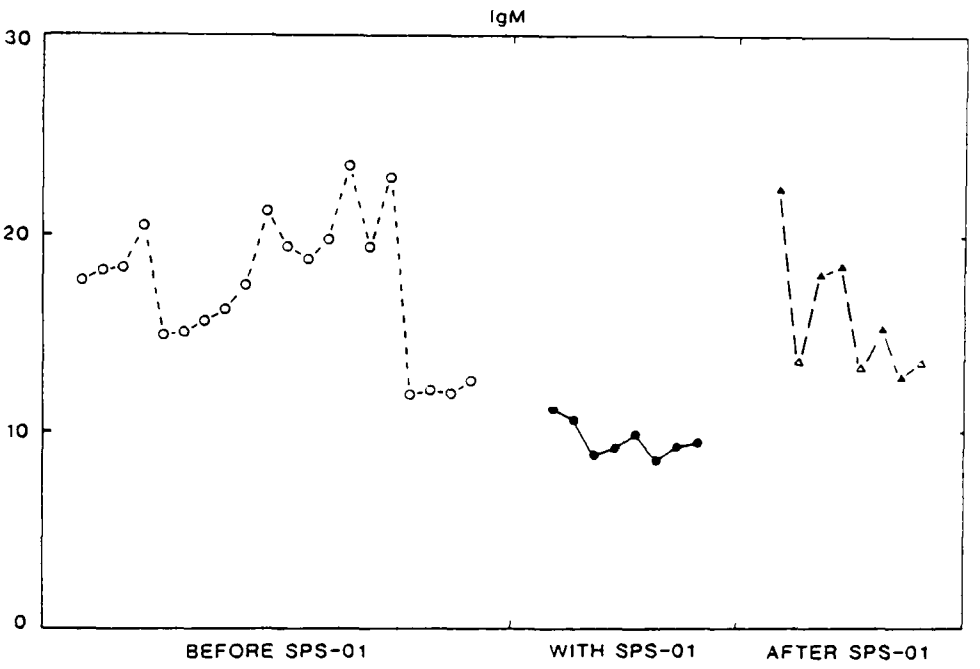
		Method group			
		Overall	**Turb**	**Neph**	**RID**
IgG	Before	11.6	10.8	12.6	11.8
	With	7.5	6.5	6.8	8.2
	p	<0.001	<0.001	<0.001	<0.001
	After	9.0	8.3	9.1	8.8
	p	<0.001	0.01	0.001	<0.001
IgA	Before	14.9	18.3	15.6	12.6
	With	8.5	7.9	8.7	9.2
	p	<0.001	<0.001	0.002	0.002
	After	11.9	12.5	12.0	9.3
	p	0.04	0.005	0.05	0.002
IgM	Before	17.8	16.5	16.3	16.1
	With	10.1	8.1	9.0	10.0
	p	<0.001	<0.001	<0.001	<0.001
	After	17.4	17.9	18.1	15.0
	p	NS	NS	NS	NS

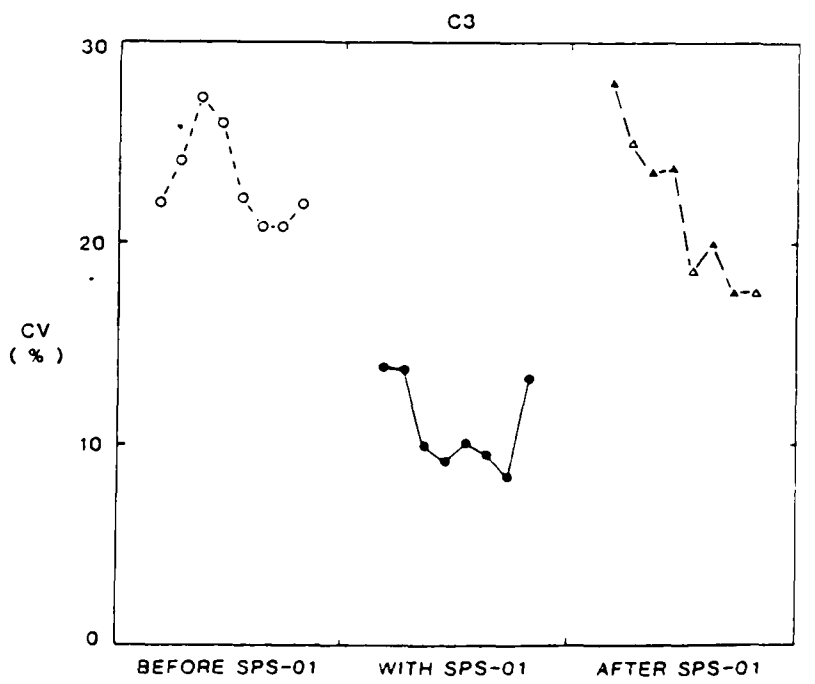
Table 13.2 Mean between-laboratory CV for C3, C4 and A1-AT before, with and after SPS-01. For explanation see Table 13.1

		Method group			
		Overall	**RID-Behring**	**RID**	**All others**
C3	Before	23.8	16.5	27.2	22.4
	With	12.0	13.8	12.6	9.3
	p	<0.001	NS	<0.001	<0.001
	After	21.9	15.3	18.0	13.7
	p	NS	NS	NS	0.002
C4	Before	18.8	15.4	18.0	17.7
	With	13.6	14.1	17.3	9.8
	p	<0.001	NS	NS	<0.001
	After	21.8	13.2	-	20.4
	p	NS	NS		NS
A1-AT	Before	23.1	15.9	21.9	26.0
	With	12.7	14.1	16.7	7.7
	p	<0.001	NS	NS	<0.001
	After	18.6	13.5	16.5	10.0
	p	0.005	NS	NS	<0.001

Figure 13.1 Interlaboratory agreement (CV) before, with and after SPS-01 for IgG, IgM and C3 in UKEQAS for Specific Proteins







magnitude of this improvement indicates the extent of differences existing between the various calibrants currently available for these proteins. The improvement for C4 was less marked (from 19% to 14%), reflecting greater consistency among calibrants for this protein.

The significant improvements observed in immunoglobulin assays (Table 13.1, Figure 13.1) are important. Recognised reference preparations for these proteins have been available for some years and commercial calibrants should be standardised against them, but these results demonstrate that considerable diversity still exists, in particular for IgM (Figure 13.1).

The results for individual method groups yield further evidence that the higher CVs before SPS-01 are due largely to inter-calibrant variation. The homogeneous groups, ie those comprising a single method and calibrant (eg for C3, C4 and A1-AT the RID group using Behring plates and reagents), showed no significant improvements. The remaining method groups are all heterogeneous with respect to calibrant and in these significantly lower CVs were observed with SPS-01; the two exceptions were the RID groups for C4 and A1-AT, which may in part reflect their homogeneity with respect to procedure. Also, commercial calibrants for C4 are known to show little variation, and before SPS-01 the RID group for A1-AT included four calibrants only: in the same group for C3, with 7 different calibrants being used initially, a significant improvement with SPS-01 was observed (Table 13.2).

Thus the major potential for improvement appears, as might be expected, to be where there is greatest diversity in the methods and calibrants used.

13.2.2 Potential confounding factors

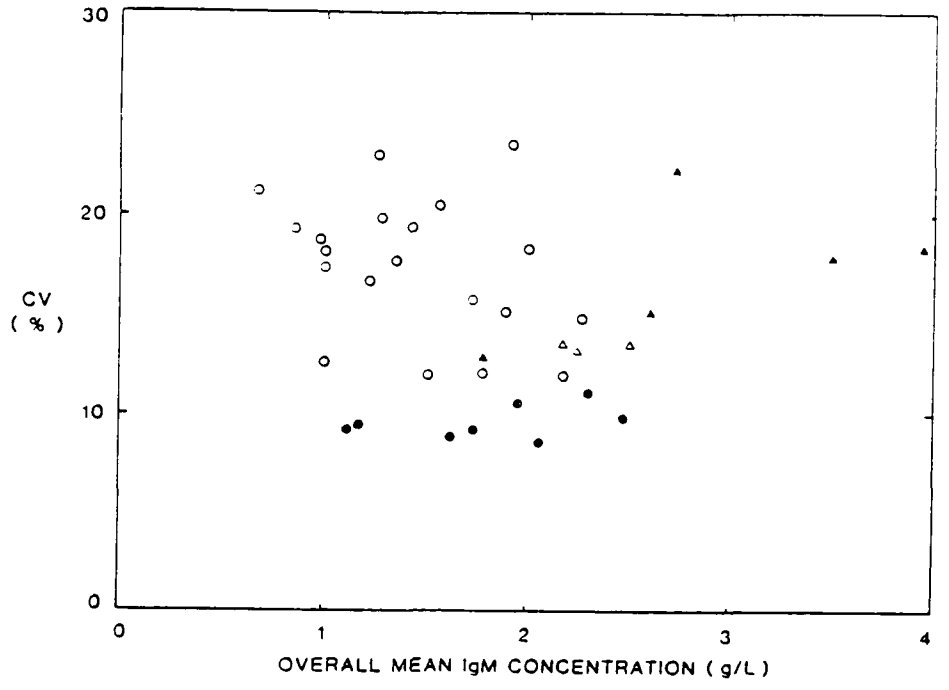
Could these improvements have been due to factors other than the use of a common calibrant? The scheme design allows such potential factors to be studied.

The lower CVs observed with SPS-01 cannot be ascribed to differences in the material distributed. All specimens before and with SPS-01 comprised pooled normal human serum, prepared in the same manner and examined by high resolution agarose gel electrophoresis to exclude protein deficiencies or the presence of abnormal proteins (e.g. paraproteins, high molecular weight immune complexes, rheumatoid factor).

Analyte concentrations may also influence between-laboratory agreement, as discussed in Chapter 6. Here, however, improvements cannot be ascribed to differences in analyte concentration since CVs were independent of concentration, as illustrated in Figure 13.2 for IgM and C3; patterns for C4 and A1-AT were similar to C3, and those for IgG and IgA to IgM. Also, no significant differences were observed apart from an increase for C4 (Table 13.3); this apparent change probably only reflects differences in assigned values between SPS-01 and other calibrants since the material distributed in both periods was normal human serum.

The average within-laboratory CVs (Appendix I.2.5; Chambers et al, 1984) for distributions with SPS-01 were in general lower than the corresponding CVs before but, apart from IgM, not significantly so. This improvement was unexpected, since within-laboratory precision should be independent of calibration procedure. It may, however, be related to the small effect that 'special treatment' of EQAS specimens may have on performance

Figure 13.2 Relationship between interlaboratory agreement (CV) and concentration for IgM and C3 in UKEQAS for Specific Proteins



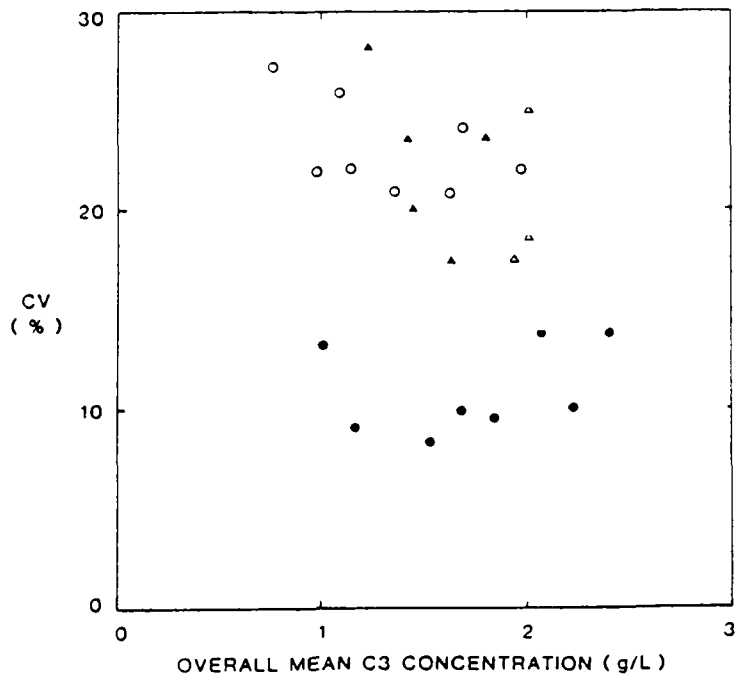


Table 13.3 Average and range of overall means before, with and after SPS-01. For explanation see Table 13.1

		Before	With	After
IgG		12.95	13.88	15.16
		5.7-20.0	8.6-18.7	10.7-19.2
	p		NS	NS
IgA		2.46	2.26	3.02
		1.4-4.8	1.3-3.3	2.6-3.4
	p		NS	0.02
IgM		1.45	1.81	2.67
		0.68-2.3	1.1-2.5	1.8-4.0
	p		NS	<0.001
C3		1.31	1.73	1.68
		0.76-2.0	1.0-2.4	1.2-2.0
	p		NS	NS
C4		0.32	0.55	0.44
		0.18-0.48	0.34-0.75	0.36-0.53
	p		0.01	0.02
A1-AT		2.29	2.16	3.05
		1.3-3.6	1.2-3.0	2.5-3.4
	p		NS	0.02

(Rumley and Roberts, 1984; Rowan et al, 1984), with the extra attention involved in the distributions including SPS-01 leading to better precision. Such marginal effects, however, are unlikely to have contributed significantly to the major improvements in interlaboratory agreement.

13.2.3 Period following SPS-01 study

Did this exercise have any educational effect on participants or manufacturers of calibration materials? Apparently so, since between-laboratory agreement for IgG, IgA and A1-AT remained significantly better after SPS-01 than before (Tables 13.1 and 13.2), reflected for IgG in Figure 13.1 by the continued lower CVs and decreased scatter during the period after SPS-01. This improvement was also maintained within the more heterogeneous method groups, particularly the 'other methods' group for A1-AT.

The CVs for C3, C4 and IgM initially reverted to levels that were not significantly different from those before SPS-01, thus the improvement with SPS-01 was not due to coincidental improvement in performance. This may have also been related to the introduction into the scheme of pathological material. IgM in particular (Figure 13.1) has shown marked method biases and significantly worse between-laboratory agreement for the majority of pathological samples, due apparently to positive interference by rheumatoid factor. However, performance for IgM and C3 (Figure 13.1) and for C4 now seems to be improving, perhaps reflecting a longer-term influence of this study on participants' routine calibration procedures and manufacturers' practices, and increasing sales of SPS-01.

13.3 Pregnancy oestrogens in urine

Studies (detailed in Appendix III.4.2) through the UKEQAS for

Urinary Pregnancy Oestrogens (Appendix I.2.4; Bullock and Wilde, 1985) illustrate two further important points regarding calibration. The first concerns the importance of commutability between specimens and calibrant, and the second the influence of between-laboratory agreement on the effectiveness of a common calibration procedure.

13.3.1 Commutability between specimens and calibrant

The initial study investigated the use of a common primary calibrant, two aqueous solutions of oestriol, to improve interlaboratory agreement of oestrogen assay (Bullock and Wilde, 1985). The results (Table 13.4) show no major improvement and in some cases a worsening of agreement. This effect might be attributable to differing responses of the several methods in use to varying oestrogen species: thus the 'unknown' specimen was pregnancy urine containing a mixture of oestrogens, whereas the calibrant consisted of a single species only. The hypothesis is supported by the known variation in response to different oestrogens, ie lack of perfect cross-reactivity, among methods (Wilde and Oakey, 1975).

Later studies therefore used a common secondary calibrant, for convenience another specimen prepared for use in the scheme. Table 13.4 also shows the results of other studies (Bullock and Wilde, 1985). In study B, where both specimens originated from the same set and therefore contained virtually identical mixtures of oestrogen, the improvement in agreement with 'calibration' was most impressive. This even stimulated suggestions that a common working calibrant be prepared for provision to UK laboratories to improve the then unsatisfactory situation (Oakey, 1980; Bullock and Wilde, 1985).

Table 13.4 Effect of 'calibration' on between-laboratory agreement for urinary pregnancy oestrogen assay. Calibrants were aqueous oestriol (Study A), or lyophilised urine from the same (Study B) or a different (Studies C and D) set of specimens

	n	Survey data	Recalculated
Study A (March 1981)			
Overall	106	14.9	12.9
Lever	43	9.4	11.4
Brombacher	41	10.3	9.6
RIA	12	11.3	7.5
Oakey	6	13.0	17.9
Study B (May 1981)			
Overall	111	17.4	8.4
Lever	47	12.0	6.4
Brombacher	41	14.4	9.4
RIA	13	17.8	7.3
Oakey	7	7.6	11.7
Study C (November 1982)			
Overall	103	9.7	7.0
Lever	57	6.5	4.5
Brombacher	31	8.3	6.5
RIA	10	15.1	6.6
Oakey	5	18.4	15.3
Study D (July 1986)			
Overall	54	20.2	21.0
Lever	37	19.9	24.5
Brombacher	10	10.0	9.1
Miscellaneous	6	29.0	22.6

Specimens from different sets, and hence probably with differing oestrogen composition, were used in study C (Table 13.4). Here the improvement in agreement was less marked, confirming the importance of similarity of composition between calibrant and clinical specimens. For this assay, however, there are certain to be differences among the clinical specimens and any calibrant must therefore represent a compromise.

13.3.2 Relationship to interlaboratory agreement

An additional factor bearing on the lesser improvement shown in study C is the improving state of the art over this period (see Figure 3.1). Thus the 'uncalibrated' (survey) CVs were lower in study C than in study B, providing correspondingly less scope for improvement. This factor is further demonstrated in the more recent study D (Table 13.4). Here calibration, again using specimens from different sets, failed to yield any improvement other than within the heterogeneous 'Miscellaneous' group.

As agreement improves the susceptibility to the confounding effects of errors such as specimen or result transposition becomes more marked. In general the need to improve agreement by means such as common calibration also becomes less urgent as agreement itself improves. The improvement (Figure 3.1) observed, which may have been related to participants' review of their calibration procedure following study B or of critical features in their method following the recommendation of a modified Lever procedure (Working Party on Urinary Pregnancy Oestrogens, 1981), thus obviated the need for production of a common calibration material for the UK.

13.4 Total urinary protein

This determination also concerns mixtures of analytes which differ from patient to patient. Methods should therefore show minimal differences in response to protein species (Dilena et al, 1983), both to yield reliable results on all specimens and to facilitate accurate calibration. Three UKEQAS surveys were conducted, with investigation of the effects of participants' own calibration procedures and of 'calibration' using another specimen (Appendix III.4.3) on agreement.

Table 13.5 summarises the survey and calibrated data, both overall and for the main method groups, for Survey 2 and 3; transposition and other errors by participants precluded reliable analysis of data from the first survey. Despite the very poor agreement in the survey data, these show negligible improvement within method groups and only a slight improvement overall. This finding was unexpected, and was not completely explicable in terms of transposition and similar errors.

Indeed the effects of the calibration material used routinely by participants are difficult to explain. For example, specimen 4 (Survey 2) comprised human albumin in normal urine yet even the subgroup of laboratories calibrating their sulphosalicylic acid (SSA) turbidimetric method against human albumin showed an overall over-recovery of about 100%; the similar specimen 6 (Survey 3) and specimen 1 (Survey 1; human albumin in saline) gave overestimations of 30% and 60%. The overestimation is somewhat greater than the 10% observed for urine from nephrotic patients (specimens 5 and 7 in Surveys 2 and 3), which contained almost exclusively albumin. Thus there appears to be some non-commutability between participants' calibrants and the specimens

Table 13.5 Effect of 'calibration' on between-laboratory agreement for urinary total protein assay in Surveys 2 and 3. The value for the reference specimen was 3.4 g/L in Survey 2, and 5.0 g/L in Survey 3; means in g/L, CV as %

		Survey 2				Survey 3			
	n	Survey		Recalculated		Survey		Recalculated	
		Mean	CV	Mean	CV	Mean	CV	Mean	CV
Overall	348	5.30	23.3	4.17	25.0	5.06	31.0	4.83	18.1
SSA	92	6.40	23.7	3.25	33.7	6.92	41.6	4.54	25.1
SSA/Na_2SO_4	47	4.91	16.7	4.30	18.3	4.78	19.9	4.85	8.1
TCA	29	4.83	13.6	5.21	9.1	4.00	18.9	4.96	7.8
Dye binding	34	4.73	19.7	4.34	9.3	4.75	17.7	4.98	6.8
Biuret	28	4.89	18.7	5.00	12.6	4.72	12.6	5.14	7.8
Direct biuret	6	5.56	20.1	4.91	15.4	6.61	24.0	5.66	11.6
Direct Coomassie	44	4.93	19.2	4.30	9.6	4.87	14.4	4.79	6.7
Benzethonium	7	4.95	12.9	4.51	7.2	4.35	7.5	4.44	6.4
Miscellaneous	7	4.97	24.0	4.57	10.8	4.78	21.6	4.86	8.6

distributed. This is not, however, necessarily a reason for rejecting the survey findings. If it is due to some ageing process (the specimens were prepared in Bristol, bottled in Sheffield, and finally distributed from Birmingham several weeks later) then susceptibility to ageing effects could still be important in the routine application of this assay.

13.5 Assays of enzyme activity in serum

Calibration studies were carried out for the enzymes shown in Table 13.6 in four UKEQAS Enzyme Surveys (Appendices I.2.2 and III.4.1; Bullock et al, 1986b). The objective was to obtain information on the applicability of calibration to these assays.

13.5.1 Calibration studies

The effects of recalculating results using one material as a 'calibrant' are exemplified in Table 13.7 for CK in Survey 17. Major improvements were seen both in within-method CVs and in the numerical agreement among the mean values for the various method groups. This is also demonstrated for AST, ALP and amylase in Figures 13.3 to 13.5. Table 13.7 also demonstrates that calibration can compensate for differences in factors such as activator (NAC or glutathione) and temperature (30^{o}C or 37^{o}C). Indeed, calibration could have particular application in CK assay, where reagent instability can cause significant day-to-day variations in accuracy. Table 13.6 demonstrates the greatly improved concordance in terms of overall agreement, irrespective of method, for all laboratories in Surveys 14-18, and the data for amylase are reviewed in Table 13.8.

At least part of the improved within-group agreement must be attributed to the elimination of variations, affecting the 'survey' and 'calibrant' specimens equally, in methodology,

Table 13.6 Overall statistics, irrespective of method, after 'calibration' (after exclusion of results more than 2 SD from the untrimmed mean) in UKEQAS Enzyme Surveys 14-18

		Survey					
		14	15	16	17	18	Average
AST	n			306	293	294	
	Mean (U/L)			99	164	98	
	SD (U/L)			7	20	7	
	CV			7.8%	12.3%	7.6%	9.2%
ALT	n				215	225	
	Mean (U/L)				120	46	
	SD (U/L)				11	8	
	CV				9.3%	16.6%	13.0%
LD	n			178	170	174	
	Mean (U/L)			1138	862	966	
	SD (U/L)			124	151	34	
	CV			10.9%	17.6%	3.6%	10.7%
CK	n		249	272	265	270	
	Mean (U/L)		225	279	520	175	
	SD (U/L)		23	45	53	32	
	CV		10.3%	16.3%	10.3%	18.3%	13.8%
ALP	n			299	286	302	
	Mean (U/L)			493	450	410	
	SD (U/L)			69	46	93	
	CV			14.0%	10.3%	22.7%	15.7%
Amylase	n	290	284	308	292	279	
	Mean (U/L)	548	397	525	547	609	
	SD (U/L)	54	32	56	57	41	
	CV	10.0%	8.2%	10.8%	10.5%	6.8%	9.3%

Overall weighted average (n=20) 11.7%

Table 13.7 Effect of 'calibration' for CK in Survey 17, October 1984. Each laboratory's result for the survey specimen was recalculated using the survey mean (582.0) for the reference specimen in conjunction with the laboratory's result for the reference specimen, ie using this material as a 'calibrant'. In each case results more than 2 SD from the untrimmed mean were excluded.

	Survey data for reference specimen	Survey specimen	
		Survey data	Recalculated
SCE NAC 37°C:			
n	149	153	151
Mean (U/L)	582.0	526.6	523.4
SD (U/L)	47.9	59.1	53.2
CV	8.2%	11.2%	10.2%
Glutathione 37°C:			
n		23	21
Mean (U/L)		470.1	520.1
SD (U/L)		100.6	33.2
CV		21.4%	6.4%
ACB/DGKC NAC 30°C:			
n		16	17
Mean (U/L)		351.2	504.1
SD (U/L)		44.1	41.4
CV		12.6%	8.2%
Other:			
n		68	65
Mean (U/L)		379.9	522.4
SD (U/L)		151.2	40.7
CV		39.8%	7.8%

Figure 13.3 Effect of 'calibration' for AST in UKEQAS Enzyme Survey 17, October 1984. The points and solid bars around the periphery represent the survey mean ±1 SD for the method groups; for each method group these are connected to a point and solid bar towards the centre which represent the mean ±1 SD for the 'calibrated' data. The star and dashed bar at the centre represent the survey mean ±1 SD for the SCE optimised 37°C method group as 'target'

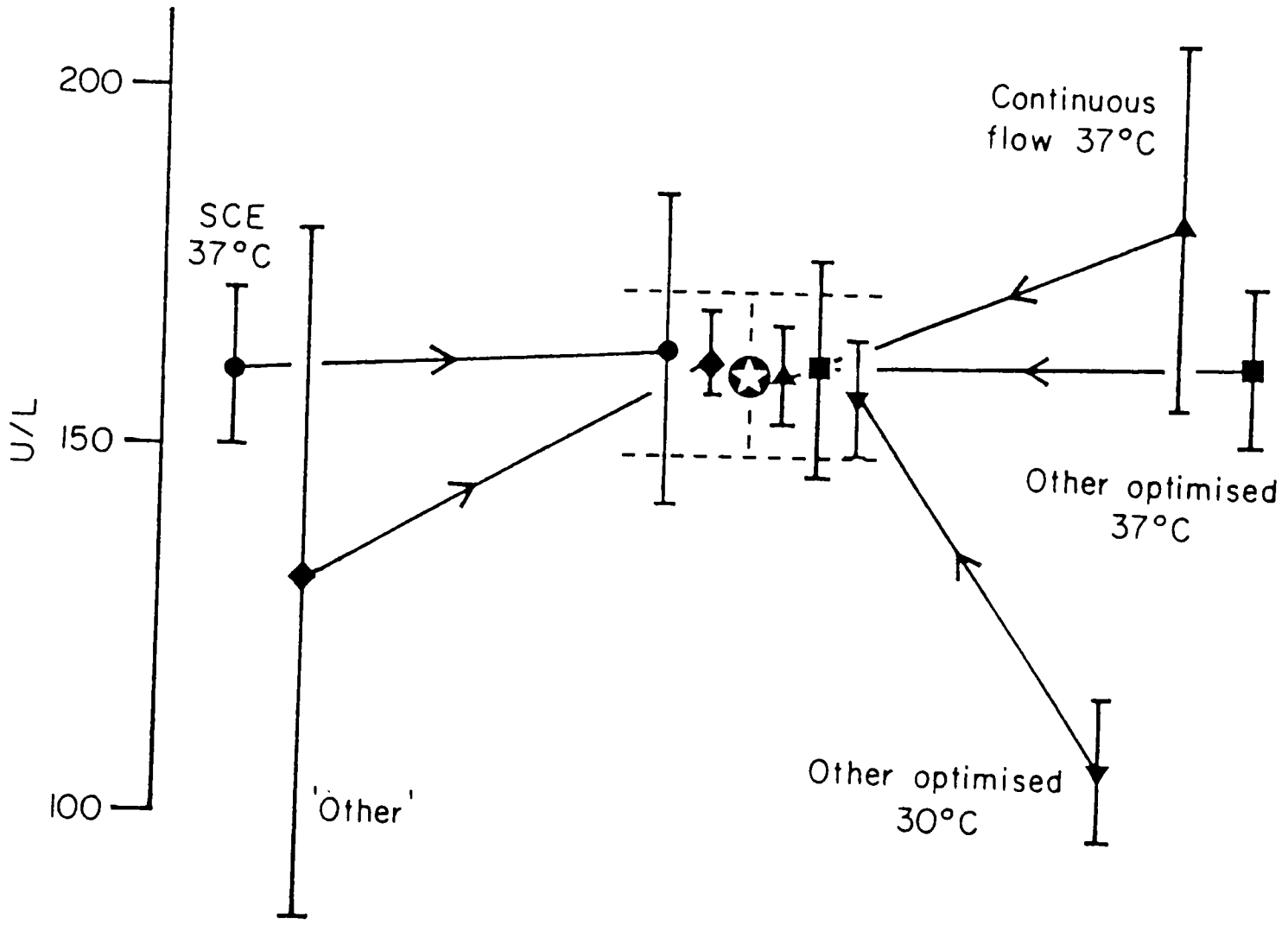


Figure 13.4 Effect of 'calibration' for ALP in UKEQAS Enzyme Survey 17, October 1984. See Figure 13.3 for explanation

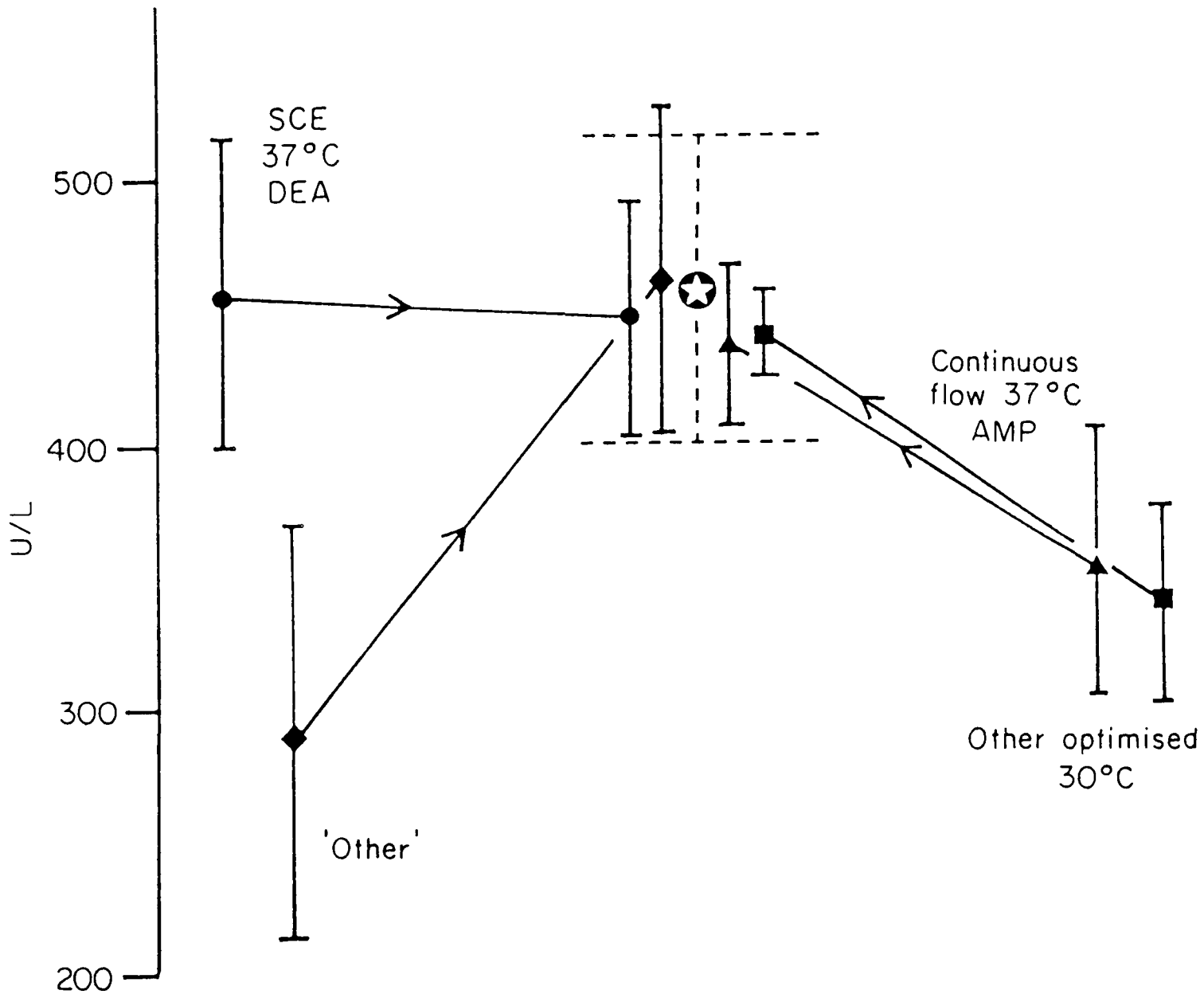


Figure 13.5 Effect of 'calibration' for amylase in UKEQAS Enzyme Survey 17, October 1984. See Figure 13.3 for explanation (Phadebas 37°C method group as 'target')

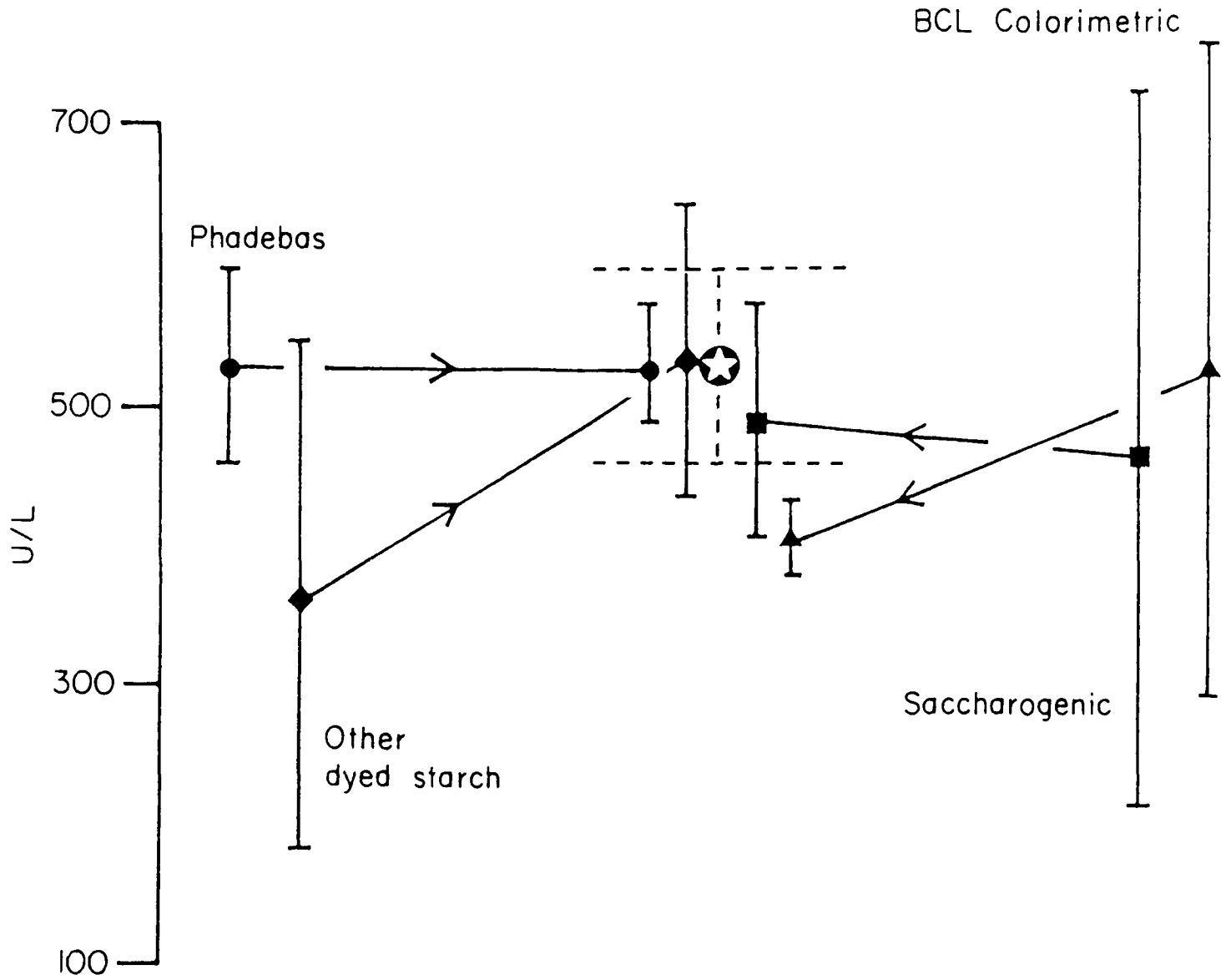


Table 13.8 Effect of 'calibration' for amylase in Surveys 14-18
All figures are CV (%); see Table 13.7 for explanation

	Phadebas (n = 238 - 252)		Other method (n = 41 - 56)	
	Reference	**Survey**	**Reference**	**Survey**
Survey 14				
Survey data	11.8	11.9	48.1	47.6
Recalculated		7.5		16.5
Survey 15				
Survey data	10.8	10.1	70.9	58.4
Recalculated		6.0		16.2
Survey 16				
Survey data	12.2	13.1	67.2	58.8
Recalculated		8.2		18.6
Survey 17				
Survey data	10.8	9.7	63.8	58.4
Recalculated		8.3		18.5
Survey 18				
Survey data	14.0	12.5	61.8	58.8
Recalculated		4.7		8.8

reagents and measurement conditions. These results are in general agreement with those of studies in the Netherlands (Jansen and Jansen, 1983) in which, as here, the major effect of calibration was seen for non-optimised methods. This is consistent with the observation above for specific protein assays that the greater the heterogeneity within a group the greater is the effect of common calibration.

13.5.2 Commutability considerations - amylase assay

Though almost all amylase assays in the UK are at present carried out by the Phadebas dyed substrate method, agreement between laboratories remains poor: the method is manual and depends upon a predetermined calibration curve. Here, the use of a common calibrant produced the expected improvement in comparability, from an average CV of 11.5% to 6.9% for the Phadebas method group in Surveys 14-18.

If only a single method is being calibrated the problem of isoenzyme bias (manifested in non-commutability) does not arise. Attempts to use a single calibrant to provide comparable results from a variety of amylase methods would, however, have to address the problem of differences between enzymes of human and animal origin (Bretaudiere et al, 1981b) and between the major isoenzymes of human amylase in their relative activities on long-chain and short-chain saccharides.

The effect of this latter difference is seen in the results for the short-chain substrate BCL colorimetric method in Survey 17 (Figure 13.5). Other differences, such as those represented within the "Other method or temperature" group, appear less important in this respect, as can be seen from the improvements

brought about by 'calibration'. The extent of improvement within this heterogeneous group varied from survey to survey (Table 13.8), reflecting the magnitude of differences in isoenzyme composition between the survey and reference specimens. For example, in Survey 14 both were human sera enhanced with porcine amylase.

Since in many cases between-laboratory concordance can be improved by 'calibration', the possibility arises that enzyme reference materials could be used to transfer values from one method (procedure and instrumentation) to another, ie for the calibration of assays. However, such a transfer is valid only within carefully defined conditions, and when methods similar in principle are considered. Thus the materials must be commutable (Fasce et al, 1973; Broughton et al, 1981; Moss et al, 1985) among the methods used, with no specimen-to-specimen variability in the relationship between the methods. For example, variations in isoenzyme composition preclude transfer between methods for amylase which differ in substrate chain-length, eg starch and oligosaccharide (Gerhardt et al, 1985). Similarly, it is not possible to transfer values between methods for AST or ALT which differ in whether pyridoxal phosphate is present (Jansen, 1985), because of serum-to-serum variations in coenzyme content. Any link between methods must be made by assaying clinical specimens by both methods, and checking that use of the reference material as a calibrant leads to full recovery of the activities for the specimens.

These and other (eg Jansen and Jansen, 1983; Gerhardt et al, 1985) calibration studies have demonstrated the potential benefit of using a calibration material in improving agreement for enzyme

activity assays. Bowers and McComb have also recently proposed a similar reference system based on IFCC reference methods (Bowers and McComb, 1984; McComb and Bowers, 1985), though their studies on EQAS data dealt only with method means rather than results from individual laboratories.

It now seems appropriate therefore to consider changing the direction of standardisation efforts in clinical enzymology, to consideration of the use of enzyme calibration materials. The introduction of suitable calibrants (Moss et al, 1985), in parallel with the adoption of reliable analytical procedures, could remove the remaining disagreements caused by the existence of several different recommended methods. This should also allow the results from any analytical system to be made comparable with those given by recommended or reference methods. Even more importantly, it could end the present sterile controversy over the choice of an agreed temperature for enzyme measurement, and be applied internationally. Such developments will, however, require critical evaluation, as discussed by Moss et al (1985) and Colinet et al (1986), and must not be used in an attempt to compensate for deficiencies in the methods used.

13.6 Summary

EQA surveys can be used to study or mimic the effects on interlaboratory agreement of using a common calibration material.

In most circumstances a substantial improvement can be demonstrated, with the extent of the improvement being directly related to the diversity in methods and calibration procedures.

Where the analytical procedures have differing specificities, any differences in properties (ie lack of commutability) between

the calibrant and clinical specimens reduce and may even negate such improvements in agreement.

In the case of enzyme activity estimations the use of suitable calibration materials, preferably but not necessarily in combination with standardisation of methods, should lead to improved numerical concordance between the results obtained in different laboratories.

Chapter 14:

THE EFFECTS OF SPECIES OF ORIGIN AND MANUFACTURING TECHNIQUE ON QUALITY CONTROL MATERIAL BEHAVIOUR

14.1 Introduction

A wide range of serum-based materials is available commercially and these products are used extensively for IQC and calibration in clinical chemistry. The choice of and quality of such materials are obviously of profound importance for the reliability and consistency of laboratory results, and a number of questions about these aspects require further study. As discussed in Chapter 3, these include:

- the species of origin of the base serum
- the production procedures used by material manufacturers
- the presentation of the material

Though it is stated both that materials based on human serum are essential for the calibration and day-to-day IQC of laboratory methods, and that animal-based materials are equally satisfactory for most inorganic and organic analytes, the issues have never been studied fully. There is thus no basis for consensus, and the reasons for selection are varied and often arbitrary since there is virtually no published information on the relative merits of these types of material. Studies within a single laboratory will be prone to error due to statistical artefacts and to the individual methods and QC materials used, whereas interlaboratory studies using a range of materials provide the basis for a more satisfactory investigation of the subject, and hence a more rational approach to material selection.

The World Health Assembly has tried to stimulate member states to become self-sufficient with regard to human blood and blood products, including QCMs. Several countries, including the UK, have made steps in this direction but it would appear that for developed countries the present demand for human blood products and QCMs cannot be met fully from blood collected within the country, and priority is rightly given to therapeutic uses. In contrast the supply of animal serum is effectively inexhaustable, which has been one of the primary reasons for encouraging its use in developing countries (Kenny and Eaton, 1981). Use of animal products should also eliminate the potential for transmission of hepatitis B (one study found that all of 22 commercial human-based QCMs were positive for one or more hepatitis virus markers; Compton et al, 1979) or acquired immunodeficiency syndrome (AIDS), though both manufacturers and health authorities have reaffirmed that it is good laboratory practice to treat all QCMs as if they were specimens from patients with potentially transmissible diseases. A demonstration that the behaviour of animal-based QCMs is as satisfactory for most common analytes as that of those derived from human sources would also have beneficial financial consequences for health care.

The serum matrix in QCMs may be considered to provide the analyte under consideration, the physicochemical milieu and a variety of interferences both defined and undefined. The potential effects of the matrix on QCM behaviour are discussed in Chapter 12, and the concept of commutability in Chapters 3, 12 and 13. When lyophilised material is considered, more modification of the physicochemical properties (and interferences) may be introduced by the freeze-drying process and by other aspects of the production procedure (eg addition of buffers or stabilisers) than

exist for example between sera from different species. Such procedures can be studied using data from EQASs which distribute a variety of materials through classification according to production method, which effectively reduces to consideration of materials grouped by manufacturer. Such matrix effects are most likely to have an influence through 'interference' effects, which are often subtle and undefined. As in the study of the influence of the species of origin, here one must rely heavily on empirical evidence gathered through EQA rather than theory (eg Stamm, 1979).

Since they were first described by Maurukas (1973), interest in serum-based QCMs incorporating ethylene glycol has been stimulated over recent years by the commercial availability of such materials. These may be advantageous because they are said to remain liquid even at usual freezer temperatures (ie -20^{o}C), and hence the degradative processes associated with freezing and thawing are avoided. Furthermore they would offer convenience and economy in use since they can be stored in the refrigerator for more prolonged periods than reconstituted lyophilised materials due to an antimicrobial action of the ethylene glycol. These factors are particularly important in developing countries, as discussed in section 3.3.2. One batch of such material was distributed in the UKEQAS for General Clinical Chemistry and a UKEQAS Enzyme Survey to assess the effects on interlaboratory agreement. This also provided a check on the efficacy of the protocol used (Bullock et al, 1979 and 1986b) in minimising the influence of post-reconstitution changes in enzyme activity.

14.2 Study of the effects of species and of material manufacturer

The objective was to assess the quality of the currently

available quality control materials for clinical chemistry by investigating the effects of the manufacturer and of the base serum's species of origin on interlaboratory agreement in the UKEQAS for General Clinical Chemistry. Since the scatter in results from participants depends upon the quality of the material distributed as well as upon the performance of individual laboratories, such questions are eminently suited to evaluation using EQA data. The 'state of the art' and its relation to factors such as analyte level must first be characterised, however, as described in Chapter 6.

Initial misgivings of the UKEQAS organisers regarding the use of such data to assess with any confidence the quality of the specimen distributed had been overcome by experience with more than 250 distributions over 13 years. Consensus values (both overall and method means) and interlaboratory coefficients of variation had proved reproducible on repeated distribution of the same (stable) material (see section 5.2.1), and CVs appeared similar for similar materials at similar analyte levels. With the relatively large number of participants (about 400 results for each distribution) there had appeared to be little fluctuation from distribution to distribution in the spread of results provided that materials distributed were of satisfactory quality. The scatter in results is indeed examined for each distribution before the results are used for assessment of laboratory performance.

Using a similar approach, assessments of QC material quality (Jansen et al, 1978; Jansen, 1980) and interactions between QC materials and analytical methods (Jansen et al, 1981) on the basis of data from smaller EQASs have been reported, and a

systematic study of UKEQAS data appeared justified. A two-year period, including 40 distributions, should have yielded a sufficiently large database for assessment without excessive complications from improvement in participants' performance. It should be noted that, though a wide range of manufacturers and types of product are covered, the materials distributed through this scheme are not fully representative of those materials used in the UK in two ways. Not all manufacturers are represented, and materials with assigned values and those intended for calibration purposes are distributed only infrequently.

The details of the studies are given in Appendix III.2, and follow the same pattern as the prior examination of relationships between analyte level and interlaboratory agreement described in Chapter 6. The basic design comprised the examination of data from a two-year study period, the conclusions drawn being reexamined in a further two-year validation period; validation was essential because previous work (Wilding et al, 1979) had demonstrated the importance of testing findings on an independent set of data. As before, (Appendix III.2.1) materials with discrepant and apparently unsatisfactory agreement were excluded from consideration for the relevant analytes. There was a uniform method classification throughout each period, though the classification was different in the study and validation periods.

14.3 The effects of species of origin

The reduced range of indicators (Table 6.1; see section 6.2.1 and Appendix III.2.2) were examined with respect to the species of origin of the base serum. The materials studied (Table 14.1) originated almost equally from human and bovine serum, with rather fewer equine sera, spread throughout the period.

Table 14.1 Classification of materials studied by species of origin and manufacturer

Code	Identification	Study period (n = 37)	Validation period (n = 40)
Species of origin:			
A	Animal (unspecified)	1	2
B	Bovine	13	18
E	Equine	7	9
H	Human	16	11
Manufacturer:			
a	Wellcome Diagnostics	6	4
b	Purce Associates/Rosslab	5	8
c	Scottish Blood Transfusion Service	5	1
d	Roche Diagnostica	4	8
e	Ortho Diagnostics	4	9
f	Technicon Instruments	4	6
g	Nyegaard	3	1
h	Boehringer Mannheim	1	-
k	Tissue Culture Services	1	-
m	Hyland Division, Travenol Laboratories	1	-
n	General Diagnostics	1	-
o	DADE Division, American Hospital Supply	1	-
p	Gibco	1	2
q	Biotrol	-	1

14.3.1 Examination of overall data

In most cases no consistent difference in performance between the types of material could be discerned, but in others two or three of the graphs of the percentage of results excluded or of VISs >200, >300 or 400 suggested a possible difference in behaviour. Thus, for example, performance for bovine-based sera appeared worse for urate and better for urea.

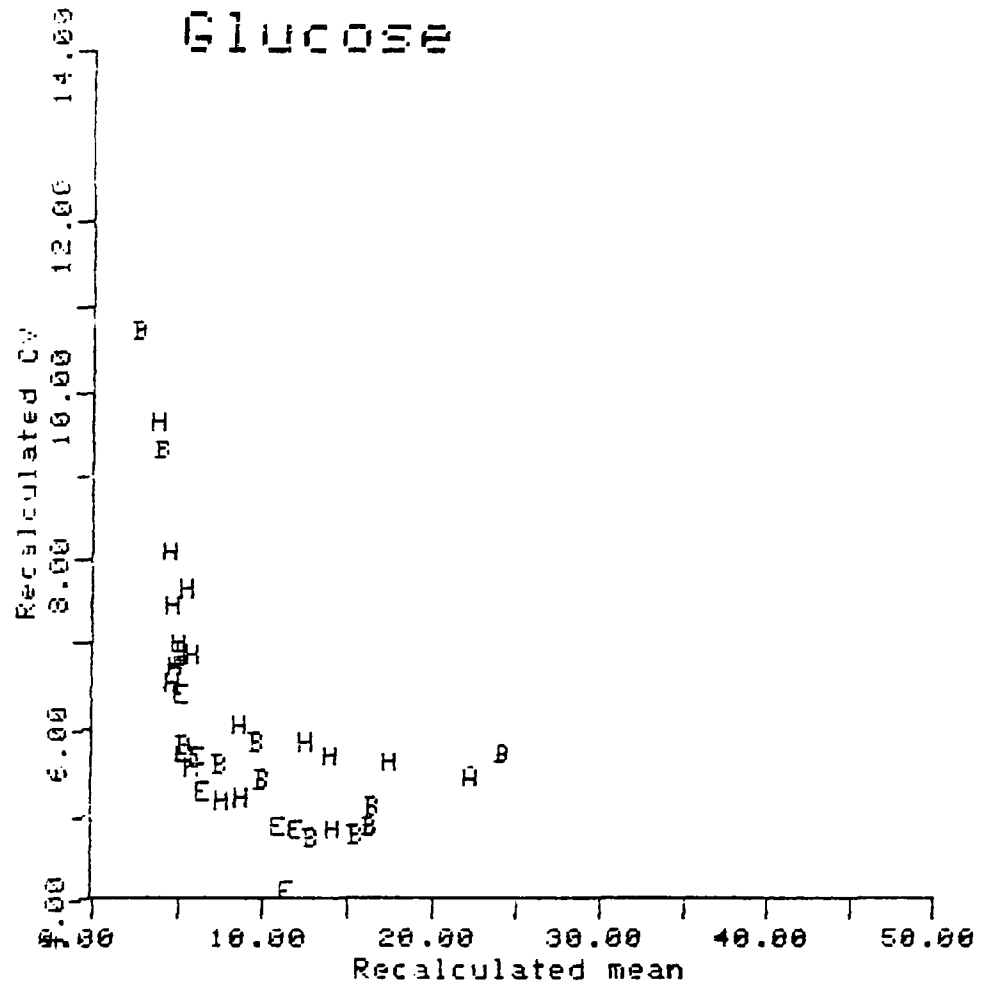
A more consistent picture, with apparent differences discernible in recalculated CV and average VIS, emerged for other analytes. For glucose, phosphate, bilirubin and total protein worse performance was seen for materials with human base serum, whereas performance for iron was worse for bovine-based sera. Figure 14.1 illustrates these effects for glucose and total protein. In no case was the situation clear-cut, however, and the differences were not supported by the graphs of all the indicators of discrepant performance.

Materials of bovine origin formed a greater proportion of those distributed during the validation period, with similar numbers of human and equine sera (Table 14.1). The poorer agreement for human-based materials was confirmed for glucose, phosphate, bilirubin (Figure 14.2) and total protein, and was suggested for cholesterol. The apparently worse agreement for iron in materials based on bovine serum was not confirmed (Figure 14.3).

14.3.2 Examination of method-related data

This confirmed in general the impressions gained from examination of the overall data. For example, Figure 14.4 shows the expected worse agreement for non-human sera for albumin by one of the BCG method groups and Figure 14.5 the exaggeration of this effect for

Figure 14.1 Relationship with recalculated CV for glucose and with average VIS for total protein, classified by species of origin



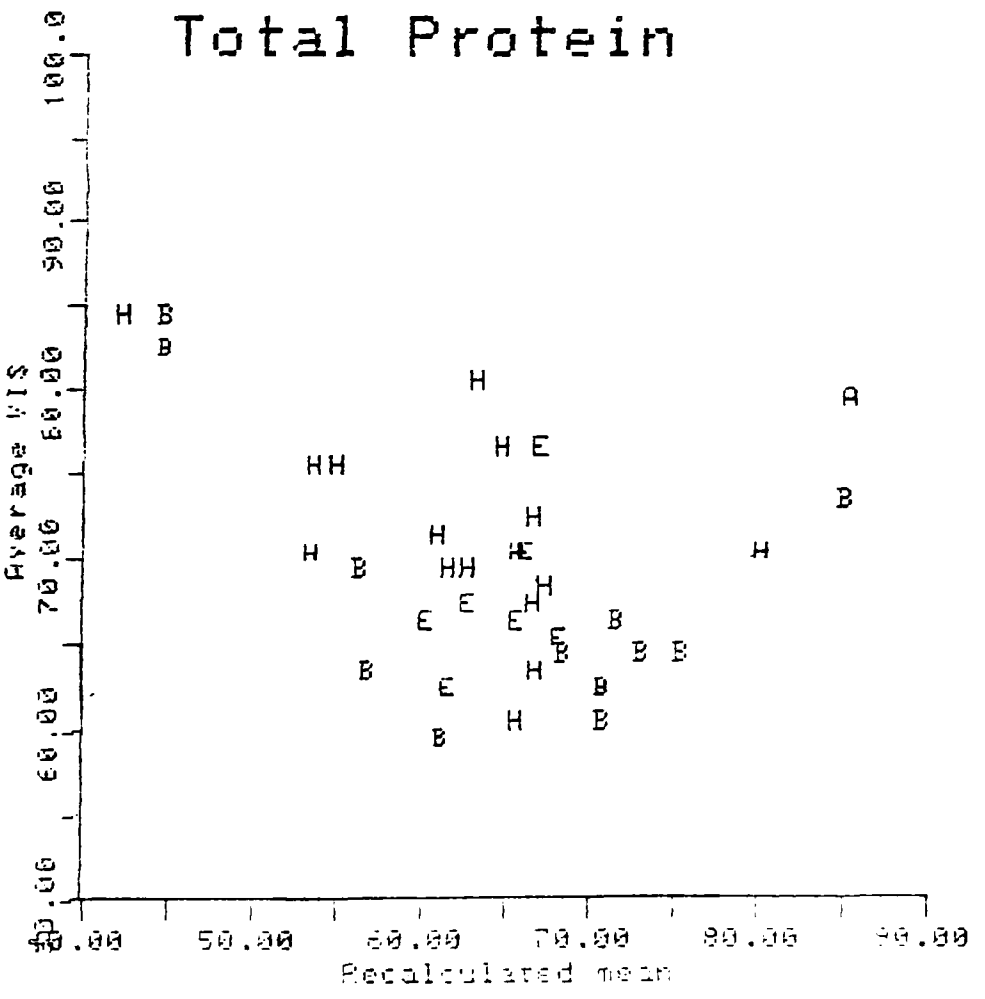
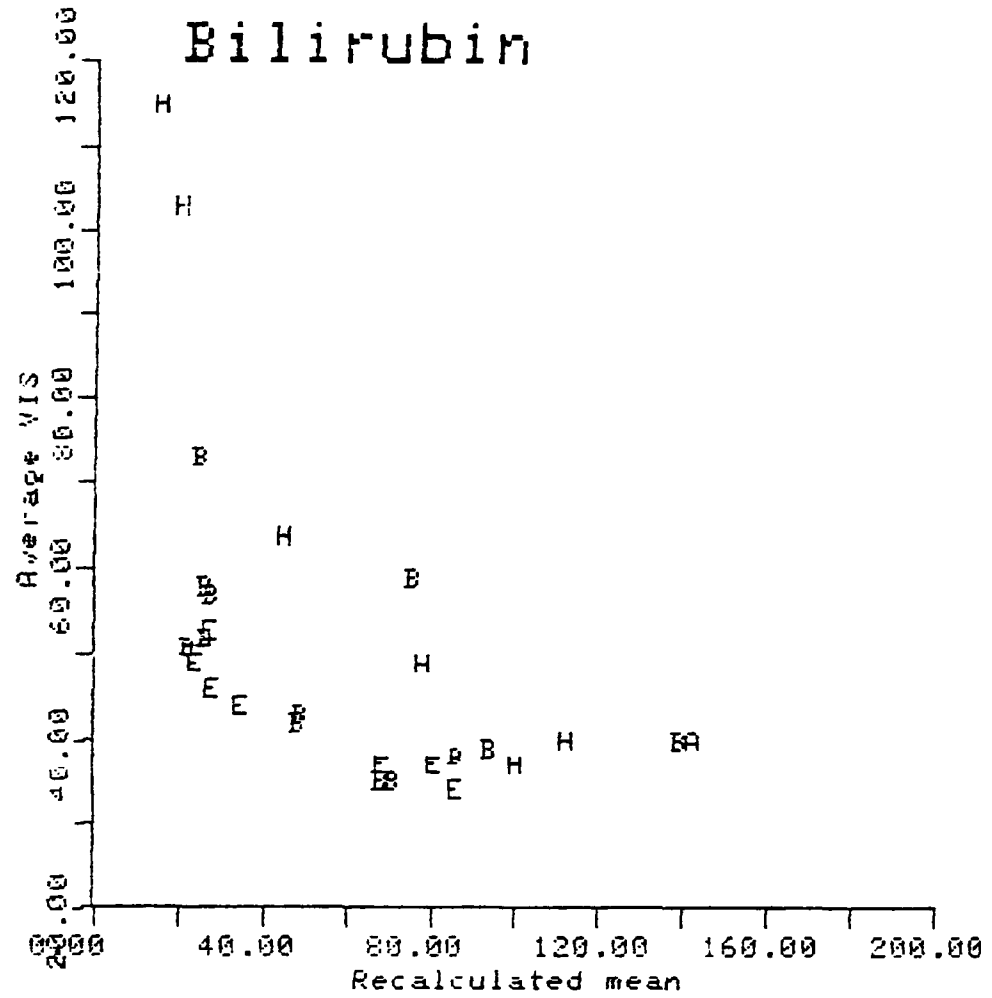


Figure 14.2 Relationship with average VIS for bilirubin, classified by species of origin, in study and validation periods

Study:



Validation:

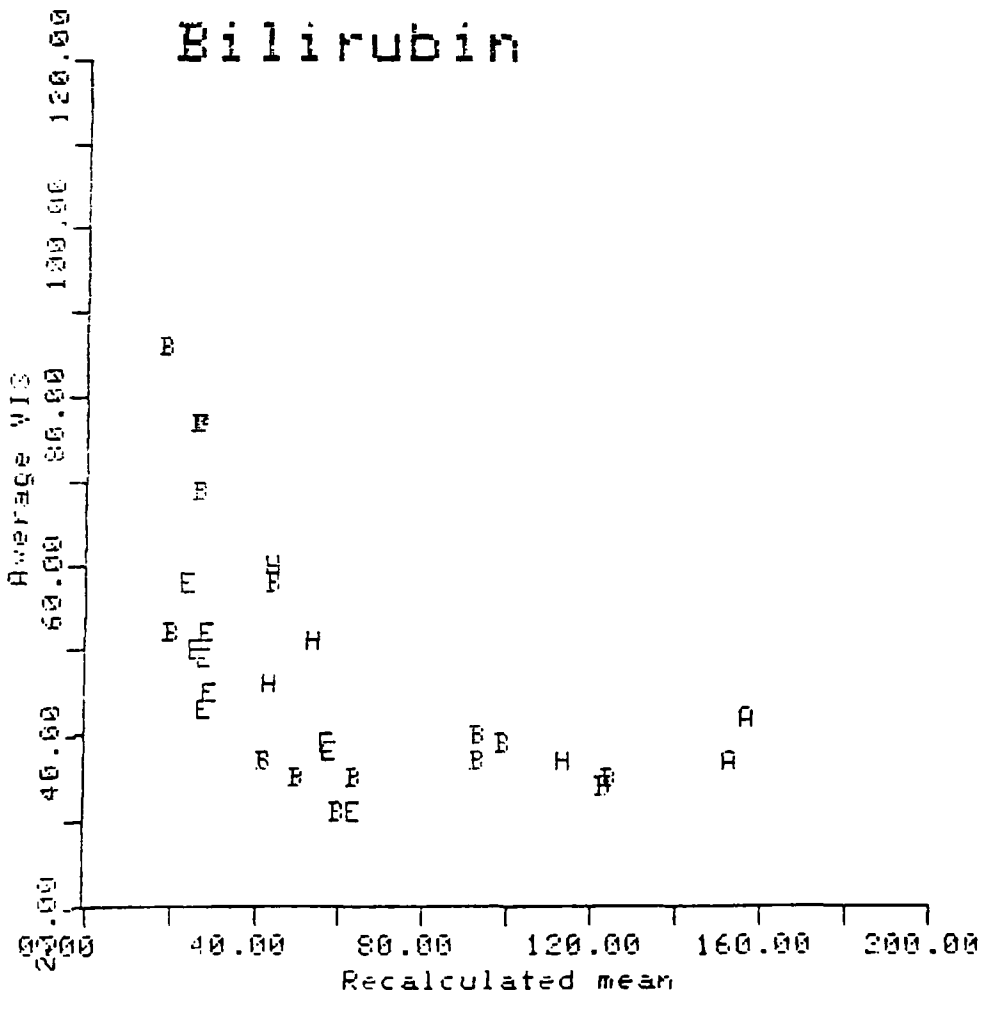
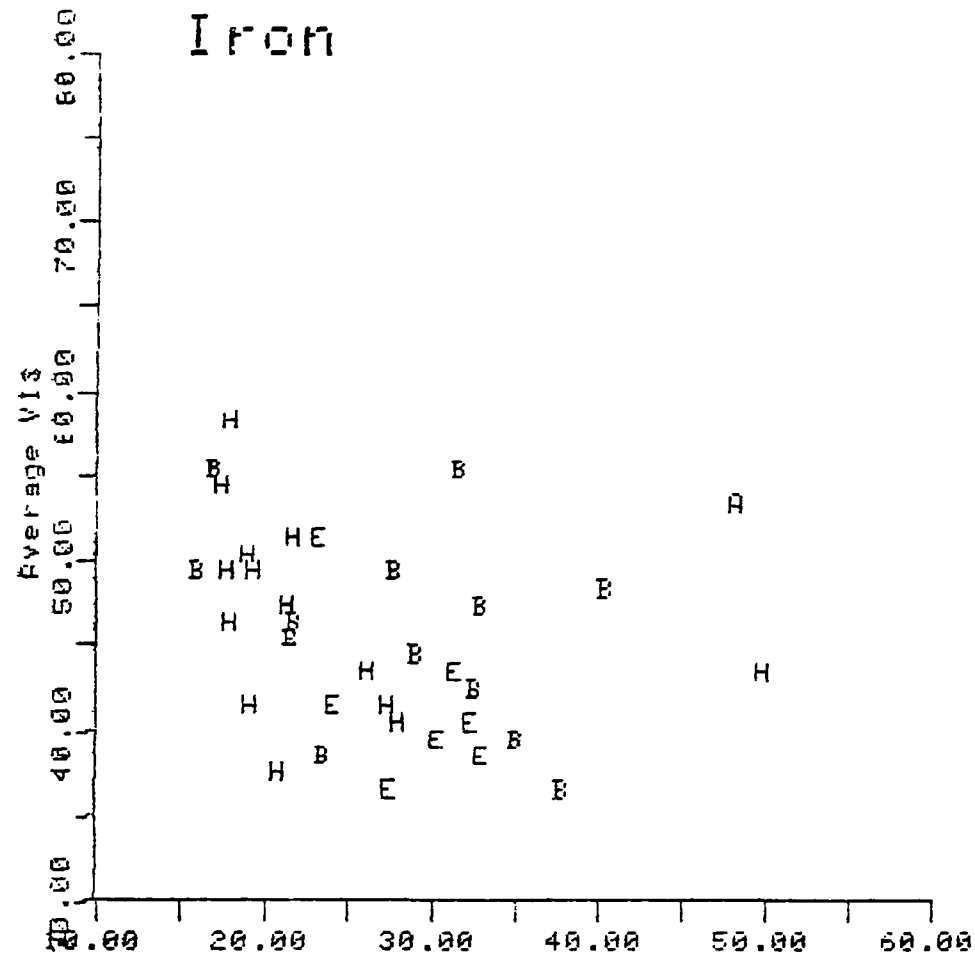


Figure 14.3 Relationship with average VIS for iron, classified by species of origin, in study and validation periods

Study:



Validation:

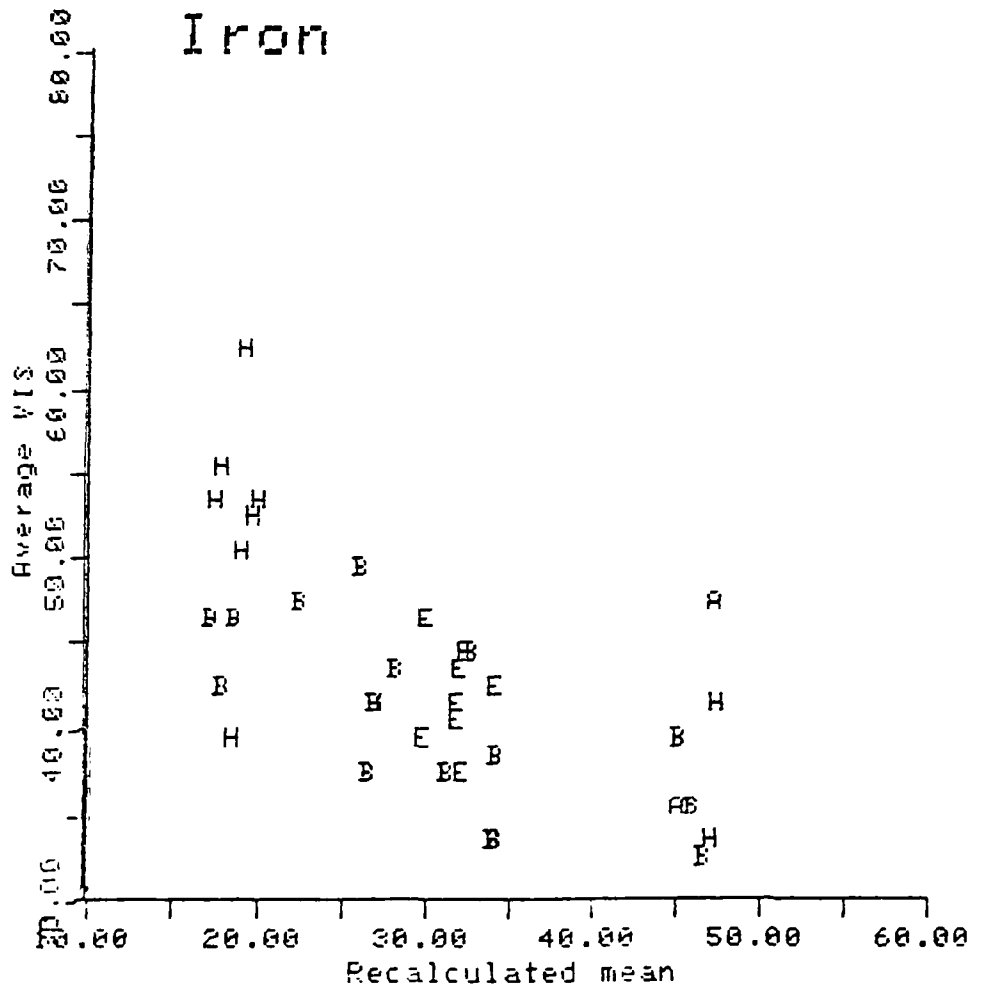
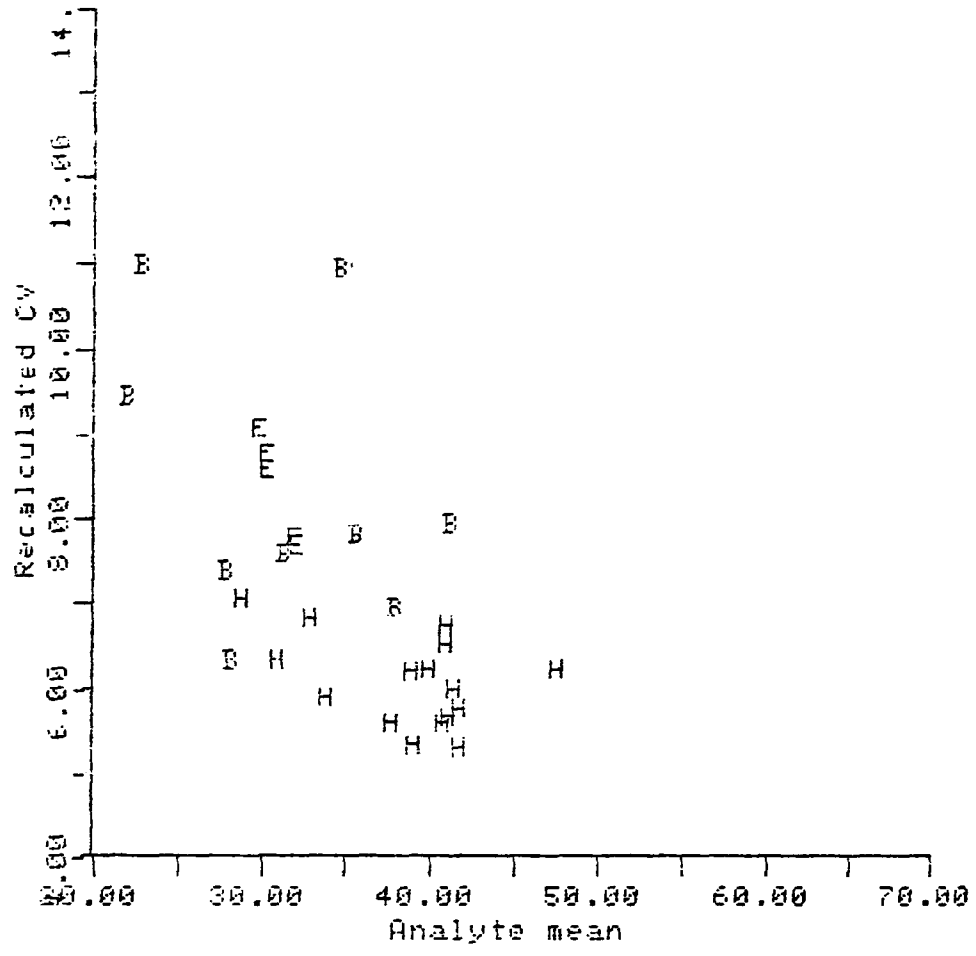


Figure 14.4 Relationship with recalculated CV for albumin by Manual BCG, classified by species of origin

Manual BCG:



AAI BCG:

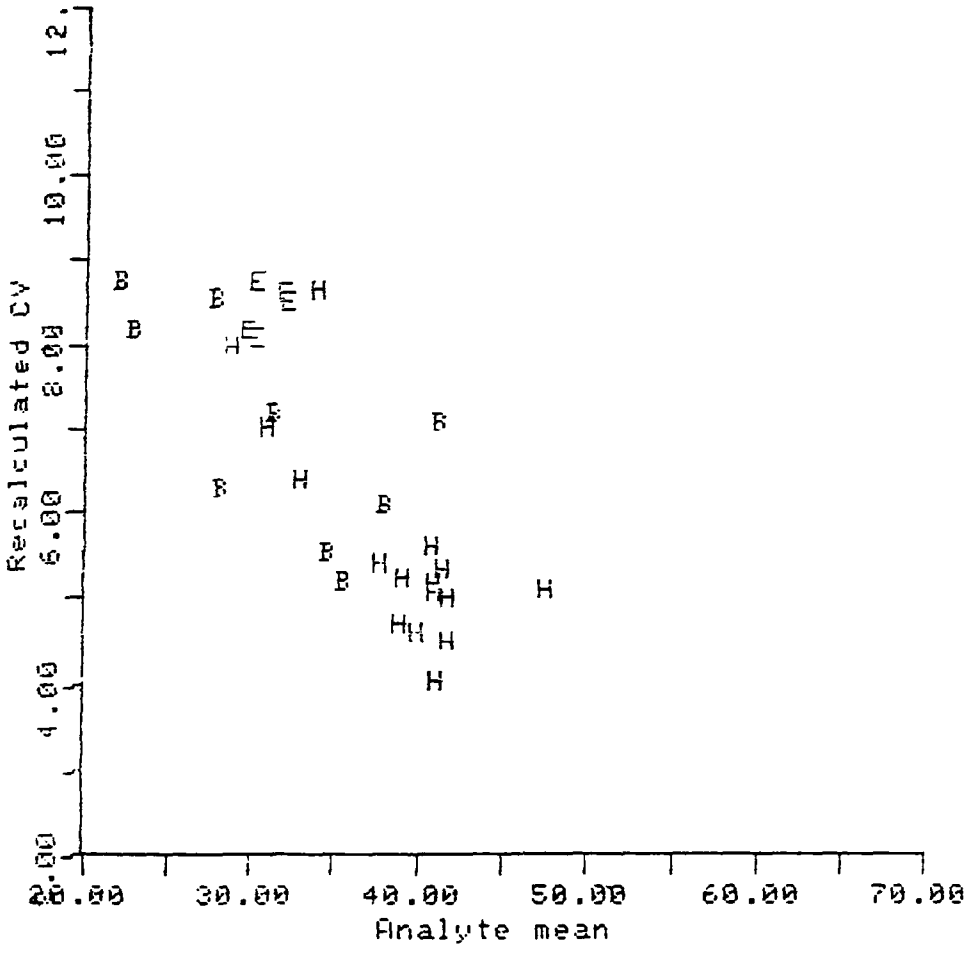
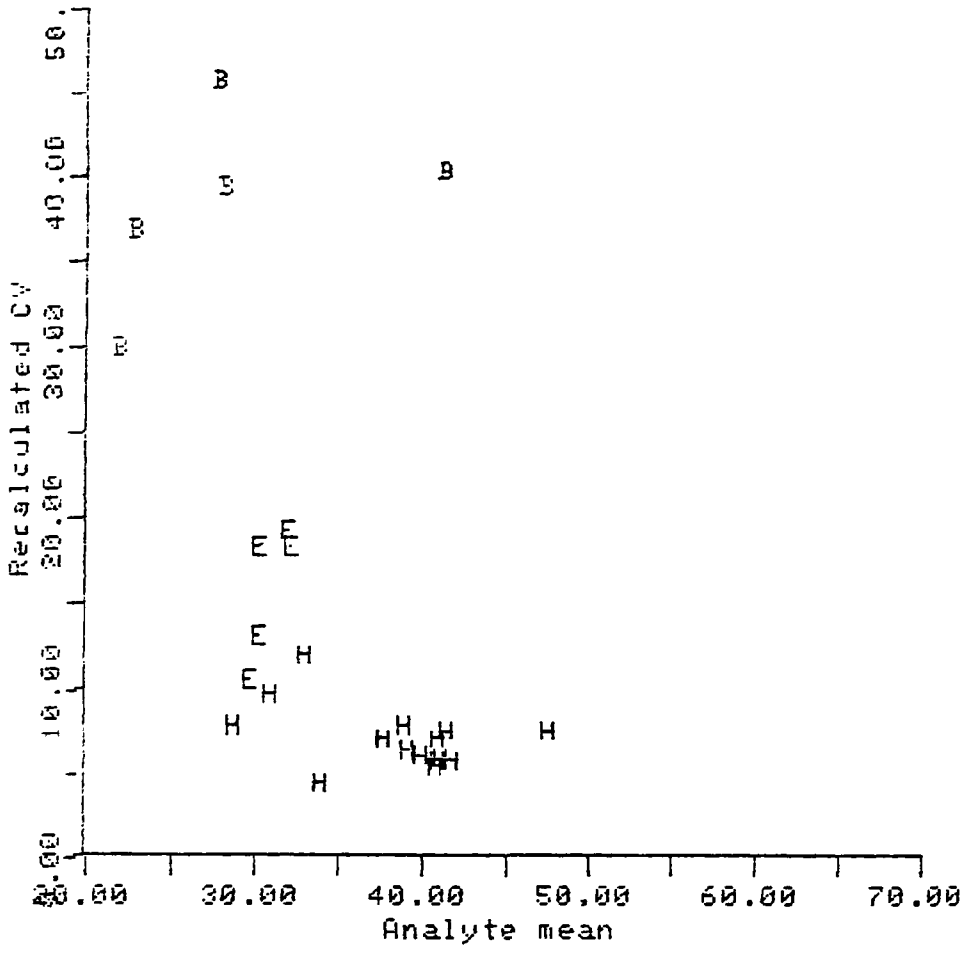
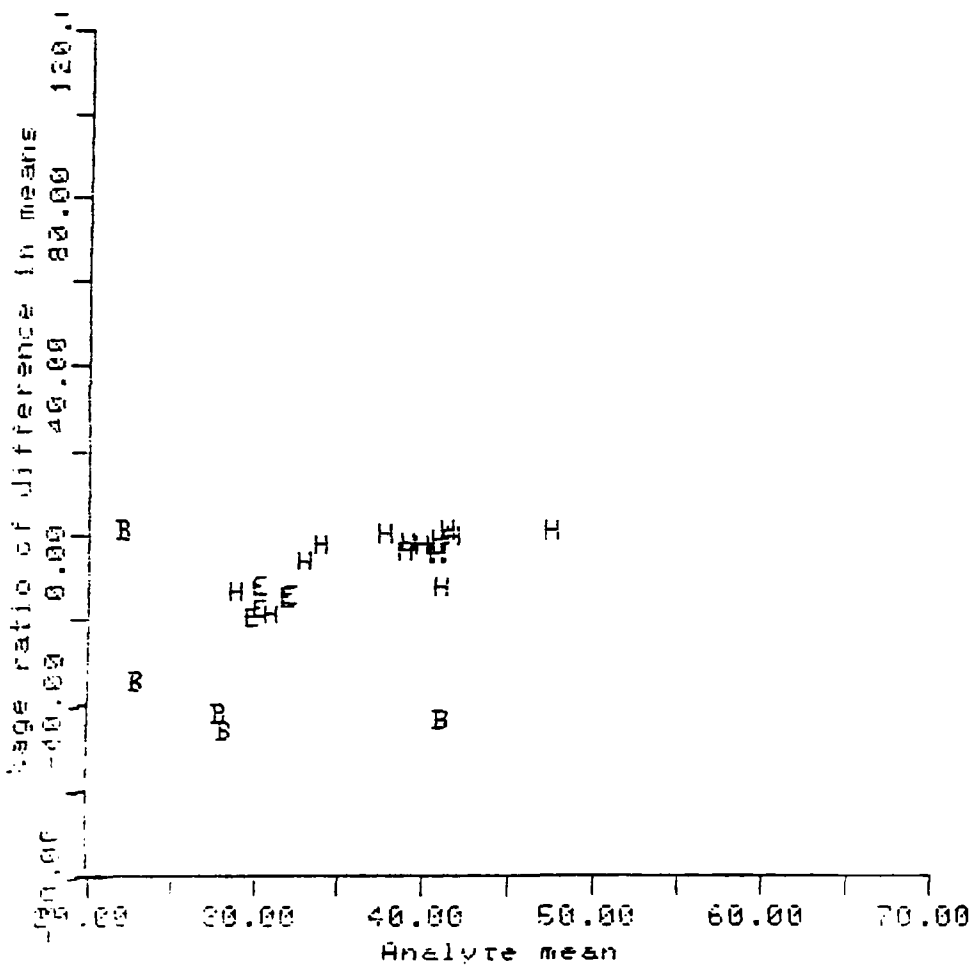


Figure 14.5 Relationship with recalculated CV and percentage difference from recalculated mean for albumin by BCP, classified by species of origin





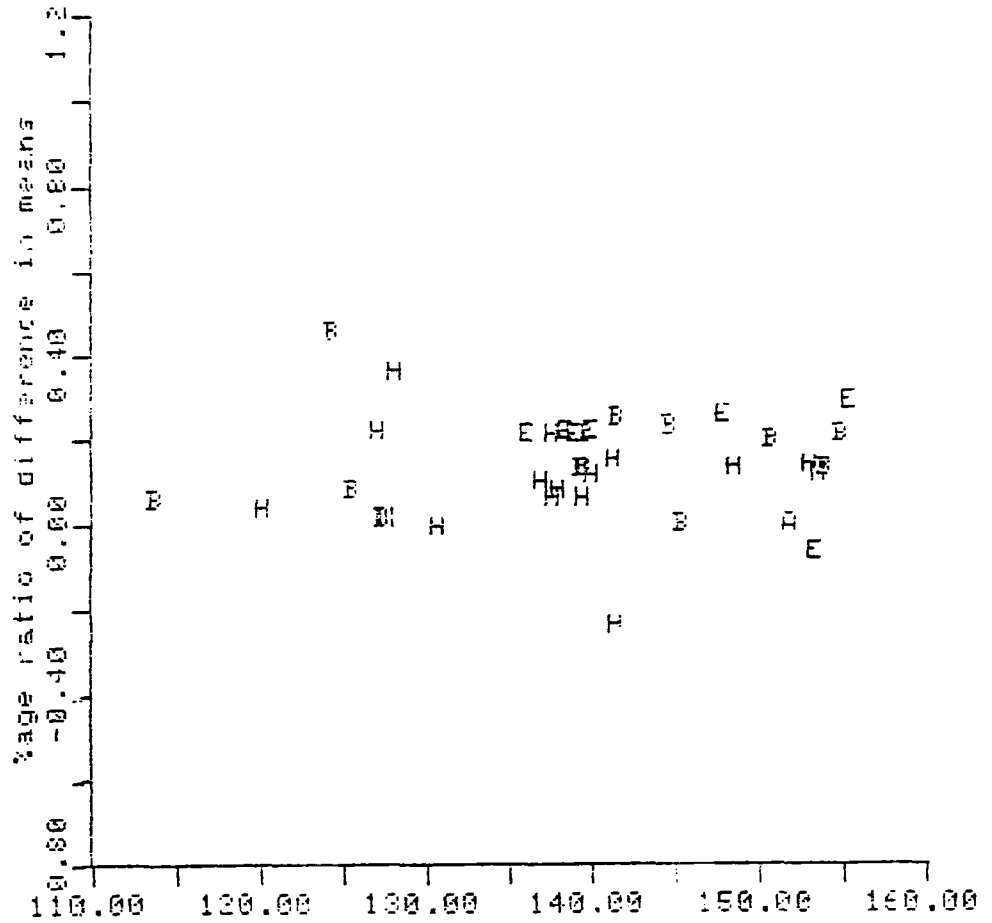
the more specific BCP method. Figure 14.5 also confirms the much lower affinity of BCP for bovine albumin.

For sodium, differences in bias relative to the overall recalculated mean were seen. There was a consistent difference at all concentrations for the Corning 430/450 flame photometer group but apparent negative bias at low concentrations and positive bias at high concentrations for the Ion-selective electrode group (Figure 14.6); these were independent of the species of origin of the base serum. The latter pattern was again seen for the Indirect ion-selective electrode method grouping in the validation period (Figure 14.7). Comparison of this pattern with that for the Direct ion-selective electrode group yields a definite difference in behaviour, and also an apparently different response of the direct-reading instruments to materials based on human serum (Figure 14.7). This may reflect the lack of dilution prior to measurement in this group, giving greater susceptibility to matrix effects.

Individual methods occasionally suggested differences in performance in the study or validation period data. Thus human-based sera appeared to give better performance for calcium by Corning titrator but worse performance for phosphate by manual methods, and bovine sera to give worse performance for iron by Bathophenanthroline and a negative bias for iron by AAII/SMA procedures. The overall apparent inferiority of human-based materials was not seen for either of the YSI or Beckman Glucose Analyzer groups, but was noted for all the other three GOD method groups. The differences in performance between materials derived from different species were largely confirmed in the validation period, as shown in Figure 14.8 for phosphate by Manual and

Figure 14.6 Relationship with percentage difference from recalculated mean for sodium by Corning-EEL 430/450 and Ion-selective electrode, classified by species of origin

EEL 430/450:



Ion-selective electrode:

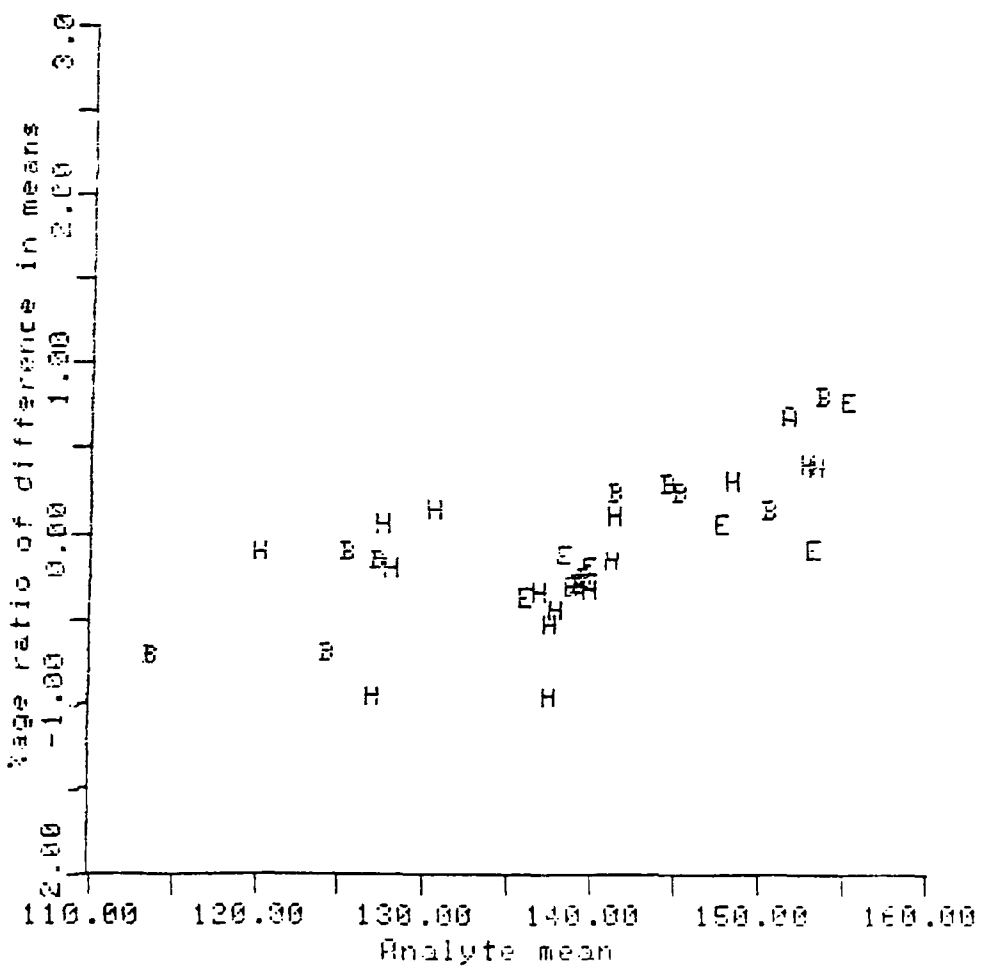
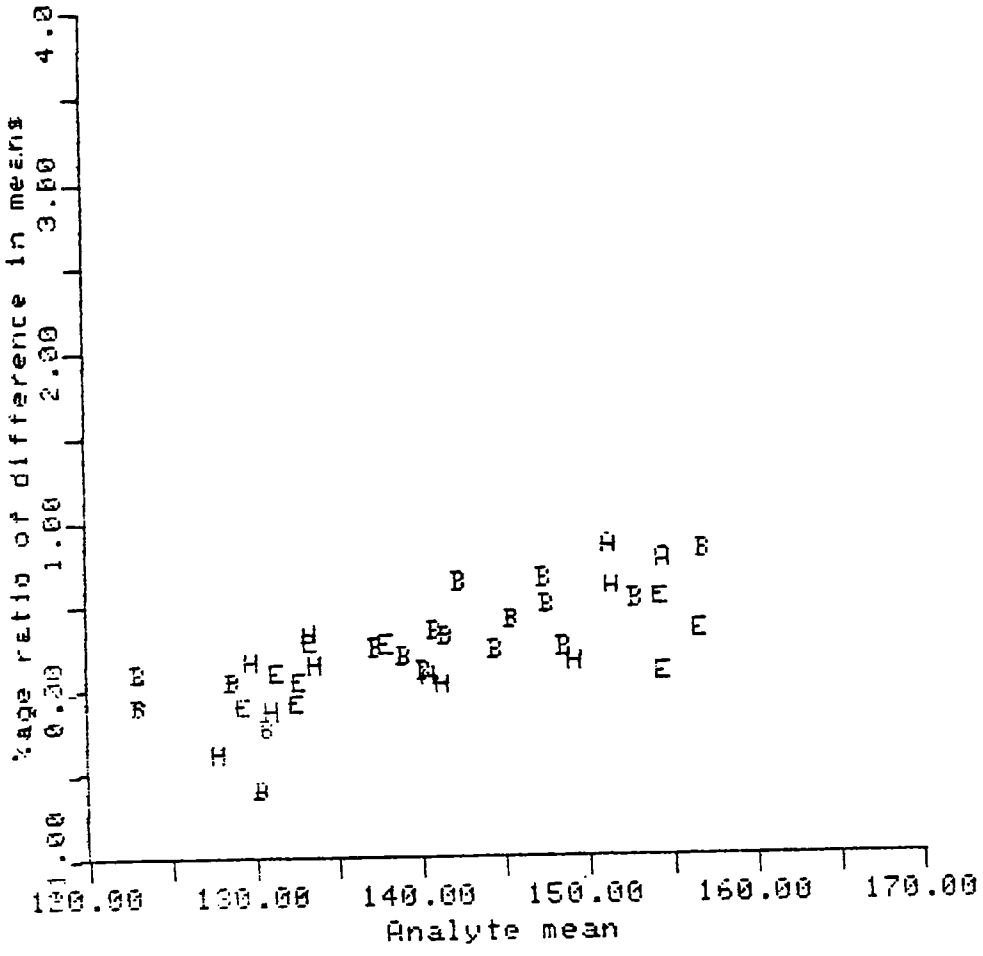


Figure 14.7 Relationship with percentage difference from recalculated mean for sodium by Indirect ion-selective electrode and Direct ion-selective electrode, classified by species of origin

Indirect ion-selective electrode:



Direct ion-selective electrode:

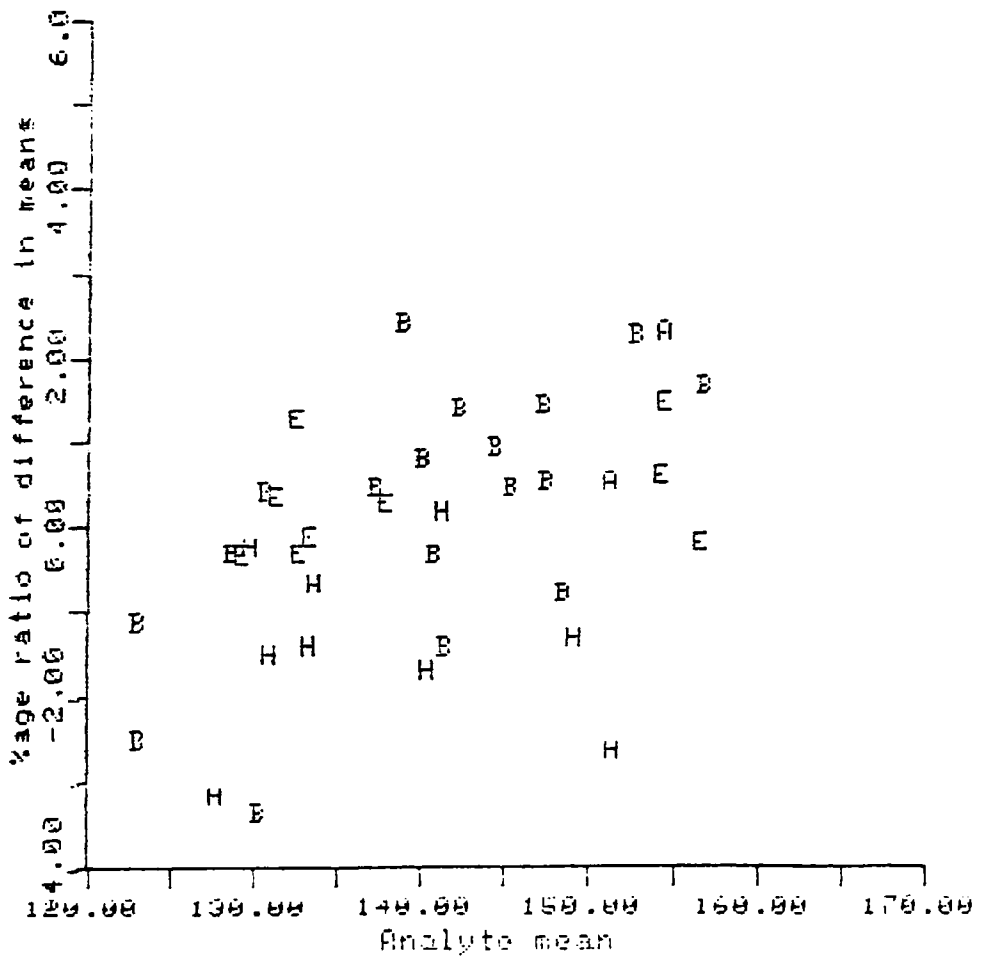
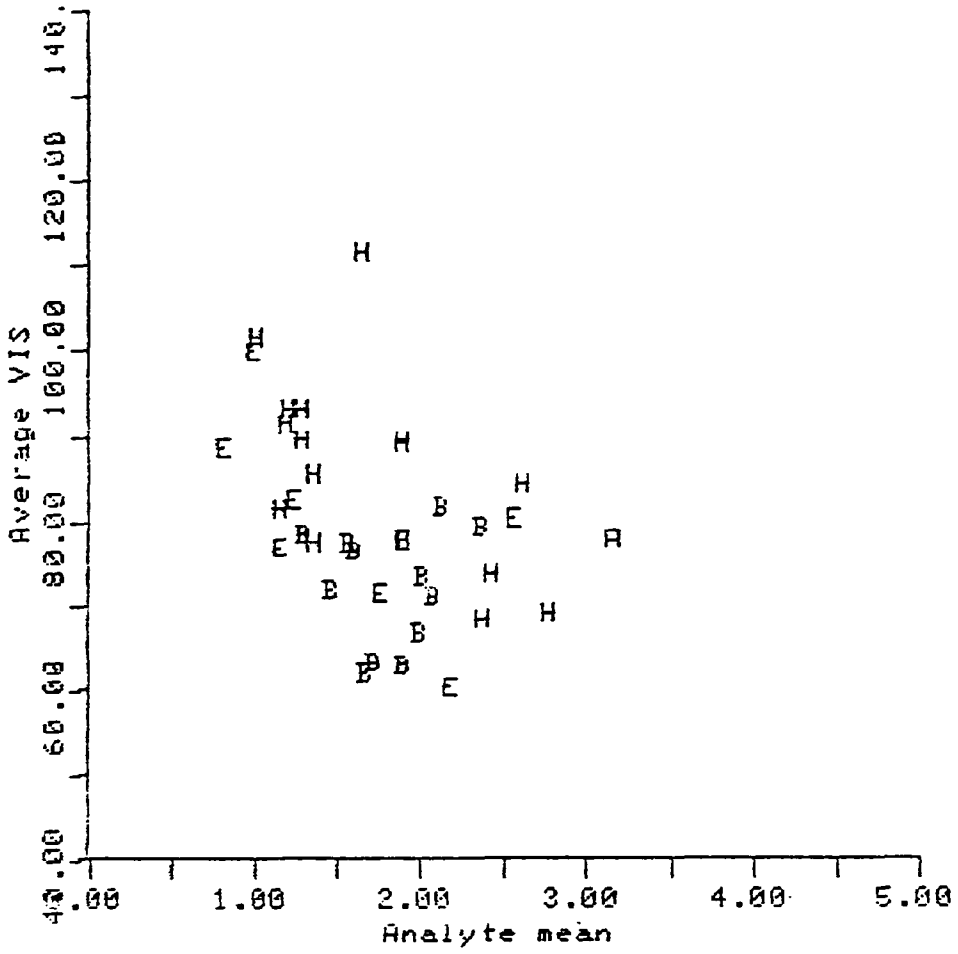
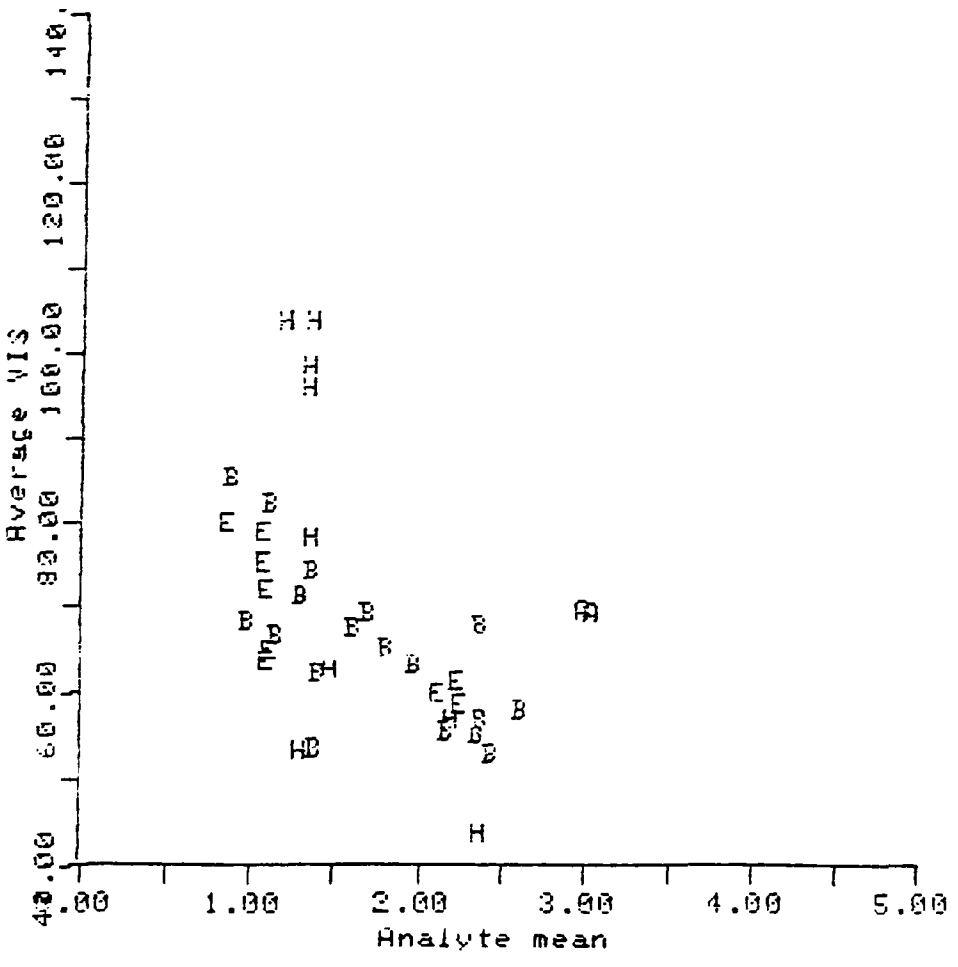


Figure 14.8 Relationship with average VIS for phosphate by Manual/discrete analyser colorimetric, classified by species of origin, in study and validation periods

Study:



Validation:



discrete analyser colorimetric procedures. Apart from these isolated instances, albumin (see above) and total protein, there was no evidence of species-related differences in behaviour.

For total protein, with an overall tendency for worse agreement for human-based materials, there was clear evidence of species-related bias. Here, relative to equine and bovine sera, there were negative biases for human-based materials for continuous flow and positive biases for manual and discrete analyser procedures, exemplified in Figure 14.9 by the AAII/SMA and Vickers M300/D300 groups respectively. In the validation period these differences seemed to be confined to (or at least much more obvious between) the continuous flow group including a serum blank and the unblanked manual and discrete analyser grouping (Figure 14.10).

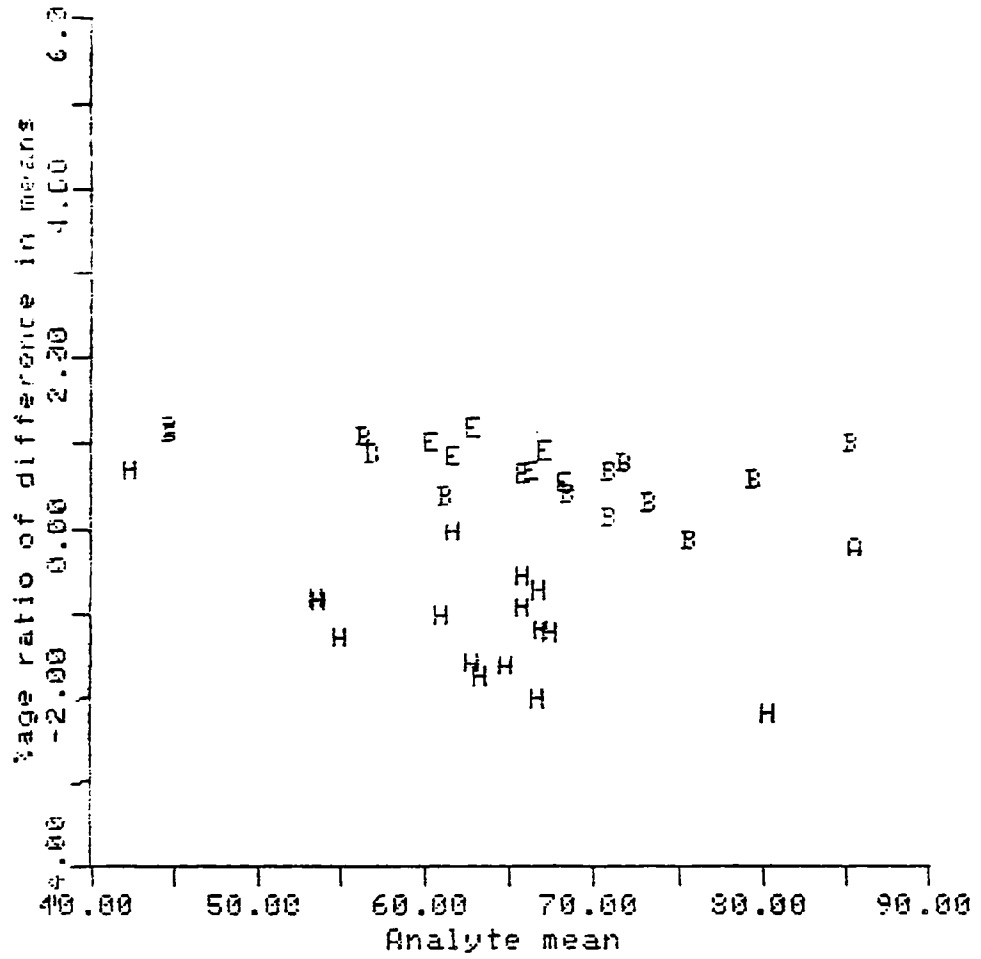
14.3.3 Appraisal of the results

Similar numbers of materials based on bovine (31) and human (27) serum were distributed during the study and evaluation periods combined, with rather fewer equine-based sera (16) and slightly different relative proportions in the two periods (Table 14.1). Reliable conclusions could therefore have been expected, but unsupported (and hence presumably artefactual) indications of differences in performance were nevertheless apparent.

This study was confined to analytes in the UKEQAS for General Clinical Chemistry, ie electrolytes and simple organic compounds; albumin (on which results are now rarely requested on non-human based sera) was assessed primarily to provide contrast. It did not include radioimmunoassays (RIAs), nor analytes (eg immunoglobulins) which are species-specific and for which human

Figure 14.9 Relationship with percentage difference from recalculated mean for total protein by AAII/SMA and Vickers M300/D300, classified by species of origin

AAII/SMA:



Vickers M300/D300:

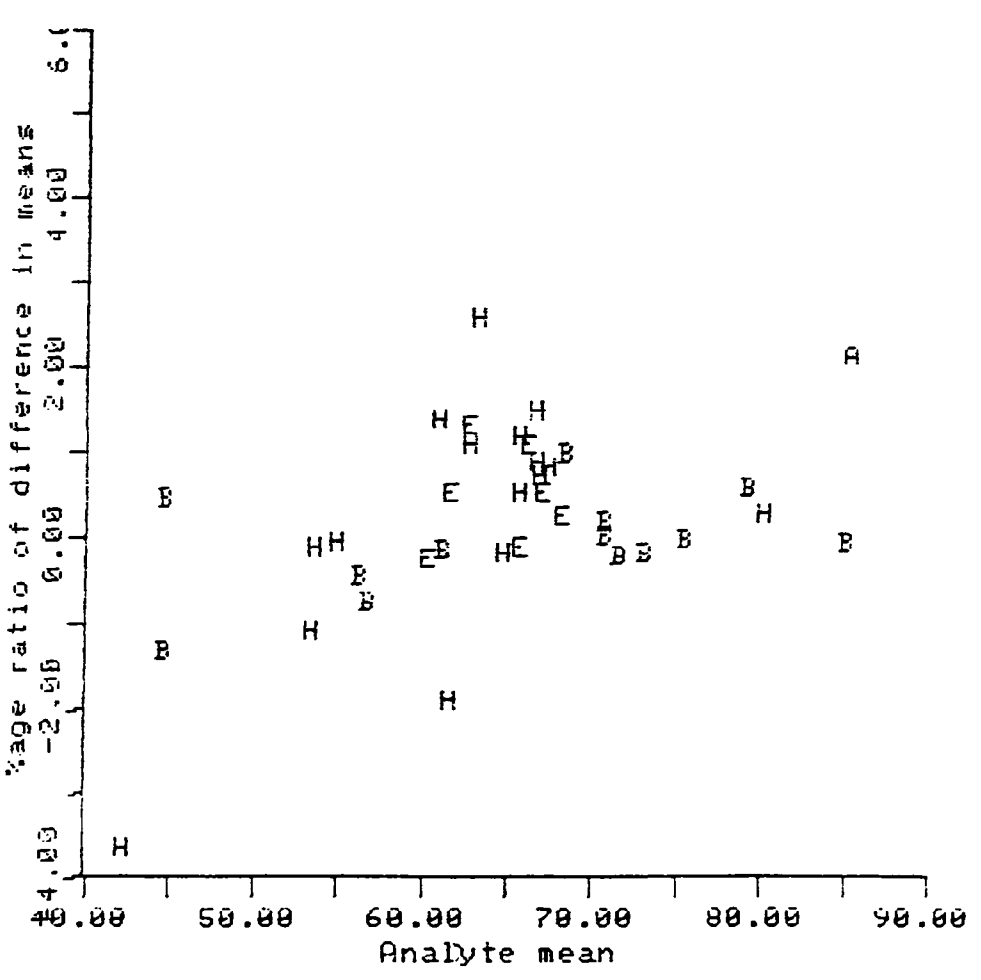
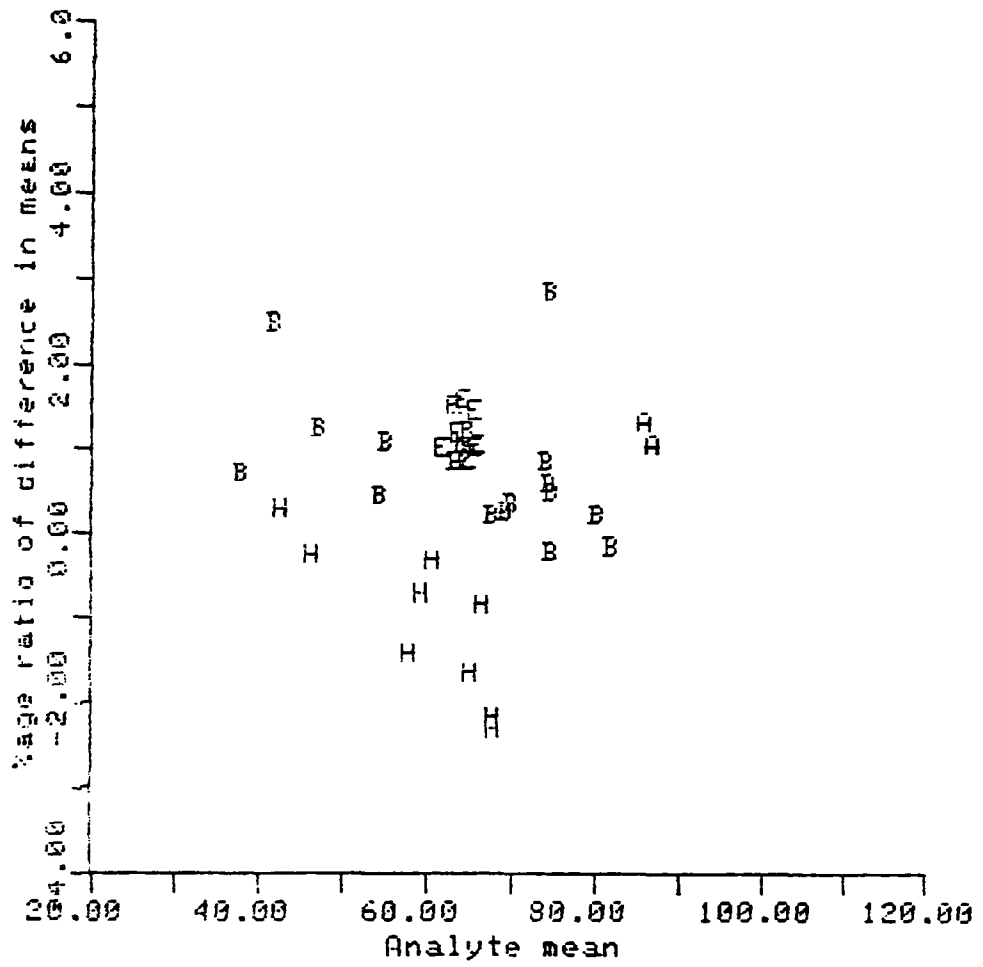
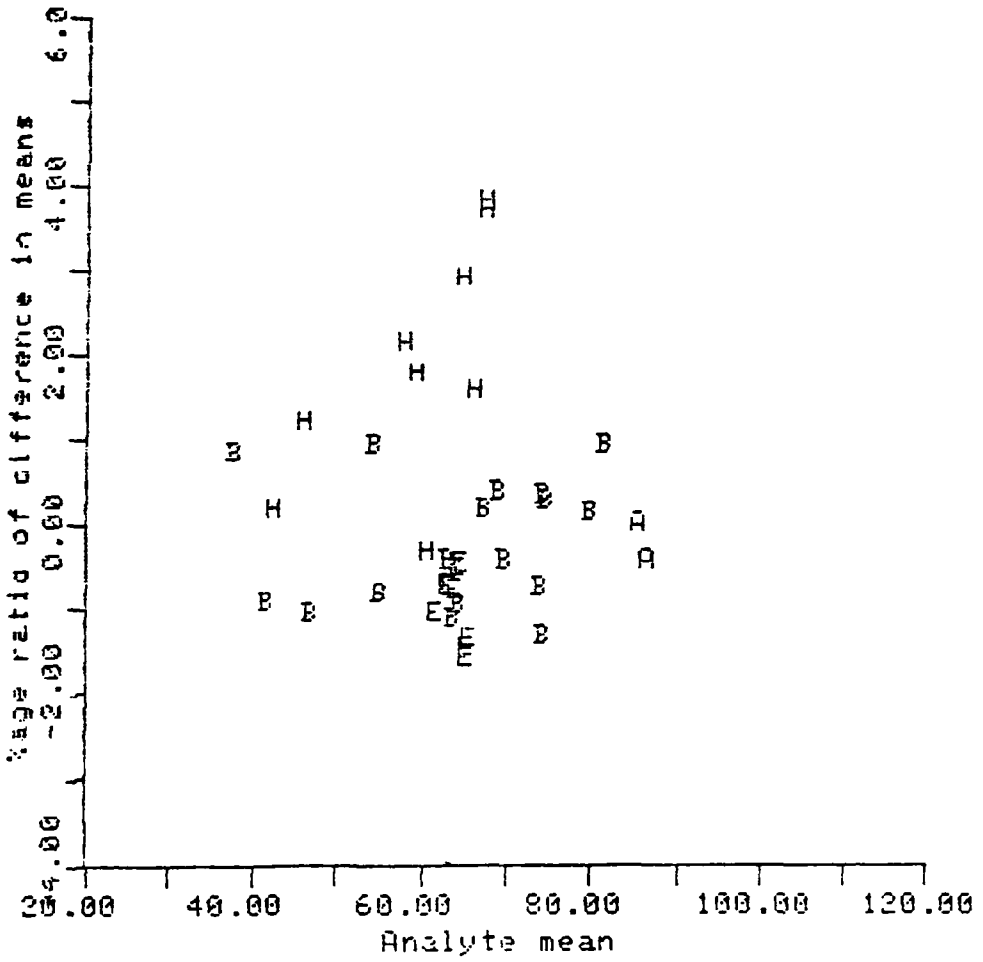


Figure 14.10 Relationship with percentage difference from recalculated mean for total protein by Continuous flow blanked biuret and Manual/discrete analyser unblanked biuret, classified by species of origin

Continuous flow blanked biuret:



Manual/discrete analyser unblanked biuret:



based sera appear indispensable a priori. In order to study subtle aspects of QC material behaviour due to species of origin, it is essential to examine data from individual method groups rather than overall statistics. The overall data, however, constituted the best practical starting point, both in providing an overall assessment and in eliminating from further analysis materials with obviously discrepant behaviour.

The study period provided definite evidence of worse interlaboratory agreement for human-based sera for four analytes (glucose, phosphate, bilirubin and total protein), corroborated by the data from the validation period. It is surprising that agreement was worse for human-based materials since it is frequently argued that, as assays are optimised for clinical specimens (ie fresh human serum), agreement between both individual laboratories and method groups is likely to be inferior for animal-based sera. It is possible that the usually higher turbidity of the human-based materials contributed to greater variability within method groups, due to differences in blanking procedures applied in participants' laboratories. Greater differences in bias between method groups for human-based sera cannot be the full explanation, since the inferiority was seen for VISs as well as CVs, but method-related data were also studied to investigate this further.

The concentration-dependent biases for sodium in the ion-selective electrode groupings are intriguing. They show a clear divergence in behaviour between indirect- and direct-reading instruments (ie those with and without dilution of the sample prior to measurement), and a dependence of the bias for the direct group upon the species of origin of the specimen.

Surprisingly, the greatest difference in mean from the other methods was seen for human-based materials, whereas the response of these instruments should be optimised for fresh human serum. It may therefore be that the effects of commercial processing and material lyophilisation predominate, and somehow counteract in part the 'deficiencies' of animal sera or alter the properties of the human-based materials.

In fact, in almost all instances of apparent species-related differences in behaviour the worst performance was seen for human-based materials. This better agreement for animal-based materials may be due to a lack of (or a lesser degree of) interfering substances or turbidity than in fresh human sera or in human-based QCMs, to an excess of such interferences in human-based lyophilised materials, or to a combination of both. Without similar data (which cannot practicably be obtained) for fresh human sera it is impossible to establish the cause with certainty, and it remains debatable which materials are reflecting more accurately the situation for clinical specimens. Ideally, to prevent acceptance of clinical results from unsatisfactory analytical batches, a material used for IQC should have slightly greater sensitivity than clinical specimens to analytical variables and the human-based sera are therefore perhaps better for that purpose.

With the exception of albumin (where species differences were expected) and total protein, such instances were very infrequent, however. It was usually difficult to discern for individual methods the differences which were apparent from examination of the overall data: most method groups are considerably smaller than the overall data set and a large number of results may be

necessary to reveal such differences. Furthermore the differences were small, and in most cases the data indicate no appreciable difference in performance for materials based on human and animal (bovine or equine) serum for the non-protein analytes studied.

For total protein, however, human-based specimens do behave differently from those based on animal serum and the differences are appreciable, with average VISs differing by 10-20. This indicates a need for further study, with consideration of whether performance should be assessed, ie VISs calculated, for total protein only when the material distributed is of human origin, as is the current practice for albumin.

14.4 The effects of material manufacturer

The examination outlined for species of origin (section 14.3) was repeated, but with the materials classified according to their manufacturer (Appendix II.2.3). However, only three manufacturers were represented by five or more materials (Table 14.1).

The proportions of manufacturers whose materials were distributed differed substantially between the study and validation periods (Table 14.1) and the conclusions concerning Nyegaard and Scottish Blood Transfusion Service materials could not be checked. The other five manufacturers were represented by larger numbers of materials, however, and interpretation was correspondingly facilitated.

14.4.1 Examination of overall data

No completely consistent evidence of superiority or inferiority of any manufacturer's products was found, though in many cases examination of the recalculated CV or average VIS graphs

suggested this. In some this impression was supported by a similar interpretation of the graph of one or other of the indicators of discrepant performance. Thus, agreement appeared better for materials from Roche Diagnostica for glucose, calcium, phosphate, iron and bilirubin (Figure 14.11), from Wellcome Diagnostics for sodium and glucose, from Technicon Instruments for urea and phosphate, and from Nyegaard for iron. Wellcome Diagnostics' materials seemed inferior for magnesium (Figure 14.12A).

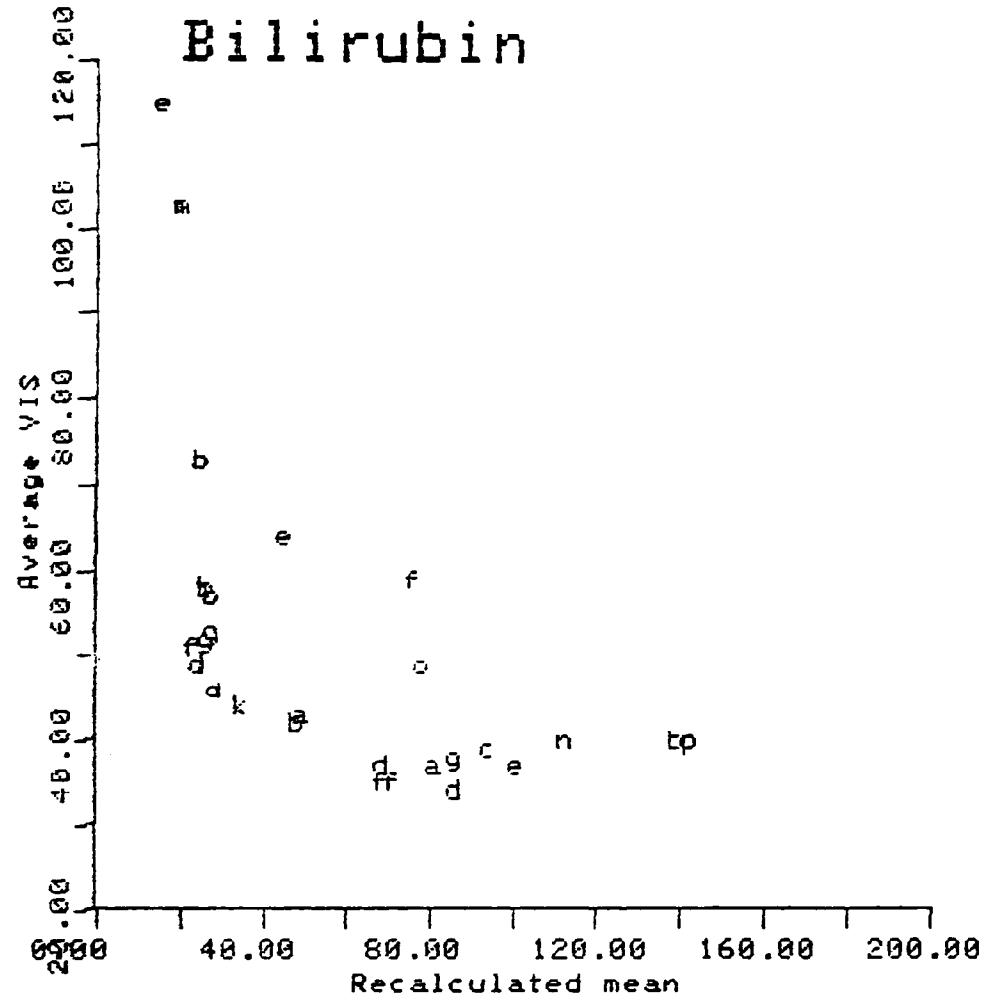
Data from the validation period suggested better interlaboratory agreement for materials from Roche for calcium, phosphate and bilirubin, from Wellcome for sodium, urate and total protein, from Technicon for creatinine and bilirubin, from Purce Associates for sodium, and from Ortho Diagnostics for chloride. Only in the first four cases did these conclusions correspond to those from the study period. The suggestion of worse agreement for magnesium with Wellcome materials was confirmed (Figure 14.12B).

14.4.2 Examination of method-related data

For sodium, the better agreement for Wellcome materials was confirmed for the flame photometric groups, but not for ion-selective electrodes (Figure 14.13).

Examination of the recalculated CVs for creatinine revealed (Figure 14.14) that although similar patterns exist for the Manual/discrete analyser endpoint Jaffe, AAII/SMA and Other (ie kinetic Jaffe) groups, the variability within the kinetic group is almost twice that within the continuous flow group (Figure 6.5). The effect of manufacturing procedures is exemplified by the Hyland material, with a CV of 35% by kinetic Jaffe but only

Figure 14.11 Relationship with average VIS and percentage of VISs >200 for bilirubin, classified by manufacturer



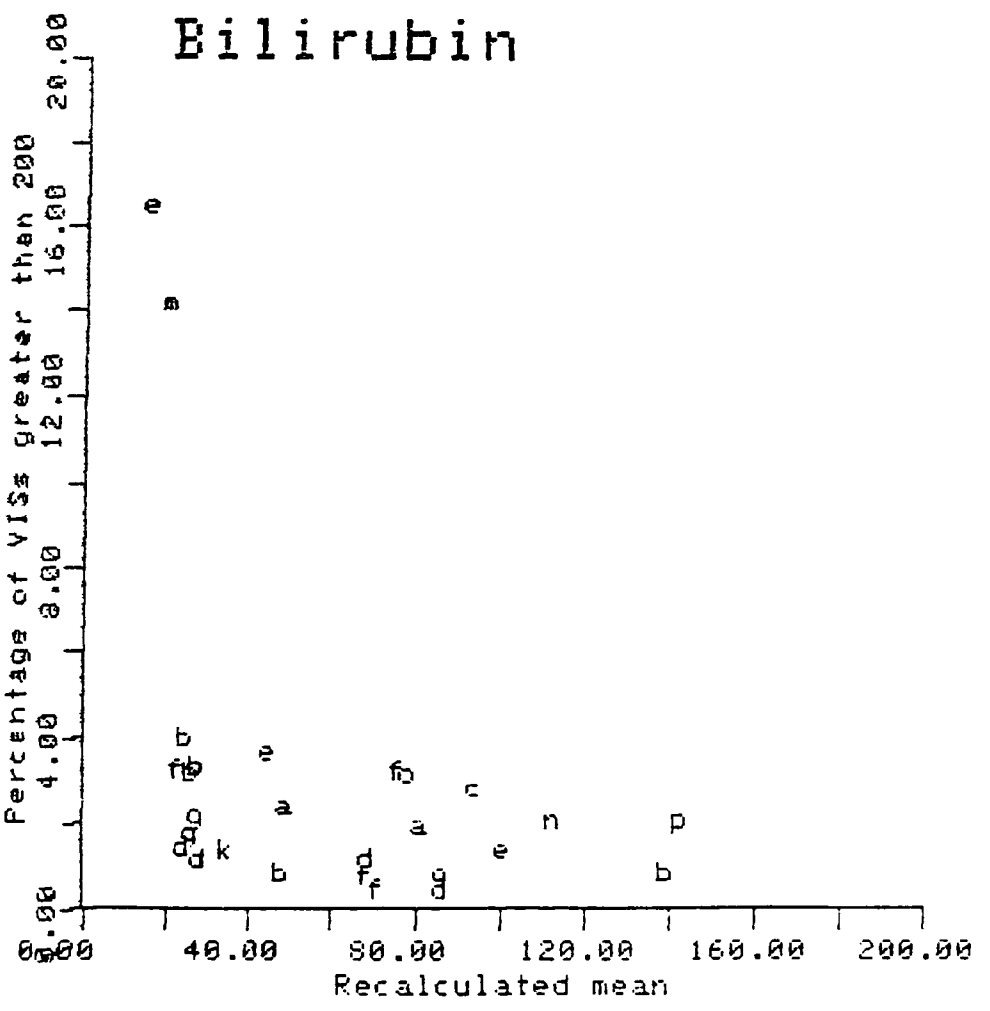
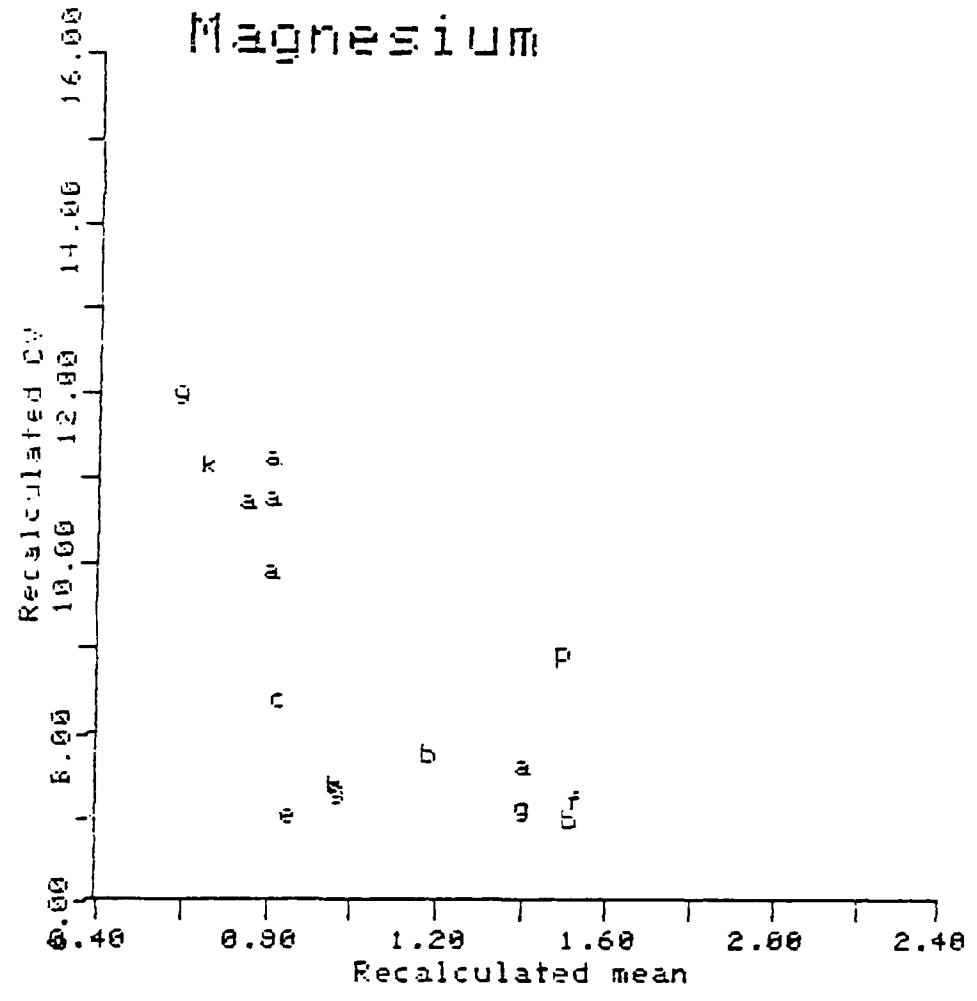


Figure 14.12 Relationship with recalculated CV and percentage of VISs >200 for magnesium, classified by manufacturer



Validation:

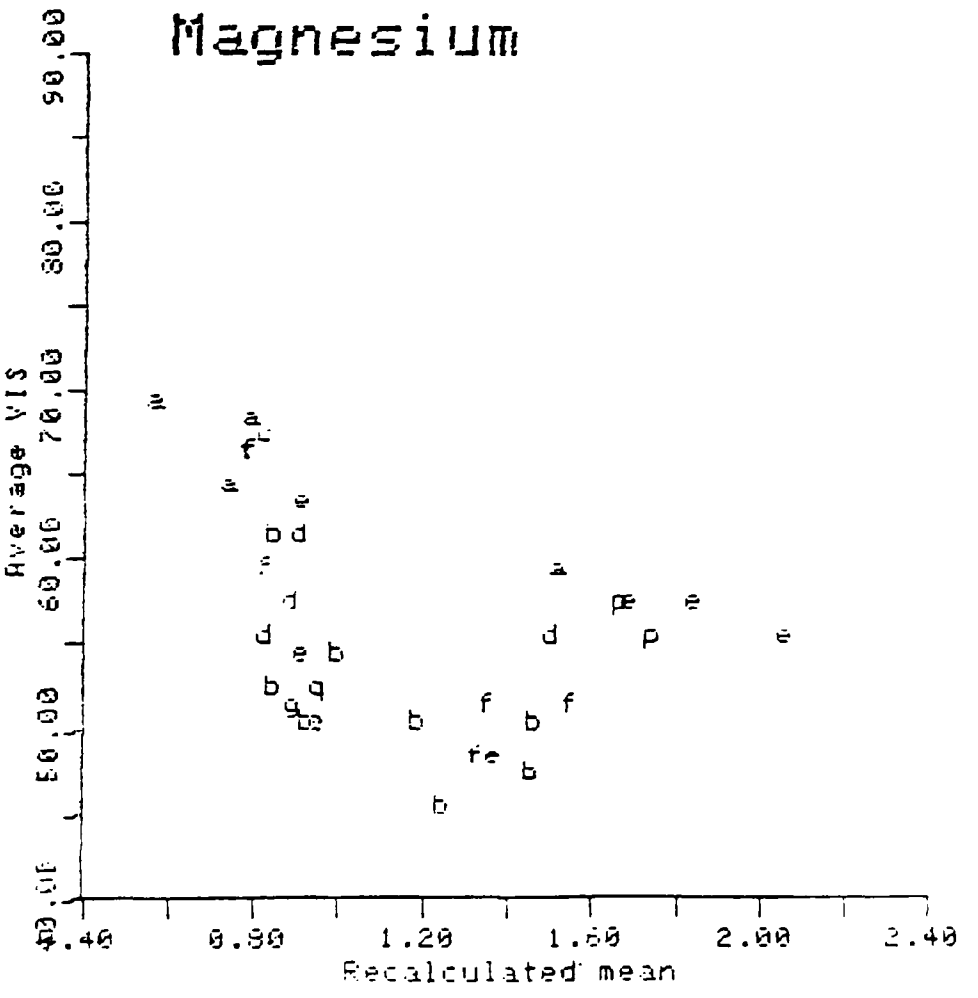
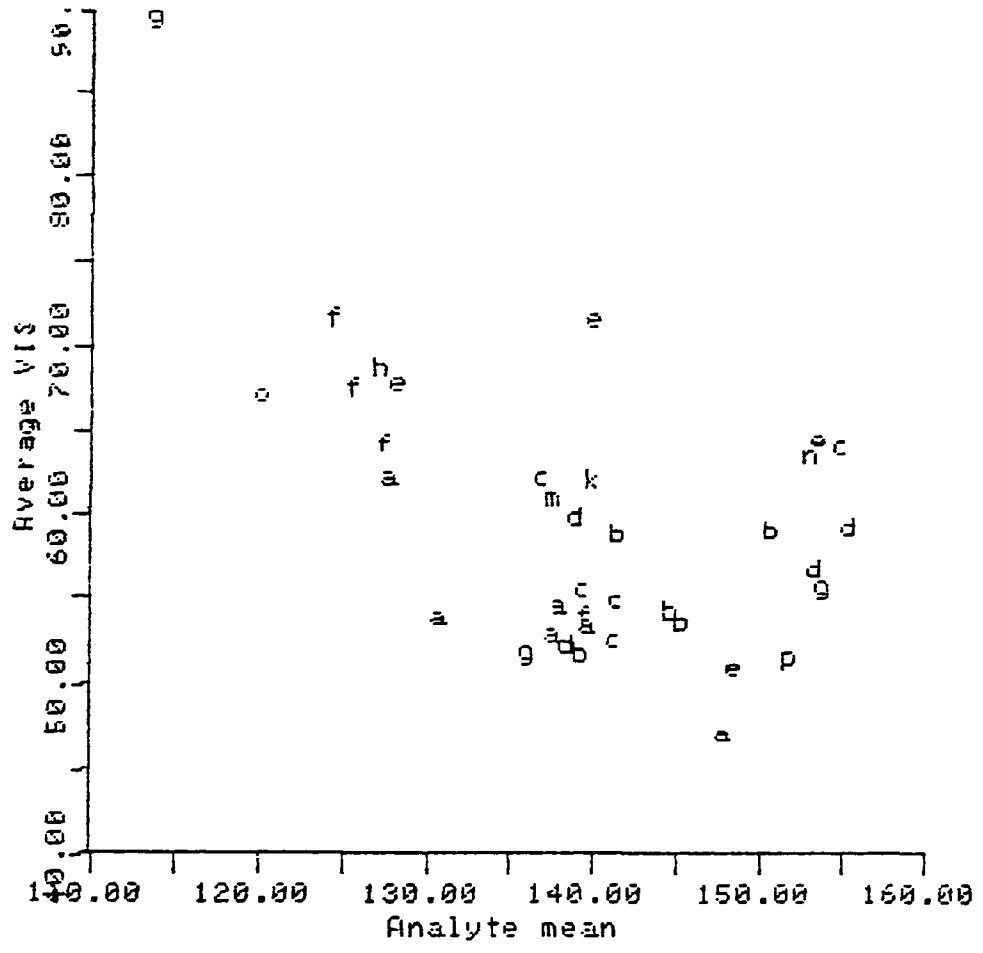


Figure 14.13 Relationship with average VIS for sodium by Continuous flow flame and Ion-selective electrode, classified by manufacturer

Continuous flow flame:



Ion-selective electrode:

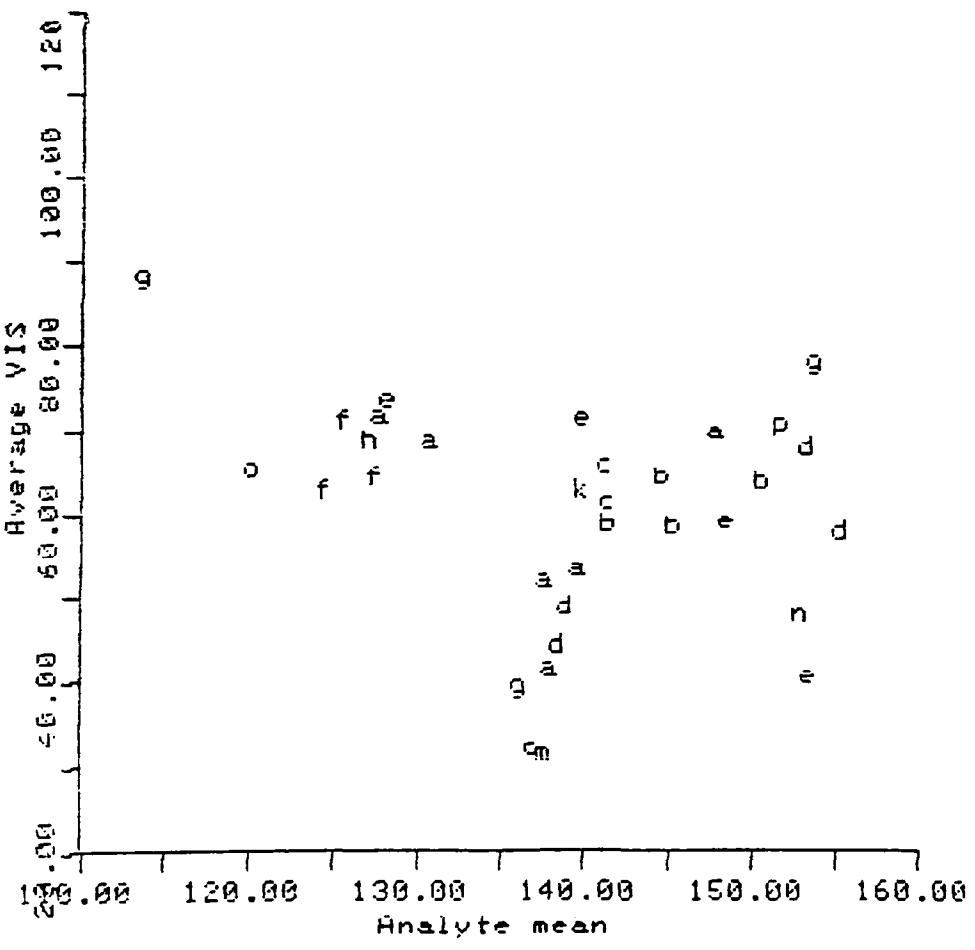
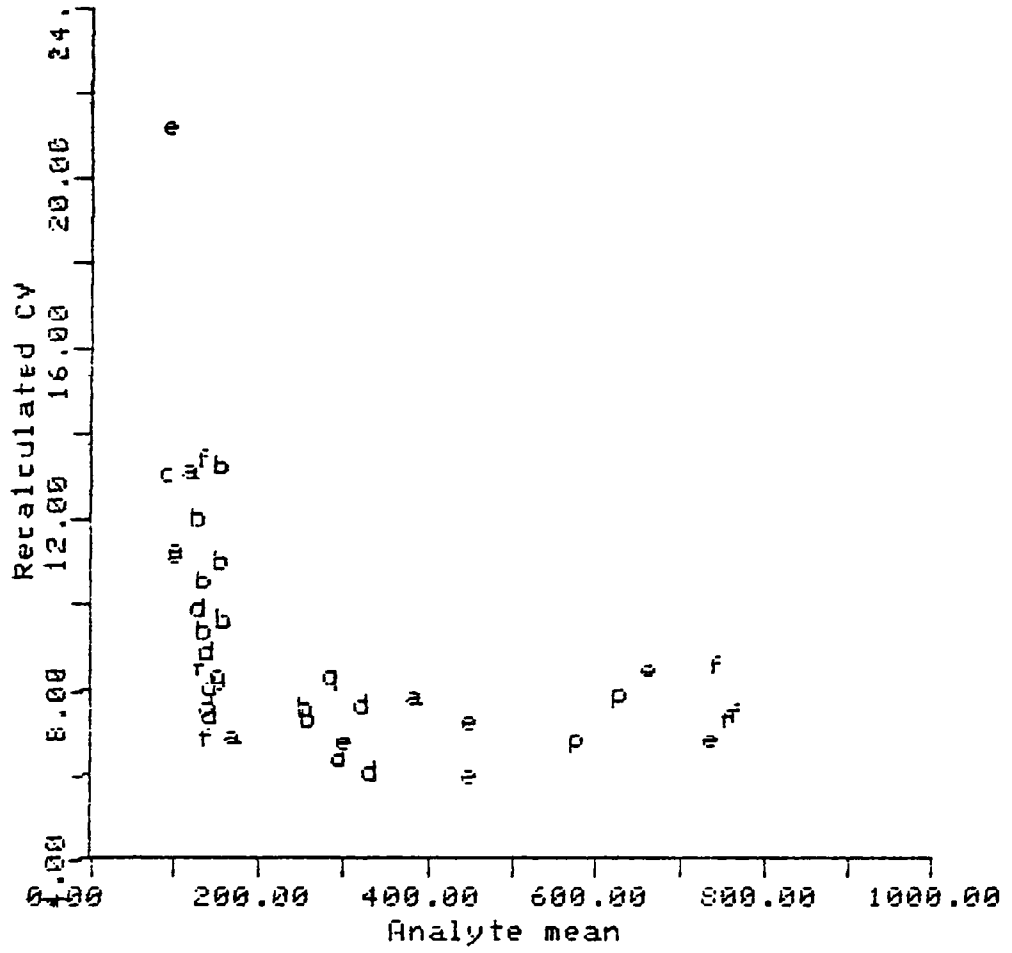
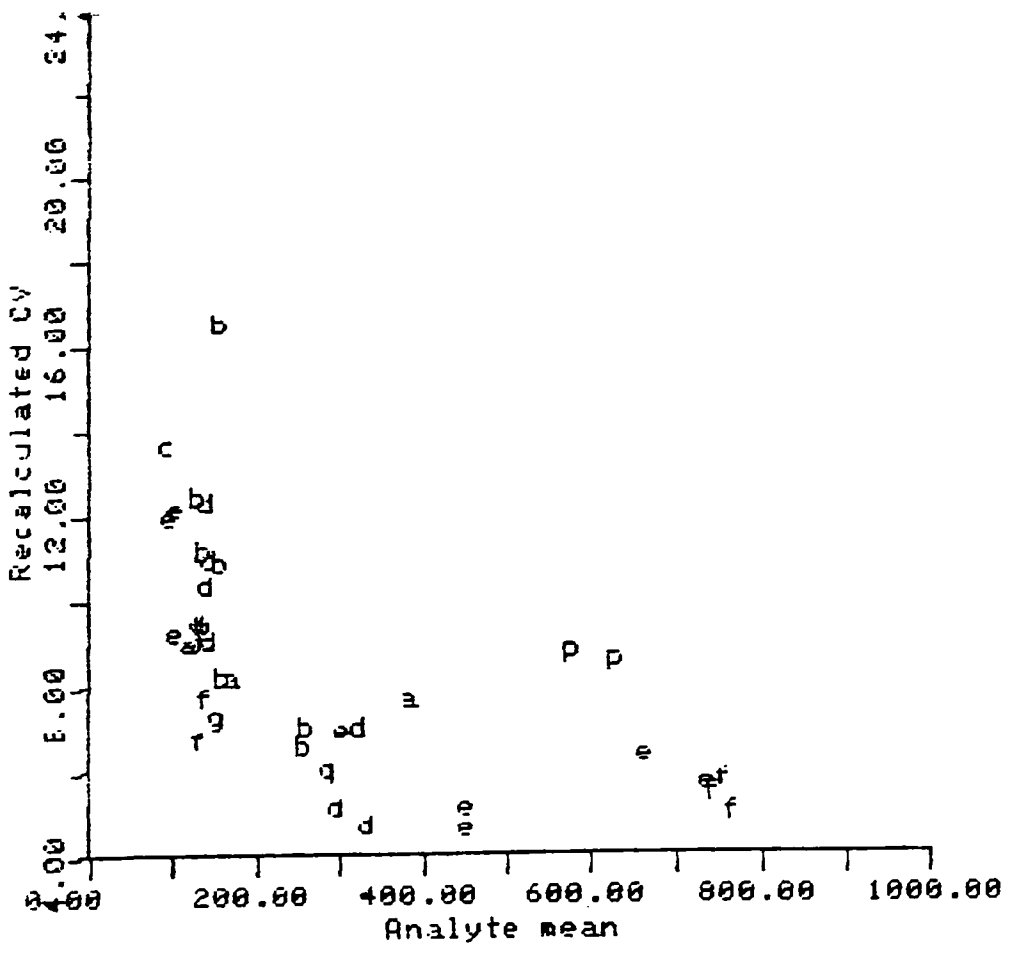


Figure 14.14 Relationship with recalculated CV for creatinine by Manual/discrete analyser endpoint and Other, classified by manufacturer

Manual/discrete analyser endpoint:



Other:



of 7.5% by AAII/SMA; Purce Associates' materials also seemed to show a similar effect, but of lesser magnitude. No material from Hyland was distributed during the validation period, so the discrepant performance could not be confirmed. There seemed to be no confirmation of the suggested worse agreement for Purce Associates' sera, which by then no longer contained Tris buffer; this interpretation would be supported by data for the Gibco sera, which do contain Tris (Figure 14.14).

14.4.3 Appraisal of the results

UKEQAS distributions do not constitute a fully representative selection of the materials available from different manufacturers, and only a limited number of manufacturers provided a sufficient number of materials to form an interpretation (Table 14.1). To be reliable, however, conclusions should obviously be based on substantially more than the two to five usually considered here. It is in this situation that the validation of conclusions by assessment of the independent set of data assumes paramount importance. It was thus more doubtful that definite conclusions on this aspect could be drawn, especially since it is UKEQAS policy not to distribute materials from manufacturers previously found to have products of inferior quality.

Nevertheless, three manufacturers were represented by 4-6 materials, ie 10-15% of the database, and several manufacturers' materials did appear to yield better interlaboratory agreement, as assessed by more than one of the performance indicators, for one or more analytes. There were fewer unsupported (and hence presumably artefactual) indications of differences from the validation than from the study period. This might have been

expected, since fewer manufacturers, each represented by more materials, were considered.

Similar conclusions were drawn from both periods, for materials from Roche Diagnostica of better agreement for calcium, phosphate and bilirubin, and for materials from Wellcome Diagnostics of better agreement for sodium and of worse agreement for magnesium. It is difficult to attribute causes to these observations, though better agreement for sodium might reflect better vial-filling precision. They are unlikely to be due to a relative lack of intermethod bias since these conclusions were drawn from average VISs as well as recalculated CVs, and might simply reflect more homogeneous (and perhaps less turbid) materials. The worse agreement for magnesium is particularly difficult to interpret as the same manufacturer's materials appeared the best for sodium, though some change in the matrix affecting magnesium chelation is possible.

14.5 Effect of addition of ethylene glycol

This was assessed by distribution in 1982 and 1983 of a commercial liquid material (Link II, Beckman Instruments) based on human serum and preserved with ethylene glycol (Maurukas, 1973) in the UKEQAS for General Clinical Chemistry and UKEQAS Enzyme Surveys.

14.5.1 UKEQAS for General Clinical Chemistry

This distribution followed the usual procedures (Appendix I.2.1), except that, on the suppliers' recommendation, participants were requested to freeze the specimen on receipt. Table 14.2 compares the overall mean values and CVs obtained with those from the preceding and following distributions of lyophilised human sera

Table 14.2 Interlaboratory agreement obtained for human serum stabilised with ethylene glycol in the UKEQAS for General Clinical Chemistry, April 1982, compared with that for lyophilised serum. Beckman Link II, lot C111047, compared with Ortho Unassayed Human Serum Level II, lot W24X02B (UKEQAS human serum, lot HIQC/5, for potassium and cholesterol)

	Serum with ethylene glycol		Lyophilised serum	
	Mean	CV (%)	Mean	CV (%)
Sodium	145.6	3.9	148.1	1.2
Potassium	5.77	4.6	4.11	2.1
Chloride	110.7	2.9	112.1	2.1
Urea	18.1	5.3	13.9	3.9
Glucose	14.2	5.8	13.8	4.5
Calcium	3.26	5.0	2.85	2.7
Phosphate	1.7	11.1	2.4	5.3
Iron	50.7	12.7	56.7	7.5
Urate	0.55	7.6	0.53	5.0
Creatinine	545	10.4	242	5.0
Bilirubin	65.1	11.9	42.1	15.0
Total protein	76.8	4.3	61.1	2.8
Albumin	47.8	6.5	41.0	5.5
Cholesterol	5.41	7.9	4.19	7.5

with similar analyte levels.

The interlaboratory agreement was obviously far inferior to that usually obtained, and thus no VISs were calculated. Investigation and further discussion with Beckman revealed that this was probably due to differing treatment of the material by participants, with some (correctly) having allowed the serum to equilibrate at ambient temperature before thoroughly mixing it and assaying it. Others had probably taken it directly from the freezer, with possibly inadequate mixing, before assay, and yet others adopted intermediate procedures.

The temperature-dependence of the material's viscosity had then interacted with the analytical procedures to generate the additional variability in results. Laboratory investigation confirmed this for sodium assay by means of a flame photometer (IL 453) incorporating a viscosity-dependent dilutor: results for successive samplings of the material taken directly from the freezer declined steadily, until aspiration stopped completely.

Thus correct specimen handling is essential for EQA data to reflect the true performance of participants and the material distributed. Such assessment is invalidated if additional variance is introduced.

14.5.2 UKEQAS Enzyme Surveys

Particular care was therefore taken to ensure that participants followed the correct procedure (to store the material at $4^{o}C$, but allow it to equilibrate to ambient temperature with adequate mixing before assay) for handling this specimen when it was distributed for enzyme activity assays. Here the primary aim was to assess the efficacy of the usual protocol (Appendix I.2.2;

Table 14.3 Interlaboratory agreement obtained for human serum stabilised with ethylene glycol in UKEQAS Enzyme Survey 13, 1983, compared with that in Enzyme Survey 14 for 'reliable' method groups. SCE optimised 37°C for all except amylase (Phadebas Tablet Test 37°C); Beckman Link II, lot C111047, compared with Ortho Unassayed Human Serum Level II, lot X40Y02B

	Serum with ethylene glycol		Lyophilised serum	
	Mean (U/L)	CV (%)	Mean (U/L)	CV (%)
AST	89	7.6	159	6.7
ALT	89	9.8	122	7.4
LD	987	9.5	881	8.5
CK	307	9.0	531	11.7
ALP	267	10.8	470	11.4
Amylase	747	11.6	546	11.9

Bullock et al, 1979 and 1986b) in minimising the effect of post-reconstitution changes in enzyme activity with lyophilised specimens.

Table 14.3 shows the interlaboratory agreement obtained for 'reliable' method groups for the 6 enzymes surveyed, with comparison data from Survey 14. The average CVs of 9.7% and 9.6% for the 6 enzymes surveyed indicate there is no substantial difference between the two types of specimen. Assuming the liquid material contributed no excess variance, the protocol for lyophilised sera has thus been successful.

14.6 Summary

External quality assessment data are an excellent means of examining properties of the materials distributed. Before such detailed investigation, however, the concentration-dependence of interlaboratory agreement must first be characterised, as discussed in Chapter 6.

Assessment of data from the UKEQAS for General Clinical Chemistry demonstrates that animal-based quality control materials are generally as satisfactory as those based on human serum for the non-protein analytes studied. In a few instances human products gave worse interlaboratory agreement, due perhaps to their greater turbidity or a relative lack of interferents in the animal sera. More detailed examination suggested that such instances may arise in some individual method groups, eg kinetic Jaffe procedures for creatinine assay.

Similar assessment failed to reveal major differences among materials from different manufacturers, perhaps reflecting the Organisers' exclusion of previously unsatisfactory material types

from future distribution. A few cases were found, however. Apart from the interference of Tris buffer in some kinetic Jaffe procedures for creatinine assay as well as its known effect on some urea assay methods employing urease, no explanations for these were apparent.

Studies of materials including ethylene glycol emphasised the importance of correct specimen handling procedures, and supported the efficacy of existing procedures to minimise changes in enzyme activity following reconstitution of lyophilised sera.

Chapter 15:

ASSESSMENT OF THE EFFECTS OF HEAT-TREATMENT ON THE BEHAVIOUR OF LYOPHILISED HUMAN SERUM

15.1 Introduction

Increasing awareness of the potential infectivity of specimens of human origin, including calibration and quality control materials, has led to greater precautions in clinical laboratory work (Advisory Committee on Dangerous Pathogens, 1986; DHSS, 1986a). Among these is the policy of UK External Quality Assessment Schemes (UKEQASs) to distribute human-based materials only if they have been tested and found negative for antibody to human immunodeficiency virus (HIV; human lymphotropic virus type III or lymphoadenopathy-associated virus) at the individual donor stage or have been adequately treated to inactivate retroviruses such as HIV. Such treatment may be heating (eg at 56^{o}C for a minimum of 30 min) or chemical (eg beta-propiolactone).

It is not always possible at present to ensure testing of individual donations in all circumstances, and inactivation treatment may therefore be needed. Before preparing large quantities of such material it is essential to study the effects of the treatment on the properties of the product, to ensure that its intended use is not compromised.

The UKEQAS for General Clinical Chemistry has since 1980 processed human serum and provided this, with assigned values derived from distribution through the UKEQASs for General Clinical Chemistry, Steroid Hormones and Thyroid-related Hormones, to National Health Service laboratories in the UK for use as a check on bias

(Appendix IV). Much of the stock of serum (collected by the National Blood Transfusion Service) for use in preparing this HIQC material was obtained, and stored in 5 litre pools, before tests for antibody to HIV were available. Thus treatment of the base serum would be necessary before these stocks could be used. Heat treatment was preferred since only the base serum need be treated, whereas chemicals would persist in the processed material and affect also the analyte supplements added during processing and interact with analytical procedures on the final product. In addition, such chemicals could interfere with the lyophilisation of the serum and are suspected to be carcinogenic. Heat treatment, however, is not an established procedure in serum processing and no purpose-designed equipment is available to accomplish this.

Preparation of a pilot-scale batch was therefore undertaken in 1986, in collaboration with Wellcome Diagnostics, who also processed the HIQC material at that time. Conventional evaluation would yield only limited information on the properties of the material, and thus an appraisal of the between-laboratory agreement obtained in EQASs constituted the main means of assessment. The objective was to assess the effects of this heat-treatment on the wide variety of analytical procedures adopted by scheme participants, ie to assess the degree to which alterations in the serum matrix affected these methods, through comparison with the performance of a batch of similar but untreated serum.

15.2 Effects of heat-treatment

The study (Bullock et al, 1986a) compared the behaviour of this batch (lot K466610) with that of preceding batches of HIQC material, and with HIQC/12 (processed two months later to a

similar specification but from serum donations which had been tested and found negative for antibody to HIV) in particular. Full details are given in Appendix III.6.

15.2.1 Evaluation data

The physical properties evaluated at WRL showed few differences of the heat-treated serum from previous batches of HIQC material (Table 15.1). The turbidity was lower, perhaps reflecting aggregation of unstable proteins which were then removed during filtration rather than aggregating during lyophilisation. A number of other factors, including pool-to-pool variation in protein (eg lipoprotein) content and serum ageing processes may, however, be involved as evidenced by the variations in turbidity among HIQC/9 to HIQC/12.

No microbiological growth was reported in vials from any of the batches studied. Lot K466610 was more stable than lot HIQC/12, which was of similar stability to the preceding batches. Thus the stability of the inorganic, organic and enzyme components studied seemed unaffected by the heat-treatment, as might be expected since the addition of exogenous analytes was made after completion of heat treatment.

15.2.2 Inorganic, organic and enzyme constituents

There was little difference in the overall performance for the analytes in the UKEQAS for General Clinical Chemistry as judged from interlaboratory CVs (Table 15.2); the figures for enzymes are for a 'reliable' method group (Bullock et al, 1986b). This impression is borne out by the average VISs (Table 15.3), which are calculated relative to the mean for each laboratory's method and thus contain no contribution from differences among method means.

Table 15.1 Comparison of pH, turbidity and vial-to-vial variability for heat-treated batch and lots HIQC/9 to HIQC/12. pH and turbidity data are the mean for three vials; vial-to-vial variability derived from sodium assay on 40 vials

	Heat-treated batch	HIQC/9 - 12	
		Average	Range
pH	8.87	8.87	8.60 - 9.07
Turbidity			
A_{400nm}	0.450	0.488	0.406 - 0.538
A_{500nm}	0.282	0.338	0.297 - 0.399
A_{600nm}	0.129	0.161	0.135 - 0.188
Vial-to-vial variability			
CV (%)	0.18	0.31	0.22 - 0.47

Table 15.2 Comparison of interlaboratory agreement (CV) for heat-treated batch and lots HIQC/9 to HIQC/12, with mean value for heat-treated batch. CVs (%) are irrespective of method after exclusion of outliers, apart from those for enzymes: SCE optimised 37°C for AST, ALT, LD, CK and ALP, Phadebas 37°C for amylase, carboxynitroanilide substrate 37°C for GGT.

Analyte	Heat-treated batch	HIQC/9 - 12	
		Average	Range
Sodium	1.2	1.3	1.2 - 1.3
Potassium	2.1	2.1	2.1 - 2.4
Chloride	1.8	1.9	1.8 - 1.9
Urea	3.9	4.0	3.6 - 4.4
Glucose	4.1	4.0	3.8 - 4.2
Calcium	3.0	2.7	2.6 - 2.8
Phosphate	5.7	5.4	4.9 - 6.1
Iron	10.6	11.2	9.8 -12.9
Urate	6.2	5.8	5.4 - 6.3
Creatinine	8.6	7.0	5.7 - 7.5
Bilirubin	8.6	9.8	9.0 -10.8
Total protein	3.5	3.6	3.3 - 4.1
Albumin	5.1	4.6	4.4 - 5.0
Cholesterol	5.7	5.5	5.2 - 5.7
Lithium	3.9	4.5	3.9 - 4.9
Magnesium	7.5	6.7	5.8 - 7.2
Osmolality	1.8	2.0	1.8 - 2.1
TIBC	19.0	20.2	13.3 -26.2
AST	9.7	9.3	8.1 -11.1
ALT	13.0	14.0	12.0 -17.8
LD	8.9	9.0	8.2 -10.2
CK	19.9	19.5	13.1 -27.1
ALP	9.6	10.4	8.4 -14.3
Amylase	11.6	11.3	8.5 -13.1
GGT	8.5	8.7	-

Table 15.3 Comparison of average Variance Index Scores (VISs) for heat-treated batch and lots HIQC/9 to HIQC/12. No VISs calculated for TIBC or GGT (CCVs not yet established)

Analyte	Heat-treated batch	HIQC/9 - 12	
		Average	Range
Sodium	54	58	56 - 64
Potassium	55	56	55 - 56
Chloride	65	68	65 - 70
Urea	50	56	50 - 62
Glucose	39	42	40 - 43
Calcium	60	52	50 - 54
Phosphate	54	54	48 - 59
Iron	55	58	53 - 68
Urate	63	58	53 - 64
Creatinine	55	51	41 - 56
Bilirubin	34	42	39 - 46
Total protein	65	69	66 - 74
Albumin	47	48	45 - 52
Cholesterol	61	57	56 - 59
Lithium	32	32	27 - 34
Magnesium	54	50	47 - 55
Osmolality	48	57	-
AST	63	66	61 - 71
ALT	58	69	65 - 72
LD	63	72	69 - 75
CK	91	85	-
ALP	55	58	52 - 64
Amylase	75	72	70 - 74

Repeated distribution of lot K466610 in October 1986 showed excellent agreement with its initial distribution 6 months earlier. The consensus values for the 14 non-enzyme analytes requested differed by no more than 0.9%, apart from a reduction in glucose of 0.1 mmol/L (1.9%); there was no indication of deterioration of any analyte. The CVs were also reproducible, and the average VISs for each analyte differed by no more than 6.

Lots HIQC/12 and K466610 were prepared to a similar analyte specification and thus the ratio of mean values should be close to 1.0, though the most important observation is the consistency of the ratio for the individual methods. Table 15.4, however, does show minor differences in behaviour among the methods for iron, but not for sodium or calcium. For all other analytes, the range of ratios was less than 0.05.

15.2.3 Thyroid-related and steroid hormones

Tables 15.5 and 15.6 give data for these hormones for lots K466610 and HIQC/12, classified according to method, with the ratio of the means for the two materials for each method group. Since the numbers in most groups are small, tests of statistical significance are of limited use. An arbitrary criterion of a 10% difference in ratio from that for the overall means was therefore used to identify discrepant method groups; these are denoted by an asterisk.

There was similar behaviour among methods for the thyroid-related hormones, apart possibly from the in-house polyethylene glycol (PEG) precipitation group for thyroxine. For the steroid hormones, however, gross discrepancies were seen for many of the method groups (Table 15.6).

Table 15.4 Comparison of method-related data for general clinical chemistry analytes for heat-treated batch and lot HIQC/12. Ratios are mean for heat-treated batch divided by mean for lot HIQC/12

	HIQC/12	Heat-treated batch			Ratio
	CV	n	Mean	CV	
Sodium (mmol/L)					
Overall	1.3	426	143.7	1.2	0.99
Indirect ISE	1.2	164	144.3	1.1	0.99
Flame with dilutor	1.0	121	143.5	1.0	0.99
Continuous flow flame	0.9	106	143.5	0.9	0.99
Direct ISE	1.9	29	140.5	1.6	0.99
Calcium (mmol/L)					
Overall	2.7	399	2.873	3.0	0.97
Manual/discrete CPC	2.7	189	2.892	3.2	0.97
Continuous flow CPC	2.3	131	2.871	2.2	0.97
Methylthymol blue	4.2	25	2.852	4.2	0.98
Corning titrator	2.3	26	2.815	2.1	0.97
Atomic absorption	2.6	23	2.827	2.6	0.98
Iron (umol/L)					
Overall	12.9	232	19.2	10.6	1.08
Manual/discrete ferrozine	13.5	107	19.1	11.4	1.04
Continuous flow ferrozine	8.0	63	19.3	6.9	1.12
Bathophenanthroline	23.2	21	18.6	12.1	1.06
Continuous flow TPTZ	5.2	20	19.1	9.1	1.06

Table 15.5 Comparison of method-related data for hormones for heat-treated batch and lot HIQC/12. Ratios are mean for heat-treated batch divided by mean for lot HIQC/12; asterisks denote ratios differing by >10% from that for the overall means

	HIQC/12	Heat-treated batch			
	CV	n	Mean	CV	Ratio
Thyroxine (nmol/L)					
Overall	8.8	135	145.4	9.8	1.17
In-house PEG	11.1	23	136.4	10.5	1.10
In-house double Ab	2.5	13	148.6	11.5	1.21
In-house double Ab/PEG	8.4	12	152.0	6.9	1.19
Amersham	7.7	11	136.8	6.6	1.15
NETRIA	6.8	11	152.2	8.0	1.20
Triiodothyronine (nmol/L)					
Overall	12.4	99	2.6	13.7	0.87
Amersham	8.3	33	2.4	6.0	0.86
Excluding Amersham	11.9	54	2.8	14.4	0.90
In-house double Ab	8.8	15	2.8	12.3	0.90
TSH (mU/L)					
Overall	30.6	200	1.8	33.6	0.95
Monoclonal IRMA	15.3	67	1.5	16.2	0.88
RIA	26.5	64	2.3	27.9	1.00
Serono MAIA IRMA	11.5	57	1.6	10.5	0.94
Amersham RIA	20.3	37	2.2	13.8	0.92
In-house RIA	48.5	24	2.6	52.5	1.18*
Corning IRMA	33.1	21	2.2	26.8	1.05*
Boots-Celltech IRMA	9.2	15	1.7	9.5	0.94
NETRIA IRMA	13.2	11	1.8	14.6	1.05*

Table 15.6 Comparison of method-related data for steroid hormones for heat-treated batch and lot HIQC/12. For explanation see Table 15.5

	HIQC/12	Heat-treated batch			
	CV	n	Mean	CV	Ratio
Cortisol (nmol/L)					
RIA	12.4	166	946	13.1	1.04
Fluorimetric	6.5	16	880	9.8	1.01
Amerlex	12	49	947	18	0.98
DPC	11	24	866	9	1.02
Farmos	10	18	1049	18	1.26*
In-house direct	11	20	982	35	1.12
Becton-Dickinson	9	11	861	9	1.02
Abbott TDx	8	11	1072	12	0.98
Travenol	6	8	1028	8	1.15*
Serono	20	6	1199	21	1.30*
NEN	7	6	793	22	0.87
Corning	4	5	1057	9	1.06
Oestradiol (pmol/L)					
Overall	19.3	47	530	15.8	0.87
Steranti/EIR	15	17	540	13	0.82
DPC	16	13	481	9	0.93
In-house extraction	23	10	558	22	0.76*
Sorin	9	5	642	9	1.10*
Progesterone (nmol/L)					
Overall	17.7	93	35.8	24.1	1.38
DPC	9	36	30.6	9	1.33
In-house direct	17	15	40.7	36	1.47
Amerlex-M	6	12	36.5	10	1.30
NETRIA pH 10	27	7	64.9	35	1.65*
Gamma-B	7	7	41.5	7	1.46
In-house extraction	13	4	33.1	19	1.28
Testosterone (nmol/L)					
Overall	18.3	48	9.7	16.1	1.17
Extraction	15.8	32	9.4	20.4	1.07
Direct	17.8	16	10.6	20.9	1.39*
In-house extraction	24	22	9.6	17	1.08
Gamma-B	19	13	11.4	20	1.44*
STRIA extraction	8	7	9.7	13	1.10
DPC	8	6	9.3	-	1.30*

15.2.4 Appraisal of the results

Materials used for calibration, for accuracy control in internal quality control or in EQA must have assigned values which are consistent with the methods used. They must either be commutable with fresh clinical specimens (see Chapters 12-14) or have method-specific assigned values. The policy of the UKEQAS Value Assignment Committee (Appendix IV.2) has been towards the assignment to HIQC materials of a single value for all methods where the similarity of means and SDs for the individual method groups justifies this.

The data presented here indicate that this approach would probably also be applicable to the majority of the analytes in material which had been heat-treated during processing, though careful examination would be required. For the steroid hormones, however, it would almost certainly be necessary to assign values for individual method groups, and the small numbers of results in many of the groups would preclude assignment of a reliable mean value. The usefulness of such material in IQC of methods which show a substantially different behaviour (in terms of a discrepant mean and/or increased variability) might also be questioned.

The results in this study differ from those in separate studies conducted by the hormone UKEQASs. These studies (Groom et al, 1986; Swift et al, 1986) used liquid sera distributed through the UKEQAS, both untreated and after heating. Such treatment did not appear to cause significant differences in behaviour for the thyroid-related hormones (Swift et al, 1986). For the steroid hormones, the only discernible changes were reductions of about

5% in cortisol using fluorimetric, Amerlex and Corning methods, and an increase in progesterone by the NETRIA pH10 assay (Groom et al, 1986).

Hormone immunoassays depend on balancing the conflicting requirements of using reagents such as 8-anilino-1-naphthalene sulphonic acid to displace completely the hormone from its serum carrier protein while maintaining binding of both hormone and labelled hormone to antibody. Such assays may not be robust and are therefore liable to perturbation by any changes in the system. Partial denaturation of the carrier protein would be one such change, and both sex hormone binding globulin and thyroxine binding globulin (TBG) are potentially unstable at 60^{o}C. The findings for lot K466610 may thus be explicable in terms of local heating above 60^{o}C at the point of steam injection into the vessel jacket. This is supported by the observation of a greater apparent free thyroxine concentration relative to total thyroxine in lot K466610 than in the other HIQC materials (Table 15.7), reflecting possible degradation of TBG.

15.3 Studies on lot HIQC/13

This batch was prepared using procedures similar to those for lot K466610, except that the bulk serum was heated at 56^{o}C for 1 h rather than 30 min. Distributions through the same UKEQASs were made, in September 1986.

The results of both the evaluation and UKEQAS distributions resembled closely those for lot K466610. Thus the only major differences in behaviour from untreated serum were for the steroid hormone assays; poor interlaboratory agreement for testosterone may have masked any effects of the heat-treatment.

Table 15.7 Total and free thyroxine content for heat-treated batch and lots HIQC/9 to HIQC/12, and lot HIQC/13 (also heat-treated)

	Total T4 (nmol/L)	Free T4 (pmol/L)	Free/total (%)
Heat-treated batch	154	60.6	0.039
HIQC/9	71	13.8	0.019
HIQC/10	100	21.5	0.022
HIQC/11	137	28.1	0.021
HIQC/12	127	23.4	0.018
HIQC/13	121	54.8	0.045

These problems prevented assignment of RIA ALTMs for the steroid hormones. The results obtained by participants using 6 commercial kits were discrepant for between one and three of these hormones (four kits for cortisol and for oestradiol, three for progesterone). These discrepancies were in general agreement with the data in Table 15.6.

The results of total and free thyroxine assays on this batch are also given in Table 15.7, demonstrating that the prolonged heating may have caused a slight further degradation of TBG.

15.4 Summary

Studies using EQAS data confirm that heat-treatment (at 56^{o}C for 30 min or 1 h) prior to lyophilisation of human serum has only minor effects on its behaviour for assays of inorganic, organic and enzyme constituents. There are similarly no effects which limit its usefulness with respect to assays of TSH and of total T4 and T3.

Changes are seen, however, in the intermethod differences for steroid hormones from those for both lyophilised and liquid native sera, and for heat-treated liquid sera. These differences preclude assignment of a single value for all methods, though the majority of UK laboratories use methods for which a useful value might be assigned. The material may still be useful for internal QC of apparently discrepant methods, using an appropriate assigned value.

Chapter 16:

GENERAL DISCUSSION

16.1 The contribution to patient care of external quality assessment in clinical chemistry

The central position of quality assurance in maintaining and improving the reliability of laboratory investigations in the diagnosis and monitoring of patients is now well-established. Clinical chemical investigations play an ever-increasing role in the detection, diagnosis and monitoring of disease and ill-health, and must be of excellent quality for the avoidance of suffering and provision of efficient and economic health care.

EQA is an essential aspect of quality assurance, being complementary to internal quality control procedures. This has been reflected in the introduction of many local, regional, national and supra-national schemes, and EQA is now practised almost universally in developed countries. The development from initial restricted EQA surveys, accompanied by the introduction of IQC, to national schemes was outlined in Chapter 1.

16.1.1 The relationship of analytical quality to patient care

Intuitively, it appears obvious that patient care will benefit from all improvements in the quality of analytical performance. Consideration (section 2.2.1) of attempts to differentiate between populations of individuals, eg normal and diseased groups, confirmed that reductions in analytical variance will improve test sensitivity, specificity and efficiency.

More recently this view has been challenged (eg Subcommittee on Analytical Goals, 1979; Fraser, 1983), with the counter-proposal

that further improvement is unnecessary when 'analytical goals' have been satisfied (section 2.2.2). Such goals are based upon biological variation, but remain arbitrary in their relation to it. Furthermore, application to monitoring intra-individual changes at elevated analyte concentrations in disease states may yield more stringent goals for some analytes. Comparison of such goals with performance achieved (sections 2.2.2 and 9.6) also indicated that they are currently not generally met, and analytical variance is not negligible relative to biological variation.

16.1.2 Consideration of the requirements, applications and benefits of external quality assessment

Despite wide application, EQA scheme design does not seem to be based on logical principles. Though their application in regulating laboratory performance is obvious, the potential roles of EQA schemes in the assessment of interlaboratory agreement, in the assessment of analytical methods, in the assessment of individual laboratory performance and in the assessment of quality control materials, have not been widely appreciated.

These applications of external quality assessment schemes and their benefits have therefore been examined in this thesis, including consideration of the basic requirements of scheme design. The main question addressed is: **What has external quality assessment to contribute to the scientific development of clinical chemistry and to patient care, besides the 'policing' of laboratory performance?**

16.2 EQA in the assessment of interlaboratory agreement

16.2.1 EQA in assessing the state of the art

Chapter 2 describes the role of EQA surveys in providing an assessment of the state of the art, ie the prevailing standard of performance, without necessarily committing the organisers to establishing a regular EQAS. The use of the information gathered in such surveys in determining the most appropriate further action is summarised in section 2.5.

Many EQASs (the requirements of which are discussed below) have evolved from initial exploratory surveys, as described in section 2.4 for assays which effectively determine clinical management where there was clear evidence of unsatisfactory performance.

Section 2.3 exemplified the application of EQA surveys to extra-laboratory assays, which are playing an increasing role in patient management. These surveys demonstrated clearly that assay quality was generally unsatisfactory. Here, however, EQA alone is unlikely to bring about improvement in the absence of adequate quality assurance and IQC procedures, and the importance of involvement in QA procedures by local laboratory staff was emphasised in section 2.3.4.

16.2.2 Fundamental requirements of scheme design

The primary consideration in EQAS design is to ensure the confidence of participants in the scheme's assessment of their performance, and thus increase the likelihood of their taking action on the conclusions drawn.

Chapter 3 emphasises the importance of certain elements and factors in interlaboratory survey design (summarised in section 3.5) which are important in engendering and maintaining this

confidence.

Though many schemes have broad similarities in design, there are individual differences. It is concluded that a scheme design employing frequent distributions of single specimens is satisfactory, provided data are cumulated over time to provide a more robust assessment of performance and changes in performance. The purposes of scheme operation also differ, as discussed in section 16.4 below.

Several aspects merit more detailed appraisal. Scoring systems were therefore examined in Chapters 4 and 9, the validity of consensus values in Chapter 5, and the properties of QC materials in Chapters 12 to 14.

16.2.3 The application of scoring systems in EQA

EQA schemes and surveys generate large amounts of numeric performance data which often lack impact, and much of the information content may be lost on the scheme participants. Data reduction is therefore essential, and Chapter 4 demonstrates the major contribution of scoring systems to the effectiveness of EQA in performance assessment.

Though all such evidence is circumstantial and a causal relationship cannot be established, experience in UKEQASs strongly indicates that scoring is linked with overall improvement. The similar patterns of rapid and sustained improvement following the introduction of scoring into schemes with little evidence of previous improvement were described in section 4.5.

Though such systems originated to simplify the assessment of performance for individual participants (Chapter 9), Chapter 4

describes their contribution to the assessment of:

- the state of the art
- the effects of analytical procedure
- progress over time
- relative performance in different schemes and countries

The unsuitability of many of the systems devised for such use was outlined in section 4.9. To be fully useful in the assessment of progress and for comparison of performance among schemes, systems must be independent of other participants' performance (section 4.7.2).

The Variance Index system (defined fully in Appendix II) appears to overcome many of these problems. Cumulation of VISs (as estimators of total error) over time gives performance indicators for individual analytes (MRVISs), and over analytes provides an index of overall performance (OMRVIS). These running scores are robust and readily interpretable, by both scheme participants and organisers, in terms of the state of the art (section 9.2).

Potential problems in application of the VI system must also be considered. Section 4.8 concluded that readjustment of CCVs to restore the 'common currency' may be necessary despite the practical difficulties entailed. The data presented in Chapter 6 confirm the need to maintain a balance of analyte concentrations in the specimens distributed; the effects of failure so to do are shown in Figure 11.1 (section 11.2.1).

Overall, application of the VI system has proved of great value in many schemes, provided care is taken to avoid these potential problems.

16.2.4 Consensus values in EQA

Chapter 5 examines the validity of consensus values (the mean of results obtained by all participants or by those using a particular method), which have long been used in EQA as target values. The deficiencies of alternative value assignment procedures using reference laboratories were detailed in section 3.4.3.

Consensus values have been shown to be highly reproducible on repeated EQAS distribution of the same material (section 5.2) and may therefore be considered reliable. Are they accurate, however? This is of great importance, since the primary aim of EQA is numerical comparability of results among laboratories through the elimination of bias.

Close agreement between consensus values obtained in even completely independent national EQA schemes (NEQASs), has been shown (sections 5.3 and 5.5). More importantly, comparisons using a variety of specimens showed close agreement of consensus values with definitive methods for many organic and inorganic analytes (section 5.4.1). Longitudinal study of values assigned using reference laboratories and methods confirmed that UKEQAS consensus values have not been subject to any 'drift' away from accuracy (section 5.4.2).

For practical reasons such studies were not systematic. The novel design of the WRL International Intercomparison Scheme (WIIS; section 5.6) was therefore used to address the question of comparability among countries. This feasibility study showed good agreement in most cases, especially for the most developed countries. UKEQAS values were confirmed to be in good agreement with the international consensus from 30 NEQASs.

Attempts to adjust this comparison to take account of participants' bias in their NEQAS led, however, to divergence in some cases. Possible reasons were discussed in section 5.6.4, including possible influence of participation in the WIIS on laboratories' performance. This was not intended, but attests to the power of EQA in improving interlaboratory comparability.

On balance, consensus values offer the best practical approach to providing target values in an EQAS. They must, however, be kept under continual scrutiny for the potential problems of inaccuracy and 'drift'.

16.2.5 The influence of analyte concentration on interlaboratory agreement

'Precision profiles', reflecting dependence of analytical precision upon analyte concentration, are useful in IQC (eg Ekins, 1983). Similar relationships would be expected in EQAS data, but have not been generally reported.

Studies in UKEQASs (sections 6.2 and 6.3), confirmed such dependence for many analytes, with interlaboratory agreement best at intermediate concentrations. The complex interrelationships between concentration-dependence and interlaboratory agreement are summarised in section 6.4.

Such concentration-dependence is important to the design of reliable EQA schemes, since its effects on performance assessment must be minimised, eg by confining scoring to a band of concentrations. These studies also underline the need to provide a balance of concentrations in the materials distributed, as discussed in section 16.2.3 above. Characterisation of such

dependence is a prerequisite for the assessments of specimen properties described in Chapter 14.

16.2.6 EQA of non-quantitative investigations

The requirements for EQA of semi-quantitative and qualitative investigations differ from those described in Chapter 3 for quantitative assays. They are considered in Chapter 7, using blood phenylalanine assay in phenylketonuria (PKU) screening and the chromatography of urinary aminoacids as examples. The differences are summarised in section 7.4.

It is important that the scheme design approaches the routine clinical situation, to avoid the appearance of an 'artificial' exercise. Performance assessment must also address the essentials of the investigation; patients are most at risk from a laboratory which fails in the basic objective, eg does not recognise that a chromatographic pattern is abnormal.

Experience shows that EQASs for such investigations can also be successful in improving the reliability of patient care, though variation between surveys in the specimens distributed may complicate performance assessment.

16.3 EQA in the assessment of analytical methods

All clinical chemists face at some time the problem of selecting an analytical procedure (method principle and instrument) for routine use. Section 8.1 outlined the deficiencies of common information sources other than EQA data.

Section 8.2 demonstrated the use of information from EQA, using as example the selection of a method for serum calcium assay. The conclusions were confirmed by re-examination of subsequent data.

The assessment of aspects of performance other than interlaboratory agreement, such as accuracy, was also considered (section 8.4.1). In schemes subject to matrix effects, however, less confidence can be attached to conclusions regarding relative bias unless the specimens are demonstrated to have properties resembling those of clinical specimens (ie to be commutable, as discussed in Chapters 12-14).

Though limited by the discrimination in the classification used, EQAS data have been shown to be invaluable in assessing the routine performance of methods already in general use.

16.4 EQA in the assessment of individual laboratory performance

The main objective of EQASs is the improvement of interlaboratory agreement, to facilitate better patient care. The role of such schemes should be primarily educational in intent, and the scheme design should emphasise this aspect in encouraging participants to appraise and improve when necessary their own performance.

The 'policing' of performance undertaken in licensing schemes has therefore not been considered. The purpose of such schemes is to determine the acceptability of laboratories for registration or reimbursement, the main examples being those organised in GFR (Bundesärztekammer, 1971) and by CDC in the USA (Boone, 1984). The adverse consequences of failure to satisfy the criteria imposed also encourage attempts to improve the performance assessment through special treatment of specimens (Rumley and Roberts, 1982; Rowan et al, 1984), and mutual confidence between scheme organisers and participants is lacking. Consequently distributions tend to be infrequent, with performance assessment by a 'pass/fail' system. The data are thus of very limited assistance to participants, who need an objective and detailed

appraisal of their performance.

16.4.1 Scoring systems for performance assessment

Scoring systems were introduced into EQASs as a powerful means of data reduction to facilitate participants' interpretation of information derived from EQA. The scoring system must allow participants to compare their current performance with that of other laboratories and with their own past performance; scores must be independent of the standard of other participants' performance, as discussed in section 16.2.3.

For optimal efficiency in interpretation the system should allow stepwise appraisal by participants of increasingly detailed information as appropriate. This was demonstrated in section 9.2 using the VI system and an example from the UKEQAS for General Clinical Chemistry, and summarised in section 9.7.

Section 9.5 outlined the use of scoring systems to detect participants experiencing performance difficulties, so assistance may be offered to them (eg Browning, 1984). The problems of excessive competition in scoring systems were also considered (section 9.3.1).

The relative merits of systems such as VI based upon the state of the art and of alternatives based on 'medical requirements' and 'analytical goals' were discussed in section 9.6. The most practical and effective approach is to use a system such as VI and take clinical requirements into account when interpreting scores.

16.4.2 Use of graphical data presentations in EQA

Chapter 10 demonstrates how graphical presentations assist both

participants and scheme organisers in interpreting EQA data.

The limitations of simple widely-used presentations such as histograms of results were considered (section 10.3). Graphical displays which facilitate assessment of progress (section 10.4) and the investigation of factors contributing to overall error for an individual analyte (section 10.5) were discussed in detail.

For resource reasons, as outlined in section 10.6, graphical presentations have been little used. Newer technology now offers the prospect of cheaper and faster systems with high quality output, combining graphical and text elements to increase the application of EQA data by participants.

16.4.3 Investigation of factors affecting performance

As shown for analytical methods in Chapter 8 and section 16.3 above, EQA data are useful in identifying factors associated with good or poor performance. These are not the only factors involved, however, and the Nuffield survey (section 11.1.1; Maclagan et al, 1980) provides a good example of a systematic study of others.

The association of large laboratory size (expressed as annual workload) with performance was examined using UKEQAS data in section 11.2, with particular consideration of the problems arising from decreasing workloads and the consequent smaller and less frequent assay batches.

The laboratory's organisation and general attitude to maintaining professional standards is also important. Section 11.3 considers the performance of UK private sector and Irish laboratories in relation to the "small laboratory syndrome" (Browning, 1984).

16.5 EQA in the assessment of quality control material suitability

Section 12.1 outlines the differing performance requirements for quality control materials (QCMs) for use in different circumstances. Thus materials for EQA and for calibration must be commutable with (ie show intermethod differences similar to those for) clinical specimens (Fasce et al, 1973; Rej et al, 1984; Moss et al, 1985), unless designated values are obtained by the same method. Problems relate to the interaction of differences in QCM composition with less than perfectly specific analytical methods, and are usually termed 'matrix effects'.

16.5.1 Suitability of QC materials for EQA

The problems in selecting suitable QCMs for HDL cholesterol assay and the need for interlaboratory comparability were outlined in section 12.2. These stimulated the formulation of a protocol to assess QCM suitability for use in EQA, described in section 12.2.1. The conclusions were validated in an EQA survey (section 12.2.2).

The inhomogeneity in EQA data for sodium and potassium using direct ISE methods, reflecting susceptibility to matrix effects and a lack of commutability, was described in section 12.4. It was concluded that commutable QCMs could not be identified a priori. Application of a protocol similar to that applied for HDL cholesterol might, however, thus permit performance assessment for direct ISEs.

Such a protocol could form the basis of a general procedure for assessing the suitability of QCMs for use in EQA.

16.5.2 Calibration effects in EQA

Calibration is a further major factor affecting laboratory performance, and Chapter 13 describes the use of EQA surveys to examine the effects on interlaboratory agreement of using a common calibration material.

Most studies demonstrated substantial improvement. The degree of improvement was directly related to the diversity in methods and calibration procedures, with the most marked improvement seen within the most heterogeneous groups (section 13.2). Indeed preliminary results from a recent European survey of specific protein assays (data not presented) indicate that differences among the commercial calibrants used largely account for the differences observed among national consensus values for these assays, since these differences can effectively be abolished by 'calibration' of the EQA data using a common assigned value.

Where analytical procedures have differing specificities, any differences in properties (ie lack of commutability) between the calibrant and clinical specimens become important. The resultant effects on improvement were demonstrated for urinary total oestrogen assay in section 13.3.1.

Problems of non-commutability and the limitations of calibration for enzyme activity estimations were considered in section 13.5, though in most cases EQA studies confirmed the favourable effects of calibration. Use of suitable calibration materials (Moss et al, 1985), preferably but not necessarily in combination with method standardisation, thus offers a means to improve greatly the unsatisfactory numerical concordance between results obtained in different laboratories and consequently contribute to more reliable patient care.

16.5.3 Effects of material source and processing

Chapter 14 demonstrates the application of EQA data in examining properties of the materials distributed, provided the concentration-dependence of interlaboratory agreement has first been characterised (Chapter 6 and section 16.2.5 above).

The main study was of the relative suitability of materials based upon human and animal serum (section 14.3). It was concluded (section 14.3.3) that animal-based materials may be used with equally satisfactory results for calibration, IQC and EQA of the non-protein analytes studied, with important economic and ethical consequences for laboratory practice.

EQA data were also used to assess the effects of manufacturing procedures on QCM behaviour (section 14.4), and failed to reveal major differences among materials.

The advantages of materials including ethylene glycol were described in section 14.5, with further characterisation of their properties through EQA distribution. These studies also confirmed the efficacy of existing procedures to minimise changes in enzyme activity following reconstitution of lyophilised sera.

16.5.4 Effects of heat-treatment

The heat-treatment of serum to inactivate viruses in pooled human serum exemplifies a non-physiological QCM production process which is necessary in some circumstances (section 15.1).

The studies using EQAS data described in Chapter 15 confirmed the minor effects of heat-treatment prior to lyophilisation upon QCM behaviour for most analytes. An explanation of the discrepant findings for steroid hormones was discussed in section 15.2.4.

16.6 Conclusion

The role of external quality assessment in 'policing' laboratory performance through the operation of licensing schemes is obvious. Apart from this, however, its contribution is mediated through the four aspects of assessment of interlaboratory agreement, assessment of analytical methods, assessment of individual laboratory performance, and assessment of quality control materials.

This thesis has presented evidence of the requirements, applications and benefits in each of these aspects, and thus established the importance of external quality assessment to the scientific development of clinical chemistry and to patient care.

REFERENCES

Advisory Committee on Dangerous Pathogens (1986) LAV/HTLV III - the causative agent of AIDS and related conditions - Revised guidelines. London: DHSS

Andrews S, Cooke PR, Went J (1983) A quality-control scheme for ward-based blood glucose estimations. Med Lab Sci **40**: 279-282

Archibald RM (1950) Criteria of analytical methods for clinical chemistry. Anal Chem **22**: 639-642

Association of Clinical Biochemists (1980) Proposed methods for determination of some enzymes in blood serum. NewsSheet Assoc Clin Biochemists **202**: 1s-39s

Bacchus RA, Chang PC, Bullock DG, Whitehead TP (1982) The organization and evolution of the Middle East external quality assessment scheme in clinical chemistry. In: Proceedings of 4th International Symposium on Quality Control, Osaka June 1981, International Congress Series no 585. Amsterdam, Holland: Excerpta Medica; 260-269

Bacchus RA, Bullock DG, Noy GA, Whitehead TP (1987) The Middle East External Quality Assessment Scheme for Clinical Chemistry. Commun Lab Med (in press)

Bacon RRA, Hunter WM, McKenzie I (1983) The UK external quality assessment schemes for peptide hormones: objectives and strategy. In: Hunter WM, Corrie JET eds. Immunoassays for clinical chemistry. 2nd edn. Edinburgh: Churchill Livingstone; 669-679

Ballantyne FC (1984) Role of the clinical biochemistry laboratory in the assessment of dyslipoproteinaemias. Ann Clin Biochem **21**: 166-175

Barnett RN (1968) Medical significance of laboratory results. Am J Clin Pathol **50**: 671-676

Becktel JM, Martinek RG, Morrissey RA (1973) Use of data from patients in assessing interlaboratory proficiency and for quality control. Clin Biochem **6**: 53-59

Belk WP, Sunderman FW (1947) A survey of the accuracy of chemical analyses in clinical laboratories. Am J Clin Pathol **17**, 853-861

Belsey R, Baer D, Sewell D (1986) Laboratory test analysis nearer the patient. Opportunities for improved clinical diagnosis and management. J Am Med Assoc **255**: 775-786

Benenson AS, Thompson HL, Klugerman MR (1955) Application of laboratory controls in clinical chemistry. Am J Clin Pathol **25**: 57-584

Blijenberg BG, Brouwer HJ, Roetering HA, Leijnse B (1984) Surveys of neonatal bilirubin: an evaluation. J Clin Chem Clin Biochem **22**: 609-612

Boerma GJM (1979) Studies in standardization - serum cholesterol analysis performed for epidemiological investigations. PhD thesis; 170pp. Rotterdam, The Netherlands: Erasmus University

Bold AM, Browning DM (1975) Quality control of thyroid function tests in vitro. J Clin Pathol **28**: 234-238

Boone DJ (1984) Proficiency testing summary analysis. Chemistry profile 1984 II. Atlanta, Georgia, USA: US Dept of Health & Human Services

Bowers GN, McComb RB (1984) A unifying reference system for clinical enzymology: aspartate aminotransferase and the International Clinical Enzyme Scale. Clin Chem **30**: 1128-1136

Bowers GN, Burnett RW, McComb RB (1975) Preparation and use of human serum control materials for monitoring precision in clinical chemistry. Clin Chem **21**: 1830-1836

Bowers GN, Bergmeyer HU, Hørder M, Moss DW (1979) IFCC methods for the measurement of catalytic concentration of enzymes, Part 1. General considerations concerning the determination of the catalytic concentration of an enzyme in the blood serum or plasma of man. Clinica Chim Acta **98**: 163F-174F

Bowyer RC, Geary TD, Penberthy LA, Thomas DW (1981) An Australian quality control scheme for clinical chemistry. Clin Biochemist Newsletter **60**: 43-49

Bretaudiere JP, Cherruau B, Boulu RD, Bailly M (1974) Are quality control sera reliable for assessment of laboratory accuracy in collaborative surveys? Consequences for laboratory accredition. Clin Chem **20**: 877

Bretaudiere J-P, Dumont G, Rej R, Bailly M (1981a) Suitability of control materials. General principles and methods of investigation. Clin Chem **27**: 798-805

Bretaudiere J-P, Rej R, Drake P, Vassault A, Bailly M (1981b) Suitability of control materials for determination of alpha-amylase activity. Clin Chem **27**: 806-815

Broughton PMG, Eldjarn L (1985) Methods of assigning accurate values to reference serum. Part 1. The use of reference laboratories and consensus values, with an evaluation of a procedure for transferring values from one reference serum to another. Ann Clin Biochem **22**: 625-634

Broughton PMG, Maas AHJ (1984) Recent developments with ion selective electrodes. IFCC News **1984/3**: 4-5

Broughton PMG, Gowenlock AH, McCormack JJ, Neill DW (1974) A revised scheme for the evaluation of automatic instruments for use in clinical chemistry. Ann Clin Biochem **11**: 207-218

Broughton PMG, Bullock DG, Carter TJN (1981) Laboratory management, quality assurance, and reference values. Clin Biochem Rev **2**: 1-30

Broughton PMG, Smith SCH, Buckley BM (1985) Calibration of direct ion-selective electrodes for plasma Na^+ to allow for the influence of protein concentration. Clin Chem **31**: 1765

Browning DM (1984) The role and activities of the National Quality Assurance Advisory Panel in clinical chemistry. In: Quality assurance and control in clinical laboratories. Farr AD ed. London: Institute of Medical Laboratory Sciences

Browning DM, Bullock DG (1987) The quality of extra-laboratory assays: evidence from external quality assessment surveys. Ann Clin Biochem **24 S1**: 171-172

Browning DM, Cowell DC, Kilshaw D, et al (1984) Clinical chemistry equipment outside laboratories. Med Lab Sci **41**: 99-107

Browning DM, Hill PG, Vazquez R Olazabal DA (1986) Preparation of stabilized liquid quality control serum to be used in clinical chemistry. (LAB/86.4) Geneva, Switzerland: World Health Organization

Buckley BM, Broughton PMG, Russell LJ, Carter TJN (1984) New ways with old ions. Ann Clin Biochem **21**: 75-77

Bullock DG (1985) UKEQAS for General Clinical Chemistry and sub-schemes. Review: April 1985. Commun Lab Med **1**: 50-52

Bullock DG (1987) Methods for paracetamol assay studied through external quality assessment scheme data. Ann Clin Biochem **24 S1**: 42-43

Bullock DG, Wilde CE (1985) External quality assessment of urinary pregnancy oestrogen assay: further experience in the United Kingdom. Ann Clin Biochem **22**: 273-282

Bullock DG, McSweeney FM, Saidi H, Whitehead TP (1979) Alkaline phosphatase activity in lyophilised quality control sera. Ann Clin Biochem **16**: 271-274

Bullock DG, Carter TJN, Cosgrove CM, Hughes SV (1980a) An assessment of the Corning electrophoretic procedure for high density lipoprotein cholesterol estimation. Ann Clin Biochem **17**: 148-152

Bullock DG, Carter TJN, Hughes SV (1980b) Applicability of various quality-control sera to assay of high-density lipoprotein cholesterol. Clin Chem **26**: 903-907

Bullock DG, Groom GV, Swift AD (1986a) The influence of heat treatment during the preparation of lyophilised human serum on interlaboratory agreement for clinical chemistry analytes. Commun Lab Med **2**: 98-102

Bullock DG, Moss DW, Whitehead TP (1986b) External quality assessment of serum enzyme activity assays and the effect of calibration on interlaboratory concordance. Ann clin Biochem **23**: 577-584

Bullock DG, Smith NJ, Whitehead TP (1986c) External quality assessment of assays of lead in blood. Clin Chem **32**: 1884-1889

Bundesärztekammer (1971) Richtlinien der Bundesärztekammer zur Durchführung von Massnahmen der statistischen Qualitätskontrolle und von Ringversuchen im Bereich der Heilkunde. Dt Ärztebl **68**: 2228-2231

Burrin JM, Williams DRR, Price CP (1985) Performance of a quality-assessment scheme for blood glucose meters in general practice. Ann Clin Biochem **22**: 148-151

Burstein M, Scholnick HR, Morfin R (1970) Rapid method for the isolation of lipoproteins from human serum by precipitation with polyanions. J Lipid Res **11**: 583-595

Büttner J, Borth R, Boutwell JH, Broughton PMG, Bowyer RC (1979a) Approved recommendation on quality control in clinical chemistry. Part 1. General principles and terminology. Clinica Chim Acta **98**: F129-F143

Büttner J, Borth R, Boutwell JH, Broughton PMG, Bowyer RC (1979b) Approved recommendation on quality control in clinical chemistry. Part 2. Assessment of analytical methods for routine use. Clinica Chim Acta **98**: F145-F162

Büttner J, Borth R, Boutwell JH, Broughton PMG, Bowyer RC (1980a) Approved recommendation (1979) on quality control in clinical chemistry. Part 3. Calibration and control materials. J Clin Chem Clin Biochem **18**: 855-860

Büttner J, Borth R, Boutwell JH, Broughton PMG, Bowyer RC (1980b) Approved recommendation on quality control in clinical chemistry. Part 6. Quality requirements from the point of view of health care. J Clin Chem Clin Biochem **18**: 861-866

Büttner J, Borth R, Broughton PMG, Bowyer RC (1983a) Approved recommendation (1983) on quality control in clinical chemistry. Part 4. Internal quality control. J Clin Chem Clin Biochem **21**: 877-884

Büttner J, Borth R, Boutwell JH, Broughton PMG, Bowyer RC (1983b) Approved recommendation (1983) on quality control in clinical chemistry. Part 5. External quality control. J Clin Chem Clin Biochem **21**: 885-892

Cali JP, Bowers GN, Young DS (1973) A referee method for the determination of total calcium in serum. Clin Chem **19**: 1208-1213

Caragher TE, Grannis GF (1978) Performance evaluation of multichannel analyzers by use of linearly-related survey specimens. Clin Chem **24**: 403-413

Castelli WP, Doyle JT, Gordon T, et al (1977) Cholesterol and other lipids in coronary heart disease: the Co-operative Lipoprotein Phenotyping Study. Circulation **55**: 767-772

Chambers RE, Whicher JT, Bullock DG (1984) External quality assessment of immunoassays for specific proteins in serum: 18 months' experience in the United Kingdom. Ann Clin Biochem **21**: 246-253

Chambers RE, Bullock DG, Whicher JT (1987) Improved between-laboratory agreement for specific protein assays in serum following introduction of a common reference preparation (SPS-01) demonstrated in an external quality assessment scheme. Clinica Chim Acta **164**: 189-200

Cohn EJ, Strong LE, Hughes WL, et al (1946) Preparation and properties of serum and plasma proteins. IV. A system for the separation into fractions of the protein and lipoprotein components of biological tissues and fluids. J Am Chem Soc **68**: 459-475

Colinet E, Siest G, Moss DW (1986) Reference materials for clinical enzymology: the work of the Community Bureau of Reference of the European Community. Ann Clin Biochem **23**: 361-363

Committee on Enzymes of the Scandinavian Society for Clinical Chemistry and Clinical Physiology (1974) Recommended methods for the determination of four enzymes in blood. Scand J Clin Lab Invest **33**: 291-306

Compton J, Bonderman D, Proksch G, Griep J (1979) HB_SAg, anti-HB_S, anti-HB_C, and anti-HAV in commercial lyophilized quality-control sera. Clin Chem **25**: 1347-1348

Cumming HS, Hazen HH, Sanford AH, et al (1935) The evaluation of serodiagnostic tests for syphilis in the United States: report of results. J Am Med Assoc **104**: 2083-2087

de Leenheer AP, Steyart HLC, Thienpont LMR, Jonckheere JA (1983) External quality control of clinical chemistry laboratories in Belgium. Clinica Chim Acta **133**: 1-14

Desmond FB (1964) A clinical chemistry proficiency survey. N Z Med J **63**: 716-720

DHSS (1986a) HIV (LAV/HTLV III - 'AIDS virus') antibody in diagnostic reagents and quality control and calibration materials. Health Notice HN(86)25. London: DHSS

DHSS (1986b) United Kingdom External Quality Assessment Schemes. Lewis SM, Jennings RD eds. Annual Report of Advisory Committee on Assessment of Laboratory Standards. London: DHSS

Dilena BA, Penberthy LA, Fraser CG (1983) Six methods for determining urinary protein compared. Clin Chem **29**: 553-557.

Dixon K, Northam BE (1970) Quality control using the daily mean. Clinica Chim Acta **30**: 453-461

Drucker RF, Williams DRR, Price CP (1983) Quality assessment of blood glucose monitors in use outside the hospital laboratory. J Clin Pathol **36**: 948-953

Ekins RP (1983) The precision profile: its use in assay design, assessment and quality control. In: Hunter WM, Corrie JET eds. Immunoassays for clinical chemistry. 2nd edn. Edinburgh: Churchill Livingstone; 76-105

Eldjarn L, Broughton PMG (1985) Methods of assigning accurate values to reference serum. Part 2. The use of definitive methods, reference laboratories, transferred values and consensus values. Ann Clin Biochem **22**: 635-649

Epton J (1979) Paracetamol: a report of a regional quality control scheme. Ann Clin Biochem **16**: 265-270

Evans AS, Kozelka FL, Moyer G (1966) Cooperative clinical chemistry evaluation program. I. Results of glucose evaluation. Wisconsin Med J **65**: 157-161

Fasce CF, Rej R, Copeland WH, Vanderlinde RE (1973) A discussion of enzyme reference materials: applications and specifications. Clin Chem **19**: 5-9

Fleck A, Colley C M (1982) Temperatures used in the determination of enzyme activity in clinical biochemistry laboratories in Britain: results of a survey. Ann Clin Biochem **19**: 405-411

Fraser CG (1983) Desirable performance standards for clinical chemistry tests. Adv Clin Chem **23**: 299-339

Fraser CG (1986) Analytical goals for glucose analyses. Ann Clin Biochem **23**: 379-389

Fraser CG, Peake MJ (1980) Problems associated with clinical chemistry quality control materials. CRC Crit Rev Clin Lab Sci **12**: 59-86

Galen RS, Gambino SR (1975) Beyond normality: the predictive value and efficiency of medical diagnoses. New York, USA: John Wiley & Sons

Gaskell SG, Brownsey BG, Groom GV (1984) Analyses for progesterone in serum by gas chromatography/mass spectrometry: target data for external quality assessment of routine assays. Clin Chem **30**: 1696-1700

Georges RJ (1985) Validity of the consensus mean as the target value for a small external quality assessment scheme. Ann Clin Biochem **22**: 283-290

Gerhardt W, Waldenstrøm J, Hørder M, et al (1985) SCE Nordic-amylase method selection and calibration study. Scand J Clin Lab Invest **45**: 397-404

German Society for Clinical Chemistry (1970) Standardisation of methods for the estimation of enzyme activity in biological fluids. J Clin Chem Clin Biochem **6**: 659-660

German Society for Clinical Chemistry (1972) Standardisation of methods for the estimation of enzyme activities in biological fluids. J Clin Chem Clin Biochem **8**: 281-291

Gilbert RK (1975a) Progress and analytic goals in clinical chemistry. Am J Clin Pathol **63**: 960-973

Gilbert RK (1975b) The accuracy of calcium analysis in the United States. Am J Clin Pathol **63**: 974-983

Gilbert RK (1976) A comparison of participant mean values of duplicate specimens in the CAP chemistry survey program. Am J Clin Pathol **66**: 184-192

Glenn GC (1980) The results of analyte enhancement and use of supplied urine protein standard in the CAP urine chemistry survey program. Am J Clin Pathol **74**: 531-534

Goldberg JM (1978) Test precision and high-density-lipoprotein cholesterol. Clin Chem **24**: 2061

Gordon T, Castelli WP, Hjortland MC, Kannel WB, Dawber TR (1977) High density lipoprotein as a protective factor against coronary heart disease. Am J Med **62**: 707-714

Gowenlock AH (1969) Results of an interlaboratory trial in Britain. Ann Clin Biochem **6**: 126-133

Grannis GF (1976) Studies of the reliability of constituent target values established in a large inter-laboratory survey. Clin Chem **22**: 1027-1036

Grannis GF (1979) Performance evaluation by interlaboratory surveys. Clin Chem **25**: 196

Grannis GF, Caragher TE (1977) Quality-control programs in clinical chemistry. CRC Crit Rev Clin Lab Sci **7**: 327-364

Grannis GF, Miller WG (1976) On the design of clinical chemistry quality control sera. Clin Chem **22**: 500-512

Gräsbeck R, Siest G, Wilding P, Williams GZ, Whitehead TP (1979) Provisional recommendation on the theory of reference values (1978) Part 1. The concept of reference values. Clin Chem **25**: 1506-1508

Groom GV (1985a) External quality assessment of hormone assays in the UK. Commun Lab Med **1**: 1-4

Groom GV (1985b) External quality assessment of hormone assays in the UK. Part II: Preparation of samples, data analysis and GC-MS targets. Commun Lab Med **1**: 27-32

Groom GV (1985c) Steroid NEQAS: the performance of commercial kits used according to protocol and with modifications. Commun Lab Med 1: 53-54

Groom GV (1985d) Interpretation of external quality assessment reports. Commun Lab Med 1: 106-117

Groom GV, Adams VM, Groom MA (1986) Effect of heat treatment on serum concentrations of steroid hormones: results of an external quality assessment study. Commun Lab Med 2: 95-97

Hainline A, Cooper GR, Olansky AS, Winn CL, Miller DT (1980) CDC survey of high density lipoprotein cholesterol measurement: a report. Atlanta, Georgia, USA: US Department of Health and Human Services

Hansert E, Stamm D (1980) Determination of assigned values in control specimens for internal accuracy control and for interlaboratory surveys. J Clin Chem Clin Biochem 18: 461-490

Hardy AV (1952) Reliability of laboratory findings. Pediatrics 10: 624-626

Harris EK (1979) Statistical principles underlying analytic goal-setting in clinical chemistry. Am J Clin Pathol 72: 374-382

Hartmann AE, Juel RD, Barnett RN (1981) Long term stability of a stabilised liquid quality control serum. Clin Chem 27: 1448-1452

Haug H, Immich H, von Klein-Wisenberg A, Müller H, Rotzler A, Sieder M (1978) Qualitätskontrolle - Kritik und Verbesserungsvorschläge. Lab Med 2 A+B: 149-153

Hayes M, Ferguson KM, Jeffcoate SL, Bacon RRA, Seth J (1985) Improved performance of plasma gonadotrophin assays using common reagents and assay protocols: evidence from the UK External Quality Assessment Scheme. Ann Clin Biochem 22: 179-184

Health and Safety Commission (1980) Control of lead at work. Approved Code of Practice. London: HMSO

Healy MJR (1979) Outliers in clinical chemistry quality control schemes. Clin Chem 25: 675-677

Healy MJR, Whitehead TP (1980) Outlying values in the National Quality Control Scheme. Ann clin Biochem 17: 78-81

Henry RJ (1959) Use of the control chart in clinical chemistry. Clin Chem 5: 309-319

Henry RJ, Segalove M (1952) The running of standards in clinical chemistry and the use of the control chart. J Clin Pathol 5: 305-311

Hoffmann RG, Waid ME, Henry JB (1961) Clinical specimens and reference samples for the quality control of laboratory accuracy. Am J Med Technol **27**:309-317

Holton JB (1982) Diagnosis of inherited metabolic diseases in severely ill children. Ann Clin Biochem **19**: 389-395

Holtz AH (1959) Klinisch-chemisch enquête-onderzoek in Nederland. Ned Tijdschr Geneesk **103**: 2366

Holtzman C, Slazyk WE, Cordero JF, Hannon WH (1986) Descriptive epidemiology of missed cases of phenylketonuria and congenital hypothyroidism. Pediatrics **78**: 553-558

Hunter WM, McKenzie I (1979) Quality control of radioimmunoassays for proteins. Ann Clin Biochem **16**: 131-146

Isherwood DM, Fletcher KA (1985) Neonatal jaundice: investigation and monitoring. Ann Clin Biochem **22**: 109-128

Jansen AP (1980) Reliability of control sera. Ann Clin Biochem **17**: 69-73

Jansen RTP (1983) Comparison of routine analytical methods in the Netherlands for seven serum constituents using pattern recognition. Ann Clin Biochem **20**: 41-51

Jansen AP (1985) Difficulties in the normalization of aminotransferase measurement with enzyme standards. J Clin Chem Clin Biochem **23**: 209-212

Jansen RTP, Jansen AP (1980) A coupled external/internal quality control program for clinical laboratories in the Netherlands. Clinica Chim Acta **107**: 185-201

Jansen RTP, Jansen AP (1983) Standards versus standardised methods in enzyme assay. Ann Clin Biochem **20**: 52-59

Jansen AP, van Kampen EJ, Leijnse B, Meijers CAM, van Munster PJJ (1977) Experience in the Netherlands with an external quality control and scoring system for clinical chemistry laboratories. Clinica Chim Acta **74**: 191-201

Jansen AP, van Kampen EJ, Meijers CAM, van Munster PJJ, Boerma GJM (1978) Quality control and the quality of commercial test sera. Clinica Chim Acta **84**: 255-258

Jansen RTP, Pijpers FW, de Valk GAJM (1981) A technique for the objective assessment of routine analytical methods in clinical laboratories using pattern recognition. Ann Clin Biochem **18**: 218-225

Jeffcoate S (1981) Efficiency and effectiveness in the endocrine laboratory. London: Academic Press

John WG (1983) Analysis of serum high density lipoproteins. MSc thesis; 186 pp. Birmingham: University of Birmingham

Jordan A (1965) The development of chemical pathology. J Clin Pathol **18**: 274-276

Kenny AP, Eaton RH (1981) Practical guidelines for the preparation of quality control sera for use in clinical chemistry. (LAB/81.4) Geneva, Switzerland: World Health Organization

Knox WE (1972) Phenylketonuria. In: Stanbury JB, Wyngaarden JB, Fredrickson DS eds. The metabolic basis of inherited disease. 3rd edn. New York, USA: McGraw Hill; 266-295

Külpmann WR, Lagemann J, Sander R, Maibaum P (1985) A comparison of reference method values for sodium, potassium and chloride with method-dependent assigned values. J Clin Chem Clin Biochem **23**: 865-874

Lawson NS, Haven GT, Ross JW (1980) Regional quality assurance for the 1980's, current status and future directions. Am J Clin Pathol **74**: 552-559

Lawson NS, Haven GT, Williams GW (1982) Analyte stability in clinical chemistry quality control materials. CRC Crit Rev Clin Lab Sci **17**: 1-50

Leblanc A, Woodford FP, Gardner PS, et al (1985a) Standard for quality assurance. Part 1: Terminology and general principles. Beckenham, Kent: ECCLS

Leblanc A, Woodford FP, Gardner PS, et al (1985b) Standard for quality assurance. Part 2: Internal quality control in microbiology. Beckenham, Kent: ECCLS

Leblanc A, Woodford FP, Gardner PS, et al (1985c) Standard for quality assurance. Part 3: External quality assessment in microbiology. Beckenham, Kent: ECCLS

Leclercq R (1975) Interlaboratory quality control using the daily means of patients' results. In: Anido G, van Kampen EJ, Rosalki SB, Rubin M eds. Quality control in clinical chemistry VI, 21-40. Berlin, GFR: W de Gruyter

Legg EF, Hurrell AE (1984) External quality assessment of quantitative urinary analysis. Ann Clin Biochem **21**: 491-493

Lever M, Munster D (1979) Performance evaluation by interlaboratory surveys. Clin Chem **25**: 194-196

Lever M, Munster DJ, Walmsley TA, Stewart AW (1981) Analytical errors in clinical laboratories as assessed by an interlaboratory survey. Ann Clin Biochem **18**: 28-36

Levey S, Jennings ER (1950) The use of control charts in the clinical laboratory. Am J Clin Pathol **20**: 1059-1066

Lewis SM (1984) The principles and methods of quality assurance in haematology. (LAB/84.3) Geneva, Switzerland: World Health Organization

Lewis SM, Burgess BJ (1969) Quality control in haematology: report of interlaboratory trials in Britain. Brit Med J **4**: 253-256

Lewis SM, Cavill I, Goguel A, et al (1986) Standard for quality assurance. Part 5: External quality assessment in haematology. Beckenham, Kent: ECCLS

Ley DCH, Ezer S (1974) The development of an interlaboratory profiency testing program for the Province of Ontario. I. A preliminary survey of clinical chemistry. Clin Biochem **7**: 223-238

Limonard CBG (1979) A quality control program to evaluate accuracy and precision of clinical chemistry determinations. Clinica Chim Acta **95**: 353-367

Logan JE (1983) Revised recommendation (1983) on evaluation of diagnostic kits. Part 2. Guidelines for the evaluation of clinical chemistry kits. J Clin Chem Clin Biochem **21**: 899-902

Maclagan NC, Kind PRN, Daly JF, et al (1980) Factors affecting analytical performance in clinical chemistry laboratories. Report of Working Party. London: Nuffield Provincial Hospital Trust

Marchandise H, Colinet E (1983) Assessment of methods of assigning certified values to reference materials. Fresenius Z Analyt Chem **316**: 669-672

Marks V (1983) Clinical biochemistry nearer the patient. Brit Med J **286**: 1166-1167

Marsters RW (1949) A survey of the accuracy of Rh antibody titrations in several hospital laboratories. Am J Clin Pathol **19**: 1032-1038

Maurukas J (1975) Biologic composition for use as a reference control in diagnostic analysis. US patent 3876375, April 8, 1975

McComb RB, Bowers GN (1985) Alkaline phosphatase and the International Clinical Enzyme Scale. Am J Clin Pathol **84**: 67-73

McSweeney FM, Bullock DG, Gregory A, Whitehead TP (1979) An interlaboratory survey of hydrogen ion and blood gas determinations. Ann Clin Biochem **16**: 249-253

Medical Research Council Steering Committee for the MRC/DHSS Phenylketonuria Register (1981) Routine neonatal screening for phenylketonuria in the United Kingdom 1964-78. Brit Med J **282**: 1680-1684

Merritt BR, McHugh RB, Kimball AC, Bauer H (1965) A two-year study of clinical chemistry determinations in Minnesota hospitals. The effect of survey participation and laboratory consultation upon accuracy and reliability. Minnesota Med **48**: 939-956

Milford Ward A, White PAE, Thompson RA, et al (1984) Preparation of a calibration material for specific protein assay (SPS-01). Ann Clin Biochem **21**: 254-256

Mitchell JA (1947) Control of the accuracy and precision of industrial tests and analyses. Anal Chem **19**: 961-967

Mollison PL, Cutbush M (1954) Haemolytic disease of the newborn. In: Gairdner D ed. Recent advances in paediatrics. London: Churchill Livingstone; 110-132

Moss DW, Brettschneider H, Bullock DG, et al (1985) ECCLS Proposed Standard for Enzyme Reference Materials, First Draft. Beckenham, UK: ECCLS

Oakey RE (1980) Oestrogen determinations in urine from pregnant women: a review of six years' quality assessment in the United Kingdom. Ann Clin Biochem **17**: 311-314

Okuda K (1984) Guidelines (1984) for listing spcifications of clinical chemical analysers. In: Saris N-E ed. IFCC recommendations. Vol 1 1978-1983. Berlin, GFR: W de Gruyter; 109-114

Passing (1981a) The inadequacy of normal distribution models for the establishment of assigned values in control sera. J Clin Chem Clin Biochem **19**: 1145-1151

Passing (1981b) Comparison of three distribution-free procedures in the establishment of assigned values in control sera. J Clin Chem Clin Biochem **19**: 1153-1166

Passing H, Bablok W, Glocke M (1981) An optimized design for the establishment of assigned values in control sera. J Clin Chem Clin Biochem **19**: 1167-1179

Peake MJ, Pejakovic M, Fraser CG (1984) The effect of instrument variables on calculation factors for standardisation of colorimetric analyses. Ann Clin Biochem **21**: 188-192

Petranyi G, Petranyi M, Scobie IN, et al (1984) Quality control of home monitoring of blood glucose concentrations. Brit Med J **288**: 757

Pope TP, Caragher TE, Grannis GF (1979) An evaluation of ethylene glycol-based liquid specimens for use in quality control. Clin Chem **25**: 413-418

Price F (1984) Right first time. Aldershot, Hants: Gower Publishing Co

Rej R, Jenny RW, Bretaudiere J-P (1984) Quality control in clinical chemistry: characterization of reference materials. Talanta **31**: 851-862

Ritchie RF, Rippey JH (1982) Performance on immunoglobulin IgG, IgA and IgM tests in CAP survey specimens. Am J Clin Pathol **Suppl. 78**: 644-650

Röhle G, Breuer H (1978) External quality control for hormone determinations in the Federal Republic of Germany. Hormone Res **9**: 450-454

Roehle G, Voigt U (1986) Concept for the evaluation of analytical results in clinical chemistry. Part I: Concentration-dependent dispersion profiles of survey results. Commun Lab Med **2**: 81-85

Roehle G, Voigt U, Siekmann L (1986) Concept for the evaluation of analytical results in clinical chemistry. Part II: The accumulated distribution-free dispersion parameter and its application. Commun Lab Med **2**: 131-138

Röschlau P, Bernt E, Gruber W (1974) Enzymatische Bestimmung des Gesamt-Cholesterins im Serum. J Clin Chem Clin Biochem **12**: 226

Rosalki SB (1972) A percentage comparison method of expressing enzyme survey results. In Rappoport AE ed. Quality control in clinical chemistry IV. Bern, Switzerland: H Huber; 71-99

Rowan RM, Laker MF, Alberti KGMM (1984) The implications of assaying external quality control sera under 'special conditions'. Ann Clin Biochem **21**: 64-68

Rowe DS, Anderson SG, Grab B (1970) A research standard for human serum immunoglobulins IgG, IgA and IgM. Bull World Health Org **42**: 535-552

Rowe DS, Grab B, Anderson SG (1972) An international reference preparation for human serum immunoglobulins G, A and M. Content of immunoglobulins by weight. Bull World Health Org **46**: 67-79

Rubin M, Barnett RN, Bayse D, et al (1984) Revised recommendation (1983) on evaluation of diagnostic kits. Part 1. Recommendations for specifications on labelling of clinical laboratory materials. In: Saris N-E ed. IFCC recommendations. Vol 1 1978-1983. Berlin, GFR: W de Gruyter; 114-118

Rumley AG, Roberts LB (1982) Effect of favourable treatment of samples on indices of performance in external quality assessment schemes. Ann Clin Biochem **19**: 171-175

Saidi HS (1979) Reliability of quality control materials in clinical chemistry. MSc thesis; 123 pp. Birmingham: University of Birmingham

Schreiner RL, Glick MR (1982) Interlaboratory bilirubin variability. Pediatrics **69**: 277-281

Shephard MDS, Penberthy LA, Fraser CG (1982) An inter-laboratory survey of qualitative urinalysis. Pathology **14**: 333-336

Shephard MDS, Penberthy LA, Fraser CG (1983) The 1982 Australasian programme for quantitative urine analysis. Clin Biochemist Rev 3: 128-132

Shewhart WA (1931) Economic control of quality of manufactured products. New York, USA: D van Nostrand

Shuey HE, Cebel J (1949) Standards of performance in clinical laboratory diagnosis. Bull US Army Med Dept **9**: 799-815

Siekmann L (1985) Determination of creatinine in human serum by isotope dilution-mass spectrometry. J Clin Chem Clin Biochem **23**: 137-144

Siekmann L, Breuer H (1982) Determination of cortisol in human plasma by isotope dilution-mass spectrometry. Definitive methods in clinical chemistry, I. J Clin Chem Clin Biochem **20**: 883-892

Simpson E, Thompson D (1978) An assessment of hospital routine urinalysis. Ann Clin Biochem **15**: 241-242

Skendzel LP, Youden WJ (1969) A graphic display of inter-laboratory test results. Am J Clin Pathol **51**: 161-165

Smith JH (1983) Laboratory staff and nurses' performance compared when using the blood glucose Reflotest-Reflomat system. Med Lab Sci **40**: 283-285

Smith I (1985) The hyperphenylalaninaemias. In: Lloyd JK, Scriver CR eds. Genetic and metabolic disease in pediatrics. London: Butterworth & Co; 166-210

Smith SCH, Buckley BM, Broughton PMG (1986) The influence of protein concentration on sodium measurement by six different direct ISE analysers. In: Maas AHJ, Boink FBJJ, Saris NEL, Sprokholt R, Wimberley PD eds. Methodology and clinical applications of ion selective electrodes. Vol 7. Copenhagen, Denmark: IFCC

Snavely JG, Golden WRC (1949) A survey of the accuracy of certain common chemical determinations. Connecticut State Med J **13**: 190-193

Snavely JG, Golden WRC (1951) The accuracy of certain common chemical determinations: second survey. Connecticut State Med J **15**: 667-669

Stamm D (1975) The determination of assigned values for control specimens. In: Anido G, van Kampen EJ, Rosalki SB, Rubin M eds. Quality control in clinical chemistry VI. Berlin, GFR: W de Gruyter; 113-130

Stamm D (1979) Reference materials and reference methods in clinical chemistry. J Clin Chem Clin Biochem **17**: 283-297

Stamm D (1981) Guidelines for a basic programme for internal quality control of quantitative analyses in clinical chemistry. (LAB/81.3) Geneva, Switzerland: World Health Organization

Stamm D (1982) A new concept for quality control of clinical laboratory investigations in the light of clinical requirements and based on reference method values. J Clin Chem Clin Biochem **20**: 817-824

Subcommittee on Analytical Goals in Clinical Chemistry of the World Association of Societies of Pathology (1979) Analytical goals in clinical chemistry: their relationship to medical care. Am J Clin Pathol **71**: 624-630

Swift AD, Jones A, Ratcliffe JG (1986) Effect of heat treatment on serum concentrations of thyroid-related hormones: results of an external quality assessment study. Commun Lab Med **2**: 92-94

Taylor RN, Fulford K (1981) Assessment of laboratory improvement by the Center for Disease Control diagnostic immunology proficiency testing program. J Clin Microbiol **13**: 356-368

Tonks DB (1963) A study of the accuracy and precision of clinical chemistry determinations in 170 Canadian laboratories. Clin Chem **9**: 217-233

Tonks DB (1982) Some faults with external and internal quality control programs. Clin Biochem **15**: 67-68

Tonks DB, Allen RH (1955) The accuracy of glucose determinations in some Canadian hospital laboratories. Canad Med Assoc J **72**: 605-607

Tydeman J, Morrison IJ, Hardwick DF, Cassidy PA (1982) The cost of quality control procedures in the clinical laboratory. Am J Clin Pathol **77**: 528-533

van Helden WCH, Visser RWJ, van den Bergh FAJTM, Souverijn JHM (1979) Comparison of intermethod analytical variability of patient sera and commercial quality control sera. Clinica chim Acta **93**: 335-347

Waid ME, Hoffmann RG (1955) The quality control of laboratory precision. Am J Clin Pathol **25**: 585-594

Walker G (1985) United Kingdom National Quality Assurance Advisory Panels. Eur Newsletter on Quality Assurance **2(3)**: 2

Ward PG, Lewis SM (1975) Inter-laboratory trials: a national proficiency assessment scheme in Britain. In: Lewis SM, Coster JF eds. Quality control in haematology. London: Academic Press; 37-51

Warnick GR, Albers JJ, Leary ET (1980) HDL cholesterol: results of interlaboratory proficiency tests. Clin Chem **26**: 169-170

Watkinson LR, St John A, Penberthy LA (1982) Investigation into paediatric bilirubin analyses in Australia and New Zealand. J Clin Pathol **35**: 52-58

Watson D (1980) Analytical investigations closer to the patient. Brit Med J **281**: 31-35

Wellcome Diagnostics (1984) Quality control in clinical chemistry. Dartford, Kent: Wellcome Diagnostics

Wernimont G (1946) Use of control charts in the analytical laboratory. Industr Eng Chem Anal Ed **18**: 587

Westgard JO, Groth T (1981) Design and evaluation of statistical control procedures: applications of a "quality control simulator" program. Clin Chem **27**: 1536-1545

Westgard JO, Groth T (1983) A predictive value model for quality control: effects of the prevalence of errors on the performance of control procedures. Am J Clin Pathol **80**: 49-56

Westgard JO, Falk H, Groth T (1979) Influence of a between-run component of variation, choice of control limits, and shape of error distribution on the performance characteristics of rules for internal quality control. Clin Chem **25**: 394-400

Westgard JO, Barry PL, Hunt MR, Groth T (1981) Proposed selected method. A multi-rule Shewhart chart for quality control in clinical chemistry. Clin Chem **27**: 493-501

Westwood A, Bullock DG, Whitehead TP (1986) An examination of the hexokinase method for serum glucose assay using external quality assessment data. Ann Clin Biochem **23**: 92-96

Whitby LG, Mitchell FL, Moss DW (1967) Quality control in routine clinical chemistry. Adv Clin Chem **10**: 65-156

Whitehead TP (1976) Principles of quality control. (LAB/76.1) Geneva, Switzerland: World Health Organization

Whitehead TP (1977) Quality control in clinical chemistry. New York, USA: John Wiley & Sons

Whitehead TP, Garvey K (1985) Quality assessment of tests performed outside the laboratory. (Broadsheet 114) London: Association of Clinical Pathologists

Whitehead TP, Morris LO (1969) Methods of quality control. Ann Clin Biochem **6**: 94-103

Whitehead TP, Woodford FP (1981) External quality assessment of clinical laboratories in the United Kingdom. J Clin Pathol **34**: 947-957

Whitehead TP, Browning DM, Gregory A (1973) A comparative survey of the results of analyses of blood serum in clinical chemistry laboratories in the United Kingdom. J Clin Pathol **26**: 435-445

Whitehead TP, Browning DM, Gregory A (1975) The role of external quality control schemes in improving the quality of laboratory results. In: Anido G, van Kampen EJ, Rosalki SB, Rubin M eds. Quality control in clinical chemistry VI. Berlin, GFR: W de Gruyter; 131-141

Whitehead TP, Bullock DG, Carter TJN, et al (1979) High density lipoprotein (HDL) cholesterol analysis. NewsSheet Assoc Clin Biochemists **190**: 7-10

Whitlow KJ, Campbell DJ (1983) Assessment of technologist workload as a factor in quality of laboratory performance. Am J Clin Pathol **79**: 609-610

WHO (1981) External quality assessment of health laboratories. Copenhagen, Denmark: WHO Regional Office for Europe

Wiener K (1980) Experiences with a regional quality control scheme for salicylate and paracetamol over a period of two years. Ann Clin Biochem **17**: 82-86

Wilde CE, Oakey RE (1975) Biochemical tests for the assessment of feto-placental function. Ann Clin Biochem **12**: 83-118

Wilding P, Bullock DG, Kricka LJ, Morriss P, Holder RL (1979) Spectral analysis of urinary reactions: a preliminary study based on a novel concept. Clin Chem **25**: 476-480

Williamson ML, Koch R, Azen C, Chang C (1981) Correlates of intelligence test results in treated phenylketonuric children. Pediatrics **68**: 161-167

Wootton IDP (1956) International biochemical trial 1954. Clin Chem **2**: 296-301

Wootton IDP (1957) Standardization in clinical chemistry. Clin Chem **3**: 401-405

Wootton IDP, King EJ (1953) Normal values for blood constituents. Inter-hospital differences. Lancet **1**: 470-471

Working Party on Pregnancy Oestrogens of the DHSS Advisory Committee on Assessment of Laboratory Standards (1981) Method for pregnancy oestrogens in urine. London: DHSS

APPENDICES

Appendix I:

EXTERNAL QUALITY ASSESSMENT SCHEMES AND SURVEYS

I.1 General aspects of EQA scheme and survey design

The EQASs organised from WRL have many features in common. Specimens and the corresponding results document are despatched together to participants by the fastest practical means. The results document identifies the specimen (with reconstitution instructions), the analytes required and the date by which results must be received at WRL. Participants write their results on the document (which specifies the units) and return it to WRL.

The results received are entered into the WRL computer system by two data clerks independently, and any discrepancies are investigated. A checklist is produced, indicating any results >1.5 SD from the untrimmed mean to be checked by UKEQAS staff against the results document received. After this data validation procedure the main processing is carried out.

Data processing involves program modules to:

- calculate the overall mean and SD, and recalculate these after exclusion of results >2 or >3 SD from the untrimmed mean
- calculate similarly means and SDs for each method group
- calculate BISs and VISs (see Appendix II) for each laboratory's results
- cumulate BISs and VISs for the calculation of an OMRVIS, and of MRVIS/MRBIS/SDBIS for each analyte
- produce reports for participants

Reports are then mailed to participants by the fastest practicable means. The reports include the participant's own results, overall and method-related consensus values and SDs, histograms of results, and VI scoring parameters.

The schemes differ primarily in the:

- distribution frequency
- number and type of specimens constituting a distribution
- analytes surveyed
- unit sets used
- truncation limits
- inclusion of VI scoring

- use of overall or method mean as designated value (DV)
- selection of VI scoring parameters reported to participants
- detailed report format

I.2 UK External Quality Assessment Schemes (UKEQASs)

Table I.1 details the current (1987) characteristics of the full schemes, the development of which has been summarised previously (Bullock, 1985; DHSS, 1986b). Specimens and reports are despatched by first class mail. Results received too late for the main computer run, but before report despatch, are scored on the basis of the main computer run and the files updated accordingly.

Descriptions of each scheme are confined to the variables mentioned above and to major changes during their development.

I.2.1 UKEQAS for General Clinical Chemistry

Vials of lyophilised quality control sera, based on human or animal serum and to be reconstituted with distilled water, from a variety of commercial sources or commissioned by UK EQAS (Table I.2) are distributed. Results received 11 days after specimen despatch are included in the main computer run. Outlier elimination is by truncation at ± 3 SD. BISs are calculated using the recalculated method mean as DV for each method group except Miscellaneous (recalculated overall mean) and Other (no scores). Participants' reports now (1987) include all VI parameters, though histograms of results were discontinued in 1984.

Prior to 1976 specimens of pooled liquid human serum were used (Whitehead et al, 1973). The analytes surveyed are listed in Table 4.3; the first 14 have been included since the scheme's inception, the others being added later. The enzymes were incorporated in 1986. Reporting was changed from mass to molar SI units in 1976. No VI parameters were reported routinely prior to 1974, and the range has been increased steadily from the OMRVIS alone, incorporating successively VISs, MRVISs, BISs, and MRBISs and SDBISs for each analyte.

I.2.2 UKEQAS Enzyme Surveys

The surveys have been described previously (Bullock et al, 1986b). Most of the specimens distributed were lyophilised QC sera; one batch of liquid serum preserved with ethylene glycol

Table I.1 Description of UKEQASs at January 1987

UKEQAS for	Participants	Distributions /year	Analytes	Established
General Clinical Chemistry	560	24	18 (from 26)	1969
Salicylate and Paracetamol	330	12	2	1984
Specific Proteins	280	12	8	1980
Lead in Blood	140	24	2	1973
Urinary Pregnancy Oestrogens	60	12	2	1980 (transferred)
Phenylketonuria Screening	35	4 (6 specimens)	1	1980

Table I.2 Materials distributed in UKEQAS for General Clinical Chemistry, 1978-1987. * denotes an Enzyme Survey, + a repeat distribution; A = animal, B = bovine, E = equine, H = human

	Date	Supplier	Material	Lot	Base	Volume (mL)
131	06.01.78	DADE	Monitrol II.X	XPT-9568	H	10
132	20.01.78	GD	Unassayed	-	H	10
133	03.02.78	Roche	WHO Reference	Serum A	H	10
134	17.02.78	GD	Validate	0389057	H	10
135	03.03.78	GD	Unassayed	0664107	B	10
136	17.03.78	Wellcome	NQCS High	K5439	H	10
137	31.03.78	Wellcome	NQCS Normal	K5438	H	10
138	14.04.78	Roche	WHO Reference	77/1	H	10
139	28.04.78	Wellcome	NQCS Low	K5437	H	10
140	12.05.78	Hyland	Q-Pak I	1779N003AB	H	5
141	26.05.78	Hyland	Q-Pak II	1778P002A	H	5
142	16.06.78	TCS	Equitrol	0468	E	10
143	30.06.78	Wellcome	+NQCS High	K5439	H	10
144	14.07.78	DADE	+Monitrol II.X	XPT-9568	H	10
145	28.07.78	GD	CCQCS	0247017	H	10 dil
146	18.08.78	Purce	Armtrol	387	B	10
147	08.09.78	Purce	Armtrol	390	B	10
148	22.09.78	TCS	Equitrol	0419a	E	10
149	06.10.78	Wellcome	+NQCS High	K5439	H	10
150	20.10.78	Hyland	+Q-Pak I	1779N003AB	H	5
151	03.11.78	DADE	Monitrol I.X	XLT-350	H	10
152	17.11.78	Roche	Control N	A2137	E	5
153	01.12.78	Roche	Control P	A0438	E	5
154	15.12.78	GD	QAS Level II	4D365	H	10 dil
155	12.01.79	Nyegaard	Seronorm	144	E	5
1	25.02.79	*Roche	Enzykon SN	X1033	A	3
156	09.02.79	Liberton	NQCS Trial	L1/78	H	10
157	09.03.79	GD	+QAS Level II	4D365	H	10 dil
2	23.03.79	*Roche	Enzykon SP	X1033	A	3
158	06.04.79	Liberton	SNBTS	12	B	10
159	27.04.79	TCS	Human	0863	H	10
160	18.05.79	Technicon	MultiSystem	X9C202	B	10
3	08.06.79	*Hyland	Q-Pak I	1840U001A	H	5
161	15.06.79	Ortho	Abnormal	9S317	H	10
162	29.06.79	Roche	Control P	A0842	E	5
163	13.07.79	Roche	Control N	A2941	E	5
164	27.07.79	Hyland	Q-Pak II	P13/R237	H	5
165	10.08.79	Hyland	Q-Pak I	1779N005B	H	5
4	24.08.79	*DADE	Monitrol II.X	XPT9568	H	10
166	07.09.79	BCL	Precinorm U	801	H	10
167	21.09.79	Miles	Mytrol	458	H	5
168	05.10.79	Technicon	MultiSystem	X9G207	B	10
169	19.10.79	Biotrol	Biotrol-00	1921	B	10
170	02.11.79	Miles	Mytrol	361	H	5
5	16.11.79	*GD	Versatol-E Plus	4F986	A	3
171	23.11.79	Wellcome	NQCS Internal QC	HIQC/1	H	10
172	07.12.79	TCS	+Human	0863	H	10
173	21.12.79	Liberton	+NQCS Trial	L1/78	H	10
174	14.01.80	Liberton	NQCS	L3/79	H	10
175	28.01.80	Roche	Control N	N1538	E	5
176	11.02.80	Roche	Control P	N2238	E	5
6	25.02.80	*DADE	Enzatrol E	ET247	S	3
177	10.03.80	Technicon	MultiSystem N	X9M111	B	10
178	24.03.80	DADE	Monitrol I.X	XLT-373	H	10

179	14.04.80	Ortho	Normal unassayed	9S219	H	10
180	28.04.80	DADE	QAP (Monitrol II)	SPXP9601	H	10
181	19.05.80	*Liberton	NQCS	L2/79	H	10
7	02.06.80	*Ortho	Abnormal unassayed	9S317	H	10
182	09.06.80	Liberton	Scottish BTS Control	13	B	10
183	23.06.80	Technicon	Alert 1	BOD363	B	10
184	07.07.80	Purce	Armtrol	547	B	10
185	21.07.80	Purce	Armtrol abnormal	488	B	10
186	04.08.80	Liberton	+NQCS Trial	L1/78	H	10
187	18.08.80	Purce	Armtrol	489	B	10
188	01.09.80	Technicon	Alert 2	BOE364	B	10
189	15.09.80	*BCL	Precipath U	805	H	5
8	29.09.80	*BCL	Precipath U	805	H	5
190	13.10.80	Nyegaard	Seronorm	154	E	5
191	27.10.80	Wellcome	NEQAS Internal QC	HIQC/2	H	10
192	10.11.80	*Hyland	Q-Pak I	1840U001AA	H	5
9	17.11.80	*Liberton	Scottish BTS Control	37	B	10
193	01.12.80	GD	QAS Level II	4D523	H	10 dil
194	15.12.80	TCS	Equitrol	1240	E	10
195	12.01.81	Roche	Control N	T0832	E	5
196	26.01.81	Roche	Control P	T1532	E	5
197	09.02.81	Liberton	NQCS Trial	L4/80	H	10
198	23.02.81	Technicon	Alert 2	BOM874	B	10
199	09.03.81	BCL	PreciFlo	11-603	B	10 dil
200	23.03.81	*Wellcome	NEQAS Internal QC	HIQC/3	H	10
10	06.04.81	*DADE	Monitrol II.X	XPT9581	H	10
201	27.04.81	Ortho	+Abnormal	9S317	H	10
202	18.05.81	Ortho	+Normal	9S219	H	10
203	01.06.81	TCS	Equitrol	1304	E	10
204	15.06.81	Purce	+Armtrol	489	B	10
205	29.06.81	Ortho	Abnormal	W24X02B	H	10
206	13.07.81	Nyegaard	Pathonorm L	18	B	5
207	27.07.81	Technicon	Alert 2	B1E511	B	10
208	10.08.81	Wellcome	NEQAS Internal QC	HIQC/4	H	10
209	24.08.81	*Nyegaard	Pathonorm H	18	B	5
11	14.09.81	*Ortho	Abnormal	9S317	H	10
210	28.09.81	Ortho	Normal	W27X02B	H	10
211	12.10.81	Liberton	+NQCS Trial	L4/80	H	10
212	26.10.81	DADE	Monitrol II.X	SPXP9613	H	10
213	09.11.81	Purce	Armtrol	551	B	10
214	23.11.81	Gibco	Gibcotrol High	159	A	10
215	07.12.81	Wellcome	Special Services	HIQC/5	H	10
12	04.01.82	*Liberton	SNBTS	68	B	10
216	18.01.82	Wellcome	Wellcomtrol II	K9122	E	10
217	01.02.82	Roche	Control N	T1734	E	5
218	15.02.82	Roche	Control P	T2440	E	5
219	01.03.82	Wellcome	Wellcomtrol I	K2690	B	10
220	15.03.82	Purce	Armtrol	650	B	10
221	29.03.82	Ortho	+Abnormal	W24X02B	H	10
222	19.04.82	Beckman	Decision	C111747G	H	Liquid
223	10.05.82	Wellcome	+NEQAS Internal QC	HIQC/5	H	10
224	07.06.82	Nyegaard	Seronorm	158	E	5
225	21.06.82	Purce	Armtrol	651	B	10
226	05.07.82	Purce	Armtrol	465	B	10
227	19.07.82	Purce	NEQAS Internal QC	660	H	10
228	02.08.82	Roche	Control N	E3031	E	5
229	16.08.82	Roche	Control P	E0641	E	5
230	06.09.82	Roche	Control N	E0340	E	5
231	20.09.82	Wellcome	NEQAS Internal QC	HIQC/6	H	10

232	04.10.82	Technicon	Alert 1	B2E289	B	10
233	18.10.82	Ortho	Level II	X40Y02B	H	10
234	01.11.82	Technicon	Alert 2	B2G290	B	10
235	15.11.82	Ortho	Level I	X39Y02B	H	10
236	29.11.82	Biotrol	Biotrol-00	1929	B	10
237	13.12.82	Gibco	Gibcotrol High	199	A	10
238	03.01.83	Wellcome	Special Services	K5164/2	B	10
239	17.01.83	*Purce	+NEQAS Internal QC	660	H	10
13	31.01.83	*Beckman	Link II	C204236	H	Liquid
240	14.02.83	Roche	Control N	P1833	E	5
241	28.02.83	Ortho	Level II	008Y01	B	10
242	14.03.83	Purce	+Armtrol	551	B	10
243	28.03.83	Ortho	Level III	009Y01	B	10
244	18.04.83	*Roche	Control P	P2533	E	5
14	09.05.83	*Ortho	Level II	X40Y02B	H	10
14A	09.05.83	*Beckman	Link II	C204236	H	Liquid
245	23.05.83	Wellcome	+Special Services	K6026	B	10
246	13.06.83	Ortho	Level I	007Y01	B	10
247	27.06.83	Liberton	+NEQAS	L4/80	H	10
248	11.07.83	Wellcome	Special Services	K6025	B	10
249	25.07.83	Roche	Control N	P1039	E	5
250	08.08.83	Ortho	Abnormal Assayed	025A01	H	5
251	22.08.83	Purce	Armtrol	726	B	10
252	12.09.83	Ortho	Normal Assayed	020A01	H	5
253	26.09.83	Technicon	Alert 1	B3G623	B	10
254	10.10.83	Roche	Control P	P2439	E	5
255	24.10.83	Technicon	Alert 2	B3G624	B	10
256	07.11.83	Ortho	+Normal	W27X02B	H	10
257	21.11.83	*Purce	NEQAS Internal QC	HIQC/8	H	10
15	05.12.83	*Purce	NEQAS Internal QC	HIQC/8	H	10
151	05.12.83	*Technicon	RA-100 Calibrator I	B3A581	B	3
258	19.12.83	Purce	Armtrol	727	B	10
259	16.01.84	Gibco	Gibcotrol High	195	A	10
260	30.01.84	Ortho	Level II	008X01	B	10
261	13.02.84	Technicon	Alert 1	B3M351	B	10 dil
262	27.02.84	Technicon	Alert 2	B3M361	B	10 dil
263	12.03.84	Roche	Control N	U0433	E	5
264	26.03.84	Ortho	+Level II	X40Y02B	H	10
265	09.04.84	Biotrol	Biotrol-00	1934	B	10
266	30.04.84	Wellcome	Wellcomtrol Two	K8609	E	10
267	14.05.84	*DADE	QAP Level II	524.01	H	10
16	04.06.84	*Wellcome	NEQAS Internal QC	K8680/HIQC/9	H	10
161	04.06.84	*Wellcome	Wellcomtrol Two	K8609	E	10
268	18.06.84	Wellcome	NEQAS Internal QC	K8680/HIQC/9	H	10
269	02.07.84	Roche	Control Serum N	U2140	E	5
270	16.07.84	DADE	QAP Level I	523.01	H	10
271	30.07.84	Roche	Control Serum P	U2840	E	5
272	13.08.84	Nyegaard	Pathonorm L	20	B	5
273	03.09.84	Technicon	TESTpoint 2	B4F361	B	10 dil
274	17.09.84	Nyegaard	Pathonorm H	20	B	5
275	01.10.84	*Technicon	TESTpoint 1	B4F351	B	10 dil
17	15.10.84	*Ortho	+Level II	X40Y02B	H	10
17A	15.10.84	*Technicon	+RA-1000 Calibrator	I B3A581	B	3
276	29.10.84	Ortho	Abnormal Assayed	025B01	H	5
277	12.11.84	Lorne	Normal	N41	H	5
278	26.11.84	Ortho	Normal Assayed	020B01	H	5
279	10.12.84	Lorne	Pathological	P41	H	5
280	31.12.84	Wellcome	NEQAS Internal QC	HIQC/10	H	10

281	21.01.85	*Roche	Control N	B1732	E	5
18	04.02.85	*Wellcome	NEQAS Internal QC	HIQC/10	H	10
18A	04.02.85	*Ortho	Level III	009B01	B	10
282	18.02.85	Ortho	Level II	008A01	B	10
283	04.03.85	Roche	Control P	B0637	E	5
284	18.03.85	Wellcome	Abnormal	K9301	B	10
285	01.04.85	Roche	Control N	B2936	E	5
286	22.04.85	Ortho	Level III	009B01	B	10
287	13.05.85	Nyegaard	Seronorm	166	E	5
288	03.06.85	Gibco	Gibcotrol High	412	A	10
289	17.06.85	*Purce	Armtrol	815	B	10
19	01.07.85	*Wellcome	NEQAS Internal QC	HIQC/11	H	10
290	15.07.85	Wellcome	NEQAS Internal QC	HIQC/11	H	10
291	29.07.85	Roche	Control N	B2941	E	5
292	12.08.85	Roche	Control P	B1442	E	5
293	02.09.85	*Wellcome	"Abnormal"	K7275	B	10
20	16.09.85	*Nyegaard	Seronorm Enzyme	802	E	3
294	30.09.85	Ortho	+Level II	008A01	B	10
295	14.10.85	Wellcome	BCZ6	K340110	B	10
296	28.10.85	Technicon	TESTpoint 1	V5J128	B	10 dil
297	11.11.85	*Technicon	TESTpoint 2	V5J072	B	10 dil
21	25.11.85	*Wellcome	"Abnormal"	K7275	B	10
298	09.12.85	Gilford	Level II	008501	B	10
299	06.01.86	Purce	Armtrol	833	B	10
300	20.01.86	Ortho	+Level III	009B01	B	10
301	03.02.86	Gibco	Gibcotrol High	451	B	10
302	17.02.86	Purce	Armtrol	848	B	10
303	03.03.86	Wellcome	+"Abnormal"	K7275	B	10
304	17.03.86	Wellcome	Wellcomtrol 0344	K4077	B	10
305	03.04.86	Wellcome	Heat-treated human	K466610	H	10
306	21.04.86	Gibco	Gibcotrol Abnormal	458	A	10
307	12.05.86	Nyegaard	Pathonorm H	21	B	5
308	29.05.86	Wellcome	NEQAS Internal QC	HIQC/12	H	10
309	16.06.86	Wellcome	WHO reference	K860210	B	10
310	30.06.86	Roche	Control N	L3231	E	5
311	14.07.86	Roche	Control P	L2132	E	5
312	28.07.86	DADE	QAP Level II	524.03	H	10
313	11.08.86	Roche	Control N	L0736	E	5
314	08.09.86	Roche	Control P	L2236	E	5
315	25.09.86	Wellcome	NEQAS Internal QC	HIQC/13	H	10
316	13.10.86	Nycomed	Seronorm	177	E	5
317	27.10.86	Wellcome	+Heat-treated human	K466610	H	10
318	10.11.86	Wellcome	Abnormal	K409650	B	10
319	24.11.86	Roche	Control N	L1541	E	5
320	08.12.86	Roche	Control P	L2941	E	5
321	05.01.87	Roche	Control N	L2838	E	5
322	19.01.87	Purce	Armtrol	896	B	10
323	02.02.87	Nycomed	Pathonorm L	21	E	5
324	16.02.87	Wellcome	[BC03]	Unlabelled	B	10
325	02.03.87	Gilford	Level III	009501	B	10
326	16.03.87	Purce	Armtrol	901	B	10
327	30.03.87	Ortho	Level I	007B01	B	10
328	13.04.87	Technicon	TESTpoint 2	V7B011	B	10 dil
329	11.05.87	Technicon	TESTpoint 1	V7B013	B	10 dil
330	01.06.87	+Wellcome	WHO reference	K860210	B	10
331	15.06.87	Gibco	Gibcotrol High	495	A	10
332	29.06.87	Gibco	Gibcotrol Abnormal	491	A	10

was distributed twice (Table I.2).

To minimise the effects of interlaboratory variations in the time between reconstitution and assay, a protocol for the handling of specimens was established; the criterion for a negligible effect was a change in activity of <5% (Bullock et al, 1979 and 1986b). This protocol specified addition of the appropriate volume of distilled water, gentle agitation at ambient temperature until dissolution was complete (20-30 min) and storage at 4^{o}C until the time of assay. CK activity was to be assayed immediately, ALP activity within 4 h, and AST, ALT, LD and amylase activities within 48 h (preferably within 24 h).

Outlier elimination was by truncation of any results >2 SD from the untrimmed method mean, and no calculations were made for any method group with <10 results. VI scoring was introduced at Survey 19, the method mean being used as DV (no scoring for the Other method or temperature group).

The method classification systems used (Bullock et al, 1986b) emphasised robust methods in wide use, primarily those based on officially recommended procedures. Those from the Scandinavian Committee on Enzymes (SCE) were regarded as 'reliable' methods for many of the studies.

Reports showed for each enzyme the statistical data classified according to method, the laboratory's result and (from Survey 19) its VIS. A histogram of the results for the laboratory's method group was also given, except for the heterogeneous Other method or temperature group.

I.2.3 UKEQAS for Lead in Blood

The preparation of the liquid haemolysates from fresh human blood for use in this scheme has been described by Bullock et al (1986c). Truncation is at $\pm$3 SD. Reports were changed from mass to molar SI units in 1979, and cadmium was included from 1982 (with scoring from 1986). The overall mean is used as DV, and reporting of VISs and MRVISs was introduced in 1979.

I.2.4 UKEQAS for Urinary Pregnancy Oestrogens

The operation of this scheme has been described by Bullock and Wilde (1985) and previously by Oakey (1980). Distribution of the lyophilised urine specimens was formerly three-weekly, but was

changed to fortnightly on transfer to WRL and later reduced to 4-weekly in 1984. Truncation is at ±3 SD, with results expressed in umol/24h assuming a urine volume of 2L. Creatinine (mmol/24h) was included fully in 1984.

The method mean is used as DV (or overall mean for the Miscellaneous group), and BIS, MRVIS, MRBIS and SDBIS have been reported since scheme transfer.

I.2.5 UKEQAS for Specific Proteins

The initial surveys in this scheme were described by Chambers et al (1984). Specimens comprise clarified liquid normal human serum, with two pairs of dilution-related specimens being distributed in initial surveys. Later surveys comprised first two and later one unrelated specimen, with a final 4-weekly frequency. Initial surveys included immunoglobulins G, A and M, with complement components C3 and C4, $alpha_1$-antitrypsin (A1-AT), orosomucoid and C-reactive protein being added later.

Truncation is at ±2 SD. VISs and MRVISs, for immunoglobulins only, have been reported since the CCVs were established in 1985. The overall mean is used as DV.

I.2.6 UKEQAS for Salicylate and Paracetamol

Initial surveys included separate specimens for the two analytes, but 4-weekly distribution of a single specimen for both drugs was selected in 1985. The liquid specimens used comprise pure drugs in sterile equine serum (Bullock, 1987). Truncation is at ±3 SD, and VISs and MRVISs based on the overall mean as DV have been reported since 1985.

I.2.7 UKEQAS for PKU Screening

Specimens are prepared by addition of phenylalanine (PheA) solutions to fresh human whole blood with CPDA as anticoagulant. After one hour's gentle agititation, blood is spotted onto filter paper cards or placed into capillary tubes, as appropriate for each participant. Laboratories are asked to classify each specimen as having a normal, intermediate or high PheA content, giving their coded action (Table 7.1) and also a quantitative or semiquantitative PheA concentration if appropriate.

There is no computer processing of the results, and reports comprise tabulations of results, actions and scores (eg Figures

7.1 and 7.2). The PheA concentration determined by two (formerly one) reference laboratories not undertaking PKU screening is used as a target. The evolution of this design and the details of the scoring system are described fully in sections 7.2.2 and 7.2.3.

I.3 Other EQASs

For the two international schemes sponsored by WHO, sets of 6 specimens are distributed, to be assayed by participants at monthly intervals. Truncation is at ±3 SD. DVs are derived from prior distribution of the same material through the UKEQAS. Reports include the laboratory's results, comparative data from the UKEQAS and scheme, and VI parameters (VISs and OMRVIS). Similar schemes have also been organised in China and Thailand.

I.3.1 International EQAS (IEQAS)

The scheme was established in 1975, and currently has about 150 participants from about 65 countries. Specimens and reports are sent by diplomatic pouch via WHO Geneva to avoid customs delays.

UKEQAS overall consensus values were used as DVs for all analytes except glucose and cholesterol, for which the UKEQAS method means were used. Since 1986 scoring has been against the appropriate UKEQAS or IEQAS method mean (overall mean for the Miscellaneous group; no score for the Other group).

I.3.2 Middle East EQAS (MEEQAS)

The MEEQAS (Bacchus et al, 1982 and 1987) was established in 1980, as a WHO-recognised scheme financed by the Riyadh Al-Kharj Programme in Saudi Arabia, who are the joint organisers with WRL. Specimens are sent to Riyadh via the Royal Saudi Air Force to avoid customs delays, for local distribution to participants. The scheme is a regional one including participants from 10 neighbouring countries as well as from Saudi Arabia, and has about 80 participants. Grouped method means from the UKEQAS are used as DVs, with no scoring for the Other group.

I.3.3 Intensive EQASs

The objective of these schemes, outlined in section 3.2.3, is a short-term intensive interaction with a limited number of participants to improve performance. There is no cumulation of data, since the laboratories should show continuing improvement. Such schemes have been operated in Colombia, Italy, Mexico, Thailand and UK.

Each laboratory receives three sera for assay together. Evaluation uses UKEQAS consensus values as DVs, and includes calculation of VISs. The reports include the VISs and an average VIS over all analytes in the distribution for the laboratory, in a table also showing the same information for other participants and average VISs for each analyte and for all laboratories.

An integral part of the report is a graph of the laboratory's results against DVs for each analyte (eg Figure 10.9), with helpful interpretive comments regarding the likely sources of error added by the scheme organiser. Participants also receive a further set of vials, so they can investigate the situation and check the effectiveness of their corrective actions before the next distribution.

I.4 WRL International Intercomparison Scheme (WIIS)

This scheme was established to give a more objective assessment of whether biases did exist between established national EQASs, and of the magnitude of any such biases. The principles of the study are described in section 5.6.1.

NEQAS organisers were requested to select three or 5 suitable laboratories, ie those with small bias and regular return rate, from their scheme. Sets of 6 specimens were distributed to participants via their NEQAS organiser, for replicate analysis at monthly intervals and return of results, in either SI or 'conventional' (mg/dL) unit sets, to WRL for processing.

Overall means were obtained after truncation at ± 3 SD, and the percentage deviation of each result from this WIIS consensus then calculated. These percentage deviations were cumulated for all the WIIS participants in each country over 4 distributions, to yield a cumulative average percentage deviation. VISs (Appendix II.2.4) were also calculated, but not reported.

Reports to participants showed the data derived from laboratories in their country, with the cumulative average percentage deviations for all countries (identified only by a country code). The participant's own results were not included. Any results received after the deadline for the main computer run were assessed and the files updated before results for the following

distribution were processed.

I.5 EQA surveys

The operating procedures for surveys are similar to those described above for EQASs. There is no scoring, nor any cumulation of data.

I.5.1 HDL cholesterol

These surveys (John, 1983) included the group of 12-20 laboratories participating the standardisation initiative (Whitehead et al, 1979). The specimens in Surveys 2, 4 and 5 comprised pooled sera from patients and freshly-drawn plasma from normal subjects; Survey 3 comprised 6 commercial QCMs (sera D, G, N, P, U and W; Table III.3).

Results for HDL were classified according to precipitation procedure (PhT, Hep or PEG). Results not obtained using enzymic methods for cholesterol assay were excluded, as were those >2 SD from the untrimmed mean.

I.5.2 Urinary total protein

Specimens comprised both normal and pathological urine specimens and aqueous solutions of salts and urea; various protein preparations (human and bovine sera and albumins) were added to some. Each distribution included one to four specimens. Procedures used by participants were classified according to method principle (Table 2.8) and calibrant. Truncation was at ± 2 SD.

I.5.3 Extra-laboratory assays

These surveys (section 2.3; Browning and Bullock, 1987) used lyophilised bovine sera distributed previously in the UKEQAS for General Clinical Chemistry (as distributions 286 and 302; Table I.2) initially, but the third survey comprised two lots of "Sugar-Chex" (Alpha Laboratories), a suspension of fixed bovine erythrocytes in an aqueous medium.

The specimens were distributed via laboratories: those responding to a previous questionnaire (Browning et al, 1984) for the first survey, then all UKEQAS participants. The laboratories were requested to reconstitute the lyophilised sera and distribute the specimens to extra-laboratory sites for analysis. Surveys 1 and 2 included sodium, potassium and glucose, and Survey 3 glucose

only; bilirubin was also requested in Survey 1.

The surveys were conducted anonymously, participants returning only their results and details of the reagents, calibrant and instrument used. Truncation was at ±2 SD, and the UKEQAS data (Surveys 1 and 2) were reprocessed similarly for comparison. VISs were calculated against the appropriate method mean from the extra-laboratory survey as DV. The method classification for glucose was based on instrument manufacturer (Table 2.4), with a sub-coding according to model. Glucose results obtained by visual reading were excluded from this analysis, but were specifically sought in Survey 3 and analysed separately.

I.5.4 Urinary aminoacids

The development and operation of these surveys is described fully in section 7.3. The specimens, distributed in pairs, comprised urine from normal subjects or patients with well-characterised clinical conditions. Participants were requested to characterise the chromatographic pattern obtained and return results from their usual 'spot tests' (Figure 7.3). Reports included details of specimen origin, a list of all laboratories' responses and summaries of the spot test results (Figures 7.4 and 7.5), and were accompanied by comments from scientific advisors.

Appendix II:

VARIANCE INDEX SCORING SYSTEM

II.1 General aspects of Variance Index (VI) scoring

Variance Index scoring was described originally by Whitehead et al (1973). This was soon changed to the more reliable version incorporating scaling by Chosen Coefficient of Variation (CCV) rather than by SD (Whitehead et al, 1975; Whitehead, 1977). Later refinements which introduced BISs and their cumulated running indices were described by Bullock and Wilde (1985). As described in Chapters 4 and 9, the system provides a simple but reliable indication of laboratory performance which has proved useful over many years, in assessing both laboratory performance and changes in this over time.

II.2 VI scoring in UKEQASs

The definitions and derivation of the parameters used are as follows:

II.2.1 Bias Index Score (BIS)

The difference between the result obtained by the laboratory (x) and the designated value (DV; see below) expressed as a percentage of the method mean, divided by the CCV (see below) for the analyte and again expressed as a percentage:

$$BIS = \frac{(x - DV)}{DV} \cdot 100 \cdot \frac{100}{CCV}$$

Any score greater in magnitude than 400 is set to 400. The BIS may therefore be in the range -400 to +400.

II.2.2 Designated Value (DV)

The DV is the 'target value' for the analyte in the specimen distributed. It is usually an overall or method-related mean (consensus value) from the scheme (see Appendix I).

II.2.3 Chosen Coefficient of Variation (CCV)

The CCV is a scaling factor for each analyte, correcting for differences in the state of the art and yielding VISs in a 'common currency'; it does **not** represent a 'clinically acceptable error'. For the original 14 general clinical chemistry analytes (sodium to cholesterol) CCVs are the best interlaboratory CVs achieved in the UKEQAS in 1972, which are still representative of the relative performance; their values are given in Table 4.3.

CCVs for other analytes were selected to yield similar VISs, using 'calibrated' data for reliable method groups (Bullock et al, 1986b; Chapter 13); their values are given in Tables 4.3 and II.1.

II.2.4 Variance Index Score (VIS)

The VIS is the absolute value of the BIS, ie ignoring its sign. Values may be in the range 0 to 400.

II.2.5 Mean Running VIS (MRVIS)

The MRVIS is the mean of the 10 most recent VISs for the individual analyte. Values may be in the range 0 to 400.

II.2.6 Mean Running BIS (MRBIS)

The MRBIS is the mean of the 10 most recent BISs for the individual analyte. Values may be in the range -400 to +400.

II.2.7 Standard Deviation of the BIS (SDBIS)

The SDBIS is the SD of the 10 most recent BISs for the individual analyte. Values may be in the range 0 to 422.

II.2.8 Overall Mean Running VIS (OMRVIS)

The OMRVIS is the mean of the 40 most recent VISs for the laboratory, **irrespective of analyte.** Values may be in the range 0 to 400.

II.3 VI scoring in other EQASs

II.3.1 International EQAS and Middle East EQAS

The system described above is also applied in these EQASs operated from WRL (Appendix I.3), with two differences:

- the DV is not derived from the scheme, but from prior distribution of the same material in the UKEQAS for General Clinical Chemistry.
- the OMRVIS is calculated from the most recent 30, rather than 40, VISs.

II.3.2 Intensive EQASs

In such schemes (Appendix I.3.3), the VISs calculated are averaged for each laboratory over all analytes and specimens in the current distribution only. There is no cumulation of scores from distribution to distribution.

II.4 Graphical presentation of VIS data

Graphs of running scores against time are produced, covering a

Table II.1 Chosen Coefficients of Variation for analytes not in UKEQAS for General Clinical Chemistry

	CCV
Salicylate	12.5
Paracetamol	15.0
IgG	11.9
IgA	14.2
IgM	15.9
Lead	15.0
Cadmium	15.0
Oestrogens	15.0
Creatinine (urine)	12.0

period of about two and a half years for a scheme with fortnightly distributions. These are plotted against distribution number and include OMRVIS alone, MRVIS alone, or MRVIS, MRBIS and SDBIS together (eg Figures 10.4, 10.6 and 10.7 respectively). Indications of the current 5th, 50th and 95th centiles of OMRVIS or MRVIS are included.

Appendix III:

STUDY DESIGNS

III.1 Studies on the validity of consensus values

III.1.1 Reproducibility

The percentage differences between successive distributions of the same material in an EQAS were calculated. The mean and SD from all pairs within the period considered were then calculated (Tables 5.1 and 5.3). The periods covered distributions 205-330 in the UKEQAS for General Clinical Chemistry (Table I.2; 13 pairs), 99-131 in the IEQAS (Table III.1; 11 pairs) and 55-83 in the MEEQAS (Table III.2; 8 pairs).

III.1.2 Ad hoc comparisons between EQASs

The UKEQAS distributions (Table I.2) considered were 185 and 187 (Table 5.4), 197 and 224 (Table 5.5), 267, 270 and 312 (Table 5.6) and 161, 179 and 295 (Table 5.7). The values for comparison were obtained from the scheme organisers in Holland and South Africa, from Broughton and Eldjarn (1985) for Norway, and from the QCM suppliers in the other cases.

III.1.3 Comparisons with values assigned by reference and definitive methods

The data for the WHO material (Table 5.8) were obtained from WHO and those for Seronorm lot 158 (Table 5.9; UKEQAS distribution 224; Table I.2) from Eldjarn and Broughton (1985).

The difference of the overall consensus value in the UKEQAS from the value assigned by DGKC (supplied by Roche) was calculated for all batches of Roche Control Sera N and P distributed in the UKEQAS from 1977 to 1986 (Table I.2) and averaged for each year (Table 5.10). No DGKC values were available for distributions 195, 196, 228, 229, 240 or 313.

III.1.4 Comparisons between EQASs

The percentage difference of the overall consensus value obtained in the IEQAS or MEEQAS from that in the UKEQAS was calculated. The mean and SD (Tables 5.11 and 5.12) were then calculated for 20 distributions: 111-130 in the IEQAS (Table III.1) and 41-60 in the MEEQAS (Table III.2).

III.1.5 Comparisons among NEQASs in the WIIS

Appendix I.4 describes the calculation of cumulative average

Table III.1 Materials distributed in International EQAS, 1984-1987. For explanation see Table I.2

	Date	Supplier	Material	Lot	Base	Volume (mL)
92	02.01.84	Liberton	NEQAS	L4/80	H	10
93	06.02.84	Purce	+Armtrol	651	B	10
94	05.03.84	Ortho	Level II	008Y01	B	10
95	02.04.84	Wellcome	NEQAS	HIQC/6	H	10
96	07.05.84	Purce	Armtrol	727	B	10
97	04.06.84	Wellcome	Special Services	K6025	B	10
98	02.07.84	Gibco	Gibcotrol High	195	A	10
99	27.08.84	Ortho	Level II	008X01	B	10
100	24.09.84	Technicon	Alert 2	B3M361	B	10 dil
101	29.10.84	DADE	QAP Level II	524.01	H	10
102	26.11.84	Wellcome	Wellcomtrol Two	K8609	E	10
103	31.12.84	Ortho	+Level III	009Y01	B	10
104	28.01.85	Biotrol	Biotrol-OO	1934	B	10
105	25.02.85	Ortho	+Level II	008X01	B	10
106	25.03.85	Purce	NEQAS	HIQC/8	H	10
107	29.04.85	Technicon	TESTpoint 2	B4F361	B	10 dil
108	27.05.85	Wellcome	+Wellcomtrol Two	K8609	E	10
109	24.06.85	Dade	+QAP Level II	524.01	H	10
110	29.07.85	Ortho	Level II	008A01	B	10
111	26.08.85	Ortho	Level III	009B01	B	10
112	30.09.85	Wellcome	Abnormal	K9301	B	10
113	28.10.85	Gibco	Gibcotrol High	412	A	10
114	25.11.85	Wellcome	NEQAS	HIQC/9	H	10
115	30.12.85	Ortho	+Level II	008A01	B	10
116	27.01.86	Purce	Armtrol	815	B	10
117	25.02.86	Ortho	+Level III	009B01	B	10
118	28.03.86	Wellcome	NEQAS	HIQC/11	H	10
119	29.04.86	Wellcome	"Abnormal"	K7275	B	10
120	27.05.86	Technicon	TESTpoint 2	V5J072	B	10 dil
121	24.06.86	Gilford	Level II	008501	B	10
122	29.07.86	Purce	Armtrol	833	B	10
123	25.08.86	Gibco	Gibcotrol High	451	A	10
124	29.09.86	Wellcome	Heat-treated human	K466610	H	10
125	27.10.86	Gibco	Gibcotrol Abnormal	458	A	10
126	24.11.86	Wellcome	+"Abnormal"	K7275	H	10
127	29.12.86	Wellcome	+NEQAS	HIQC/11	H	10
128	26.01.87	Purce	Armtrol	848	B	10
129	23.02.87	Wellcome	+Abnormal	K930150	B	10
130	30.03.87	Gibco	+Gibcotrol High	451	A	10
131	27.04.87	Wellcome	+Heat-treated human	K466610	H	10
132	25.05.87	Wellcome	BCZ6	K340110	B	10
133	29.06.87	Wellcome	HIQC/13	K641010	H	10

Table III.2 Materials distributed in Middle East EQAS, 1983-1987.
For explanation see Table I.2

	Date	Supplier	Material	Lot	Base	Volume (mL)
33	20.01.83	DADE	Monitrol II.X	SPXP9613	H	10
34	17.02.83	Purce	Armtrol	650	B	10
35	17.03.83	Ortho	+Normal	9S219	H	10
36	21.04.83	Wellcome	+Wellcomtrol I	K2690	B	10
37	26.05.83	Ortho	Level II	X40Y02B	H	10
38	23.06.83	Purce	Armtrol	651	B	10
39	21.07.83	Gibco	Gibcotrol High	199	A	10
40	25.08.83	Purce	+Armtrol	551	B	10
41	29.09.83	DADE	+Monitrol II.X	SPXP9613	H	10
42	27.10.83	Biotrol	Biotrol-00	1929	B	10
43	01.12.83	Ortho	Level II	008Y01	B	10
44	05.01.84	Wellcome	Special Services	K6026	B	10
45	02.02.84	Ortho	+Normal	W27X02B	H	10
46	01.03.84	Ortho	Level III	009Y01	B	10
47	05.04.84	Wellcome	Special Services	K6025	B	10
48	03.05.84	Purce	Armtrol	726	B	10
49	07.06.84	Liberton	+NEQAS Trial	L4/80	H	10
50	05.07.84	Wellcome	Special Services	K6026	B	10
51	09.08.84	Gibco	Gibcotrol High	195	A	10
52	06.09.84	Ortho	Level II	008X01	B	10
53	11.10.84	Purce	Armtrol	727	B	10
54	08.11.84	Ortho	+Level II	X40Y02B	H	10
55	13.12.84	Ortho	+Level III	009Y01	B	10
56	17.01.85	Technicon	Alert II	B3M361	B	10 dil
57	14.02.85	Wellcome	Special Services	K8680	H	10
58	14.03.85	Biotrol	Biotrol-00	1934	B	10
59	18.04.85	Wellcome	Wellcomtrol II	K8609	E	10
60	16.05.85	DADE	QAP Level II	524.01	H	10
61	27.06.85	Ortho	Level II	008A01	B	10
62	25.07.85	Purce	UKEQAS Internal QC	HIQC/8	H	10
63	29.08.85	Ortho	Level III	009B01	B	10
64	26.09.85	DADE	+QAP Level II	524.01	H	10
65	24.10.85	Gibco	Gibcotrol High	412	A	10
66	28.11.85	Wellcome	Abnormal	K9301	B	10
67	09.01.86	Ortho	+Level III	009Y01	B	10
68	06.02.86	Purce	Armtrol	815	B	10
69	06.03.86	Technicon	TESTpoint 2	B4F361	B	10 dil
70	10.04.86	Wellcome	UKEQAS Internal QC	HIQC/10	H	10
71	08.05.86	Wellcome	Wellcomtrol	K7275	B	10
72	12.06.86	Gibco	+Gibcotrol High	412	A	10
73	17.07.86	Purce	Armtrol	833	B	10
74	14.08.86	Ortho	+Level III	009B01	B	10
75	18.09.86	Purce	Armtrol	848	B	10
76	16.10.86	Wellcome	+Wellcomtrol	K7275	B	10
77	20.11.86	Gibco	Gibcotrol High	451	A	10
78	01.01.87	Gilford	Level II	008501	B	10
79	05.02.87	Wellcome	BCZ6	K340110	B	10
80	05.03.87	Gibco	+Gibcotrol High	451	A	10
81	02.04.87	Wellcome	Heat-treated human	K466610	H	10
82	07.05.87	Gilford	+Level II	008501	B	10
83	11.06.87	Purce	+Armtrol	833	B	10

percentage deviations for each country. The mean of these parameters at distributions 104, 108 and 112 were calculated (Table 5.14), representing all distributions between 101 and 112. The specimens were the same as those in the IEQAS (Table III.1).

NEQAS organisers were asked to provide information on the average deviation of the laboratories participating in the WIIS from their NEQAS over the same period (October 1984 - September 1985; Table 5.16). These deviations were used as described in section 5.6.3 to estimate the net differences of each NEQAS from the WIIS consensus (Tables 5.17 and 18).

III.2 Studies of interlaboratory agreement

Data for UKEQAS distributions 181-220 (Table I.2) were accessed, including the results for the 16 analytes (sodium to magnesium; Table 4.3) then covered by the scheme. The conclusions were validated by testing them on the independent database provided by the 43 distributions (221-263) in the following two years.

III.2.1 Relationship with analyte level

The data were assessed initially by plotting against the recalculated mean (after truncation at $\pm$3 SD) the indicators of overall performance and of discrepant performance listed in Table 6.1. These graphs suggested that the overall CV was very variable and hence of little utility in assessing the quality of the specimens distributed, as were the numerical presentations of the excluded results and high VISs. These indicators were not pursued further, and attention was concentrated on the recalculated CV, average VIS and the percentage presentations.

The graphs also confirmed that the data from some distributions were at variance with the rest of the database, largely due to poor between-laboratory agreement. Further examination explained most: 4 materials containing Tris buffer (an inhibitor of urease) gave high CVs for urea, a material with a Tris/caesium/carbonate buffer as diluent and intended primarily as a calibrant for SMA systems yielded poor agreement for other methods and hence poor overall CVs, and one material with bacterial contamination and another which was subsequently confirmed by the manufacturer to have been dispensed by a defective machine gave discernibly high CVs for most analytes. These three completely unsatisfactory materials were excluded completely from further data processing,

and the others for the specific analytes.

Distributions 222, 235 and 238 from the validation period were similarly excluded; these materials had been found to be unsatisfactory for comparison purposes (one liquid serum stabilised by ethylene glycol for which incorrect handling instructions had been provided, one material for which vials from two lots had been labelled with the same lot number, and one with an apparently unacceptable filling precision). Three materials with Tris buffer were excluded from consideration for urea, as were materials with (unexplained) poor interlaboratory agreement for urate and bilirubin. Distributions 221 and 223 were excluded from the examination of data classified according to method, in order to give a uniform method classification throughout the period; this classification was different from that used in the study period.

Data classified according to method were later examined, for a more limited range of 11 analytes (the 16 studied apart from potassium, chloride, bilirubin, lithium and magnesium). The same indicators (Table 6.1) were again plotted, for each method, against the recalculated mean (for all participants), and with the same distributions as before excluded.

III.2.2 Relationship with species of origin

The reduced range of indicators were re-examined having classified the materials with respect to the species of origin of the base serum (Table 14.1).

III.2.3 Relationship with manufacturer

The reduced range of indicators were re-examined having classified the materials with respect to their manufacturer (Table 14.1). As only three manufacturers were represented by 5 or more materials the 7 manufacturers represented by more than two were considered.

The proportions of manufacturers whose materials were distributed through the NEQAS differed substantially between the study and validation periods, and the conclusions concerning Nyegaard and Scottish Blood Transfusion Service materials could therefore not be validated. The other 5 manufacturers were represented by larger numbers of materials (Table 14.1).

III.3 Studies in the UKEQAS for PKU Screening

In 1984 participants completed a questionnaire requesting details of their annual workload for 1983, the average number of assay batches per week, and their current method.

III.3.1 Relationship with workload

The average score (section 7.2.3) per survey over Surveys 15-19 (1985-1986) was calculated for each laboratory. The average turnround time, estimated as the period between reported dates of specimen receipt and analysis (see section 8.4.2 for explanation), was also determined. Both performance parameters were plotted against the laboratory's 1983 workload (Figure 11.2).

III.3.2 Relationship with analytical method

For maximum reliability of conclusions, the average score over Surveys 6-19 (1983-1986) was calculated as above. Turnround time in 1985-1986 was similarly assessed, and both parameters related to the analytical method used (Figures 8.4 and 8.5). The 'Chromatography' group included both thin-layer and paper procedures.

III.4 Calibration studies

On several occasions two or more specimens were distributed together. Participants were requested to assay the materials together in the same analytical batch.

Following the usual data processing and report production stages, a further data set was generated from the results returned. Each laboratory's result for the 'survey' specimen was recalculated, using its result for and the value assigned to the other 'reference' or 'calibrant' specimen:

$$\text{Calibrated result} = \text{Survey result} \cdot \frac{\text{Assigned value}}{\text{Reference result}}$$

The usual statistical parameters, both overall and classified according to method, were calculated, but no individual reports were generated for participants.

III.4.1 UKEQAS Enzyme Surveys

Surveys 14 to 18 each comprised two specimens (Table I.2). Participants reconstituted and assayed the materials together for selected enzymes (Table 13.6; Bullock et al, 1986b). The assigned

value was the mean value for the SCE optimised 37°C method group for all enzymes except amylase, for which the Phadebas 37°C group mean was used.

III.4.2 UKEQAS for Urinary Pregnancy Oestrogens

These studies (Table 13.4) have been described by Bullock and Wilde (1985). In study A participants received two aqueous solutions containing 40 and 60 umol/L oestriol with specimen 171 (Lever method group mean 110 umol/24h); each laboratory's mean result for these calibrants was then used in conjunction with an assigned value of 100 umol/24h (urine volume 2L).

In studies B, C and D the calibrant was another lyophilised urine specimen. This came from the same set of linearly-related specimens in study B, and from a different set in studies C and D. The Lever group means for survey specimen and calibrant were 44 and 82, 113 and 156, and 40 and 186 umol/24h respectively.

III.4.3 Urine protein surveys

Surveys 2 (November 1985) and 3 (May 1986) each included two specimens for calibration assessment (Table 13.5). The survey specimens comprised pooled urine from patients with nephrotic syndrome, and the calibrants human albumin in normal urine. In each case the assigned value was the weighed-in concentration of protein (3.4 and 5.0 g/L respectively).

III.4.4 UKEQAS for Specific Proteins

The procedure used in this scheme differed, since many methods do not have a linear calibration relationship between the measured property, eg RID immunoprecipitin ring diameter, and analyte concentration. Therefore some distributions included a calibration material with assigned values, which participants were requested to use as a calibrant in their assays (Chambers et al, 1987). This was reference preparation SPS-01, prepared and calibrated for general use in the UK for specific protein assays (Milford Ward et al, 1984).

Five distributions (Chambers et al, 1984), each of two sample pairs related by dilution (range of ratios 0.67 to 0.83), were made before this study. These yielded 20 sets of results for immunoglobulins and 8 for C3, C4 and A1-AT. Two distributions, each again comprising two dilution pairs (total 8 specimens), were made with SPS-01 calibrant; all assays were calibrated

directly or indirectly against SPS-01.

Four distributions, consisting of two unrelated specimens only (total 8 specimens), made after the SPS-01 study were considered. Three specimens were of normal human serum as before. The other 5 were pooled human sera derived from patients with inflammatory disease; these immune complexes and rheumatoid factor, and had increased concentrations of acute phase proteins.

The mean between-laboratory CV for each distribution period was calculated from pooled individual means and variances. Within-laboratory precision was estimated from the differences for each laboratory within each pair of specimens, after correction for dilution. The significance of differences in mean CVs was determined by the Mann-Whitney U test.

III.5 Studies on HDL cholesterol assay

These studies have been described previously (Bullock et al, 1980b).

III.5.1 Materials and methods

The 25 sera studied are identified in Table III.3, which also shows the species of origin of the base serum. All but serum V (a frozen liquid preparation) were lyophilised materials, obtained commercially or commissioned by UKEQAS. Pooled and individual sera from patients were also studied.

The PhT precipitation procedure was modified from that of Burstein et al (1970). Lipoproteins other than HDL were precipitated from 1 mL serum by adding 100 uL of phosphotungstate reagent (phosphotungstic acid, 45 g/L; sodium hydroxide, 160 mmol/L) and 25 uL of magnesium chloride (2.0 mmol/L). The Hep precipitation procedure was also modified from that of Burstein et al (1970). To 1 mL serum 50 uL of heparin (7500 kilo-USP units/L in 150 mmol/L saline; Grade I, cat no H3125, Sigma) and 50 uL of manganous chloride (2.02 mmol/L) were added. For both procedures, after incubation at ambient temperature (22-25^{o}C) for 30 min the precipitate was removed by centrifugation and the cholesterol content of the supernatant determined using a fully enzymic procedure (product no 187313, BCL; Röschlau et al, 1974, modified). After mixing 40 uL of supernatant with 1 mL of reagent and incubation at ambient temperature for at least 35 min, the

Table III.3 Identity of the QC materials assessed for HDL cholesterol assay All materials lyophilised except serum V (frozen liquid presentation); * denotes material not available commercially

Serum	Origin	Supplier	Material
A	Human	BCL	Precilip, 763
B	Human	BCL	Precinorm U, 801
C	Human	Bio-Rad	ECS Lypho-Check Elevated Lipid, 21196
D	Human	DADE	Moni-trol I.X, XLT-350
E	Human	GD	*Clinical Chemistry, 7109
F	Human	GD	Quality Assurance Serum Level II, 4D365
G	Human	Hyland	Q-Pak I, 1779N003AA
H	Equine	Nyegaard	Seronorm, 144
I	Bovine	Nyegaard	Seronorm Lipid, 55
J	Human	Ortho	Abnormal Unassayed, 9S317
K	Human	Ortho	Elevated Lipids, 2S906
L	Human	Liberton	*NQCS Trial, L1/78
M	Bovine	Liberton	*Clinical Chemistry QC, 12
N	Bovine	Purce	Armtrol, 390
P	Equine	Roche	Control Serum N, A2137
Q	Human	Roche	Lipid Control Serum, A0740
R	Human	SKI	Target Normal, 626011
S	Human	SKI	Target Elevated Lipid, 593311
T	Animal	Technicon	MultiSystem ABN, X9C202
U	Equine	TCS	Equitrol Lyophilised, 0419a
V	Equine	TCS	Equitrol Frozen, 0819
W	Bovine	Wellcome	Autoset H, K2847
X	Human	Wellcome	*NQCS Special Services, K5437
Y	Bovine	Wellcome	Wellcomtrol 3, K6147
Z	Human	Roche	*WHO Tentative Reference, 77/1

absorbance at 500 nm was measured and multiplied by 4.37 (PhT) or 4.27 (Hep) to obtain the serum HDL cholesterol concentration (mmol/L).

III.5.2 Study protocol

Within-batch precision (Table 12.1) was determined by replicate analysis (n=10; n=20 for the pooled sera) of each serum by both procedures. Between-day precision (Table 12.2) was also determined for 6 QCMs of varying origin and within-batch performance (sera D, G, N, P, U and W; Table III.3) by analysis on 10 successive working days. Reconstituted sera were stored (at $4^{o}C$) for no longer than 7 days before assay.

III.6 Studies on heat-treated human serum

These studies have been described previously (Bullock et al, 1986a).

III.6.1 Preparation of materials

Lots HIQC/9 to HIQC/12 (see Appendix IV) were prepared using Wellcome Diagnostics' usual manufacturing procedures. In summary, the constituent 5 litre pools were bulked and maintained at $4^{o}C$, with continual mixing, during filtration, addition of inorganic, organic, enzyme and hormone components, and sterile filtration. The filled vials of serum were then lyophilised using conventional techniques, stoppered in an atmosphere of nitrogen, capped, labelled and stored at $4^{o}C$.

After pooling, the bulk serum used in preparation of the heat-treated batch (lot K466610) was heated to $56^{o}C$ in a well-stirred vessel by gradual injection of steam into the vessel jacket. After 30 min the final temperature was attained and maintained for 30 min (the maximum temperature of the bulk serum was $57.0^{o}C$), before cold water was introduced into the vessel jacket to cool the serum to $4^{o}C$ over 5 h. The remainder of the processing was as described above for lots HIQC/9 to HIQC/12. Lot HIQC/13 was prepared similarly, except that the temperature of $56^{o}C$ was maintained for 1 h.

III.6.2 Evaluation at WRL

Measurements of pH, turbidity (absorbance at 400, 500 and 600nm of a 1 in 10 dilution of serum in 154 mmol/L saline) and microbiological contamination were each made on three vials. Vial-to-vial variability was assessed by determination of sodium

on the reconstituted contents of 40 individual vials.

Stability was assessed at 25°C and 37°C for both unreconstituted and reconstituted material. Duplicate vials were reconstituted and analysed (for sodium, potassium, urea, glucose, calcium, urate, bilirubin, total protein, albumin and AST, on an SMA 12/60; Technicon Instruments Corp) each week for 13 weeks. The reconstituted material was similarly analysed over 29 hours following reconstitution.

III.6.3 UKEQAS distributions

The materials constituted normal distributions in the UKEQAS for General Clinical Chemistry and UKEQAS Enzyme Surveys (Table I.2). For lots HIQC/9 to HIQC/11 these were separate, but lots K466610, HIQC/12 and HIQC/13 constituted distributions 305, 308 and 315, following incorporation of enzymes into the UKEQAS for General Clinical Chemistry. Lot K466610 was redistributed as 317, 6 months after its initial distribution. The materials were also distributed through the UKEQASs for Steroid Hormones and Thyroid-related Hormones (Groom, 1985a; DHSS, 1986b) as additional specimens; participants' performance was not assessed on these.

Appendix IV:

UKEQAS HUMAN SERUM FOR INTRALABORATORY QUALITY CONTROL

IV.1 Serum processing and production procedures

This material is prepared from serum collected by the National Blood Transfusion Service from male and female volunteer donors whose blood is unsuitable for therapeutic purposes. The blood is collected without anticoagulant, and the serum separated after clotting at ambient temperature. Each donation is tested for HB_SAg and (since October 1985) antibody to HIV, and those found negative are pooled and stored frozen (-20 or -30^oC).

Processing is carried out by Wellcome Diagnostics (Purce Associates for lots HIQC/7 and HIQC/8) using their usual production procedures, outlined in Appendix III.6.1 and by Bullock et al (1986a). Only inorganic and simple organic analytes were added to the initial batches, but the levels of bilirubin, enzymes, thyroid-related hormones and steroid hormones were enhanced in later batches. Analyte concentrations have not been increased greatly (to discourage use of this material for calibration), and no preservatives or stabilisers are added.

Sample vials are provided to WRL for evaluation and for distribution through the appropriate UKEQASs (Appendix III.6.2 and III.6.3). After value assignment (see below), packs of 5 or 6 vials of the remaining material are prepared and provided to UK clinical laboratories at a rate equivalent to one 10mL vial per week (or fortnight, if the laboratory needs less) to permit an occasional but regular check on the accuracy and stability of their analytical systems.

Each pack includes a package insert (Figures IV.1 and IV.2). As well as giving the assigned values, this describes the purpose, preparation and properties of the material and the value assignment procedures.

IV.2 Value assignment procedures

Evidence from the evaluation and UKEQAS distributions are considered by a Value Assignment Committee, comprising nominees from the professional bodies in clinical chemistry (Figure IV.1) and WRL staff. Provided the material is of satisfactory quality and properties, values based on the UKEQAS data are assigned.

Figure IV.1 Package insert for lot HIQC/12 - description of purpose, preparation and value assignment procedures

U.K. External Quality Assessment Scheme for General Clinical Chemistry

Human Serum for Intralaboratory Quality Control

Lot: HIQC/12

UKEQAS
Wolfson Research Laboratories
Queen Elizabeth Medical Centre
Birmingham B15 2TH

U.K. External Quality Assessment Scheme for General Clinical Chemistry*

Human Serum for Intralaboratory Quality Control, HIQC/12

INTRODUCTION In recent years the World Health Assembly has passed resolutions, the object of which is to stimulate countries to become self-sufficient with regard to human blood products, including quality control materials of human origin. The distribution of the enclosed product represents our first step in this direction. Human serum has been collected by the National Blood Transfusion Service in the United Kingdom, and has been processed for distribution within the UK for internal quality control in clinical chemistry laboratories. This project is supported by the DHSS.

This lyophilised preparation has been processed with as little modification as possible to provide, after reconstitution, a product with properties similar to those of normal fresh human serum. The object is to enable laboratories to check their assay results, occasionally but regularly, against the values assigned to the serum. Please note that the serum is intended as a check on calibration and stability of the analytical system and not as a calibration material in itself; any discrepancies between observed and assigned values should therefore stimulate the user to investigate his assay system, *not* simply to alter the values assigned to his calibration material

Human serum is scarce and difficult to collect, and most internal quality control procedures in a laboratory should use animal serum or previously assayed pooled patients' serum. Occasionally, however, a cross-check against independently assayed human serum is desirable, and our intention is to distribute sufficient lyophilised serum to provide 10 ml per laboratory per week.

ORIGIN The base of this product is human serum, collected by the National Blood Transfusion Service from volunteer male and female donors who would normally donate for therapeutic purposes but whose blood was unsuitable for transfusion. The blood was allowed to clot at ambient temperature and the exuded serum pooled after centrifugation and stored at −20°C, all operations being performed aseptically. Individual 5 litre pools which were shown to be negative for HB_sAg, were thawed and combined, and supplemented where necessary, including cortisol, oestradiol, progesterone, thyroxine and triiodothyronine. The processing, vial filling and freeze drying were carried out by Wellcome Diagnostics.

RECONSTITUTION AND STORAGE The material should be stored at 4°C, and reconstituted by adding 10·0 ml of distilled water. The material should be gently mixed for 30 minutes at ambient temperature and then stored in the dark at 4°C until used. The product should be used **on the day of reconstitution.** The material has been tested carefully **at the individual donation stage** for HB_sAg and for antibody to HTLV III (LAV, HIV), and found to be negative

*UKEQAS, formerly UK National Quality Control Scheme

STABILITY The long term stability of the unreconstituted serum has not yet been established, and a conservative estimate of 18 months at 4°C for organic and inorganic constituents has been made. The label therefore bears the recommendation that the serum is used before 30th November 1987. It is anticipated, however, that these constituents will be stable for a further year, with the possible exception of glucose and enzymes.

VALUE ASSIGNMENT The assignment of values to the material has been undertaken by the UKEQAS with the advice of the Organisers of the relevant UKEQASs (Dr. G.V. Groom for cortisol, Prof. J.G. Ratcliffe for thyroxine and triiodothyronine) and an interprofessional Value Assignment Committee consisting of representatives of relevant professional bodies. These were Dr. H. Worth (Association of Clinical Biochemists), Dr. C. Hobbs (Association of Clinical Pathologists), Mr. M. Nicholson (Institute of Medical Laboratory Sciences) and Dr. M. Rinsler (Royal College of Pathologists).

The Committee considered the evidence regarding the quality of the material and considered it to be a satisfactory batch, though slightly more turbid than might be desirable for some analyses. The Committee decided that the values assigned to the material should be the consensus values obtained by distribution through the UKEQASs in May 1986 (participants will have received detailed reports on these distributions), excluding those for some methods which were no longer considered satisfactory, for enzymes, values were assigned only for widely-used, apparently robust methods, generally optimised according to recommendations by the Scandinavian (SCE) or German (DGKC) Societies. No values were assigned for calcium by methylthymol blue, for creatine kinase (CK), iron or for total iron-binding capacity, due to excessive variability in results; variability in magnesium by colorimetric procedures was noted.

These consensus values were adopted with the assurance that there is no evidence of bias, not explicable in terms of the analytical methods used, between UKEQAS results and those of the Central Reference Institution of the German Society for Clinical Chemistry, on checking twice yearly. This cross-check is, however, carried out on equine sera, and the consensus values for this human material have therefore also been checked against the results of replicate analysis (on 5 separate days) by at least 8 UK value monitoring laboratories using a variety of calibration materials; these laboratories were chosen as consistent performers over several years in the UKEQAS rather than the current best performers.

The assigned value and SD given are those obtained by recalculation after exclusion of results lying more than 3 SD (2 SD for enzymes) from the mean, for thyroxine and triiodothyronine see the footnote to the table. These values imply that 2 out of 3 results from participating laboratories were within one SD of the assigned value. The numbers of significant figures in the assigned values are appropriate to the number of results, and often in excess of those needed in laboratory practice. Furthermore, the differences between method means are generally small, and of uncertain significance at present.

COMMENTS As the continued distribution of material of this type is still under consideration, any comments on the material and the project would be appreciated. Assigned values have been provided for 28 analytes and suitability for others is being assessed, but you may wish to use the material for additional analytes. If you do so, please report the values obtained with a brief summary of the methods used; such values and comments will be of great assistance in developing materials for future distribution.

J.G. Ratcliffe Director
D.G. Bullock Organiser, UKEQAS
Wolfson Research Laboratories

8th July 1986

Figure IV.2 Package insert for lot HIQC/12 - assigned values

Human Serum for Intralaboratory Quality Control

Assigned Values, lot HIQC/12

		Method	Assigned value	S.D.	Number of results
Alanine transaminase - ALT	(U/L)	SCE optimised, 37°C	56	8	103
		DGKC optimised, 37°C	64	8	89
Albumin	(g/L)	Bromcresol green	39.4	1.9	353
		Bromcresol purple	37.6	2.3	43
Alkaline phosphatase - ALP	(U/L)	SCE/DGKC optimised, 37°C	460	39	134
		4NPP-AMP, 37°C	245	31	142
Amylase	(U/L)	*Phadebas, 37°C	860	94	234
		*BCL colorimetric 37°C	1232	76	41
Aspartate transaminase - AST	(U/L)	SCE/DGKC optimised, 37°C	35	9	255
Bilirubin	(umol/L)	Diazotisation	54.4	5.4	405
Calcium	(mmol/L)	Atomic absorption/Automatic titrator	2.90	0.07	53
		Cresolphthalein	2.97	0.08	355
Chloride	(mmol/L)	Colorimetric/Coulometric	100.5	1.9	257
Cholesterol	(mmol/L)	Enzymic	4.45	0.23	334
Cortisol	(nmol/L)	GCMS reference method	+ 832		
		Fluorimetric	975	57	22
		Radioimmunoassay/ competitive protein binding	907	113	179
Creatinine	(umol/L)	Continuous flow Jaffe	138.6	7.5	168
		Manual/discrete (including kinetic) Jaffe	130.8	10.2	254
* Glucose	(mmol/L)	Glucose oxidase	8.22	0.30	374
		Hexokinase	8.37	0.44	89
Lactate dehydrogenase - LD	(U/L)	SCE/DGKC optimised, 37°C	952	101	144
Lithium	(mmol/L)	Flame photometry/Atomic absorption	1.32	0.06	276
Magnesium	(mmol/L)	Atomic absorption	0.93	0.04	122
		Colorimetric	0.99	0.07	116
Oestradiol	(pmol/L)	GCMS reference method	+ Not yet available		
		Radioimmunoassay	610	118	53
Osmolality	(mosmol/kg)	Freezing point/Vapour pressure	292	5.2	249
Phosphate	(mmol/L)	Colorimetric/Phosphomolybdate 340nm	1.59	0.10	383
Potassium	(mmol/L)	Flame photometry/Indirect ion-selective electrode	5.75	0.12	439
Progesterone	(nmol/L)	GCMS reference method	+ Not yet available		
		Radioimmunoassay (excluding NETRIA pH10)	26.0	4.6	104
Sodium	(mmol/L)	Flame photometry/Indirect ion-selective electrode	145.4	1.7	440
Testosterone	(nmol/L)	GCMS reference method	+ Not yet available		
		Radioimmunoassay	8.3	1.5	59
Thyroid stimulating hormone TSH	(mU/L)	Radioimmunoassay	○2.3	○0.5	66
		Monoclonal IRMA	○1.7	○0.3	80
Thyroxine	(nmol/L)	Immunoassay	○124	○11	106
Total Protein	(g/L)	Blanked Biuret	62.7	2.5	249
		Unblanked Biuret	64.8	2.2	114
Triiodothyronine	(nmol/L)	Radioimmunoassay	○3.0	○0.4	93
Urate	(mmol/L)	Uricase/Colorimetric	0.589	0.037	335
Urea	(mmol/L)	Urease/Diacetylmonoxime	3.67	0.32	465

* Please note reconstitution and storage instructions
+ Two assays in duplicate
○ Statistical calculations based on log normal distributions with outlier rejection by the method of Healy (Clin. Chem. 25, 675-677 1979)
* Pharmacia Diagnostics AB
: Boehringer Corporation (London) Ltd

For early batches individual method means were assigned. With increasing experience, the Committee's confidence in these materials has allowed the grouping of method-related data. For many analytes this policy yields a single assigned value (Figure IV.2), though data are not combined if there are indications of differences in means or SDs among the methods involved. GCMS reference method values for steroid hormones (Groom, 1985b) have also been assigned where these were available.

Appendix V:

SUPPLIERS OF QUALITY CONTROL MATERIALS AND REAGENTS

Abbott	Abbott Laboratories Ltd, Wokingham, Berks RG11 2QZ
Alpha	Alpha Laboratories Ltd, Eastleigh, Hants SO5 4NU
Amersham	Amersham International plc, Aylesbury, Bucks HP20 2TP
BCL	Boehringer Corporation London Ltd, Lewes, E Sussex BN7 1LG
Beckman	Beckman-RIIC Ltd, High Wycombe, Bucks HP12 4JL
Becton-Dickinson	Becton-Dickinson (UK) Ltd, Oxford OX4 3LY
Bio-Rad	Bio-Rad Laboratories Ltd, Watford, Herts WD1 8RP
Biotrol	Scientific Hospital Supplies Ltd, Liverpool L7 3JG
Boots-Celltech	Boots-Celltech Diagnostics Ltd, Slough, Berks SL1 4ET
Corning	Ciba-Corning Diagnostics, Halstead, Essex CO9 2DX
DADE	Travenol Laboratories Ltd, Newbury, Berks RG16 0QW
DPC	Diagnostic Products (UK) Ltd, Wallingford, Oxon OX10 9DA
GD	General Diagnostics, Chandlers Ford, Hants
Gibco	Gibco Ltd, Paisley, Scotland PA3 4EF
Gilford	Ciba-Corning Diagnostics, Halstead, Essex CO9 2DX
Hyland	Travenol Laboratories Ltd, Newbury, Berks RG16 0QW
Liberton	Protein Fractionation Centre, Scottish National Blood Transfusion Service, Edinburgh EH17 7QT
Lorne	Lorne Diagnostics Ltd, Bury St Edmunds, Suffolk IP32 7DF
Miles	Miles Laboratories Ltd, Slough Sl1 1YT
NETRIA	NE Thames Regional Immunoassay Unit, London EC1A 7BE
Nycomed	Nycomed (UK) Ltd, Birmingham B26 3EA
Nyegaard	Nycomed (UK) Ltd, Birmingham B26 3EA
Ortho	Ortho Diagnostics Ltd, High Wycombe, Bucks HP10 9UF
Purce	Purce Associates Ltd, Co Antrim, N Ireland BT41 1AB
Roche	Roche Diagnostica Division, CH-4002 Basel, Switzerland
Serono	Seronon Diagnostics Ltd, Woking, Surrey GU21 5JY
SKI	Beckman-RIIC Ltd, High Wycombe, Bucks HP12 4JL
Sigma	Sigma Chemical Co, Poole, Dorset BH17 7NH
Sorin	CIS (UK) Ltd, High Wycombe, Bucks HP12 3RD
TCS	Tissue Culture Services Ltd, Slough, Berks SL1 4XX
Technicon	Technicon Instruments Corp, Tarrytown, New York 10591, USA
Travenol	Travenol Laboratories Ltd, Newbury, Berks RG16 0QW
Wellcome	Wellcome Diagnostics, Dartford, Kent DA1 5AH
WRL	Wolfson Research Laboratories, Birmingham B15 2TH