

META-ANALYTIC APPROACHES FOR SUMMARISING AND
COMPARING THE ACCURACY OF MEDICAL TESTS

by

YEMISI TAKWOINGI

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

Public Health, Epidemiology and Biostatistics
Institute of Applied Health Research
College of Medicine and Dental Sciences
University of Birmingham
March 2016

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

Abstract

Medical tests are essential for patient care. Evidence-based assessment of the relative accuracy of competing diagnostic tests informs clinical and policy decision making. This thesis addresses questions centred on assessing the reliability and transparency of evidence from systematic reviews and meta-analyses of comparative test accuracy, including validity of meta-analytic methods.

Case studies were used to highlight key methodological issues, and provided rationale and context for the thesis. Published systematic reviews of multiple tests were identified and used to provide a descriptive survey of recent practice. Availability of comparative accuracy studies and differences between meta-analyses of direct (head-to-head) and indirect (between-study) comparisons were assessed. Comparative meta-analysis methods were reviewed and those deemed statistically robust were empirically evaluated. Using simulation, performance of hierarchical methods for meta-analysis of a single test was investigated in challenging scenarios (e.g. few studies or sparse data) and implications for test comparisons were considered.

Poor statistical methods and incomplete reporting threatens the reliability of comparative reviews. Differences exist between direct and indirect comparisons but direct comparisons were seldom feasible because comparative studies were unavailable. Furthermore, inappropriate use of meta-analytic methods generated misleading results and conclusions. Therefore, recommendations for use of valid methods and a reporting checklist were developed.

For Moti and Mary

Acknowledgements

This PhD was mainly funded by a National Institute for Health Research Doctoral Fellowship.

I am extremely grateful to my supervisors Professor Jon Deeks and Professor Richard Riley for their guidance and support. Without Jon's confidence in me and his encouragement, I would not have applied for the Fellowship that provided a fantastic opportunity to pursue my research and to develop professionally. Richard's advice both as a supervisor and a colleague have been invaluable. Thank you both very much.

I am indebted to Professor Richard Gray and Mark Nixon who gave me my first job in applied health research and supported my decision to develop a career in medical statistics. To Professor Alex Sutton whose passion for meta-analysis fuelled my interest in the subject, I owe thanks for helpful advice over the years, especially when I was unsure about starting a career in test evaluation.

My PhD has been quite a journey, along which I have been privileged to make many new friends. Special thanks to the colleagues that I have had the privilege of working with on the examples used as case studies in this thesis, and Boliang Guo and Mariska Leeftang for their contributions. I appreciate everyone who showed interest in my PhD and wellbeing, especially Professor Petra Macaskill, Karla Soares-Weiser and Alice Sitch.

I could not have maintained my sanity (still questionable) without the love and support of Moti and Mary, my daughters, and Yohanna my dear husband. You have been amazing! Last and most importantly, to my heavenly father in whom I find hope and joy, courage to face challenges and the determination to strive to give the best I can. To God be the glory.

Contents

RESEARCH DISSEMINATION	i
List of publications.....	ii
Original manuscripts (direct output).....	ii
Related manuscripts (case studies)	ii
Oral presentations	iv
Poster presentations	v
Online publications	vi
1 INTRODUCTION.....	1
1.1 Test evaluation and thesis overview	1
1.1.1 Types of medical tests	3
1.1.2 Phases of test evaluation.....	3
1.2 Role of test accuracy	7
1.3 Analysis of primary studies of a single test	9
1.3.1 Types of data	9
1.3.2 Types of measures	10
1.4 Research synthesis and methods for meta-analysis of a single test	15
1.4.1 Heterogeneity	16
1.4.2 Univariate pooling methods	17
1.4.3 Summary receiver operating characteristic regression.....	18
1.4.4 Hierarchical models.....	20

Contents

1.5 Comparisons of two or more tests.....	27
1.5.1 Diagnostic pathways.....	29
1.5.2 Study designs for test comparisons	31
1.5.3 Analysis of test comparisons in primary studies	37
1.5.4 Meta-analysis of test comparisons	38
1.6 Challenges in assessing comparative accuracy in systematic reviews.....	44
1.7 Research questions and thesis outline	48
1.7.1 Research questions	48
1.7.2 Thesis outline	49
2 METHODOLOGICAL CHALLENGES IN META-ANALYSES OF TEST	
COMPARISONS.....	53
2.1 Introduction	53
2.2 Synopsis of case studies	54
2.2.1 Rapid diagnostic tests for uncomplicated malaria in endemic countries	56
2.2.2 Screening tests for bipolar spectrum disorders.....	58
2.2.3 Diagnostic tests for common bile duct stones	58
2.2.4 Antenatal screening for Down’s syndrome	59
2.2.5 Overview of analysis methods	61
2.3 Methodological issues	62
2.3.1 What test comparisons should be performed?.....	63
2.3.2 Are meta-analytic methods that compare both SROC curves and points needed?	73
2.3.3 Should a common shape be assumed for SROC curves across tests?.....	80

Contents

2.3.4 Is the assumption of equal variances across tests appropriate?.....	87
2.3.5 Is comparative meta-analysis feasible with few studies or sparse data?.....	93
2.4 Summary.....	97
3 IDENTIFYING SYSTEMATIC REVIEWS AND META-ANALYTIC METHODS FOR TEST COMPARISONS	100
3.1 Introduction.....	100
3.2 Identification of systematic reviews.....	100
3.2.1 Terminology	101
3.2.2 Data sources and searches	102
3.2.3 Selection of reviews	103
3.3 Identification of methods for meta-analysis of comparative accuracy	105
3.3.1 Data sources and searches	105
3.3.2 Selection of methodology studies.....	107
3.4 Search results for systematic reviews	107
3.5 Search results for meta-analytic methods	109
3.6 Discussion.....	111
3.6.1 Summary of findings	111
3.6.2 Strengths and limitations	111
3.6.3 Conclusions	112
4 REVIEW OF PUBLISHED SYSTEMATIC REVIEWS OF COMPARATIVE TEST ACCURACY.....	113
4.1 Introduction.....	113

Contents

4.2 Methods ...	115
4.2.1 Review selection and data extraction	115
4.2.2 Data analysis.....	116
4.3 Results	117
4.3.1 General characteristics	118
4.3.2 Statistical characteristics	120
4.3.3 Presentation and reporting.....	127
4.4 Discussion.....	132
4.4.1 Principal findings	132
4.4.2 Comparison with other studies	136
4.4.3 Implications for research and practice.....	137
4.4.4 Strengths and limitations	141
4.4.5 Conclusions	143
5 EMPIRICAL EVIDENCE OF THE IMPORTANCE OF COMPARATIVE STUDIES OF DIAGNOSTIC TEST ACCURACY	145
5.1 Introduction.....	145
5.2 Methods.....	147
5.2.1 Review selection and data extraction	147
5.2.2 Data synthesis and analysis	149
5.3 Results	154
5.3.1 Availability of studies with comparative designs.....	154
5.3.2 Characteristics of reviews included in comparison of meta-analytic findings.....	157

Contents

5.3.3 Evidence of difference in meta-analyses of comparative and non-comparative studies	164
5.4 Discussion	170
5.4.1 Possible explanations	170
5.4.2 Comparison with existing evidence	172
5.4.3 Implications for research and practice	173
5.4.4 Strengths and limitations	174
5.4.5 Conclusions	176
6 METHODOLOGICAL REVIEW OF STATISTICAL METHODS FOR COMPARATIVE META-ANALYSIS	177
6.1 Introduction	177
6.2 Identification of comparative meta-analysis methods	178
6.3 Statistical methods for comparative meta-analysis of test accuracy	179
6.3.1 Methods for comparing summary points	182
6.3.2 Methods for comparing SROC curves	189
6.3.3 Methods comparing pooled estimates between meta-analyses	201
6.3.4 Methods for comparative meta-analysis of correlated (paired) data	203
6.4 Summary of comparative meta-analysis methods	213
6.5 Conclusions	216
7 EMPIRICAL ASSESSMENT OF COMPARATIVE META-ANALYSIS METHODS	217
7.1 Introduction	217

Contents

PART I: METHODS AND DESCRIPTION OF EMPIRICAL DATA	219
7.2 Methods	219
7.2.1 Selection of systematic reviews and data extraction	219
7.2.2 Selection of comparative meta-analysis models	220
7.2.3 Data analysis.....	221
7.3 Description of cohort of systematic reviews	229
7.3.1 Types of test comparisons	235
7.3.2 Frequency of zero cells.....	235
PART II: ASSESSMENT OF MODELLING ASSUMPTIONS IN HIERARCHICAL MODELS	237
7.4 Assessment of modelling assumptions based on separate meta-analyses of tests in test comparisons	237
7.4.1 Are hierarchical models stable in meta-analyses of individual tests?.....	238
7.4.2 Is heterogeneity in test performance similar between tests in test comparisons?	248
7.4.3 Is the shape of SROC curves similar between tests in test comparisons?.....	250
7.5 Summary of assessment of modelling assumptions in hierarchical models	251
PART III: IMPACT OF DIFFERENT MODELS ON FINDINGS.....	254
7.6 Impact of different modelling complexity on findings	255
7.6.1 Do test results need to be clustered within comparative studies in bivariate meta- regression models?.....	255
7.6.2 Is it important to allow variances to differ by test in bivariate meta-regression models?	258

Contents

7.6.3 Is it important to allow shape of SROC curves to differ between tests in HSROC meta-regression models?.....	263
7.7 Performance of different comparative meta-analysis methods.....	266
7.7.1 Comparison of bivariate and univariate meta-regression.....	266
7.7.2 Comparison of unweighted and weighted Moses SROC meta-regression.....	270
7.7.3 Comparison of HSROC and unweighted Moses SROC meta-regression.....	275
7.7.4 Comparison of HSROC and weighted Moses SROC meta-regression.....	280
7.8 Discussion	282
7.8.1 Summary of findings	282
7.8.2 Implications for research and practice.....	286
7.8.3 Strengths and limitations	288
7.8.4 Conclusions	289
8 PERFORMANCE OF METHODS FOR META-ANALYSIS WITH FEW STUDIES OR SPARSE DATA	291
8.1 Introduction.....	291
8.2 Model parsimony	293
8.3 Motivating examples of meta-analyses of a single test.....	296
8.3.1 Non-contrast computed tomography for diagnosing appendicitis	296
8.3.2 Computed tomography for diagnosing scaphoid fractures.....	296
8.3.3 Results from reanalysis of the two example datasets	297
8.4 Simulation study methods	301
8.4.1 Generation of simulated data.....	302

Contents

8.4.2	Meta-analytic models fitted to each dataset	306
8.4.3	Facilitating convergence of hierarchical models	307
8.4.4	Assessment of model convergence and stability	308
8.4.5	Assessment of performance of meta-analytic models	308
8.5	Simulation results	309
8.5.1	Estimability	309
8.5.2	Bias	314
8.5.3	Model accuracy	315
8.5.4	Coverage	317
8.5.5	Summary of simulation results and application to motivating examples	318
8.6	Application to test comparisons	320
8.7	Discussion	321
8.7.1	Principal findings and recommendations	321
8.7.2	Comparison with previous research	324
8.7.3	Strengths and limitations	325
8.7.4	Conclusions	326
9	THESIS DISCUSSION AND CONCLUSIONS	328
9.1	Overview of thesis	329
9.1.1	Are methods and reporting of comparative reviews adequate?	330
9.1.2	Are comparative accuracy studies essential for test comparisons?	331
9.1.3	What methods are available for comparative meta-analyses?	332

Contents

9.1.4 Do comparative meta-analysis methods give the same results?.....	332
9.1.5 How should meta-analyses be undertaken with few studies or sparse data?	333
9.2 Validity of test comparisons	333
9.3 Strengths and limitations.....	336
9.4 Implications of thesis findings for scientific research and practice	337
9.5 Future research	340
9.5.1 Sources of bias and variation in comparative studies.....	340
9.5.2 Heterogeneity in relative test performance	341
9.5.3 Performance of comparative meta-analysis methods with few studies or sparse data	341
9.5.4 Comparing test accuracy across multiple thresholds per study.....	342
9.5.5 Evaluation of Bayesian comparative meta-analysis methods	343
9.6 Conclusions.....	343
APPENDICES	345
Appendix A: Software programs	347
A.1 SAS program for fitting HSROC model with common or separate variance parameters across tests for Type 1 and Type 4 RDTs.....	347
A.2 Stata program for fitting the bivariate model with different covariance structures to ERCP versus IOC data.....	350
Appendix B: Forms, statistical methods and examples for Chapter 4.....	355
B.1 Screening form for reviews	355
B.2 Data extraction form for reviews.....	356

Contents

B.3 Statistical methods used for test comparisons	359
B.4 Summary of methodological and reporting characteristics of five exemplar comparative reviews	364
Appendix C: Sensitivity analysis for one-sided contour-enhanced funnel plot of the ratio of relative diagnostic odds ratio	365
Appendix D: Datasets and additional figures for Chapter 7	366
D.1 Characteristics of meta-analyses for empirical evaluation of methods.....	366
D.2 Datasets with convergence and estimation issues in HSROC models applied to individual tests	371
D.3 Estimates from bivariate models with and without assumption of equal variances across tests	373
D.4 Comparison of bivariate models with different covariance structures fitted to direct test comparisons.....	378
D.5 Estimates of relative sensitivity from HSROC models with common and different shape between tests for SROC curves	379
D.6 Comparison of relative sensitivity and relative specificity from bivariate and univariate models with equal variances	381
D.7 Estimates of relative accuracy from bivariate and univariate models with unequal variances	382
D.8 Estimates of variance and correlation parameters from bivariate and univariate models with unequal variances.....	384
D.9 Estimates from unweighted and weighted Moses SROC meta-regression models ..	386

Contents

D.10 Estimates from unweighted Moses SROC and HSROC meta-regression models...	388
Appendix E: Additional simulation results	390
E.1 Performance of all meta-analytic models in estimating sensitivity for scenarios with a DOR of 231	390
E.2 Performance of all meta-analytic models in estimating specificity for scenarios with a DOR of 231	392
REFERENCES	394

List of Tables

Table 1.1 Phases of test evaluation.....	5
Table 1.2 Cross classification of index test and reference standard results.....	8
Table 1.3 Joint classification of paired index tests and reference standard results	34
Table 2.1 Evidence profile of seven published reviews	55
Table 2.2 Types of rapid diagnostic tests for detecting malaria	57
Table 2.3 Direct comparisons of sensitivity of nine first trimester serum test strategies at the 5% false positive rate.....	69
Table 2.4 Summary estimates from direct and indirect comparisons of HRP-2 based RDTs versus pLDH based RDTs for P falciparum malaria.....	71
Table 2.5 Accuracy of the BSDS, HCL-32 and MDQ for detection of any type of bipolar disorder in mental health centre settings	79
Table 2.6 Comparison of the accuracy of BSDS, HCL-32 and MDQ for detection of any type of bipolar disorder in mental health centre settings.....	82
Table 2.7 Parameter estimates for asymmetric and symmetric HSROC models for the BSDS with and without an outlier.....	83
Table 2.8 Parameter estimates for asymmetric and symmetric HSROC models for the HCL-32 with and without an outlier.....	84
Table 2.9 Comparison of estimates from models based on different comparative approaches and assumptions.....	91
Table 2.10 Parameter and summary estimates for ERCP and IOC from models with different variance-covariance structure	96
Table 4.1 Summary of information extracted from review cohort.....	116

List of Tables

Table 4.2 Descriptive characteristics of 127 reviews of comparative accuracy and multiple tests	119
Table 4.3 Outcome measures and test comparison strategy in the reviews.....	121
Table 4.4 Meta-analysis methods used in the reviews	123
Table 4.5 Test comparison strategy and comparative meta-analysis methods.....	125
Table 4.6 Investigations of heterogeneity in the reviews	126
Table 4.7 Reporting and presentation characteristics of the reviews	130
Table 5.1 Characteristics of reviews included in the assessment of availability of comparative studies	156
Table 5.2 Characteristics of included reviews and test comparisons in each meta-analysis .	159
Table 6.1 Meta-analytic methods for comparing test accuracy	180
Table 6.2 Parameter and summary estimates from bivariate models with increasing complexity of the variance-covariance structure.....	188
Table 6.3 Comparison of unweighted and weighted Moses SROC meta-regression models	196
Table 6.4 Parameter and summary estimates from HSROC models with increasing complexity of the variance-covariance structure.....	200
Table 6.5 Discordant test results in the diseased and non-diseased groups for study i	203
Table 6.6 Probability of each combination of test results for two tests in study i	208
Table 6.7 Observed counts of test results cross-classified for two tests in study i	209
Table 6.8 Summary estimates and 95% credible intervals from alternative meta-analysis models for comparing accuracy of short femur and short humerus for Down syndrome screening.....	212
Table 6.9 Summary of comparative meta-analysis methods	214

List of Tables

Table 7.1 Characteristics of empirical dataset.....	230
Table 7.2 Parameter estimates from unstable HSROC models	239
Table 7.3 Bivariate and univariate (unequal variance) models with statistically significant differences in model fit.....	269
Table 7.4 Qualitative differences between unweighted and weighted Moses SROC meta-regression models (shape allowed to differ by test)	275
Table 7.5 Qualitative differences between HSROC and unweighted Moses SROC meta-regression models (shape allowed to differ by test)	280
Table 8.1 Summary accuracy measures obtained from different meta-analytic models applied to the two motivating examples.....	298
Table 8.2 Scenarios evaluated in the simulation	306
Table 8.3 Convergence and estimability of the complete HSROC model applied to 10 000 datasets in 36 different scenarios.....	311
Table 8.4 Performance of all meta-analytic models in estimating the log DOR for scenarios with a DOR of 231	312

List of Figures

Figure 1.1 Framework for the evaluation of <i>in vitro</i> medical tests.....	6
Figure 1.2 Relationship between sensitivity, specificity and test positivity threshold.....	11
Figure 1.3 Relationship between diagnostic odds ratios and symmetric ROC curves.....	14
Figure 1.4 Relationship between diagnostic odds ratios and asymmetric ROC curves.....	14
Figure 1.5 SROC plot of the MDQ at a common threshold of 7 for detection of bipolar disorder in mental health centre settings.....	24
Figure 1.6 SROC plot of the MDQ at different thresholds for detection of bipolar disorder in mental health centre settings.....	27
Figure 1.7 Roles of tests in diagnostic pathways.....	29
Figure 1.8 Simplified illustration of the role of existing tests in the diagnostic pathway for common bile duct stones.....	31
Figure 1.9 Study designs for comparing test accuracy.....	32
Figure 1.10 Robust study designs for comparing test accuracy.....	33
Figure 1.11 Approaches for comparing test accuracy.....	45
Figure 1.12 Types of test comparisons in a comparative accuracy meta-analysis.....	47
Figure 2.1 Hierarchy of rapid diagnostic tests evaluated against microscopy.....	65
Figure 2.2 Comparison of the nine selected first trimester serum test strategies.....	68
Figure 2.3 Comparison of summary points on SROC plots.....	70
Figure 2.4 Comparison of summary curves on SROC plots.....	72
Figure 2.5 SROC plot of rapid diagnostic tests for non-falciparum malaria.....	74
Figure 2.6 Sensitivity (detection rate) at a 5% false positive rate for the 9 selected test strategies.....	76

List of Figures

Figure 2.7| Forest plot of screening tests for detection of any type of bipolar disorder (BD type I, BD type II or BD NOS) in mental health centre settings..... 78

Figure 2.8| Scatterplot of D (log odds ratio) against S (implicit threshold) for the BSDS, HCL-32 and MDQ 81

Figure 2.9| Forest plot of BSDS, HCL-32 and MDQ for detection of bipolar disorder type II in mental health centre settings..... 85

Figure 2.10| Summary ROC plot of the BSDS, HCL-32 and MDQ for detection of bipolar disorder type II in mental health centre settings 86

Figure 2.11| Comparison of heterogeneity in test performance for Type 1 and Type 4 rapid diagnostic tests..... 88

Figure 2.12| Forest plot of endoscopic retrograde cholangiopancreatography (ERCP) and intraoperative cholangiography (IOC) for diagnosis of common bile duct stones..... 94

Figure 3.1| Flowchart of selection of systematic reviews 108

Figure 3.2| Flowchart of selection of reports of meta-analytic methods 110

Figure 4.1| Cohort of reviews included in the review of reviews..... 118

Figure 4.2| Reporting characteristics of 127 reviews 128

Figure 5.1| Flowchart of review and meta-analysis selection..... 158

Figure 5.2| Ratio of relative diagnostic odds ratios (with 95% confidence intervals)..... 165

Figure 5.3| One-sided contour-enhanced funnel plot of the ratio of relative diagnostic odds ratio..... 167

Figure 5.4| Absolute differences in sensitivity and specificity between tests (with 95% confidence intervals)..... 169

Figure 6.1| Number of citations per year for original publications of the Moses model, bivariate model and HSROC model for test accuracy meta-analysis..... 181

List of Figures

Figure 6.2 SROC plot of dermoscopy and unaided eye for diagnosis of melanoma.....	187
Figure 6.3 Forest plot of RIA and ELISA for diagnosis of congestive heart failure.....	192
Figure 6.4 Comparison of summary curves from Moses SROC meta-regression models	194
Figure 7.1 SROC plots of meta-analytic datasets with mixed thresholds	232
Figure 7.2 SROC plots of meta-analytic datasets with common thresholds (1).....	233
Figure 7.3 SROC plots of meta-analytic datasets with common thresholds (2).....	234
Figure 7.4 Distribution of number of primary studies for the index and comparator tests in the test comparisons	235
Figure 7.5 Number of studies with zero cells in each test comparison	236
Figure 7.6 Distribution of number of studies in each meta-analysis of a single test	238
Figure 7.7 Estimates of the correlation parameter from meta-analyses of individual tests in 57 test comparisons	241
Figure 7.8 Comparison of estimates of variances from bivariate and univariate meta-analyses	242
Figure 7.9 Forest plot of sensitivity and specificity for the dataset with the largest difference between univariate and bivariate meta-analyses	243
Figure 7.10 Comparison of estimates of standard errors of mean logit sensitivities and mean logit specificities from univariate and bivariate meta-analyses.....	244
Figure 7.11 Comparison of estimates of beta from HSROC models with estimates derived from bivariate models.....	245
Figure 7.12 Estimates of beta and its standard error against number of studies in each meta- analysis	246
Figure 7.13 Estimates of beta and their 95% confidence intervals from HSROC models	247

List of Figures

Figure 7.14| SROC plots of three meta-analyses with 12 included studies and substantial uncertainty in estimation of beta248

Figure 7.15| Variance estimates for random effects for logit sensitivity and logit specificity from meta-analysis of each test in the test comparisons249

Figure 7.16| Variance estimates for random effects for accuracy and threshold from meta-analysis of each test in the test comparisons250

Figure 7.17| Estimates of beta (shape parameter) from separate meta-analyses of pairs of tests in 52 test comparisons251

Figure 7.18| Differences in estimates from bivariate meta-regression models with and without clustering of test results in comparative studies256

Figure 7.19| Comparison of relative sensitivity and relative specificity from bivariate meta-regression models with and without clustering of test results in comparative studies257

Figure 7.20| Differences in estimates from bivariate meta-regression models with equal and unequal variances259

Figure 7.21| Comparison of relative sensitivity and relative specificity from bivariate meta-regression models with equal and unequal variances260

Figure 7.22| Comparison of bivariate meta-regression models with different covariance structures fitted to direct test comparisons262

Figure 7.23| Differences in estimates from HSROC meta-regression models with common and different shape for SROC curves.....264

Figure 7.24| Comparison of relative sensitivity from HSROC meta-regression models with common and different shape between tests for SROC curves265

Figure 7.25| Differences in estimates from bivariate and univariate meta-regression models with unequal variances267

List of Figures

Figure 7.26 Comparison of relative sensitivity and relative specificity from bivariate and univariate models with unequal variances.....	268
Figure 7.27 Differences in estimates from unweighted and weighted Moses SROC meta-regression (same shape) models	270
Figure 7.28 Comparison of unweighted and weighted Moses SROC meta-regression (same shape) models	271
Figure 7.29 Differences in estimates from unweighted and weighted Moses SROC (different shape) meta-regression models.....	273
Figure 7.30 Comparison of unweighted and weighted Moses SROC meta-regression (different shape) models	274
Figure 7.31 Differences in estimates from HSROC and unweighted Moses SROC meta-regression (same shape) models	276
Figure 7.32 Comparison of unweighted Moses SROC and HSROC meta-regression (same shape) models	277
Figure 7.33 Differences in estimates from HSROC and unweighted Moses SROC (different shape) meta-regression models.....	278
Figure 7.34 Comparison of unweighted Moses SROC and HSROC meta-regression models	279
Figure 7.35 Differences in estimates from HSROC and weighted Moses SROC meta-regression (same shape) models	281
Figure 7.36 Differences in estimates from HSROC and weighted Moses SROC meta-regression (different shape) models.....	282
Figure 8.1 Forest plot of sensitivity and specificity estimates from studies included in the two motivating examples.....	297

List of Figures

Figure 8.2 Profile log-likelihood function of the covariance parameter in the bivariate model applied to the appendicitis example	300
Figure 8.3 Underlying bilogistic distribution of diseased and non-diseased used in the simulation	303
Figure 8.4 Proportion of meta-analyses that successfully converged for the complete HSROC model in 36 different scenarios with heterogeneity in accuracy and threshold.....	310
Figure 8.5 Bias in the estimated diagnostic odds ratio for 36 scenarios with heterogeneity in accuracy and threshold	315
Figure 8.6 Mean square error for the estimated log diagnostic odds ratio for nine different scenarios with heterogeneity in accuracy and threshold and diagnostic odds ratio of 231	316
Figure 8.7 Coverage of 95% confidence intervals for the log diagnostic odds ratio for nine different scenarios with heterogeneity in accuracy and threshold and a diagnostic odds ratio of 231	318

List of Boxes

Box 2.1| SAS Proc NLMIXED code for each model.....89

Box 4.1| Criteria for reporting test comparisons in systematic reviews of test accuracy.....139

Box 8.1| Recommendations for selecting alternative models when bivariate or HSROC
models fail.....323

RESEARCH DISSEMINATION

Findings from the thesis have been published as research articles in peer reviewed journals and presented at local and international meetings as oral or poster presentations. Although supervisors and external collaborators have contributed by helping to refine some of the research ideas, revise and critique the manuscripts, the author is the primary contributor of direct outputs of the thesis and the conference contributions listed below. Other publications related to the thesis, such as the systematic reviews used as case studies in Chapter 2, were undertaken in collaboration with clinical colleagues and other methodologists. The author planned and conducted the statistical analyses in the reviews, and contributed substantially to drafting and revising the manuscripts.

Chapters 5 and 8 have been published as peer reviewed papers. The introductory sections of these papers have contributed material to the introductory Chapter of the thesis. Sections of Chapter 1 on types of data, types of measures and meta-analysis (using the bipolar disorder review presented in Chapter 2 as a working example) have also been published as a peer reviewed paper. Data sources and searches reported in the papers based on Chapters 5 and 8 have contributed to Chapter 3. Manuscripts are being prepared for the content of Chapters 4, 6 and 7.

Preliminary findings from Chapters, 4, 5 and 8 were presented as oral presentations at Cochrane Colloquia. One of the case studies in Chapter 2 (section 2.3.4.1) was also presented as a poster. The presentation based on Chapter 5 was awarded the Thomas Chalmers Prize at the 19th Cochrane Colloquium. Full results from Chapter 5 were presented as an oral presentation at an international conference of the Royal Statistical Society. The author has

also been involved in developing online distance learning modules and given workshops on topics related to the content of this thesis. Details of all publications are given below grouped according to publication type.

List of publications

Original manuscripts (direct output)

1. **Takwoingi Y**, Guo B, Riley R, Deeks J. Performance of methods for meta-analysis of diagnostic test accuracy with few studies or sparse data. *Stat Methods Med Res*. Epub ahead of print June 26 2015.
2. **Takwoingi Y**, Riley R, Deeks J. Meta-analyses of diagnostic accuracy studies in mental health. *Evid Based Ment Health* 2015; 18:103-109.
3. **Takwoingi Y**, Leeflang MM, Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. *Ann Intern Med* 2013;158(7):544-54.

Related manuscripts (case studies)

1. **Takwoingi Y**, Abba K, Garner P. Rapid diagnostic testing for *Plasmodium vivax* and non-falciparum malaria in endemic areas. *JAMA* 2015;314(10):1065–66.
2. Alldred SK, **Takwoingi Y**, Deeks JJ, Guo B, Pennant M, Alfirevic Z, Neilson JP. First trimester serum tests for Down's syndrome screening. *Cochrane Database Syst Rev* 2015;11:CD011975.

3. Gurusamy KS, Giljaca V, **Takwoingi Y**, Higgle D, Poropat G, Štimac D, Davidson BR. Endoscopic retrograde cholangiopancreatography versus intraoperative cholangiography for diagnosis of common bile duct stones. *Cochrane Database Syst Rev* 2015;2:CD010339.
4. Giljaca V, Gurusamy KS, **Takwoingi Y**, Higgle D, Poropat G, Štimac D, Davidson BR. Endoscopic ultrasound versus magnetic resonance cholangiopancreatography for common bile duct stones. *Cochrane Database Syst Rev* 2015;2:CD011549.
5. Gurusamy KS, Giljaca V, **Takwoingi Y**, Higgle D, Poropat G, Štimac D, Davidson BR. Ultrasound versus liver function tests for diagnosis of common bile duct stones. *Cochrane Database Syst Rev* 2015;2:CD011548.
6. Carvalho AF, **Takwoingi Y**, Sales PMG, Soczynska JK, Köhler CA, Freitas TH, Quevedo J, Hyphantis TN, Roger S. McIntyre RS, Vieta E. Screening for bipolar spectrum disorders: a comprehensive meta-analysis of accuracy studies. *J Affect Disord* 2014;172C:337-346.
7. Abba K, Kirkham AJ, Olliaro PL, Deeks JJ, Donegan S, Garner P, **Takwoingi Y**. Rapid diagnostic tests for diagnosing uncomplicated non-falciparum or Plasmodium vivax malaria in endemic countries. *Cochrane Database Syst Rev* 2014;12:CD011431.

8. Abba K, Deeks JJ, Olliaro PL, Naing CM, Jackson SM, **Takwoingi Y**, Donegan S, Garner P. Rapid diagnostic tests can extend access of diagnostic services for uncomplicated Plasmodium falciparum malaria. *Int J Epidemiol* 2012;41(3):607-608.
9. Abba K, Deeks JJ, Olliaro P, Naing CM, Jackson SM, **Takwoingi Y**, Donegan S, Garner P. Rapid diagnostic tests for diagnosing uncomplicated P. falciparum malaria in endemic countries. *Cochrane Database Syst Rev* 2011;7:CD008122.

Oral presentations

1. Reporting and methods in systematic reviews of comparative accuracy. 21st Cochrane Colloquium, Quebec, Canada. September 2013.
2. Empirical assessment of the validity of uncontrolled comparisons of the accuracy of diagnostic tests. Royal Statistical Society's International Conference, Telford. September 2013.
3. Reporting and methods in systematic reviews of comparative accuracy. 3rd International Symposium on Methodology for Evaluating Medical Tests. University of Birmingham. July 2013.
4. An empirical assessment of the validity of uncontrolled comparisons of the accuracy of diagnostic tests. 2nd International Symposium on Methodology for Evaluating Medical Tests. University of Birmingham. July 2010.

Dissemination

5. An empirical assessment of the validity of uncontrolled comparisons of the accuracy of diagnostic tests. 17th Cochrane Colloquium, Singapore. October 2009 (Awarded the Thomas C Chalmers Prize).
6. Performance of methods for meta-analysis of diagnostic test accuracy studies when few studies are available. 17th Cochrane Colloquium, Singapore. October 2009.

Poster presentations

1. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. Health and Population Sciences Poster day. University of Birmingham. October 2012.
2. Discrepancy in relative test performance due to modelling strategy in comparative diagnostic meta-analysis: a case study. Health and Population Sciences Poster day. University of Birmingham. October 2012.
3. Performance of methods for meta-analysis of diagnostic tests. Health and Population Sciences Poster day. University of Birmingham. October 2012.
4. Discrepancy in relative test performance due to modelling strategy in comparative diagnostic meta-analysis: a case study. 19th Cochrane Colloquium, Madrid, Spain. October 2011.

5. Performance of methods for meta-analysis of diagnostic test accuracy studies when few studies are available. 2nd International Symposium on Methodology for Evaluating Medical Tests. University of Birmingham. July 2010.

Online publications

1. **Takwoingi Y**, Leeflang MM, Davenport CF, Deeks JJ. Analysis with RevMan for Cochrane DTA reviews. Lesson 7.2: Cochrane Collaboration DTA Online Learning Materials. The Cochrane Collaboration, September 2014. Videocast (49 slides, 16 min, sound, colour). Available at <http://training.cochrane.org/authors/dta-reviews/distance-learning>.
2. **Takwoingi Y**. Meta-analysis of test accuracy studies in Stata: a bivariate model approach. Version 1.0. November 2013. Available from: <http://dta.cochrane.org/>.

1 INTRODUCTION

A paper based partly on the content of this chapter has been published.

Citation: **Takwoingi Y**, Riley R, Deeks J. Meta-analyses of diagnostic accuracy studies in mental health. *Evidence Based Mental Health* 2015; 18:103-109.

1.1 Test evaluation and thesis overview

Medical tests are necessary to resolve uncertainty about the health status of an individual. A clinically effective test should reduce ambiguity in clinical decision making, lead to prompt and appropriate intervention, and ultimately improve patient outcomes. For example, tuberculosis (TB) was declared a global emergency by the World Health Organization (WHO) in 1993.¹ The spread of HIV and the emergence of multidrug-resistant TB (MDR-TB) have worsened the impact of TB in terms of morbidity and mortality.² As such rapid diagnosis of MDR-TB is a global priority for TB control. The WHO recommends the use of the Xpert® MTB/RIF system as the initial diagnostic test in individuals suspected of having MDR-TB or HIV-associated TB.³ The information obtained from the test enables quicker decisions about treatment and infection control compared to culture.⁴ Since testing is pivotal to health care, tests should only be recommended for routine clinical use based on evidence of their clinical performance (i.e. diagnostic accuracy) and clinical effectiveness (i.e. benefits and harms) derived from relevant, high quality primary studies and systematic reviews. Well conducted systematic reviews of relevant, high quality primary research studies are generally considered to be the highest level of evidence.⁵

Diagnostic accuracy is the ability of a test to correctly identify or exclude a target condition.

A systematic review of diagnostic test accuracy (DTA) aims to identify and summarise

Chapter 1: Introduction

evidence from multiple studies addressing the same question, including an assessment of the quality and consistency of the evidence. The review may include a meta-analysis by applying statistical methods to combine the results of multiple studies. A DTA review may summarise the accuracy of one or more tests individually or compare their accuracy in a meta-analysis. This thesis focuses primarily on improving comparisons of test accuracy in systematic reviews and meta-analyses to ensure valid methods are used in identifying the most accurate test(s) from a number of competing tests. Comparative accuracy is an area of increasing importance and relevance to health technology assessment. This will be discussed further in sections 1.5 and 1.6.

This chapter provides the background and foundation for the work in the remainder of the thesis. In particular, it gives an understanding of key concepts, demonstrates the role of diagnostic accuracy within the wider context of test evaluation, and then defines the scope of the thesis questions in more detail. The first part of this chapter is designed to promote awareness of the complexities of test evaluation. Sufficient detail is given about test types to enable an understanding of the characteristics of the systematic reviews described in the thesis. The second part focuses on the role of test accuracy, analysis of primary studies of a single test and their synthesis in systematic reviews. The rationale for systematic reviews and the common meta-analytic methods for pooling test accuracy studies, which will be evaluated in later chapters, are introduced. Finally, the third part deals with test comparisons, including study design, roles of tests within a diagnostic pathway and the challenges in assessing comparative accuracy in systematic reviews and meta-analyses. In addition, justification for the questions considered in subsequent chapters is summarised.

For simplicity, the term patient will be used throughout the thesis instead of participant. The only exception will be where specific reference is made to individuals suspected of a particular target condition that is clearly not a disease e.g. screening pregnant women for fetal aneuploidy. The terms target condition and disease will be used interchangeably as appropriate.

1.1.1 Types of medical tests

There is a wide range of test types. Tests may be performed outside a living body (*in vitro* diagnostics) or directly applied on or inside the body (*in vivo* diagnostics). Some tests are specific for a particular target condition or group of related conditions (e.g. amniocentesis for detection of fetal aneuploidy), while other tests may be used for a number of diseases (e.g. CA-125, a biomarker for certain types of cancers). Tests may be as simple as individual elements of clinical/physical examination and history taking or highly sophisticated technological devices. Information from two or more tests may be combined to derive multivariable diagnostic models, clinical prediction rules or algorithms e.g. Alvarado score for diagnosis of appendicitis.⁶ Several test types will appear in examples and the cohorts of systematic reviews used in this thesis. The wide range of test types leads to different types of data and measures for quantifying test accuracy which in turn have implications for meta-analytic methods. The next section provides a synopsis of test evaluation, a necessary precursor for placing diagnostic accuracy into context.


1.1.2 Phases of test evaluation

The process of test evaluation is multifaceted and challenging, requiring a clear definition of the intended use and role of a test for a specific population within the context of a clinical

pathway. Since test evaluation is multifaceted, evaluation cannot be accomplished in a single study but requires a sequence of studies that address different aspects of test performance. The studies are often undertaken in an order which reflects increasing expense, embedding the tests deeper in clinical pathways, and an appreciation of the resource implications of implementation in clinical practice.

Various frameworks have been developed for the process of test evaluation, sometimes designed for specific test types such as laboratory tests,⁷ genetic tests,⁸ biomarkers⁹ and imaging tests.¹⁰ Lijmer et al systematically reviewed the literature on schemes for the evaluation of medical tests and identified 19 schemes published between 1978 and 2007.¹¹ Each scheme consists of between four and seven phases. The six common phases and research questions they address are summarised in Table 1.1. At the lowest level of the hierarchy is the evaluation of technical performance and at the highest level is societal efficacy expressed in terms of resource use and societal benefits.

Table 1.1| Phases of test evaluation



Phase	Type of research question
Analytical performance	Does the test give usable information (reliable and reproducible)?
Diagnostic accuracy	How well does the test distinguish between diseased and non-diseased individuals?
<i>Phase I</i>	<i>Do test results in patients with the target condition differ from those in healthy people?</i>
<i>Phase II</i>	<i>Are patients with certain test results more likely to have the target condition than patients with other test results?</i>
<i>Phase III</i>	<i>Does the test result distinguish between patients with and without the target condition in a clinically relevant population?</i>
Diagnostic thinking	Does the test change diagnostic reasoning and decisions?
Therapeutic efficacy	Does the test change patient management?
Clinical effectiveness	Do patients who undergo the test have better clinical outcomes than those who were not tested?
Societal efficacy	Is the test resource-efficient and beneficial for society?

This summary was derived from Lijmer et al.¹¹ The diagnostic accuracy phase shows the multiphase model proposed by Sackett and Haynes.¹²

Lijmer et al concluded that test evaluation is most likely a cyclic and repetitive process rather than a stepwise linear one.¹¹ More recently, Horvath et al⁹ proposed a dynamic cycle driven by the clinical pathway as shown in Figure 1.1. The cycle consists of five components related to the phases in Table 1.1—analytical performance, clinical performance (diagnostic accuracy), clinical effectiveness, cost effectiveness and broader impact (societal efficacy). Cost effectiveness is frequently assessed along with clinical effectiveness and is termed societal efficacy when costs are considered from a societal perspective.



Figure 1.1| Framework for the evaluation of *in vitro* medical tests

This cyclical approach to test evaluation is driven by the clinical pathway and shows the dynamic inter-relationship between the different components of the process. The original purpose and role of the test can be redefined if an alternative use for the test becomes apparent during the evaluation process.⁹ This new role for the test will inform the design of subsequent evaluations. (Adapted from Horvath et al 2014⁹)

As mentioned earlier in section 1.1, the value of a test ultimately depends on its clinical effectiveness, i.e., effect on patient outcomes. Test accuracy can potentially be linked to the accuracy of clinical decision making through the downstream consequences of true positive, false positive, false negative and true negative test results, but benefits and harms to patients may also be driven by other factors too. According to Ferrante di Ruffano et al¹³, testing represents the first step of a test-plus-treatment pathway and changes to components of this pathway following the introduction of a new test could trigger changes in health outcomes. They identified several mechanisms which they summarised under direct effects of testing, changes to diagnostic and treatment decisions or timeframes, and alteration of patient and clinician perceptions.

Following the overview of the test evaluation process, the remaining sections of this chapter will now focus on diagnostic accuracy. First single test evaluations will be considered and then test comparisons.

1.2 Role of test accuracy

Assessment of diagnostic accuracy is an integral part of test evaluation as shown in Table 1.1. Diagnostic accuracy describes the ability of a test to classify individuals, typically simplified into a dichotomy by applying a criteria (referred to as thresholds, cut-offs or cut-points) to define test negatives and test positives. Several authors have proposed multiple phases in the evaluation of diagnostic accuracy to distinguish between early assessment of test performance (i.e. proof-of-concept or exploratory studies) in a population of known cases and non-cases (case control study), and later assessment in a representative population in an appropriate clinical setting (prospective cross-sectional study of suspected cases).^{12,14-16}

Test accuracy is estimated by comparing results of an index test (a new or existing test of interest) with a reference standard, sometimes known as a ‘gold’ standard, as shown in Table 1.2. The reference standard is used to verify the presence or absence of the target condition, and may be a single test or a combination of tests and clinical information not routinely available in practice.

Table 1.2| Cross classification of index test and reference standard results

	Reference standard positive	Reference standard negative	Total
Index test positive	<i>a</i> (true positives)	<i>b</i> (false positives)	<i>a + b</i> (test positives)
Index test negative	<i>c</i> (false negatives)	<i>d</i> (true negatives)	<i>b + d</i> (test negatives)
Total	<i>a + c</i> (disease positives)	<i>b + d</i> (disease negatives)	<i>a + b + c + d</i> (study total)

In the United Kingdom, the National Institute for Health and Care Excellence (NICE) recommend that evidence for the appraisal of diagnostic technologies should normally incorporate evidence on the accuracy of the diagnostic technology.¹⁷ According to the Australian Medical Services Advisory Committee (MSAC) guideline, in the absence of RCTs of test effectiveness, i.e. direct evidence, linked evidence can be used to infer test effectiveness when the transferability of results between studies of test accuracy and studies of treatment effectiveness can be reasonably justified.¹⁸ Also, European Medicines Agency (EMA) guidance recommends that where it is already known that intervention following the use of a diagnostic test leads to a clinical benefit, there is no need to repeat this proof for every subsequent diagnostic test in the same setting. However, evidence of an adequate reasoning that a clinical benefit is expected should be provided.¹⁹

Assessment of diagnostic accuracy has limitations. Improved test accuracy does not guarantee better clinical outcomes, does not convey decisions made based on test results, and the link between test accuracy and patient outcomes is complex. Lord et al²⁰ suggested a framework to help decide whether a new diagnostic test can be adopted based only on evidence of test accuracy or if studies of clinical effectiveness are needed. The test-treatment randomized controlled trial (RCT) is regarded as the ideal study design for evaluating clinical effectiveness or impact on patient outcomes.²¹ In this design, patients are randomized between

new and existing tests, followed by appropriate management or intervention based on test results, and finally patient outcomes are measured and assessed. However, the cost and duration of test-treatment RCTs often make them unrealistic, and so they are rare.²² An alternative is to assess health outcomes using decision analysis (or the MSAC linked evidence approach¹⁸) to model the consequences of testing and alternative management strategies. This will be discussed briefly in section 1.5.

1.3 Analysis of primary studies of a single test

In the following subsections, types of data and summary measures that can be used to determine diagnostic accuracy are described. Several measures used in primary studies are also used in meta-analysis to quantify test accuracy. Therefore, an introduction to the commonly used measures provides useful background prior to discussing meta-analytic models in section 1.4 and other chapters of the thesis.

1.3.1 Types of data

The diversity of test type leads to different types of data for computing test accuracy. Data may be nominal (binary), ordinal (ordered categories), discrete (count) or continuous. Standard methods for computing test accuracy demand binary classification of the results of the index test and the reference standard (Table 1.2). As such for non-binary data, thresholds are needed to dichotomise the data. The threshold may be an explicit numeric value or may be implicit based on subjective visual interpretation or judgement.

1.3.2 Types of measures

Several measures are used to quantify test accuracy. These may be paired measures or single indicators of test performance. Where a test is measured on a continuum, paired measures relate to test performance conditional on a particular threshold. Some single measures are also threshold specific while others are global, assessing performance across all possible thresholds. The different types of measures are explained below.

1.3.2.1 Paired measures of test accuracy

Paired measures—sensitivity and specificity, positive and negative predictive values, and positive and negative likelihood ratios (LR+ and LR−)—are typically used to quantify test performance because of the need to distinguish between the presence and absence of the target condition. Sensitivity and specificity are the most commonly reported measures.

Sensitivity is the probability that those with the target condition are correctly identified as having the condition (i.e. $P[Y = 1|D = 1]$), while specificity is the probability that those without the target condition are correctly identified as not having the condition (i.e. $P[Y = 0|D = 0]$). D and Y are binary variables that denote disease status and test result respectively.

Sensitivity is also known as the true positive rate (TPR), true positive fraction (TPF) or detection rate, and specificity as the true negative rate (TNR) or true negative fraction (TNF).

The false positive rate (FPR) or false positive fraction (FPF), 1 -specificity, is sometimes used instead of specificity.

Figure 1.2 (panels A to E) show the effect of varying test threshold on sensitivity and specificity. For a test where disease increases the value of a biomarker, as the threshold decreases, sensitivity increases while specificity decreases, and vice versa. This trade-off

between sensitivity and specificity across thresholds is graphically depicted in a receiver operating characteristic (ROC) plot as shown in Figure 1.2 (panel F).

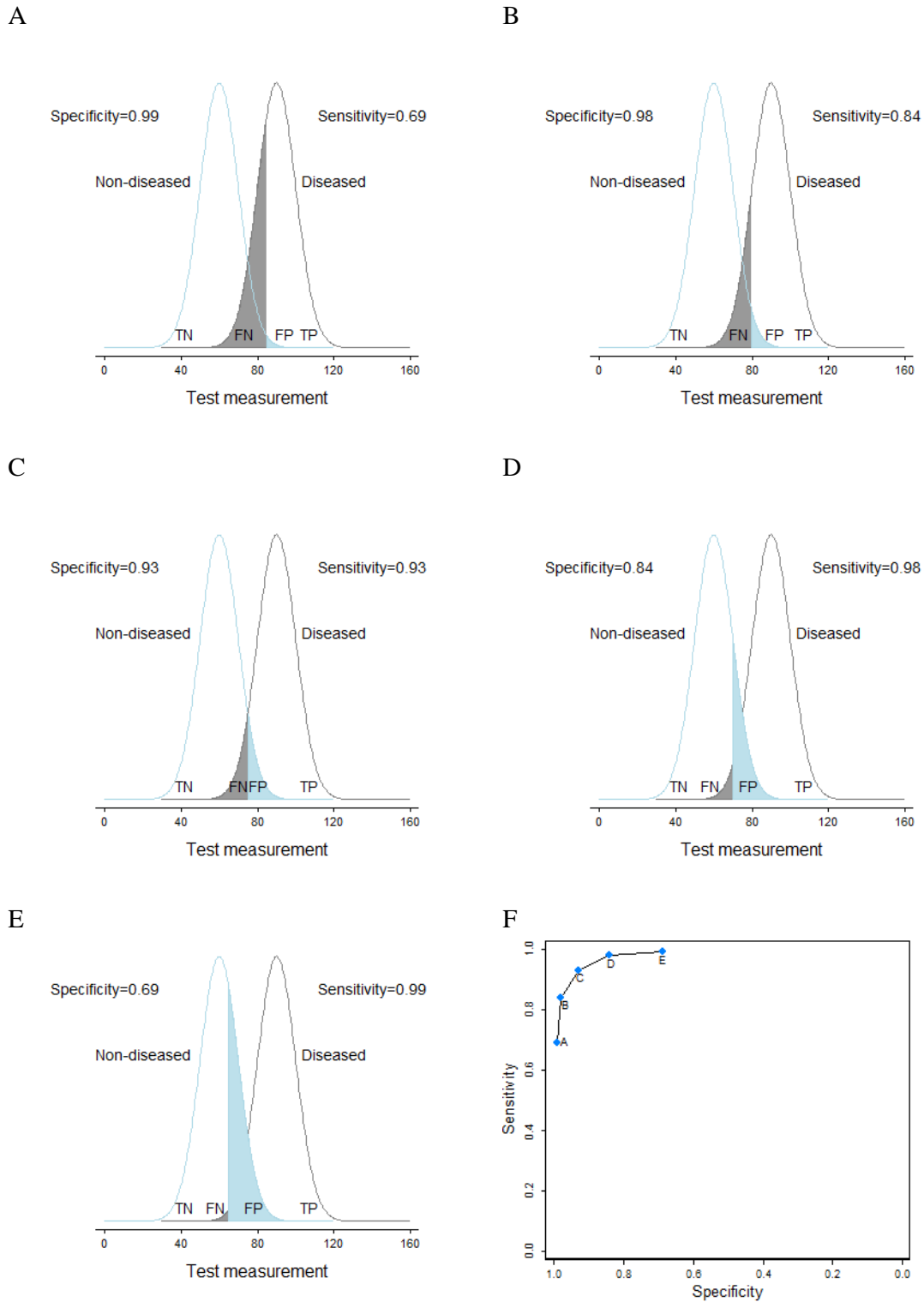


Figure 1.2| Relationship between sensitivity, specificity and test positivity threshold
 FN = false negative; FP = false positive; TN = true negative; TP = true positive.
 (Adapted from Macaskill et al 2010²³)

Traditionally, the ROC plot is a plot of sensitivity against 1-specificity. However, as in Figure 1.2 (panel F), it is possible to plot sensitivity against specificity by reversing the labelling of the specificity axis. The position of the ROC curve depends on the discriminatory ability of the test which is illustrated as the degree of overlap of the distributions of test measurements for the diseased and non-diseased groups (Figure 1.2 panels A to E); the more accurate the test is, the closer the curve to the upper left hand corner of the ROC plot. The ROC curve can also be used to evaluate the discriminative ability of multivariable diagnostic and prognostic models.²⁴

1.3.2.2 Single and global measures of test accuracy

As alluded to earlier, some single measures of test accuracy can be regarded as global measures of accuracy because they assess test performance across all possible thresholds. This includes the diagnostic odds ratio (DOR) and the area under the curve (AUC). Other single measures like Youden's index (sensitivity + specificity-1) and probability of a correct result (also known as accuracy) are threshold specific. Single measures are not always relevant in clinical practice because there is no information on the error rates in the diseased (false positives) and non-diseased groups (false positives). These error rates are important for judging the extent and likely impact of downstream consequences.²⁵

Due to its relevance for meta-analysis as will be shown in section 1.4, only the DOR is described here. The DOR is defined as the ratio of the odds of positivity in those who have the target condition relative to the odds of positivity in those without the condition.²⁶

Therefore, a test with high TPR and low FPR will have a high DOR. The DOR is calculated as

$$\text{DOR} = \frac{\text{TP} \times \text{TN}}{\text{FP} \times \text{FN}},$$

where TP, TN, FP, and FN are the number of true positives, true negatives, false positives and false negatives. The DOR can also be expressed in terms of sensitivity and specificity, or likelihood ratios as follows:

$$\text{DOR} = \frac{\frac{\text{sensitivity}}{1 - \text{sensitivity}}}{\frac{1 - \text{specificity}}{\text{specificity}}} = \frac{\text{LR}+}{\text{LR}-}.$$

The ROC curves shown in Figure 1.3 are all symmetrical about the line of symmetry (sensitivity is equal to specificity at each point on this line). An ROC curve is symmetric when the test results for the diseased and non-diseased groups have logistic distributions with equal variances. When the variances are unequal, the ROC curve is asymmetric (Figure 1.4). Symmetric ROC curves that arise from bilogistic distributions can be summarised by a constant DOR as shown for each of the six curves in Figure 1.3. In contrast, asymmetric ROC curves cannot be summarised by a single DOR because the DOR varies with change in threshold (Figure 1.4 panel B).

The DOR can be a useful measure when comparing tests, particularly if there is interest in global performance and no preference for either superior sensitivity or specificity. ROC curves can also be used to compare tests. Such comparisons account for the difference in test accuracy across the range of possible thresholds for each of the tests.

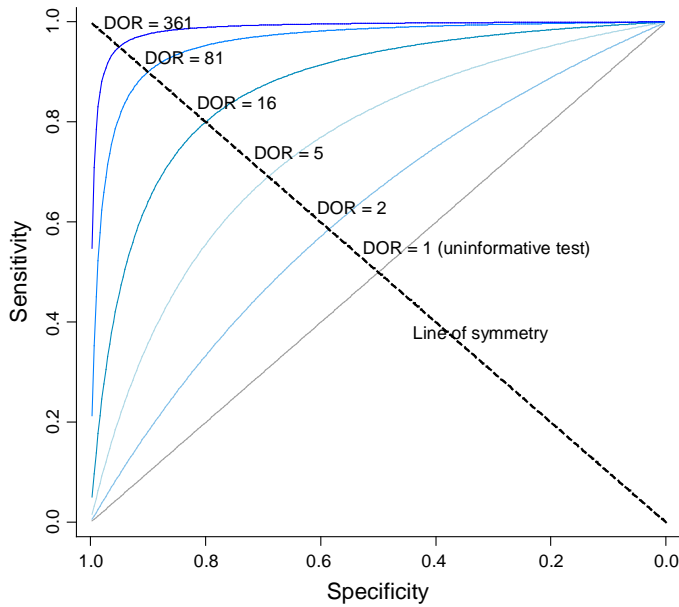


Figure 1.3| Relationship between diagnostic odds ratios and symmetric ROC curves

DOR = diagnostic odds ratio.

Six symmetric ROC curves are shown on the figure. Each point on a curve corresponds to the common diagnostic odds ratio shown on the curve. Sensitivity is equal to specificity at each point on the line of symmetry (downward diagonal).

(Adapted from Macaskill et al 2010²³)

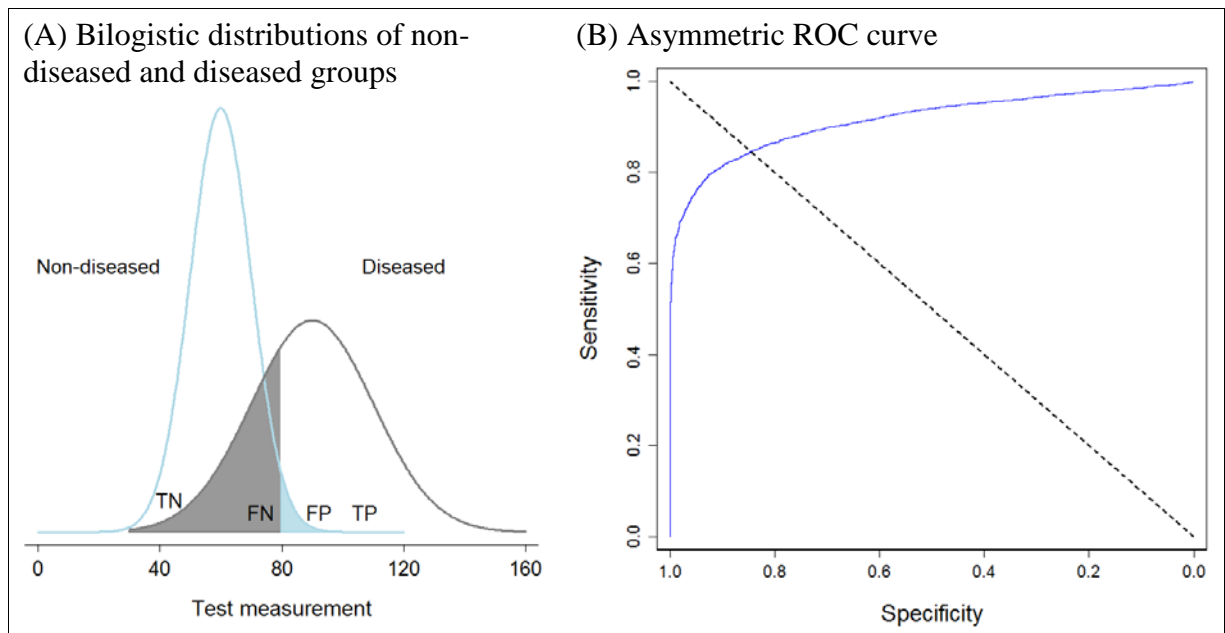


Figure 1.4| Relationship between diagnostic odds ratios and asymmetric ROC curves

DOR = diagnostic odds ratio.

(Adapted from Macaskill et al 2010²³)

1.4 Research synthesis and methods for meta-analysis of a single test

Systematic reviews and meta-analyses of test accuracy studies summarise the available evidence on the accuracy of a test for a given question. Their rate of publication has risen considerably since the 1990s.²⁷ Many test accuracy studies are small,²⁸ and even when studies are large, the number of cases may be small due to low prevalence of the target condition. A meta-analysis is an efficient approach for pooling the results of these studies to obtain more precise estimates of test performance. The extent of heterogeneity can be quantified in a meta-analysis and formal investigations of potential sources of heterogeneity may also be performed in order to explain why results differ between studies.

Test accuracy reviews can be used to inform evidence based clinical practice guidelines and health care policy,^{29,30} and are also useful for identifying research gaps. It is also important for the findings of a new study to be appraised within the context of relevant existing evidence.

According to Mulrow,

*"the hundreds of hours spent conducting a scientific study ultimately contributes only a piece of an enormous puzzle. The value of any single study is derived from how it fits with and expands previous work, as well as from the study's intrinsic properties. Through systematic review the puzzle's intricacies may be disentangled".*³¹

The methodology underpinning reviews has an impact on their validity, and several stages of the process for DTA reviews present greater challenges than reviews of interventions. These include study identification, methodological quality appraisal of studies and meta-analysis.³²

Recognising the complexity of DTA reviews, the Cochrane Collaboration—the world's largest producer of systematic reviews—delayed introducing this review type into the Cochrane Library until there was sufficient development and understanding of methodology to support their implementation and production. The first Cochrane DTA review was

published in 2008, 12 years after the formal registration of the Cochrane Screening and Diagnostic Test Methods Group.³²

Methods for conducting test accuracy meta-analyses have evolved over the last two decades.³² A number of approaches have been used ranging from simple univariate pooling methods to highly sophisticated hierarchical methods that include information from multiple thresholds used in each study. These hierarchical methods are more complex than those routinely used for synthesising the effects of interventions. The thesis focuses on methods applicable in the common situation where a single 2x2 table of the results of an index test cross classified with a reference standard is available, or can be derived for each study included in the analysis. The issue of heterogeneity is briefly introduced below before describing the commonly used meta-analytic methods.

1.4.1 Heterogeneity

It is the norm to observe heterogeneity in test accuracy between studies that is much more than would be expected from within-study sampling error alone. Measures of test accuracy are not fixed properties of a test and are not usually transferable across different populations and settings.³³ Other factors such as threshold, characteristics of the test and its conduct (including skill and experience of assessors or practitioners), and definition of the target condition may also affect test performance. The spectrum of disease in a population is dependent on prevalence, disease severity, clinical setting and prior testing. While all measures are susceptible to disease spectrum, measures such as predictive values are directly affected by prevalence.³⁴

1.4.2 Univariate pooling methods

Univariate fixed effect or random effects meta-analytic methods pool sensitivity and specificity separately, ignoring any correlation that may exist between the two measures. Fixed effect models assume homogeneity while random effects models enable the possibility of variability in test accuracy beyond sampling error alone by allowing each study to have its own test accuracy, i.e., the model includes a between-study variance component τ^2 or σ^2 . Although τ^2 is a generally used notation in traditional random effects models, σ^2 is used throughout the thesis for consistency with the notation of the bivariate model that will be described in section 1.4.4.1. Heterogeneity in test accuracy is expected and so fixed effect models that estimate the underlying common test accuracy are inappropriate and random effects models are recommended.^{23,35}

If the observed logit sensitivity and logit specificity from the i th study are $\hat{\mu}_{Ai}$ and $\hat{\mu}_{Bi}$ respectively, then the general random effects model for sensitivity and specificity can be specified as

$$\begin{aligned} \hat{\mu}_{Ai} &\sim N(\mu_{Ai}, s_{Ai}^2), \hat{\mu}_{Bi} \sim N(\mu_{Bi}, s_{Bi}^2); \\ \mu_{Ai} &\sim N(\mu_A, \sigma_A^2), \mu_{Bi} \sim N(\mu_B, \sigma_B^2), \end{aligned} \quad (1.1)$$

where μ_{Ai} and μ_{Bi} are the true underlying estimates of logit sensitivity and logit specificity, and s_{Ai}^2 and s_{Bi}^2 their variances for the i th study ($i = 1, 2, \dots, n$). μ_A and μ_B represent the average logit sensitivity and logit specificity, and σ_A^2 and σ_B^2 their variances which quantify the degree of heterogeneity between studies. Ideally, a binomial likelihood should be used to model within-study variability but the normal approximation shown in equation (1.1) is commonly used in practice.

The simplest and most commonly used random effects method is the DerSimonian and Laird approach which uses a normal distribution to model within-study variability and a moments estimator for the between-study variance. Logit transformed sensitivity or specificity and the within-study variance is undefined when there are zero cells in a study's two by two table (Table 1.2). A continuity correction (typically 0.5) is applied, leading to a downward bias in test accuracy.³⁶

1.4.3 Summary receiver operating characteristic regression

The summary receiver operating characteristic (SROC) curve approach, developed by Moses et al,³⁷ accounts for potential heterogeneity in threshold. It uses a logistic transformation of the true positive and false positive rates and linear regression to model the relationship between test accuracy, D , and the proportion test positive (related to threshold), S .

$$D_i = a + bS_i + e_i \quad (1.2)$$

with

$$e_i \sim N(0, \sigma^2)$$

where a is the intercept, b is the regression coefficient for S , and e_i is the random error. From each study i , D_i and S_i are computed as

$$D_i = \ln\left(\frac{TPR_i}{1-TPR_i}\right) - \ln\left(\frac{FPR_i}{1-FPR_i}\right) \quad (1.3)$$

and

$$S_i = \ln\left(\frac{TPR_i}{1-TPR_i}\right) + \ln\left(\frac{FPR_i}{1-FPR_i}\right). \quad (1.4)$$

The difference in the logits, D , is the log of the diagnostic odds ratio. S increases as the proportion of true and false positives increase, and is therefore considered a proxy for threshold. If accuracy does not depend on threshold, i.e. $b = 0$, the SROC curve is symmetric and can be described by a constant DOR given by the exponent of the intercept a .

A single summary estimate of sensitivity and specificity cannot be obtained from the Moses model. In computing the model variables D and S , both of which are functions of sensitivity and specificity, information is lost on the observed estimates of sensitivity and specificity from each study. However, the expected sensitivity for a given value of specificity can be estimated by

$$E(\text{sensitivity}) = \left(1 + \frac{1}{\exp(a/(1-b))} \times \left(\frac{1-\text{specificity}}{\text{specificity}} \right)^{\frac{1+b}{1-b}} \right)^{-1}, \quad (1.5)$$

where a is the intercept and b is the slope from the regression equation in (1.2).

The AUC and Q^* are sometimes used to quantify the SROC curve. The AUC is a global measure of test accuracy with values ranging between 0 and 1. The AUC can be calculated as

$$\text{AUC} = \int_0^1 \frac{\exp\left(\frac{a}{1-b}\right)\left(\frac{x}{1-x}\right)^{(1+b)/(1-b)}}{1 + \exp\left(\frac{a}{1-b}\right)\left(\frac{x}{1-x}\right)^{(1+b)/(1-b)}} dx, \quad (1.6)$$

where x is FPR. There is no closed form for the integral in (1.6) and so the AUC is obtained numerically.³⁸ For symmetric SROC curves, the AUC can be expressed in terms of the DOR as

$$\text{AUC} = \frac{\text{DOR}}{(\text{DOR}-1)^2} [(\text{DOR}-1) - \ln\text{DOR}]. \quad (1.7)$$

The usefulness of the AUC is limited because certain regions of ROC space may lack data and so the SROC curve extends beyond the range of the data. In addition, SROC curves with the same AUC may have different symmetry properties. As the AUC does not provide information on sensitivity and specificity, it is not possible to assess test accuracy at different thresholds, and the consequences of test errors cannot be inferred at any threshold.

Moses et al proposed the Q^* as an alternative to the AUC.³⁷ The Q^* statistic is the point on the SROC curve where sensitivity is equal to specificity, i.e. the intersection of the curve and the line of symmetry. The line of symmetry is the negative diagonal and sensitivity = specificity at every point on the line. Q^* can be computed as

$$Q^* = \frac{1}{1+e^{-a/2}} . \quad (1.8)$$

Based on equation 1.4, it can be deduced that $S = 0$ at Q^* . The use of Q^* is not recommended as it can be misleading if SROC curves are asymmetric, or study points lie away from the line of symmetry.

The SROC model is usually fitted using unweighted least squares linear regression, or weighted by study size or the inverse of the variance of D . The SROC approach is a fixed effect method in which variation is attributed solely to threshold effect (and sampling error if study estimates are weighted). The approach has methodological limitations which lead to inaccurate standard errors, thus rendering formal statistical inference invalid.^{39,40} Similar to the DerSimonian and Laird approach, zero cell corrections may be required.

1.4.4 Hierarchical models

Hierarchical models (also known as mixed or multilevel models) take into account correlation between sensitivity and specificity across studies while also allowing for variation in test performance between studies through the inclusion of random effects. The bivariate model⁴¹ and the hierarchical summary receiver operating characteristic (HSROC) model⁴² are the two main approaches recommended for meta-analysis when a sensitivity and specificity pair is available for each study.^{23,35,43} The two approaches differ in parameterizations, but the models

are mathematically equivalent when no covariates are included.⁴⁴ The choice of approach is often determined by variation in the thresholds reported in the included studies and the focus of inference—a summary point (summary sensitivity and specificity) or a SROC curve. Due to their shared statistical properties, SROC curves can be computed from bivariate models and average operating points from HSROC models, so the choice of model when there is only a single test is academic.^{23,44,45} However, when there are comparisons between tests or subgroups, the choice of model is important as will be explained in section 1.5.4 and illustrated in section 2.3.2 using three case studies.

For the summary sensitivity and specificity of a test to have a clinically meaningful interpretation, the analysis should be based on data at a given threshold. For the estimation of a SROC curve, data from all studies, regardless of threshold, can be included but only one threshold per study is selected for inclusion in the analysis. Methods have been proposed which allow inclusion of data from multiple thresholds per study,⁴⁶⁻⁴⁸ but these are rarely used in practice due to their complexity and limitations.⁴⁹ A SROC plot of sensitivity against specificity is usually used to display the results of the included studies as points in ROC space. The plot can also show meta-analytic summaries such as SROC curves or summary points with corresponding confidence and/or prediction regions to illustrate uncertainty and heterogeneity, respectively (see Figure 1.5 and Figure 1.6).

Both classical and Bayesian hierarchical methods are available but the latter is rarely used by meta-analysts.²⁷ Furthermore, many test accuracy meta-analyses include small numbers of studies and so prior distributions required by Bayesian methods are likely to be influential. Consequently, empirically based prior distributions for variances and correlations are needed

to inform the use of Bayesian models. Such empirical work has been done for binary outcomes for treatment effects⁵⁰ but identifying suitable priors for test accuracy requires investigation.⁵¹ Therefore to be of practical relevance to most meta-analysts, the thesis focuses on methods within the classical framework.

1.4.4.1 Bivariate random effects model

van Houwelingen et al⁵² proposed a bivariate approach to meta-analysis that was adapted by Reitsma et al⁴¹ for test accuracy meta-analysis. This bivariate model is a linear mixed model with two levels corresponding to within (level one) and between (level two) study variability. Level one is based on an approximate normal distribution for the observed logit sensitivity ($\hat{\mu}_{Ai}$) and logit specificity ($\hat{\mu}_{Bi}$) of study i , given as

$$\begin{pmatrix} \hat{\mu}_{Ai} \\ \hat{\mu}_{Bi} \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_{Ai} \\ \mu_{Bi} \end{pmatrix}, C_i \right) \text{ with } C_i = \begin{pmatrix} s_{Ai}^2 & 0 \\ 0 & s_{Bi}^2 \end{pmatrix}, \quad (1.9)$$

where s_{Ai}^2 and s_{Bi}^2 are the observed variances of the estimated logit transformed sensitivity and specificity. At level two, the model enables joint analysis of the logit sensitivities and logit specificities by combining two correlated normal distributions and takes the form,

$$\begin{pmatrix} \mu_{Ai} \\ \mu_{Bi} \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}, \Sigma \right) \text{ with } \Sigma = \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix}. \quad (1.10)$$

The model assumes a bivariate normal distribution with mean μ_A and variance σ_A^2 for the logit sensitivities, mean μ_B and variance σ_B^2 for the logit specificities, and σ_{AB} the covariance between μ_{Ai} and μ_{Bi} across studies. Instead of the covariance, the model can be parameterized using the between-study correlation, ρ_{AB} . Therefore, the basic bivariate model has the following five parameters: μ_A , μ_B , σ_A^2 , σ_B^2 , and σ_{AB} (or ρ_{AB}).

Chu and colleagues^{53,54} have shown that a binomial likelihood should be used for modelling within-study variability (especially when data are sparse) as follows:

$$y_{Ai} \sim \text{Binomial}(n_{Ai}, g^{-1}(\mu_{Ai})), y_{Bi} \sim \text{Binomial}(n_{Bi}, g^{-1}(\mu_{Bi})), \quad (1.11)$$

where y_{Ai} and y_{Bi} represent the number of true positives and true negatives, n_{Ai} and n_{Bi} the number of diseased and non-diseased subjects, and $g^{-1}(\mu_{Ai})$ and $g^{-1}(\mu_{Bi})$ the sensitivity and specificity in the i th study. The logit link $g(\cdot)$ is commonly used but other link functions can be applied.^{40,54} The random effects in this generalized linear mixed model also follow a bivariate normal distribution as in (1.10).

If this bivariate model is simplified by assuming the covariance or correlation is zero (i.e. an independent variance-covariance structure), the model reduces to two univariate random effects logistic regression models for sensitivity and specificity as follows:

$$\begin{pmatrix} \mu_{Ai} \\ \mu_{Bi} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}, \Sigma\right) \text{ with } \Sigma = \begin{pmatrix} \sigma_A^2 & 0 \\ 0 & \sigma_B^2 \end{pmatrix}. \quad (1.12)$$

Bivariate meta-analysis of likelihood ratios and predictive values are alternatives to bivariate meta-analysis of sensitivity and specificity. Zwinderman and Bossuyt⁵⁵ have highlighted the difficulty in bivariate meta-analysis of likelihood ratios. Leeftang et al³⁴ stated the main disadvantage of their proposed meta-analysis of predictive values as the interpretation of the results and translation into practice. The effect of prevalence is also likely to be greater on predictive values than on sensitivity and specificity. For these reasons, and because sensitivity and specificity are the test accuracy measures most commonly used in meta-analyses,²⁷ only methods for synthesis of sensitivities and specificities are considered in this thesis. Other

measures such as likelihood ratios can be derived from functions of the parameters of bivariate or HSROC models.

As an illustration of application of the bivariate model, Figure 1.5 shows an average operating point (summary sensitivity and specificity) for the mood disorder questionnaire (MDQ) for detection of bipolar disorder in mental health centre settings.⁵⁶ The summary point is surrounded by a 95% confidence region and 95% prediction region. The meta-analysis included only studies that used a common threshold of 7. Due to this restriction, only 19 of the 30 eligible studies were included in the meta-analysis.

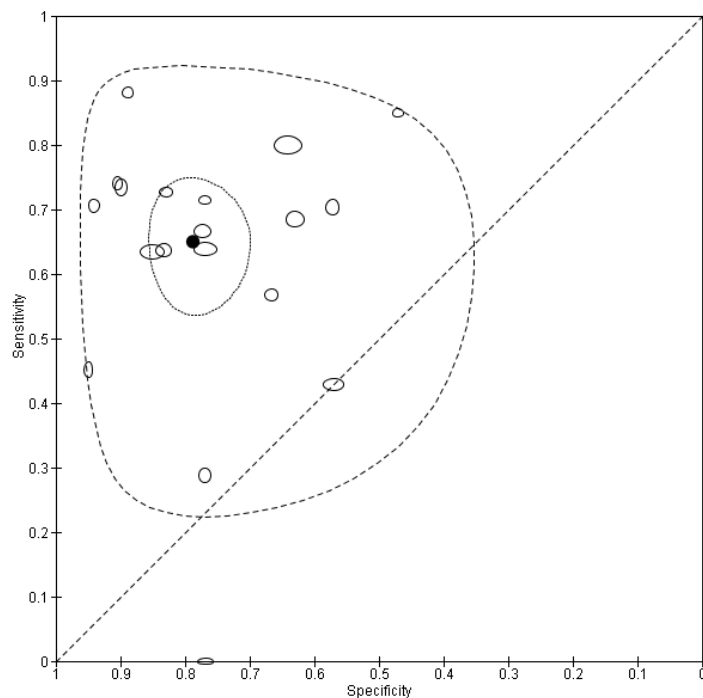


Figure 1.5| SROC plot of the MDQ at a common threshold of 7 for detection of bipolar disorder in mental health centre settings

Each study point was scaled according to the precision of sensitivity and specificity in the study; the greater the height or width of a study point relative to other study points, the greater the precision of the study's estimated sensitivity or specificity respectively. The solid circle (summary point) represents the summary estimate of sensitivity and specificity for the mood disorder questionnaire (MDQ). The summary point is surrounded by a dotted line representing the 95% confidence region and a dashed line representing the 95% prediction region (the region within which one is 95% certain the results of a new study will lie).

1.4.4.2 Hierarchical summary receiver operating characteristic model

The Rutter and Gatsonis HSROC model⁴² represents a general framework for meta-analysis of test accuracy studies and can be viewed as an extension of the Moses SROC approach in which the TPR and FPR for each study is modelled directly.⁵⁷ The HSROC model is a nonlinear generalized mixed model in which the number of test positives from the i th study, y_{ij} , is assumed to follow a binomial distribution

$$y_{ij} \sim \text{Binomial}(n_{ij}, \pi_{ij}).$$

For the non-diseased group $j = 0$ and for the diseased group $j = 1$, and n_{ij} is the number in group j . The probability of a positive test result in group j is modelled as

$$\text{logit}(\pi_{ij}) = (\theta_i + \alpha_i \text{dis}_{ij}) \exp(-\beta \text{dis}_{ij}), \quad (1.13)$$

where π_{ij} is the proportion of test positives, true or false positives depending on disease status. Disease status is represented by dis_{ij} which is coded -0.5 for the non-diseased group and 0.5 for the diseased group in the i th study. The implicit threshold θ_i (threshold parameter or positivity criteria) and diagnostic accuracy α_i (accuracy parameter which is the log of DOR) for each study are modelled as random effects with independent normal distributions $\theta_i \sim N(\Theta, \sigma_\theta^2)$ and $\alpha_i \sim N(\Lambda, \sigma_\alpha^2)$ respectively.

The parameter θ_i (equivalent to $S_i/2$ in the Moses SROC model) estimates the average log odds of a positive test result for the diseased and non-diseased groups at the threshold for a positive test result in the study.⁵⁷ The model also includes a shape or scale parameter β which enables asymmetry in the SROC curve by allowing accuracy to vary with implicit threshold (i.e. allowing true positive and false positive fractions to increase at different rates as θ_i increases). Therefore, the SROC curve is symmetric if $\beta = 0$ or asymmetric if $\beta \neq 0$. Each study contributes a single point in ROC space and so the estimation of β requires information

from all studies included in the meta-analysis. Thus β is modelled as a fixed effect. The HSROC model has the following five parameters: Λ , Θ , β , σ_{α}^2 and σ_{θ}^2 . The model reduces to a fixed effect model if $\sigma_{\alpha}^2 = 0$ and $\sigma_{\theta}^2 = 0$.

Other specifications for SROC curves based on functions of the bivariate model have been proposed^{39,58} but the focus in this thesis is only on the more established and commonly used Rutter and Gatsonis HSROC model. Figure 1.6 shows the SROC curve for the MDQ estimated from an HSROC model. The curve was drawn within the range of specificities (0.47 to 1.00) from the 30 included studies to avoid extrapolating beyond the data. Given the relationship between the bivariate and HSROC model mentioned earlier,⁴⁴ it is possible to estimate an average operating point on the SROC curve by performing another HSROC analysis for studies that report data at a common threshold, similar to that shown in Figure 1.5.

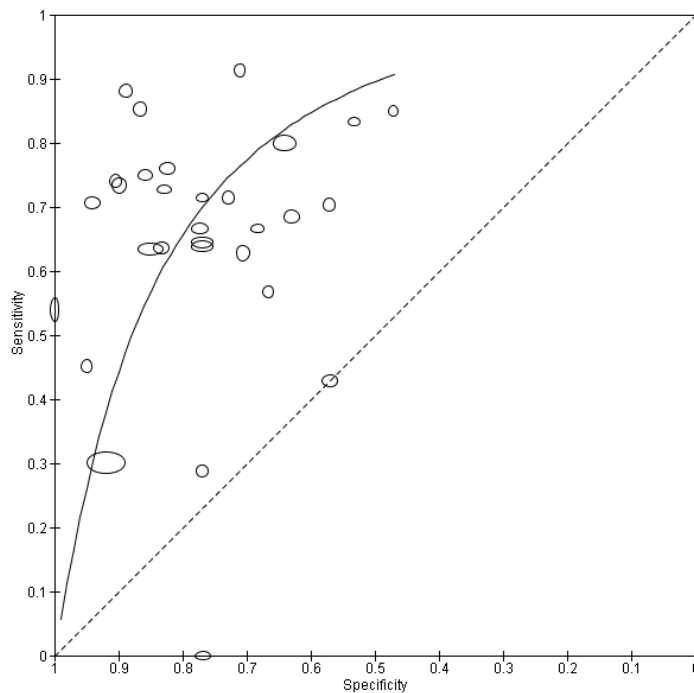


Figure 1.6| SROC plot of the MDQ at different thresholds for detection of bipolar disorder in mental health centre settings

Each study point was scaled according to the precision of sensitivity and specificity in the study. The summary curve was drawn restricted to the range of specificities (0.47 to 1.00) from the 30 studies included in the evaluation of the mood disorder questionnaire (MDQ).

1.5 Comparisons of two or more tests

This thesis focused on the comparison of test accuracy for two or more tests when evidence from multiple primary studies is available. Ideally in primary studies, the diagnostic accuracy of competing alternative tests should be compared in the same study population. Such head-to-head evaluations may compare tests to identify the best performing test(s) or assess the incremental gain in accuracy of a combination of tests relative to the performance of one of the component tests.⁵⁹

Historically, test evaluations have focused on the accuracy of a single test without making comparisons with alternative tests that can be used at the same point in the diagnostic pathway.⁶⁰ The lack of regard for comparative evidence is exemplified by separate

publications of tests evaluated in the same study. For example, Grandjean et al conducted a large multicentre study to assess the accuracy of short femur and nuchal fold measurement in the second trimester of pregnancy for detection of trisomy 21 (Down's syndrome).^{61,62} However, results were published separately for the two ultrasonographic markers. The prevalence of such publications in the literature is unknown. Single test evaluations clearly have a place in the early evaluation of the diagnostic accuracy of a test (as indicated in Table 1.1), and can be useful in identifying its potential role and usefulness. However, the clinical relevance of such evaluations is questionable when alternative tests are available, and the position of a test in the clinical pathway and its superiority in terms of clinical performance is uncertain.

Well-designed comparative studies are invaluable for clinical decision making because they can facilitate evaluation of new tests against existing testing pathways and guide test selection. These studies enable unbiased comparisons and increase confidence in the validity of the evidence. Evidence from studies of comparative accuracy can also be used in decision modelling to infer the relative effectiveness of tests when direct evidence from RCTs of test effectiveness is unavailable.⁶³ Modelling can help in understanding the expected benefits, risks, and costs of implementing newer tests, by considering improvements in accuracy as well as potential shifts in the disease spectrum for positive diagnoses.⁶⁴ While modelling is promising, it is subject to limitations such as poor data and non-transferability of test performance and/or therapeutic efficacy, and validity of modelling assumptions. Modelling is beyond the scope of this thesis.

In the following subsections, roles of tests within diagnostic pathways and their importance in designing comparative accuracy studies are illustrated. Study designs are discussed in detail, including the merits and limitations of each type of design. The analysis of comparative accuracy studies is also considered in terms of measures used to express relative accuracy.

1.5.1 Diagnostic pathways

Diagnostic or clinical pathways are used to guide evidence-based healthcare and should be at the core of test evaluation as illustrated in Figure 1.1. A new test may replace an existing one (replacement), be used before the existing test (triage) or after the existing test (add-on) as shown in Figure 1.7.⁶⁰ The test may be simultaneously assessed for more than one of these roles.

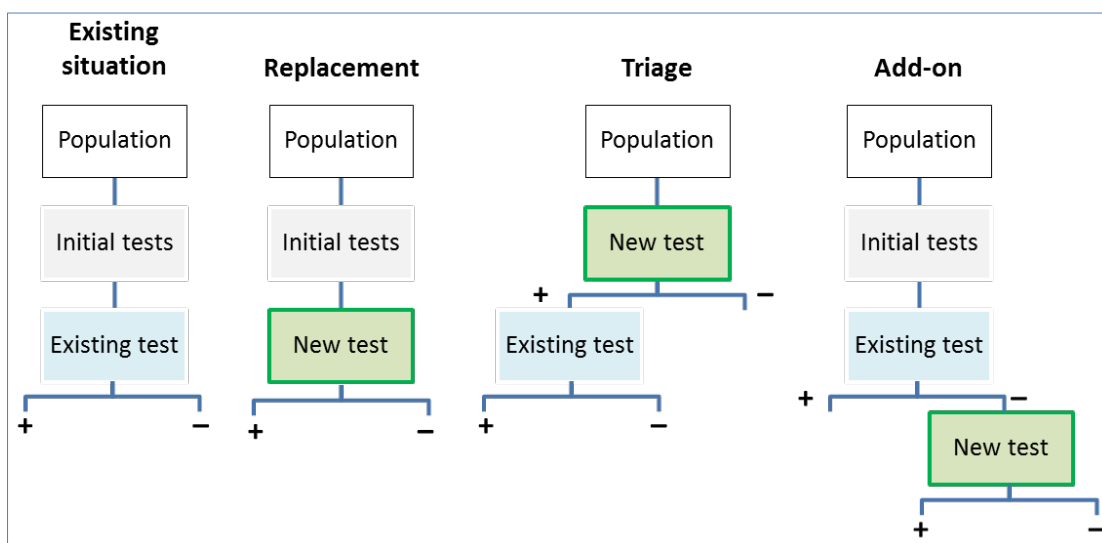


Figure 1.7| Roles of tests in diagnostic pathways

(Adapted from Bossuyt et al 2006⁶⁰)

The intended role of a test within a diagnostic pathway is important because it provides a framework for assessing test accuracy and guides study design, including the choice of a comparator. For example, consider CT and ultrasound for diagnosis of acute appendicitis in

children without classic signs and symptoms of appendicitis. If ultrasound is the initial imaging modality and standard of care (comparator), the intended role of CT should inform study design—CT and ultrasound applied to all children in the study or CT applied after a negative or inconclusive ultrasound such as in the study by Garcia Peña et al.⁶⁵ The former design can be used to address a replacement or add-on question but the latter can only be used to address an add-on question. Alternatively, children may be randomized to receive ultrasound or CT to address a replacement question, or randomized to receive ultrasound or ultrasound followed by CT to address an add-on question (as well as a replacement question if randomization is ignored) such as in the study by Kaiser et al.⁶⁶ Nevertheless, it is unlikely that CT can replace ultrasound because of the potential risks associated with use of contrast media, ionizing radiation exposure and cost. According to the American College of Radiology, US is the preferred initial examination in children because it is nearly as accurate as CT in this population and avoids use of ionizing radiation.⁶⁷ Since CT is potentially more accurate and harmful, there is no role for CT as a triage test in this diagnostic pathway.

The role of several existing tests at different points in the diagnostic pathway may also be evaluated. For example, recommended tests for diagnosis of common bile duct stones are liver function tests (e.g. alkaline phosphatase, bilirubin), abdominal ultrasound, endoscopic ultrasound (EUS), magnetic resonance cholangiopancreatography (MRCP), endoscopic retrograde cholangiopancreatography (ERCP), and intraoperative cholangiography (IOC).⁶⁸⁻⁷⁰ Figure 1.8 shows a potential diagnostic pathway for diagnosis of common bile duct stones. People at risk or suspected of having common bile duct stones initially undergo liver function tests and abdominal ultrasound. Usually both tests are used as triage tests before patients are subjected to further tests such as MRCP or EUS in the next step of the pathway. Therefore,

MRCP and EUS may be regarded as add-on tests in patients with positive abdominal ultrasound or liver function tests. If positive, the tests are followed by ERCP or IOC. In this scenario, MRCP and EUS may be regarded as triage tests. ERCP and IOC are the final diagnostic tests prior to therapeutic intervention. In general, patients with negative test results at a point in the pathway do not undergo further testing. The sequence of tests is a logical order that reflects increasing access, cost and invasiveness of the tests.

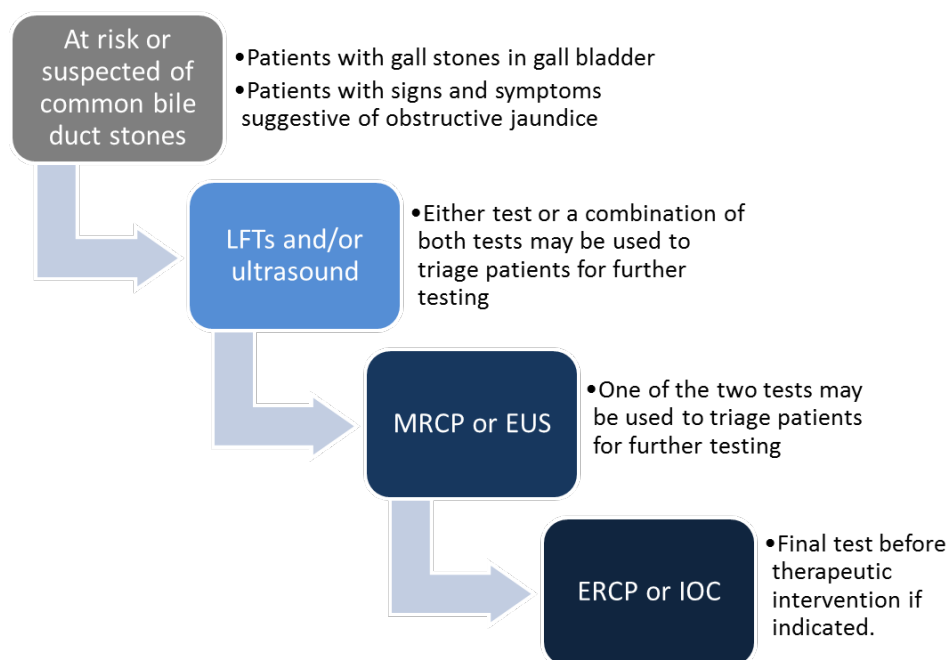


Figure 1.8| Simplified illustration of the role of existing tests in the diagnostic pathway for common bile duct stones

ERCP = endoscopic retrograde cholangiopancreatography; EUS = endoscopic ultrasound; IOC = intraoperative cholangiography; LFTs = liver function tests; MRCP = magnetic resonance cholangiopancreatography.

(Adapted from Gurusamy et al 2015⁶⁸)

1.5.2 Study designs for test comparisons

There is no standard terminology for different types of test accuracy studies. In this thesis, the term ‘non-comparative’ is used to describe a primary study that evaluated a single test (uncontrolled study) or only one of the tests being evaluated in a systematic review, and

‘comparative’ to describe a study that evaluated and compared at least two of the tests in the same population. A number of study designs, varying in methodological rigour, are used for comparison of test accuracy as illustrated in Figure 1.9. Generally, there are two comparative study designs—within-subject and between-subject designs.

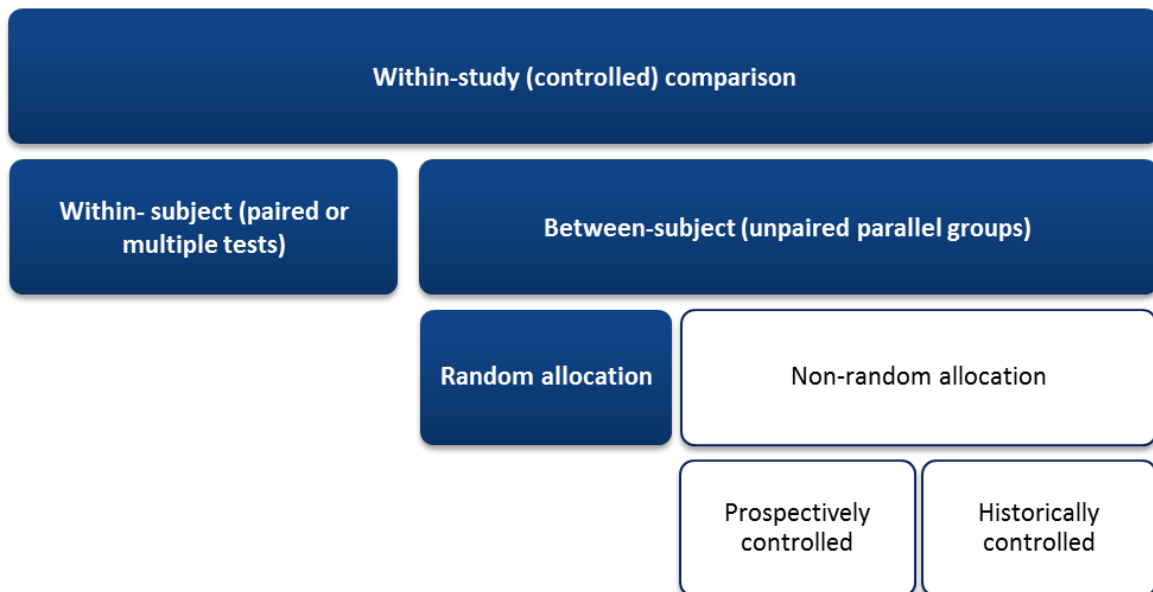


Figure 1.9| Study designs for comparing test accuracy
White boxes represent less methodologically robust designs.

Robust comparative studies of diagnostic test accuracy use either a within-subject multiple test (sometimes called "paired" or "crossover") design, in which all patients undergo all tests, together with a reference standard, or, more rarely, a between-subject randomized (unpaired or parallel group) design in which all patients undergo the reference standard test but are randomly assigned to have only one of the other tests (Figure 1.10).^{43,71} Such designs ensure validity by comparing like-with-like (either within patients or between randomized groups), thus avoiding confounding by factors such as population characteristics and study methods.

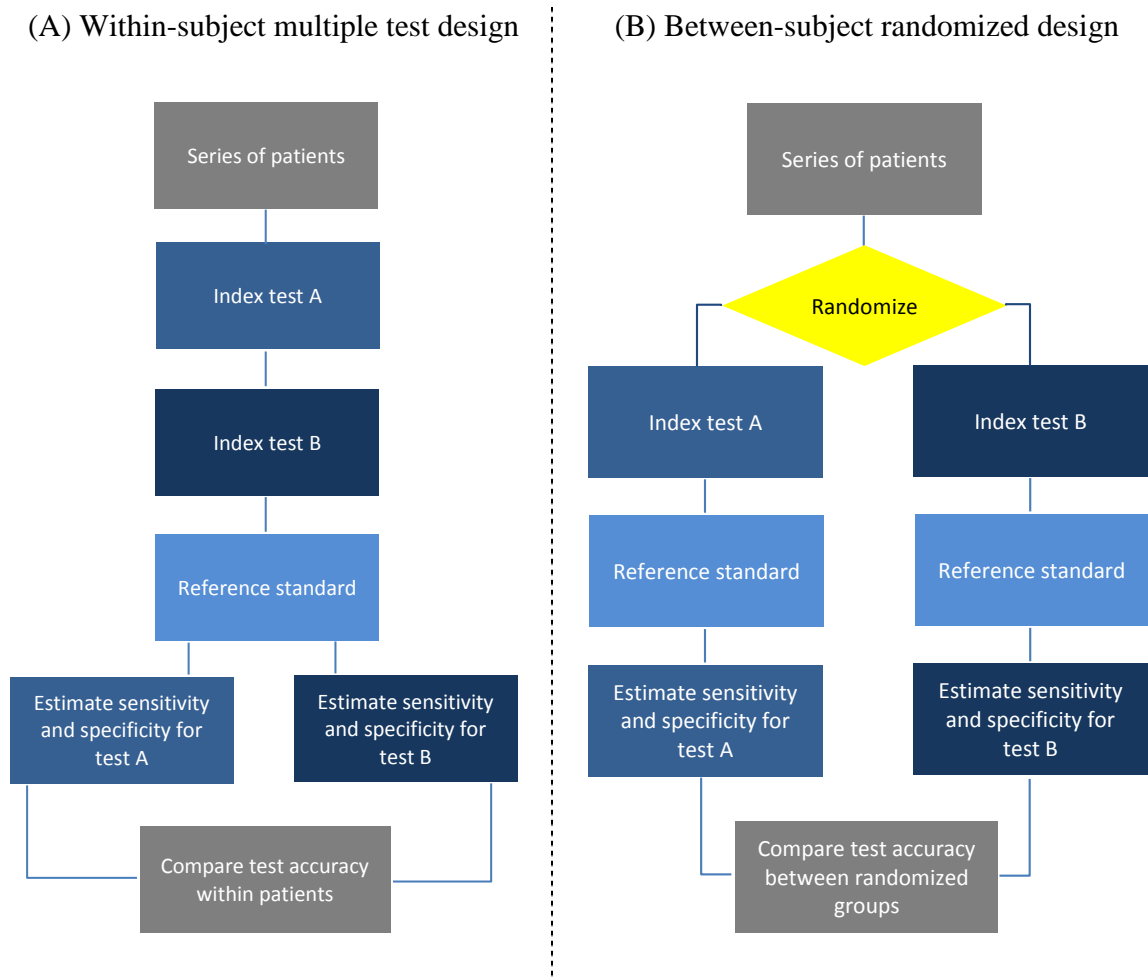


Figure 1.10| Robust study designs for comparing test accuracy

In (A) all patients undergo all index tests while in (B) patients are randomly assigned to only one of the index tests. In both (A) and (B), all patients receive the reference standard. Both designs are valid. (Adapted from Takwoingi et al 2013⁷¹)

In the following subsections, examples from the literature on imaging for diagnosis of appendicitis are used to illustrate the designs shown in Figure 1.9 and Figure 1.10.

1.5.2.1 Within-subject design

In a within-subject design, all patients undergo all tests and so each patient is their own control. Such designs are potentially resource efficient depending on the extent to which the tests are conditionally dependent. The design minimises between-subject variability and also allows estimation of the accuracy of combinations of tests. The sequence of testing may be

randomized to avoid bias.⁷² Table 1.3 shows the joint classification of the results of a paired comparison.

Table 1.3| Joint classification of paired index tests and reference standard results

	Reference standard positive			Reference standard negative		
	Test A positive	Test A negative	Total	Test A positive	Test A negative	Total
Test B positive	y_{11}^D	y_{01}^D	n_{B1}^D	$y_{11}^{\bar{D}}$	$y_{01}^{\bar{D}}$	$n_{B1}^{\bar{D}}$
Test B negative	y_{10}^D	y_{00}^D	n_{B0}^D	$y_{10}^{\bar{D}}$	$y_{00}^{\bar{D}}$	$n_{B0}^{\bar{D}}$
Total	n_{A1}^D	n_{A0}^D	n_1	$n_{A1}^{\bar{D}}$	$n_{A0}^{\bar{D}}$	n_0

For two tests labelled *A* and *B* in Table 1.3, y_{11}^D is the number of patients with the target condition for whom both tests are positive; y_{00}^D is the number of patients with the target condition for whom both tests are negative; y_{10}^D is the number of patients with the target condition for whom test *A* is positive but test *B* is negative; and y_{01}^D is the number of patients with the target condition for whom test *B* is positive but test *A* is negative. For those without the target condition, the counts $y_{11}^{\bar{D}}$, $y_{00}^{\bar{D}}$, $y_{10}^{\bar{D}}$ and $y_{01}^{\bar{D}}$ can be interpreted in a similar manner.

For tests *A* and *B*, n_{A1}^D and n_{B1}^D are the numbers of true positives; $n_{A0}^{\bar{D}}$ and $n_{B0}^{\bar{D}}$ are the numbers of true negatives; n_{A0}^D and n_{B0}^D are the numbers of false negatives; and $n_{A1}^{\bar{D}}$ and $n_{B1}^{\bar{D}}$ are the numbers of false positives. If n_1 and n_0 represent the number of individuals with and without disease respectively, then the sensitivity and specificity of the two tests can be estimated from the marginal frequencies as follows:

$$\text{Sensitivity}(A) = n_{A1}^D/n_1; \text{Specificity}(A) = n_{A0}^{\bar{D}}/n_0$$

$$\text{Sensitivity}(B) = n_{B1}^D/n_1; \text{Specificity}(B) = n_{B0}^{\bar{D}}/n_0 .$$

Although a table of the joint classification of the results of two tests against those of the reference standard is potentially useful, studies do not often present results in this format but rather give a separate 2x2 table comparing each test with the reference standard. For example, although Poortman et al⁷³ and Pickuth et al⁷⁴ both compared ultrasound and CT for diagnosis of acute appendicitis by applying both tests to all patients, it was only possible to obtain a separate 2x2 table for each test compared with the reference standard from the results reported.

Notwithstanding the advantages of a within-subject design, such a comparative study may be unethical if the burden of testing is unacceptable due to the invasive nature or risk of complications associated with the tests.^{72,75} Also, if one test adversely affects the performance or application of another test—akin to carryover effects in crossover trials of therapeutic effectiveness—a within-subject design will be inappropriate. The design is also inappropriate if there is potential for disease progression during the interval between administering the tests, or if multiple testing is likely to delay initiation of therapy. Since within-subject test comparison designs can be likened to crossover therapeutic trials, the same principles of good study design for trials apply.⁷⁵

1.5.2.2 Between-subject design

In a between-subject design, the allocation of tests to patients should ideally be randomized. Principles of good design for RCTs of therapeutics, also suggested in the preceding section, can be applied.⁷⁵ The randomized design is a valid alternative in situations where a paired design is inappropriate for reasons such as those stated above. A disadvantage of this design is that a larger sample is typically required and test combinations cannot be explored unless

patients receive all tests in one of the arms of the study. For example, Kaiser et al⁶⁶ compared the accuracy of ultrasound alone and abdominal CT performed in addition to ultrasound for diagnosis of childhood appendicitis by randomly allocating children to one of the two arms. The authors presented the results of this comparison as well as the joint classification of results from the arm that received both tests.

Patients may also be allocated to groups by using stratified randomization. Stratified randomization is a two-stage procedure in which patients are first grouped into strata, according to clinical features that may influence the outcome, and then followed by randomization within each stratum.⁷⁶ Tsai et al⁷⁷ evaluated ultrasound and CT by prospectively stratifying patients to CT or ultrasound based on body mass index (BMI ≥ 30 had CT and BMI < 30 had ultrasound). The objective was to determine if stratifying patients improved the accuracy of each modality because choice of ultrasound or CT may depend on body habitus. Clearly, the design of this study cannot answer the research question.

Although non-randomized studies are prone to bias, the randomized design is seldom used and non-randomized studies, prospectively or historically controlled, are more common. For example, in a study by Sivit et al⁷⁸, the decision to give a patient ultrasound, CT, or both was based on the clinical judgment of the referring surgeon or emergency department physician. In contrast, Lowe et al⁷⁹ compared both tests by evaluating CT in a prospective cohort but ultrasound was assessed using a historical cohort that was retrospectively identified from computerized hospital records.

1.5.3 Analysis of test comparisons in primary studies

In comparative test accuracy studies, absolute differences can be computed for sensitivity and specificity whilst relative differences can be computed for most of the measures discussed in section 1.3.2.⁸⁰ Assuming the newer or experimental test is test *A* and the older test or standard practice is test *B*, then the absolute difference in sensitivity and specificity can be written as:

$$\text{Absolute difference in sensitivity} = \text{sensitivity}(A) - \text{sensitivity}(B)$$

$$\text{Absolute difference in specificity} = \text{specificity}(A) - \text{specificity}(B).$$

The relative probabilities can be written as:

$$\text{Relative sensitivity} = \text{sensitivity}(A)/\text{sensitivity}(B)$$

$$\text{Relative specificity} = \text{specificity}(A)/\text{specificity}(B),$$

and the odds ratios as

$$\text{Odds ratio (true positive)} = \frac{\text{sensitivity}(A)}{1-\text{sensitivity}(A)} / \frac{\text{sensitivity}(B)}{1-\text{sensitivity}(B)}$$

$$\text{Odds ratio (true negative)} = \frac{\text{specificity}(A)}{1-\text{specificity}(A)} / \frac{\text{specificity}(B)}{1-\text{specificity}(B)}.$$

It should be noted that these odds ratios are not equivalent to the DOR. Here the ratios compare the same measure (sensitivity or specificity) for one test to that of another test while the DOR compares two groups (diseased and non-diseased) for one test. The DORs of two tests can be compared and expressed as the relative DOR (rDOR). If logistic regression models are used to compare the sensitivity and specificity of multiple tests, odds ratios are the natural output. Nevertheless, absolute differences and relative probabilities are more familiar quantities and are straightforward to interpret. Odds ratios do not have an intuitive interpretation. The approximate relationship between odds ratios and relative risks that is

exploited in epidemiology when events are rare⁸¹ is invalid in test research because events (i.e. true positives or true negatives) are common.

The remainder of this chapter is devoted to comparative systematic reviews and meta-analyses, and the rationale for the research undertaken in this thesis.

1.5.4 Meta-analysis of test comparisons

Regression-based meta-analytic methods such as those described in section 1.4.3 (Moses SROC model) and section 1.4.4 (hierarchical models) are straightforward to extend to investigate association between test performance and potential sources of heterogeneity; study level explanatory variables or covariates are added as indicator (dummy) variables to a regression equation.^{37,41,42} This meta-regression approach can also be used to compare test accuracy by using test type as the covariate. For example if there are N tests, $N-1$ indicator variables which take the value zero or one are added to the model. Thus the effect of test type on model parameters can be estimated; the regression coefficients estimate the performance of one test relative to that of the test used as the reference category (note this test is not the reference standard but another index test or comparator test) for the test type covariate.

The meta-regression approach is flexible, allowing the comparison of multiple tests; test comparisons are not limited to a pair of tests as will be illustrated using case studies in section 2.3.1. Since hierarchical models are recommended for meta-analysis of test accuracy,^{23,35,43,44} only hierarchical meta-regression methods are discussed in the next two sections. Methods based on the Moses SROC approach, and other methods identified from the searches described in Chapter 3 will be examined in Chapters 6 and 7.

Test comparisons based on hierarchical models may be a comparison of summary points and/or SROC curves as will be shown using two examples in section 2.3.1.3. The choice of summary points or curves should ideally be driven by the research question (i.e. focus of inference on points or curves) but is sometimes influenced by the nature of the available data (availability of data at a common cut-off or only across mixed cut-offs) and its effect on the interpretation of summary findings, software capability, and expertise of the team. In section 1.5.4.1, the comparison of summary points using a bivariate model is explained while section 1.5.4.2 details the comparison of summary curves using a HSROC model. Bivariate and HSROC meta-regression models do not fully account for dependence between tests in studies where two or more tests were compared within the same patients (i.e. within-subject design). This issue is addressed in section 2.3.4.2 and also in Chapter 6.

1.5.4.1 Bivariate random effects meta-regression model for comparing test accuracy

The approach outlined here follows from the bivariate model summarised in section 1.4.4.1 (see equations 1.10 and 1.11). A covariate for test type, t , can be used in a bivariate model to investigate whether the expected sensitivity and specificity differs between the tests. If t is indexed by k , the two levels of the bivariate model can be expressed as follows:

Level 1: Within-study likelihood

$$y_{Aik} \sim \text{Binomial}(n_{Aik}, g^{-1}(\mu_{Aik})), y_{Bik} \sim \text{Binomial}(n_{Bik}, g^{-1}(\mu_{Bik})), \quad (1.14)$$

where y_{Aik} and y_{Bik} represent the number of true positives and true negatives, n_{Aik} and n_{Bik} the number of diseased and non-diseased subjects, and $g^{-1}(\mu_{Aik})$ and $g^{-1}(\mu_{Bik})$ the sensitivity and specificity for the k th test within the i th study. The logit link, $g(\cdot)$, is often used.

Level 2: between-study likelihood

$$\begin{pmatrix} \mu_{Aik} \\ \mu_{Bik} \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_A + v_A t_k \\ \mu_B + v_B t_k \end{pmatrix}, \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix} \right). \quad (1.15)$$

where μ_{Ai} and μ_{Bi} are the logit sensitivity and logit specificity for each test within the i th study and t_k represents the study level covariate for test type. For simplicity, assume a binary covariate distinguishing between two index tests, i.e. $k = 1, 2$. As such t is an indicator variable for test type coded 0 for the index test used as the reference category (i.e. $t = 0$ if $k = 1$), and 1 for the second index test (i.e. $t = 1$ if $k = 2$), then μ_A estimates the expected logit sensitivity for index test 1 and $\mu_A + v_A$ estimates the expected logit sensitivity for index test 2. Thus, $\exp(v_A)$ gives the estimated odds ratio for the sensitivity of test 2 relative to that of test 1. The same applies to specificity where μ_B is the expected logit specificity for test 1 and $\mu_B + v_B$ estimates the expected logit specificity for test 2. The variances are σ_A^2 and σ_B^2 for the logit sensitivities (μ_{Ai1} and μ_{Ai2}) and logit specificities (μ_{Bi1} and μ_{Bi2}), and σ_{AB}^2 is the covariance between the logits across studies.

The model can be further extended by including additional terms to allow for unequal variances of the random effects for logit sensitivity and specificity between tests. Thus, the covariance matrix in (1.15) can also depend on test type as follows:

$$\begin{pmatrix} \mu_{Aik} \\ \mu_{Bik} \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_A + v_A t_k \\ \mu_B + v_B t_k \end{pmatrix}, \begin{pmatrix} \sigma_{Ak}^2 & \sigma_{ABk} \\ \sigma_{ABk} & \sigma_{Bk}^2 \end{pmatrix} \right). \quad (1.16)$$

Bivariate meta-regression models will be applied to case studies in Chapter 2 to illustrate comparison of summary points between tests.

1.5.4.2 HSROC meta-regression model for comparing test accuracy

HSROC meta-regression models—extensions of the HSROC model in (1.13) described in section 1.4.4.2—can be used to assess the effect of test type on the accuracy, threshold and/or shape parameters. Assuming t is an indicator variable for test type and γ , λ , and δ are estimated as a fixed effect, the model can be written as

$$\text{logit}(\pi_{ij}) = \left((\theta_i + \gamma t_i) + (\alpha_i + \xi t_i) \text{dis}_{ij} \right) \exp(-(\beta + \delta t_i) \text{dis}_{ij}), \quad (1.17)$$

and the distributions of the random effects for threshold and accuracy as

$$\theta_i \sim N(\Theta, \sigma_\theta^2) \text{ and } \alpha_i \sim N(\Lambda, \sigma_\alpha^2). \quad (1.18)$$

In this model, γ assesses whether the underlying threshold differ between tests, ξ assesses whether test accuracy differ between tests, and δ assesses whether the shape of the curves differ by test. The HSROC model defines test accuracy in terms of the DOR (see section 1.4.4.2). Therefore, if the summary curves being compared are symmetrical (i.e. β and δ are equal to zero) or the curves are assumed to have the same shape (i.e. $\delta = 0$), then the ratio of two DORs or rDOR (exponent of ξ) provides a summary estimate of the relative accuracy of two tests. For two index tests, where test 1 is the referent test, an rDOR greater than one indicates that test 2 is superior to test 1; an rDOR less than one indicate that test 1 is superior to test 2.

However, if each SROC curve is allowed to have its own shape and there is evidence of a difference in shape between tests, then the rDOR cannot be used quantify relative test accuracy because accuracy varies with threshold. To numerically express test performance based on the estimated SROC curves, the expected logit sensitivity at one or more values of specificity (e.g. median and interquartile range derived from the included studies for each test) can be estimated using the following equations for test 1 and test 2:

$$\begin{aligned} \text{logit}(\text{sensitivity } 1) &= [\Lambda \times \exp(-0.5\beta) + \text{logit}(1 - \text{specificity}) \times \exp(-\beta)] \\ \text{logit}(\text{sensitivity } 2) &= [(\Lambda + \gamma) \times \exp(-0.5(\beta + \delta)) + \text{logit}(1 - \text{specificity}) \times \\ &\exp(-(\beta + \delta))] . \end{aligned} \quad (1.19)$$

Similar to the bivariate model, the variance parameters of the HSROC model can be allowed to differ between tests. Allowing the difference in threshold and accuracy between tests to vary gives

$$\text{logit}(\pi_{ij}) = ((\theta_i + \gamma_i t_i) + (\alpha_i + \xi_i t_i) \text{dis}_{ij}) \exp(-(\beta + \delta t_i) \text{dis}_{ij}) \quad (1.20)$$

$$\begin{pmatrix} \theta_i \\ \gamma_i \end{pmatrix} \sim N \left(\begin{pmatrix} \Theta \\ \Gamma \end{pmatrix}, \begin{pmatrix} \sigma_\theta^2 & \sigma_\theta \sigma_\gamma \\ \sigma_\theta \sigma_\gamma & \sigma_\gamma^2 \end{pmatrix} \right) \text{ and } \begin{pmatrix} \alpha_i \\ \xi_i \end{pmatrix} \sim N \left(\begin{pmatrix} \Lambda \\ \Xi \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \sigma_\alpha \sigma_\xi \\ \sigma_\alpha \sigma_\xi & \sigma_\xi^2 \end{pmatrix} \right). \quad (1.21)$$

Correlation is assumed between the random effects for accuracy (α_i and ξ_i) and between the random effects for threshold (θ_i and γ_i). $\sigma_\alpha \sigma_\xi$ is the covariance between the random effects for accuracy and $\sigma_\theta \sigma_\gamma$ is the covariance between the random effects for threshold. The variances of the random effects for the referent test ($t = 0$) are σ_α^2 and σ_θ^2 for accuracy and threshold. The variances of the random effects for the index test ($t = 1$) can be obtained using the expression for the variance of a sum which is $\sigma_\alpha^2 + \sigma_\xi^2 + 2\sigma_\alpha \sigma_\xi$ for accuracy, and $\sigma_\theta^2 + \sigma_\gamma^2 + 2\sigma_\theta \sigma_\gamma$ for threshold.

HSROC meta-regression models will be applied to case studies in Chapter 2 to illustrate the comparison of summary curves between tests, including the impact of assumptions about the shape of the curves and the impact of assumptions about the variances of the random effects.

1.5.4.3 Relationship between bivariate and HSROC meta-regression models

As highlighted in section 1.4.4, the bivariate and HSROC models are closely related and are mathematically equivalent when no covariates are included.⁴⁴ This relationship is exploited when simplifying hierarchical models for meta-analysis of a single test in section 8.2. Harbord et al⁴⁴ also showed that a bivariate model in which one or more covariates affect both sensitivity and specificity is equivalent to an HSROC model in which the same covariates are allowed to affect both the accuracy and threshold parameters. Therefore, if covariate terms for test type are added to the accuracy and threshold parameters of the HSROC model but a common underlying shape is assumed, the model is akin to a bivariate model where the covariate affects both sensitivity and specificity. In both models, the covariance matrix of the random effects does not depend on test type.

1.5.4.4 Assessment of statistical significance of differences in test performance

The statistical significance of differences in test performance can be assessed using Wald tests or likelihood ratio tests.²³ Likelihood ratio tests compare models with and without covariate terms. The likelihood ratio chi-squared statistic is the difference in the -2Log likelihood when a covariate is added or removed from a model. The degrees of freedom used along with the chi-squared statistic to obtain a P value is the difference in the number of parameters fitted in the two models being compared. For example, the fit of the bivariate model with and without the additional parameters ν_A and ν_B can be used to assess whether sensitivity and/or specificity differ between tests. Likelihood ratio tests can also be used to assess the significance of additional variance terms for the random effects parameters in either the bivariate or HSROC model to determine whether assumption of common variances is justified or separate variances for each test is needed. It should be noted that as the models become

increasingly complex, the number of additional parameters to estimate increases, and can be difficult to fit especially when there are few studies.

1.6 Challenges in assessing comparative accuracy in systematic reviews

While systematic reviews have generally focussed on the evaluation of the accuracy of a single test, reviews comparing the accuracy of two or more tests are increasingly being published. For comparing health care interventions, properly conducted RCTs are regarded as the most valid source of evidence of comparative effectiveness, although in their absence studies with non-randomized designs may also be considered.⁸²⁻⁸⁵ Both RCTs and systematic reviews of RCTs are available to guide intervention selection for many conditions. However, clinical investigators and funders do not seem to have demanded similarly rigorous standards in the creation of reliable evidence for selecting diagnostic tests.

Systematic reviews of comparative accuracy often undertake separate meta-analyses for each test, and then compare their results as shown in Figure 1.11 (panel A). Comparing separate meta-analyses of each test can be likened to making comparisons of single arms of RCTs of interventions, or between case series.⁸⁶ This indirect between-study (uncontrolled) test comparison uses a different set of studies for each test and so does not ensure like-with-like comparisons; the difference in accuracy is prone to confounding due to differences in patient groups and study methods.

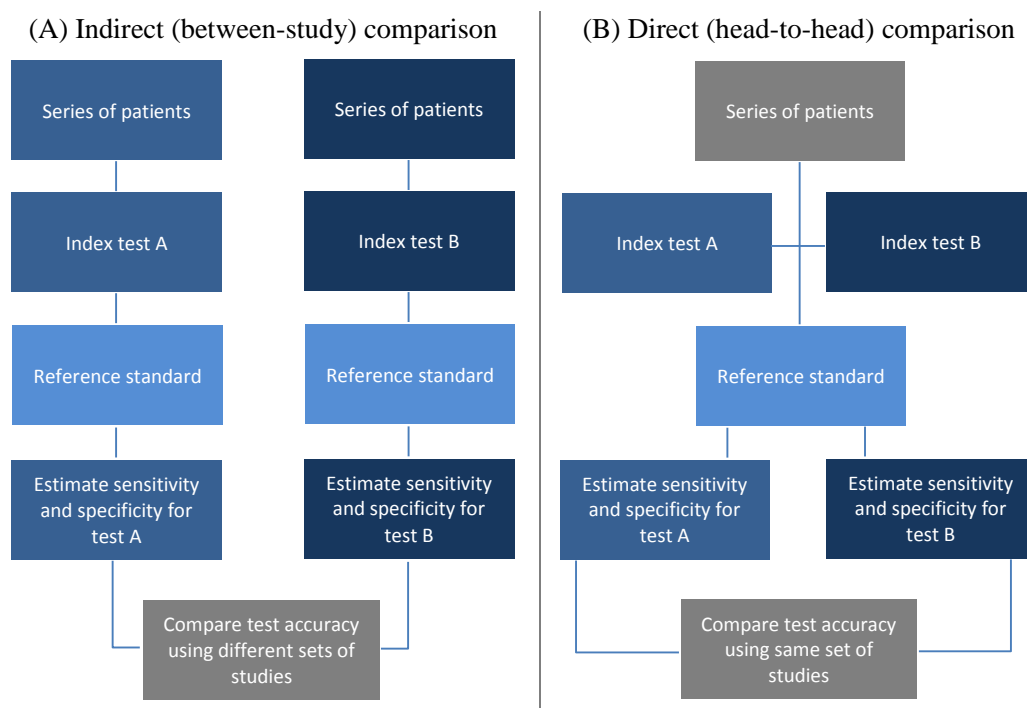


Figure 1.11| Approaches for comparing test accuracy

Systematic reviews of comparative accuracy often undertake separate meta-analyses for each test and then compare their results as shown in (A) using one study for each test. The second figure (B) shows a direct comparison using a comparative study (comparison within patients or between randomized groups).

Although direct comparisons based on only comparative studies are likely to ensure an unbiased comparison and enhance validity (Figure 1.11 panel B), such analyses may not always be feasible due to limited availability of comparative studies.^{71,87} For example, four modalities—duplex ultrasonography (DUS), computed tomography angiography (CTA), magnetic resonance angiography (MRA), and digital subtraction angiography (DSA)—are used to image steno-occlusions in lower limb peripheral arterial disease (PAD). In routine clinical practice, DSA is not used as an alternative to DUS, CTA or MRA, but is the reference standard used to determine the diagnostic accuracy of the other three modalities. For certain clinical questions, a reference standard may also be the alternative test used in practice, i.e., also a comparator test. Consequently, the reference standard is sometimes incorrectly referred to as the comparator.

DUS, CTA and MRA have specific benefits and limitations but the clinical indications for CTA are very similar to those of MRA.⁸⁸ Nonetheless, no study has directly compared the diagnostic accuracy of MRA and CTA.⁸⁹ In a systematic review of the diagnostic accuracy of CTA, MRA and DUS that included 58 studies, Collins et al⁹⁰ only found two studies that compared MRA and DUS in the same study population. One of the two studies performed all tests in all patients while the other study only performed MRA in a subset of patients. The authors concluded that MRA has better overall diagnostic accuracy than CTA or DUS.

An indirect comparison uses all eligible studies that have evaluated at least one of the tests of interest thus maximizing use of the available data (Figure 1.12). Therefore, an indirect test comparison can be regarded as a mixed test comparison when both comparative and non-comparative studies of one or more tests are included (i.e. direct and indirect comparisons are combined). However, a distinction is not usually made between both approaches and they are simply termed indirect comparisons. This is unlike indirect comparisons and mixed treatment comparisons (or network meta-analyses) of interventions where there is a clear distinction between the two methods of obtaining indirect evidence of treatment effectiveness.⁹¹⁻⁹³ Furthermore, unlike indirect comparisons of interventions, these *naïve* indirect test comparisons do not use a common control to adjust for differences in average test positivity (outcome) rates between studies.

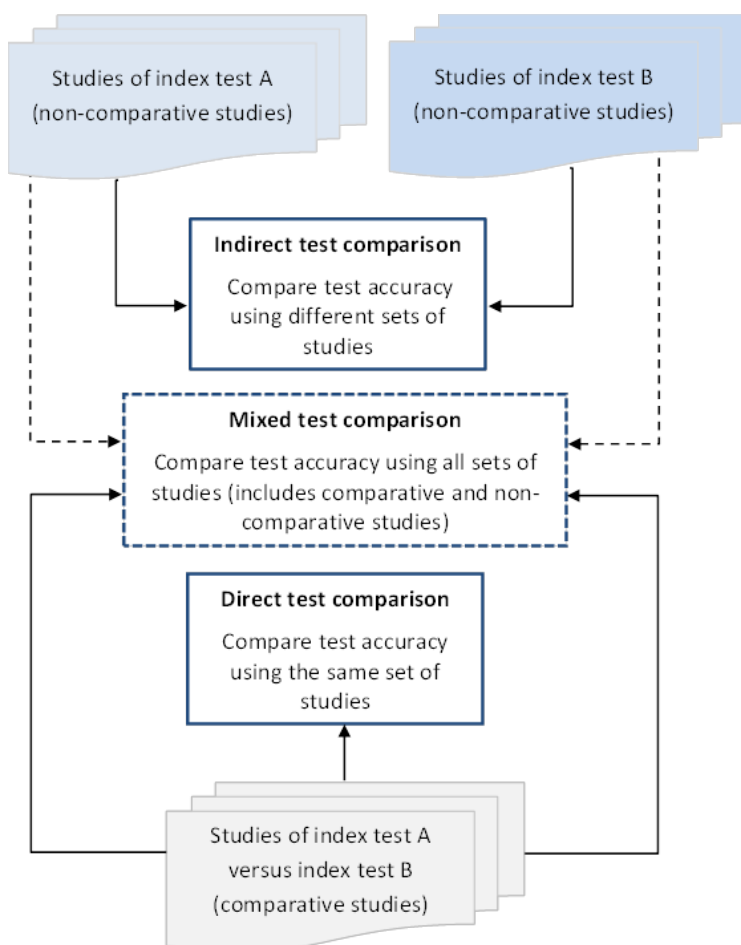


Figure 1.12| Types of test comparisons in a comparative accuracy meta-analysis

For simplicity, a pairwise comparison is shown but a test comparison may include more than two tests. The mixed test comparison includes head-to-head studies of both index tests as well as evaluations of either of the two tests. The mixed test comparison is also an indirect comparison.

The high degree of heterogeneity in estimates of sensitivity and specificity commonly observed in meta-analyses of test accuracy⁹⁴ raises concern that decisions based on the results of comparisons made between meta-analyses may be unreliable because the likelihood of confounding is high. The likely magnitude of any bias is unknown, but if substantial, robustly designed comparative test accuracy studies (see Figure 1.10) should be more routinely undertaken and preferred for evidence to guide test selection. Therefore, the availability of comparative studies and the feasibility of direct comparisons warrant investigation given the

potential value of comparative accuracy reviews as a surrogate in the absence of direct evidence about test effectiveness.

There is also concern that suboptimal methods are being used to estimate test accuracy and to make conclusions about the superiority of one test compared to another. Hierarchical meta-analytic methods have been increasingly adopted over time, but their use still remains limited.²⁷ In contrast to meta-analyses of interventions, recommended methods may sometimes be impractical due to the limited number of included studies compared to the complexity of the methods. It is unclear whether simpler methods can be an appropriate alternative in such situations. Therefore meta-analytic methods for valid comparisons of tests are needed to avoid misleading conclusions and recommendations.

1.7 Research questions and thesis outline

1.7.1 Research questions

Given the challenges and uncertainties highlighted above, the primary aim of this thesis is to assess the reliability and transparency of evidence derived from systematic reviews and meta-analyses of comparative accuracy, including the validity of the meta-analytic methods used to synthesise the evidence. To achieve this objective, the following research questions will be addressed:

1. How have comparisons of test accuracy been performed and reported in published systematic reviews and meta-analyses? What are the statistical methods used and are the methods and findings well reported?

2. How feasible are direct comparisons (i.e. how often are comparative studies available)?

Are meta-analyses of comparative test accuracy studies more reliable than meta-analyses of non-comparative studies?

3. How should meta-analyses be undertaken to compare test accuracy

- a. when both tests are evaluated in different sets of studies (indirect uncontrolled comparisons);
- b. when studies directly compare both tests (head-to-head comparisons);
- c. when there is a mixture of non-comparative and comparative studies (mixed comparison);
- d. when there are more than two tests (multiple test comparison)?

Do methods give the same results in each of the situations above?

4. How should meta-analyses be undertaken when there are few studies or sparse data?

1.7.2 Thesis outline

Chapter 2 introduces several examples of published systematic reviews comparing the accuracy of two or more tests. These examples are from some of the reviews the author has worked on in the past five years during the course of this PhD research. This chapter aims to highlight key methodological issues by providing practical demonstrations using real life case studies. The chapter also shows how the problems were addressed in the reviews and underscores the need for research undertaken in Chapters 4 to 8.

Chapter 3 focuses on the search methods for identifying systematic reviews that assessed the accuracy of at least two tests, and the methods for comparative meta-analysis investigated in the thesis. Thus the aim of this chapter is to collate the data sources and search strategies used

Chapter 1: Introduction

to obtain methods and materials for the thesis. The chapter also contains the search results for the primary cohort of reviews from which the subset of reviews used in chapters 4 to 6 were derived by applying eligibility criteria relevant to the objectives of each chapter.

Chapter 4 provides an overview of data synthesis methods in published reviews that assessed at least two tests. The aim of the chapter is to provide a descriptive survey of recent practice, and to identify shortcomings in methods and reporting with a view to making recommendations for improvements. Recent reviews selected from the primary cohort of reviews identified in Chapter 3 are examined in detail, and their general, methodological and reporting characteristics are summarised. The chapter describes how reviews handled test comparisons in terms of the strategies (indirect and/or direct comparisons) and meta-analytic methods used, presentation of results, and how review findings were interpreted in the context of the strength of the evidence. Examples of good practice are identified to aid future review authors in preparing these reviews.

Chapter 5 describes the assessment of the availability of comparative test accuracy studies, and compares meta-analyses of comparative studies (direct comparisons) and meta-analyses of non-comparative studies (indirect comparisons). This chapter aims to determine the existence and magnitude of discrepancies between meta-analyses of direct and indirect comparisons, and to provide empirical evidence of the importance of comparative accuracy studies. The meta-analyses used in this empirical evaluation are derived from the cohort of reviews identified in Chapter 3.

Chapter 6 considers methods for comparative meta-analysis that were identified from searches reported in Chapter 3. Chapter 6 aims to describe the properties of the methods such as model specification, modelling assumptions, and the advantages and limitations of each method. This methodological review will provide an overview of the available comparative meta-analysis methods and will gaps where appropriate methods are lacking.

In Chapter 7, the performance of methods identified in Chapter 6 that are deemed to be methodological rigorous or are frequently used in practice will be empirically evaluated by using a subset of the reviews identified in Chapter 3. The evaluation includes investigating the impact of alternative methods and modelling assumptions on conclusions in order to identify the most appropriate methods and to provide practical guidance for meta-analysts.

Chapter 8 investigates the performance of hierarchical meta-analytic methods in situations with few studies or sparse data for a single index test. This is a common problem faced by meta-analysts and the aim of this chapter is to identify situations where complex hierarchical methods are likely to give model fitting problems and misleading results, and to suggest simpler appropriate meta-analytic methods. The meta-analyses in two reviews are used as motivating examples and simulation is used to generate datasets that reflect realistic scenarios for meta-analyses of test accuracy studies. The performance of seven hierarchical models incorporating increasing simplifications is investigated and the chapter concludes with recommendations for practice.

Although the main emphasis of this thesis is on test comparisons, the simulation study only considers the meta-analysis of a single test. This comprehensive simulation has wide

Chapter 1: Introduction

implications and is a necessary preliminary step for establishing the validity of simpler models in the context of the evaluation of a single test prior to considering applications to test comparisons. This is also important because a comparative meta-analysis is not always possible with few studies of one or more of the tests being compared. As such separate meta-analyses may be required for some tests in addition to the comparative meta-analysis of tests with sufficient data. The chapter includes a section on how the results may be generalised to test comparisons.

Chapter 9 concludes the thesis by summarising the key findings from Chapters 2 to 8. The chapter draws together the various issues addressed in the thesis and seeks to provide a coherent summary of both the problems raised and the recommendations that were developed based on the thesis findings. The chapter also offers suggestions for future research, and the overall limitations and conclusions of the thesis.

2 METHODOLOGICAL CHALLENGES IN META-ANALYSES OF TEST COMPARISONS

A paper based partly on the content of this chapter has been published.

Citation: **Takwoingi Y**, Riley R, Deeks J. Meta-analyses of diagnostic accuracy studies in mental health. *Evidence Based Mental Health* 2015; 18:103-109.

2.1 Introduction

The challenges in assessing comparative accuracy in systematic reviews were introduced in the previous chapter. Planning and conducting comparative analyses is more complex than the analysis of a single test. Comparative meta-analyses require careful consideration of several methodological issues such as test selection strategy (i.e. which tests should be compared) if a plethora of tests is included in a review; test comparison strategy (direct, indirect or both); choice of hierarchical model to use (estimation of summary points or curves); and modelling assumptions such as whether the variance parameters for random effects should be allowed to depend on test type.

The effect of different modelling assumptions on meta-analytic findings and conclusions about the relative accuracy of tests has not been investigated previously. Therefore, using seven published systematic reviews^{56,68-70,95-97} of five target conditions—*Plasmodium falciparum* malaria, *Plasmodium vivax* or non-falciparum malaria, bipolar disorder, Down's syndrome and common bile duct stones— this chapter describes applications of meta-analysis to test comparisons. The seven reviews were chosen to exemplify key methodological issues that will be examined further in chapters 4 to 8.

Only data and information relevant for context, and the analyses required for highlighting the methodological issues are given in this chapter. Although the author conducted the statistical analyses reported in the reviews, each review was a collective effort with other individuals providing methodological and clinical expertise. An overview of each case study is provided in section 2.2. This includes a description of the target condition, review rationale and objectives, index tests, and reference standard(s). In section 2.3 each issue is illustrated using one or more case studies. Section 2.4 concludes the chapter with a summary of the issues raised and sets out the rationale for subsequent chapters.

2.2 Synopsis of case studies

A brief evidence profile of each review is given in Table 2.1. Each target condition and the reviews used as case studies are summarised in the subsections below. The two malaria reviews differ only in terms of the target condition and so both reviews are described in the same section (section 2.2.1) to avoid duplication.

Table 2.1 | Evidence profile of seven published reviews

Target condition	Reference standard	Index tests	Number of comparative studies in review	Total number of studies in review (study cohorts)*	Number of patients in review	Settings
<i>Plasmodium falciparum</i> malaria ⁹⁵	Microscopy or PCR	RDTs	10	74 (85)	48,007	Ambulatory healthcare settings in <i>Plasmodium falciparum</i> malaria endemic areas
<i>Plasmodium vivax</i> or non-falciparum malaria ⁹⁶	Microscopy or PCR	RDTs	7	37 (47)	22,862	Ambulatory healthcare settings in non-falciparum malaria endemic areas
Bipolar disorder (any type) ⁵⁶	MINI, SADS, SCAN or SCID	BSDS, MDQ, and HCL-32	12	53	21,542	Mental health care centres, primary care and general population
Common bile duct stones ⁶⁸⁻⁷⁰	Surgical or endoscopic extraction of stones, or clinical follow up	Liver function tests and ultrasound	1	5	523	Secondary care
Down's syndrome (Trisomy 21) ⁹⁷	Amniocentesis, chorionic villus sampling or postnatal macroscopic inspection	MRCP and EUS	2	18	2,366	Secondary care
		ERCPC and IOC	0	10	972	Secondary care
		Serum tests in the first trimester	27	56	204,759	Routine screening and high risk referrals for invasive testing

*Some studies in the malaria reviews included multiple cohorts and data were presented separately for each population in the study. BSDS = bipolar spectrum diagnostic scale; ERCPC = endoscopic retrograde cholangiopancreatography; EUS = endoscopic ultrasound; HCL-32 = hypomania checklist; IOC = intraoperative cholangiography; MDQ = mood disorder questionnaire; MINI = mini-international neuropsychiatric interview; MRCPC = magnetic resonance cholangiopancreatography; PCR = polymerase chain reaction; RDTs = rapid diagnostic tests; SADS = schedule for affective disorders and schizophrenia; SCAN = schedules for clinical assessment in neuropsychiatry; SCID = structured clinical interview.

2.2.1 Rapid diagnostic tests for uncomplicated malaria in endemic countries

Malaria is a life-threatening infectious disease caused by the parasitic protozoan *Plasmodium*.

P. falciparum and *P. vivax* are the two most common species infecting humans.

Approximately 40% of the world's population is at risk for *P. vivax* malaria, although infection with *P. falciparum* is associated with the highest mortality among persons with malaria.⁹⁸ Resistance to chloroquine and other antimalarials is more likely for *P. falciparum* than other *Plasmodium* species, and species identification is important to select appropriate treatment.⁹⁹ Immunochromatographic rapid diagnostic tests (RDTs) are alternatives to microscopic diagnosis which is current practice and the reference standard. Timely, high quality microscopy (by examination of thick and thin blood films) may be unavailable in resource-poor settings and remote areas, but RDTs offer potential benefits through extension of rapid diagnosis to such areas.

RDTs use different types of antibody or antibody combinations to detect *Plasmodium* antigens (Table 2.2). Some antibodies aim to detect a particular species while others are pan-malarial aiming to detect all *Plasmodium* species. Pan-specific RDTs distinguish *P. falciparum* (or mixed) infections from infections with only non-falciparum species; differentiation between non-falciparum species (*P. vivax* from *Plasmodium ovale* and *Plasmodium malariae*) is not possible. More recently developed, vivax-specific RDTs can detect *P. vivax* mono-infection or co-infection.⁹⁹

A single review evaluating RDTs for detecting all species of malaria in people living in malaria endemic areas with symptoms of malaria was planned. However, it became apparent during the systematic review process that such a publication would be very large and so the

review was split into two reviews to enhance readability.^{95,96} One review assessed the accuracy of RDTs for detecting *P. falciparum*⁹⁵ while the other review assessed *P. vivax* and non-falciparum.⁹⁶ The *P. falciparum* review included 74 unique studies evaluating one or more RDTs in a consecutive series of patients, or a randomly selected series of patients. Seven studies reported multiple cohorts within a study and presented data separately for each different population, giving a total of 85 study cohorts. The non-falciparum review included 37 unique studies of a similar design as stated above for the *P. falciparum* review. One of the studies included 10 separate cohorts and another study included two cohorts. Altogether, there were 47 study cohorts. In both reviews, a few studies assessed the accuracy of RDTs using polymerase chain reaction (PCR) as the reference standard, as well as assessing RDTs against microscopy. Only evaluations using microscopy as the reference standard were considered in this thesis.

Table 2.2| Types of rapid diagnostic tests for detecting malaria

RDT	RDT target antigen
Type 1	HRP-2 (<i>P. falciparum</i> specific)
Type 2	HRP-2 (<i>P. falciparum</i> specific) and aldolase (pan-specific)
Type 3	HRP-2 (<i>P. falciparum</i> specific) and pLDH (pan-specific)
Type 4	pLDH (<i>P. falciparum</i> specific) and pLDH (pan-specific)
Type 5	pLDH (<i>P. falciparum</i> specific) and pLDH (<i>P. vivax</i> specific)
Type 6	HRP-2 (<i>P. falciparum</i> specific), pLDH (pan-specific) and pLDH (<i>P. vivax</i> specific)
Type 7	Aldolase (pan-specific)

HRP-2 = histidine-rich protein-2; pLDH = plasmodium lactate dehydrogenase; RDT = rapid diagnostic test.

RDTs use different types of antibody or antibody combinations to detect *Plasmodium* antigens. (Adapted from Abba et al 2011⁹⁵)

2.2.2 Screening tests for bipolar spectrum disorders

Bipolar disorder is a complex chronic condition characterized by periods of mania and depression. Bipolar spectrum disorders are frequently under recognized and/or misdiagnosed in various settings leading to a long delay in diagnosis from the initiation of affective symptoms.^{100,101} The use of self-report screening instruments for bipolar disorder that are both time- and cost-effective may aid in timely diagnosis. The review compared the diagnostic accuracy of three screening questionnaires—the bipolar spectrum diagnostic scale (BSDS), the hypomania checklist (HCL-32) and the mood disorder questionnaire (MDQ).⁵⁶

Based on the Diagnostic and Statistical Manual of Mental Disorders 4th Edition (DSM-IV) criteria, the reference standards used were the Structured Clinical Interview (SCID), Mini-International Neuropsychiatric Interview (MINI), Schedules for Clinical Assessment in Neuropsychiatry (SCAN), and Schedule for Affective Disorders and Schizophrenia (SADS). Fifty three studies conducted in different settings (mental health care and primary care/general population) in general adult psychiatric populations were included. Three target conditions were considered; defined as bipolar disorder in general (i.e. any type of bipolar disorder), bipolar disorder type II (BD-II) and bipolar disorder not otherwise specified (BD-NOS). Separate analyses were conducted for each setting and target condition.

2.2.3 Diagnostic tests for common bile duct stones

Biliary stones are conglomerates of precipitated bile salts that form in the gallbladder or the common bile duct. The term 'gallstones' generally refer to the stones in the gallbladder while 'common bile duct stones' refer to stones in the common bile duct. Acute cholangitis is a

dangerous complication of common bile duct stones, caused by an ascending bacterial infection of the common bile duct and the biliary tree along with biliary obstruction.

As described earlier in section 1.1.2 (also see Figure 1.8), tests recommended for diagnosis of common bile duct stones include laboratory liver function tests (LFTs) and imaging tests such as abdominal ultrasound, EUS, MRCP, ERCP and IOC.⁶⁸⁻⁷⁰ Of these tests, IOC can only be done during an operation because the test requires surgical cannulation of the common bile duct during cholecystectomy. The other tests may be used preoperatively or postoperatively. The reference standards used were surgical or endoscopic exploration or extraction of stones if present, or symptom-free follow up for at least six months in those with a negative index test result.

Since the six index tests are likely to be used at different points in the diagnostic pathway as illustrated in Figure 1.8, three reviews were performed to compare the accuracy of the pair of tests used at each point in the pathway. Five studies were included in the ultrasound versus LFTs review,⁶⁸ 18 studies in the EUS versus MRCP review,⁶⁹ and 10 studies in the ERCP versus IOC review.⁷⁰

2.2.4 Antenatal screening for Down's syndrome

Down's syndrome, also known as Down's or Trisomy 21 (T21), is a genetic disorder due to having three, rather than two, copies of chromosome 21. The condition is characterised by significant physical and mental health problems, and disabilities. As there is no cure for Down's, antenatal diagnosis allows for preparation for the birth and subsequent care of a baby with Down's, or for the offer of a termination of pregnancy. The most accurate tests for

Down's involve testing fluid from around the baby (amniocentesis) or tissue from the placenta (chorionic villus sampling (CVS)) for the abnormal chromosomes associated with Down's. However, both tests are invasive and associated with a risk of miscarriage and are not suitable for routine screening of pregnant women. Instead, non-invasive tests based on serum, urine or ultrasound markers measured in the first and/or second trimester of pregnancy, can be used to identify 'high risk' women to be referred for definitive invasive testing. Older women are known to have a higher chance of carrying a baby with Down's syndrome,¹⁰² and so risk calculations are often based on combinations of the screening tests and maternal age.

A review identifying all screening tests for Down's syndrome used in clinical practice, or evaluated in a research setting was planned. The aim was to identify the most accurate test(s) available, and to provide clinicians, policy makers and women with robust and balanced evidence on which to base decisions about interpreting test results and implementing screening policies to triage the use of invasive diagnostic testing. Subsequently, the review became a suite of five reviews based on a single generic protocol.¹⁰³ This was done to allow for greater ease of reading and accessibility of data, and also to allow the reader to focus on separate groups of tests. An overview review comparing the best tests for antenatal Down's syndrome screening from amongst commonly used strategies and the best tests from each review was planned.

In this chapter, only the review of serum screening tests used in the first trimester of pregnancy (up to 14 weeks' gestation) is considered.⁹⁷ The review assessed the following 18 individual markers; a disintegrin and metalloprotease 12 (ADAM12), alpha-fetoprotein

(AFP), inhibin, pregnancy associated plasma protein A (PAPP-A), invasive trophoblast antigen (ITA), free beta human chorionic gonadotrophin (β hCG), placental growth factor (PIGF), Schwangerschafts protein 1 (SP1), total hCG, progesterone, unconjugated estriol (uE3), growth hormone binding protein (GHBP), placental growth hormone (PGH), hyperglycosylated hCG, proform of eosinophil major basic protein (ProMBP), human placental lactogen (hPL), free alpha human chorionic gonadotrophin (α hCG), and free β hCG to AFP ratio. These markers can be used individually, in combination with maternal age, and also in combination with each other. Twelve different cut-offs were used. Altogether, 78 test strategies were assessed in 56 included studies.

2.2.5 Overview of analysis methods

Comparative meta-analyses were conducted using the meta-regression approach described in section 1.5.4.1 for bivariate models and section 1.5.4.2 for HSROC models. The effect of the covariate terms for test type on the parameters of the bivariate or HSROC model was investigated. Details of the meta-analysis of each case study will be given as appropriate in later sections.

The hierarchical models were fitted using the NLMIXED procedure in the SAS software, version 9.2 (SAS Institute, Cary, North Carolina); the SAS macro (MetaDAS) which is a wrapper for NLMIXED developed by the author;¹⁰⁴ or the *xtnlogit* command in Stata versions 10 to 13 (Stata-Corp, College Station, Texas, USA). The NLMIXED procedure fits nonlinear and generalized linear mixed models while the *xtnlogit* command can only fit linear mixed effects logistic regression models. Stata does not have a command for fitting nonlinear models and so the HSROC models were always fitted in SAS while the bivariate

models were fitted using either Stata or SAS. Irrespective of the model fitted or the software program used, adaptive Gaussian quadrature was used for the maximum likelihood estimation. The default optimization technique in each software program was used—a quasi-Newton technique in SAS and a Newton-Raphson technique in Stata. The distribution of the random effects was always assumed to be normally distributed because both Stata and SAS do not have options for alternative distributions.

When the HSROC model was used and asymmetric curves were fitted, estimates of sensitivities were obtained from the curves at fixed values of specificity by using equation 1.19 and the ESTIMATE statement in NLMIXED. The ESTIMATE statement computes additional estimates as a function of parameter values and produces standard errors and confidence intervals using the delta method. Likelihood ratio tests were used to assess the fit of alternative models as described earlier (section 1.5.4.4). Individual study estimates of sensitivity and specificity were presented in forest plots and/or SROC plots. Review Manager version 5 (The Nordic Cochrane Centre, The Cochrane Collaboration, 2014) was used to produce all forest plots and some of the SROC plots.

2.3 Methodological issues

The case studies focus on the following methodological questions:

1. What test comparisons should be performed?
2. Are meta-analytic methods that compare both SROC curves and points needed?
3. Should a common shape be assumed for the SROC curves across tests?
4. Is the assumption of equal variances across tests appropriate?
5. Is meta-analysis of test comparisons feasible with limited data?

These questions represent key issues that may be encountered when performing a comparative meta-analysis. Each question will be considered in turn and illustrated using at least one of the reviews summarised above.

2.3.1 What test comparisons should be performed?

Test comparisons are not limited to a pair of tests. Therefore, the number of tests assessed in a comparative review can be overwhelming, and a decision needs to be made about the tests to include in a meta-analysis, and whether more than one meta-analysis is needed. The approach adopted in the Down's review is used to illustrate test selection from a multitude of tests (section 2.3.1.1), and the *P. falciparum* review is used to demonstrate an approach for structuring test comparisons (section 2.3.1.2). The availability of comparative studies and the feasibility of direct comparisons are considered collectively across the seven reviews (section 2.3.1.3).

2.3.1.1 Selecting tests and studies to include in a comparative meta-analysis

Example: First trimester serum tests for Down's syndrome screening

Due to the large number of test strategies in the Down's review (total of 78), test strategies were selected for further investigation if they were evaluated in four or more studies or, if there were two or three studies, but individual study results indicated performance was likely to be superior to a sensitivity of 70% and specificity of 90%.⁹⁷ A sensitivity of 70% was chosen because tests used in practice were known to be at least 70% sensitive and a test that performed worse than this would be unacceptable. A specificity of 90% (i.e. FPR = 10%) was chosen because a test with poorer specificity would lead to more women being incorrectly offered invasive testing which carries a risk of miscarriage—national screening policy

required tests to have a maximum 5% FPR. Six test strategies were evaluated by at least four studies and three test strategies were evaluated by two or three studies but had sensitivity greater than 70% at a 5% FPR. Therefore, these nine test strategies were selected for the main test comparison. This test comparison will be considered further in section 2.3.1.3.

2.3.1.2 Structuring test comparisons

Example: Rapid diagnostic tests for uncomplicated *Plasmodium falciparum* malaria

In order to provide a coherent description of the studies that contributed to each analysis, the results were structured by grouping studies according to their commercial brand within test type and then within antibody type (Figure 2.1).⁹⁵ As such, the test comparisons and corresponding meta-analyses in this review were viewed as a hierarchy.

At the lowest level, commercial brands were compared within test types (Types 1, 2, 3, 4, 5 and 6; see Table 2.2). At the next level, test types were compared. Primary studies reported data according to commercial brands and so where more than one brand of the same test type was reported, one brand was selected at random to avoid bias due to inclusion of the same patients more than once in the analysis. The highest level comparison compared tests according to antibody type. Each antibody type was formed by grouping test types into the two groups: Types 1, 2, 3 and 6 were classified as HRP-2 antibody-based tests while Types 4 and 5 were classified as pLDH antibody-based tests. The analytical strategy thus compared the test accuracy of commercial brands within each test type before making comparisons between test types, and then between antibody types.

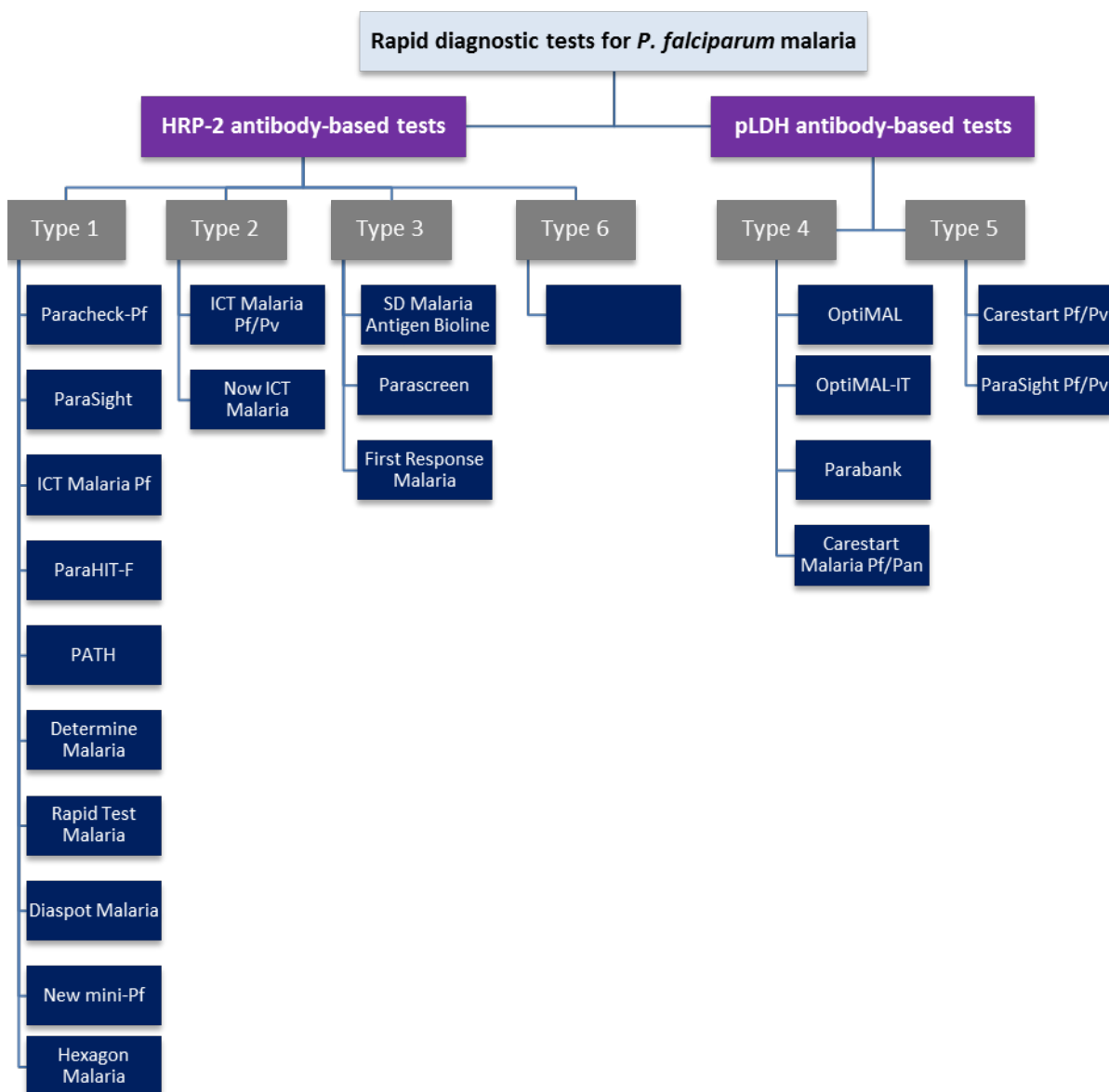


Figure 2.1| Hierarchy of rapid diagnostic tests evaluated against microscopy
 The brand box is blank for Type 6 RDTs because none of the studies evaluated Type 6 RDTs against microscopy.

There is reasonable justification for the grouping described above. Evidence about relative test performance, in addition to other factors such as local malaria epidemiology and resources, can be used to inform decision making for procurement of RDTs. The *Plasmodium* species to be detected is an important consideration in choosing an RDT (see Table 2.2) and so knowledge of differences in accuracy between HRP-2 based tests and pLDH based tests, and the types within each antibody group, is useful for guiding test selection. HRP-2 based

tests are more temperature stable than pLDH based tests and so in environments with uncontrolled temperature, the choice of which type or brand of HRP-2 tests to use is of practical importance. Over 120 RDTs are commercially available from over 60 companies.¹⁰⁵ Differences in test performance may occur between different brands of RDTs within the same type due to differences in format (e.g. cassettes, dipsticks and cards) and ease of use. There are also differences in cost, and cassette format RDTs tend to be more expensive than dipstick RDTs although they are simpler to use.¹⁰⁶ Therefore, information about the relative accuracy of different brands within a type is essential for evaluating cost effectiveness.

The issues of selecting tests and studies, and structuring test comparisons were used to highlight the complexity of test comparisons and decisions that may need to be made before a comparative meta-analysis can be performed. Careful planning is required prior to conducting meta-analysis because decisions made at the planning stage will influence the complexity of the models to be fitted. For instance, if there are many tests in the meta-analysis, it may not be possible to assess the effect of test type on certain model parameters leading to modelling assumptions that may be untenable as will be shown later in sections 2.3.3 and 2.3.4. In Chapter 4, the characteristics of comparative reviews will be examined using a cohort of reviews.

2.3.1.3 Availability of comparative studies and feasibility of direct comparisons

A systematic review may include indirect, direct or both types of comparisons (Figure 1.12). In each of the seven reviews, the number of comparative studies was small (see Table 2.1). Of the seven reviews, both types of analyses were performed in three reviews^{56,95,97} while only an indirect comparison was possible in three reviews.^{69,70,96} The remaining review included only

five studies, one of which was comparative.⁶⁸ Therefore, a formal comparison in a meta-analysis was not possible but a narrative summary was provided. Where both comparative and non-comparative studies were included in a review, the indirect comparison was performed as the main analysis and a direct comparison as secondary analysis. For a review that included more than two tests, a direct comparison was done separately for each pair of tests with sufficient data (pairwise comparison).

For the Down's review, the main analysis was an indirect comparison of the nine selected test strategies (see section 2.3.1.1). This test comparison included all 22 studies with relevant data, and was followed by pairwise direct comparisons using only studies that compared tests within the same participants. Although a separate model was used for each pairwise comparison, the comparisons are summarised in the network plot (Figure 2.2) to show the evidence base for direct comparisons, as well as indirect comparisons where there was a common comparator. Only 13 of the 22 studies were included in the plot. The remaining nine studies only evaluated one of the nine test strategies and so were not included in any of the pairwise direct comparisons. Thus direct comparisons did not make full use of the available data. Both the nodes and edges in Figure 2.2 are weighted according to the number of studies that evaluated each direct comparison. The size of the nodes indicates that the maternal age, PAPP-A, and free β hCG test strategy was the most frequent comparator across studies.

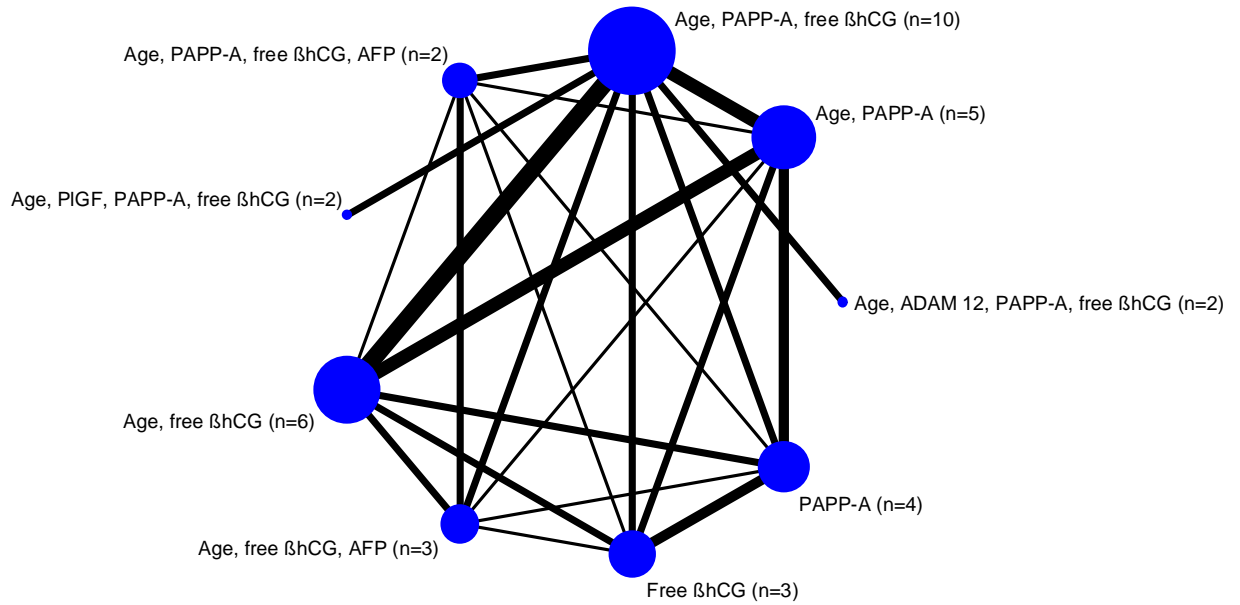


Figure 2.2| Comparison of the nine selected first trimester serum test strategies
n = number of studies that compared a particular test to other tests in the network.
The network plot shows the direct comparisons and only indirect comparisons where there was a common comparator.

Results for all possible pairwise comparisons of the nine strategies are shown as a half-matrix in Table 2.3. Altogether, 36 pairwise comparisons were possible but no comparative study was available for 13 comparisons while 19 included only one or two studies.

Table 2.3| Direct comparisons of sensitivity of nine first trimester serum test strategies at the 5% false positive rate

Ratio of sensitivity (95% CI), P-value for comparison (number of studies)	Free β hCG	PAPP-A	Age, free β hCG	Age, PAPP-A	Age, PAPP-A, free β hCG	Age, free β hCG, AFP	Age, ADAM 12, PAPP-A, free β hCG	Age, PAPP-A, free β hCG, AFP
PAPP-A	1.78 (1.10–2.88), P = 0.02 (2)							
Age, free β hCG	1.67 (1.11–2.50), P = 0.013 (2)	0.94 (0.68–1.29), P = 0.70 (2)						
Age, PAPP-A	2.15 (1.37–3.38), P = 0.001 (2)	1.20 (0.86–1.67), P = 0.29 (3)	1.26 (1.02–1.57), P = 0.034 (4)					
Age, PAPP-A, free β hCG	2.62 (1.77–3.87), P < 0.001 (2)	1.47 (1.09–2.00), P = 0.012 (2)	1.61 (1.31–1.98), P < 0.001 (5)	1.26 (1.04–1.52), P = 0.02 (4)				
Age, free β hCG, AFP	2.19 (1.31–3.64), P = 0.002 (1)	0.71 (0.52–0.98), P = 0.03 (1)	1.08 (0.80–1.46), P = 0.62 (2)	0.61 (0.46–0.82), P < 0.001 (1)	0.63 (0.47–0.86), P = 0.004 (2)			
Age, ADAM 12, PAPP-A, free β hCG	—	—	—	—	1.04 (0.85–1.26), P = 0.71 (2)			
Age, PAPP-A, free β hCG, AFP	3.94 (2.49–6.23), P < 0.001 (1)	1.29 (1.03–1.60), P = 0.024 (1)	1.91 (1.42–2.56), P < 0.001 (1)	1.11 (0.91–1.34), P = 0.31 (1)	1.02 (0.88–1.20), P = 0.77 (2)	1.62 (1.19–2.19), P = 0.002 (2)	—	
Age, PlGF, PAPP-A, free β hCG	—	—	—	—	1.03 (0.91–1.17), P = 0.61 (2)			—

— indicates no comparative study was available for the pair of tests.

Direct comparisons were made only using data from studies which compared each pair of tests on the same women. Where there were at least two studies, meta-analysis was performed to summarise and compare the sensitivities. The ratio of sensitivities was computed by division of the sensitivity for the row by the sensitivity for the column. If the ratio of sensitivity is greater than one then the sensitivity of the test for the row is higher than that for the column, if less than one the sensitivity of the test in the row is lower than in the column. The ratio of sensitivities for a test comparison from a single study was calculated as a ratio of two proportions.

(Adapted from Allred *et al* 2015⁹⁷)

Figure 2.3 illustrates a comparison of summary points using the *P. falciparum* malaria review.⁹⁵ The indirect comparison shown in panel A included 75 HRP-2 based studies and 19 pLDH based studies. Of these, nine studies compared both tests in the same patients and were included in the direct comparison shown in panel B. Table 2.4 gives the summary sensitivities and specificities from the direct and indirect comparisons. Both sets of results consistently shows that HRP-2 based tests were more sensitive than pLDH based tests, and pLDH based tests were more specific than HRP-2 based tests. However, the results of the direct comparison are less precise than the indirect comparison due to the limited number of studies.

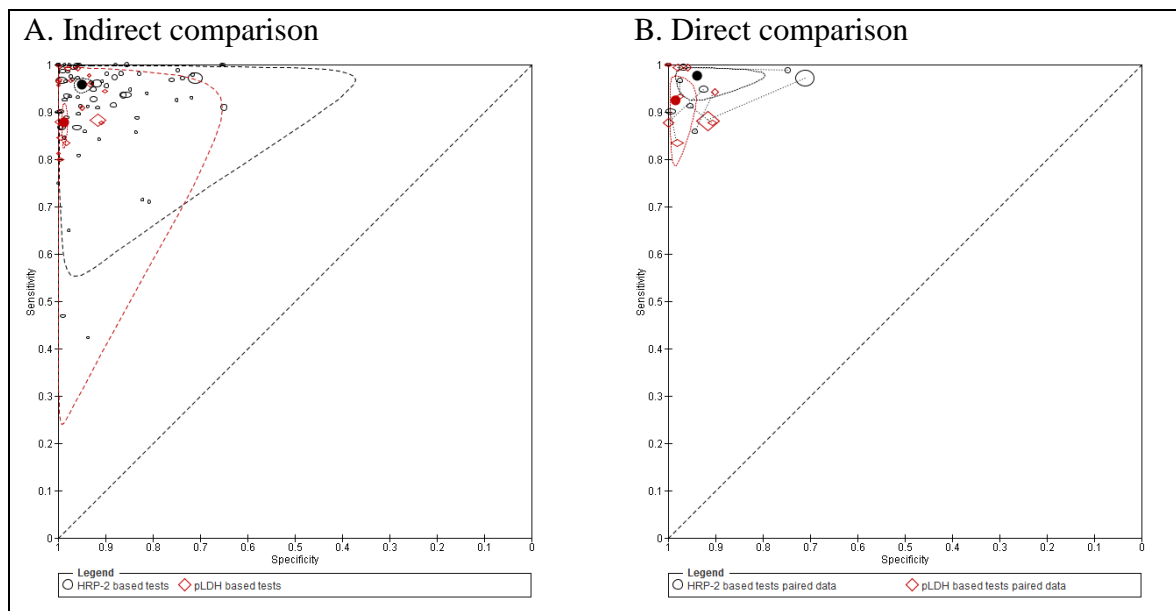


Figure 2.3| Comparison of summary points on SROC plots

HRP-2= histidine-rich protein-2; pLDH= plasmodium lactate dehydrogenase.

For each test on a SROC plot, each symbol represents the pair of sensitivity and specificity from a study. The size of each symbol was scaled according to the precision of sensitivity and specificity in the study. The solid circles (summary points) represent the summary estimates of sensitivity and specificity for each test. Each summary point is surrounded by a dotted line representing the 95% confidence region and a dashed line representing the 95% prediction region (the region within which one is 95% certain the results of a new study will lie). The indirect comparison included all studies that evaluated any of the tests while the direct comparison included only studies that compared both tests in the same patients.

(Adapted from Abba et al 2011⁹⁵)

Table 2.4| Summary estimates from direct and indirect comparisons of HRP-2 based RDTs versus pLDH based RDTs for P falciparum malaria

	Number of studies	Number of patients	Number of cases	Sensitivity (95% CI)	Specificity (95% CI)	Test*
Indirect comparison						
HRP-2 based RDTs	75	43,307	12,857	95.0 (93.5–96.2)	95.2 (93.4–96.6)	
pLDH based RDTs	19	14,787	4,674	93.2 (88.0–96.2)	98.5 (96.7–99.4)	
Ratio (95%CI); P-value				0.98 (0.94–1.02); P = 0.34	1.03 (1.02–1.05); P <0.001	P = 0.01
Direct comparison						
HRP-2 based RDTs	9	10,626	3,672	95.6 (90.0–98.1)	95.8 (84.7–98.9)	
pLDH based RDTs	9	10,623	3,672	94.8 (84.1–98.2)	98.1 (87.8–99.7)	
Ratio (95%CI); P-value				0.99 (0.94–1.04); P = 0.60	1.02 (0.98–1.07); P = 0.22	P = 0.35

*Statistical significance of the difference in test performance was assessed using a likelihood ratio test comparing models with and without covariate terms for test type. Sensitivity and specificity are presented as percentages.

(Adapted from Abba et al 2011⁹⁵)

Figure 2.4 illustrates a comparison of SROC curves using the bipolar disorder review.⁵⁶ The indirect comparison (panel A) included the 44 studies that evaluated the diagnostic accuracy of the MDQ (30 studies), the BSDS (8 studies) and the HCL-32 (17 studies) for detection of any type of bipolar disorder in a mental setting while the direct comparison of the MDQ and HCL-32 included only eight studies (panel B). Three studies directly compared the BSDS and MDQ. No study directly compared the HCL-32 and BSDS in a mental health setting.

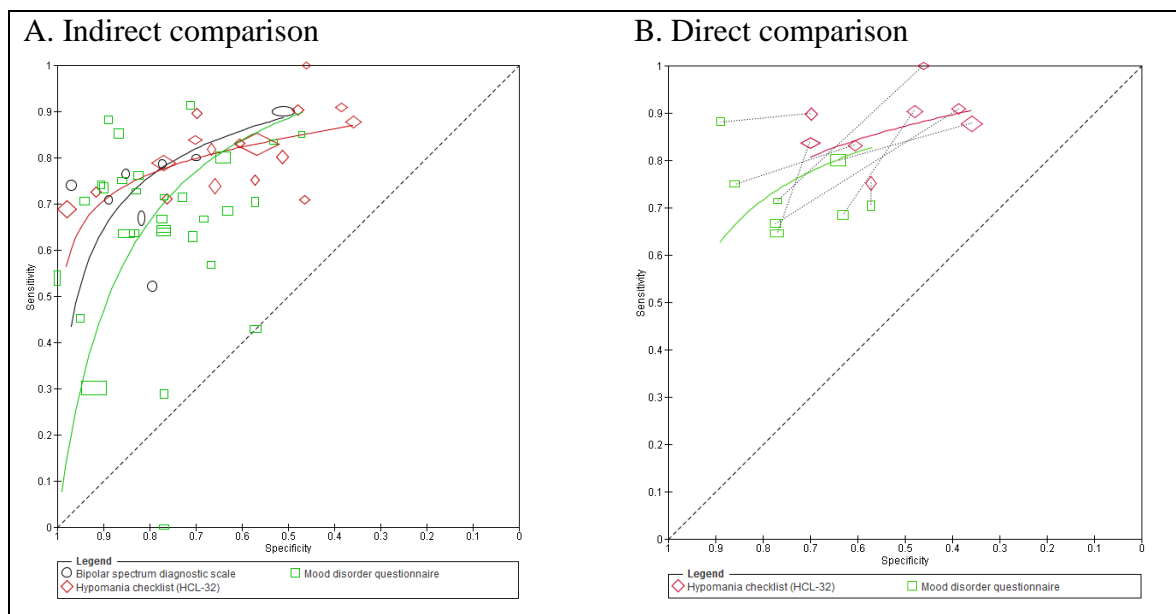


Figure 2.4| Comparison of summary curves on SROC plots

BSDS= bipolar spectrum diagnostic scale; HCL-32= hypomanic checklist; MDQ= mood disorder questionnaire.

For each test on a SROC plot, each symbol represents the pair of sensitivity and specificity from a study. The size of each symbol was scaled according to the precision of sensitivity and specificity in the study. Each summary curve was drawn restricted to the range of specificities from included studies that evaluated the test. The indirect comparison included all studies that evaluated any of the tests while the direct comparison included only studies that compared both tests in the same patients.

(Adapted from Carvalho et 2014⁵⁶)

For diagnosis of common bile duct stones, only an indirect comparison was possible in the ERCP versus IOC review. Five studies evaluated each test. For the comparison of EUS and MRCP, 11 studies evaluated EUS alone, five studies evaluated MRCP alone, and two studies evaluated both tests. For diagnosis of non-falciparum malaria, only an indirect comparison of 28 studies (38 study cohorts) of types 2, 3, and 4 RDTs was performed. Five of the 28 studies directly compared tests but meta-analyses restricted to direct comparisons were not possible because three studies compared type 2 and 3 RDTs, one study compared type 2 and 4 RDTs, and one study compared type 3 and 4 RDTs. In the whole review, only seven of the 37 studies (47 study cohorts) evaluated more than one RDT brand; one compared four brands, three compared three brands and three compared two brands.

The various examples illustrated in this section show the paucity of comparative evidence and even when comparative studies were available, meta-analysis based solely on comparative studies was not always feasible because of limited data. Therefore, the systematic reviews mainly relied on indirect comparisons for making inferences about the relative accuracy of competing tests. The availability of comparative studies will be extensively investigated in Chapter 5 in order to provide empirical evidence of the importance of comparative accuracy studies. Furthermore, an empirical assessment of the impact of study design on summary estimates of test performance will be conducted to determine the existence and magnitude of differences between meta-analyses of direct and indirect comparisons.

2.3.2 Are meta-analytic methods that compare both SROC curves and points needed?

Test comparisons may be based on a comparison of summary points and/or SROC curves as shown in section 2.3.1.3. In the following examples, the choice was influenced by the research question and available data.

2.3.2.1 Summary points only

Example: Rapid diagnostic tests for uncomplicated non-falciparum malaria

RDTs give a binary test result based on a colour change (visible test line) on a strip to indicate the presence of antigens produced by malaria parasites in the blood of infected individuals. This is a binary outcome, therefore it is reasonable to focus on the estimation of summary sensitivities and specificities (summary points). Also, because a common threshold for the judgement of a colour change is assumed, the summary estimates are meaningful and clinically applicable. Overall accuracy (measured by the DOR) is not of interest here because consequences for missed malaria cases outweigh those for false positives.

A bivariate model that included a covariate for test type was used to investigate the association of test type with sensitivity and specificity (equations 1.14 and 1.15 in section 1.5.4.1). The bivariate model was chosen because it directly models sensitivity and specificity unlike the HSROC model. Figure 2.5 shows the summary points for the three RDT types on a SROC plot. Each summary point is surrounded by a 95% confidence region to show the uncertainty around the point estimate, as well as a 95% prediction region to visually illustrate the extent of between-study heterogeneity for each test.

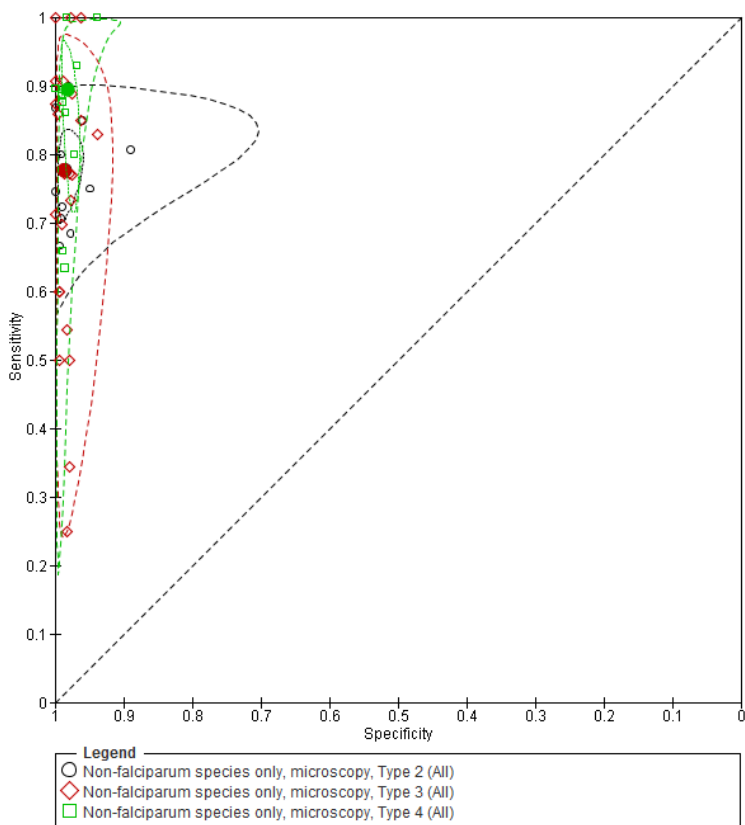


Figure 2.5| SROC plot of rapid diagnostic tests for non-falciparum malaria

The solid circles are the summary estimates of sensitivity and specificity for each RDT type, and are shown with a 95% confidence region (dotted line) and a 95% prediction region (dashed line) around each summary point. The summary point for Type 2 and the 95% confidence region for Type 3 are not visible because Type 2 and Type 3 have identical summary estimates and 95% confidence regions but their 95% prediction regions differ. The size of the symbols for study specific estimates was shrunken to make the summary points visible. (Adapted from Abba et al 2014⁹⁶)

2.3.2.2 Summary points at fixed specificity

Example: First trimester serum tests for Down's syndrome screening

Although studies reported results at different thresholds, it is common in this clinical field for studies to report sensitivity (detection rate) at a fixed specificity (usually a 5% FPR). The chosen FPR level is determined as the FPR deemed acceptable in a particular screening programme. Since all specificities are the same value, there is no need to account for correlation between sensitivity and specificity across studies in a hierarchical meta-analytic model. The main meta-analysis comparing test accuracy included only studies that used a 5% FPR threshold. A univariate random effects logistic regression model (a bivariate model reduced to two univariate models as explained in section 1.4.4.1) that allowed for a separate variance term for the random effects of logit sensitivity for each test was used.⁹⁷ Equation 1.16 was simplified to a univariate model as

$$(\mu_{Aik}) \sim N((\mu_A + v_A t_k), \sigma_{Ak}^2) \quad (2.1)$$

where μ_{Aik} is the logit sensitivity for the k th test within the i th study; t_k represents the k th test; μ_A estimates the expected logit sensitivity for the index test used as the reference category (referent test and not reference standard), $\mu_A + v_A t_k$ estimates the expected logit sensitivity for the k th test, and σ_{Ak}^2 is the variance of logit sensitivity for the k th test.

Based on all available data for the nine test combinations described above, Figure 2.6 shows the point estimates, including confidence intervals, of detection rates for a 5% FPR. For example, the plot shows that for the double test with a marker combination of free β hCG, AFP and maternal age (labelled G), the estimated detection rate at a 5% FPR was 49% (95% CI 39% to 60%) based on data from three studies with 157 affected cases out of 2,992 participants. The test combinations in Figure 2.6 were ordered according to decreasing

detection rates. The single test strategies with and without maternal age (PAPP-A alone; free β hCG alone, PAPP-A and maternal age, and free β hCG and maternal age) have the worst performance, whereas, the triple test strategies (ADAM 12, PAPP-A, free β hCG and maternal age; PAPP-A, free β hCG, AFP and maternal age) have the highest performance.

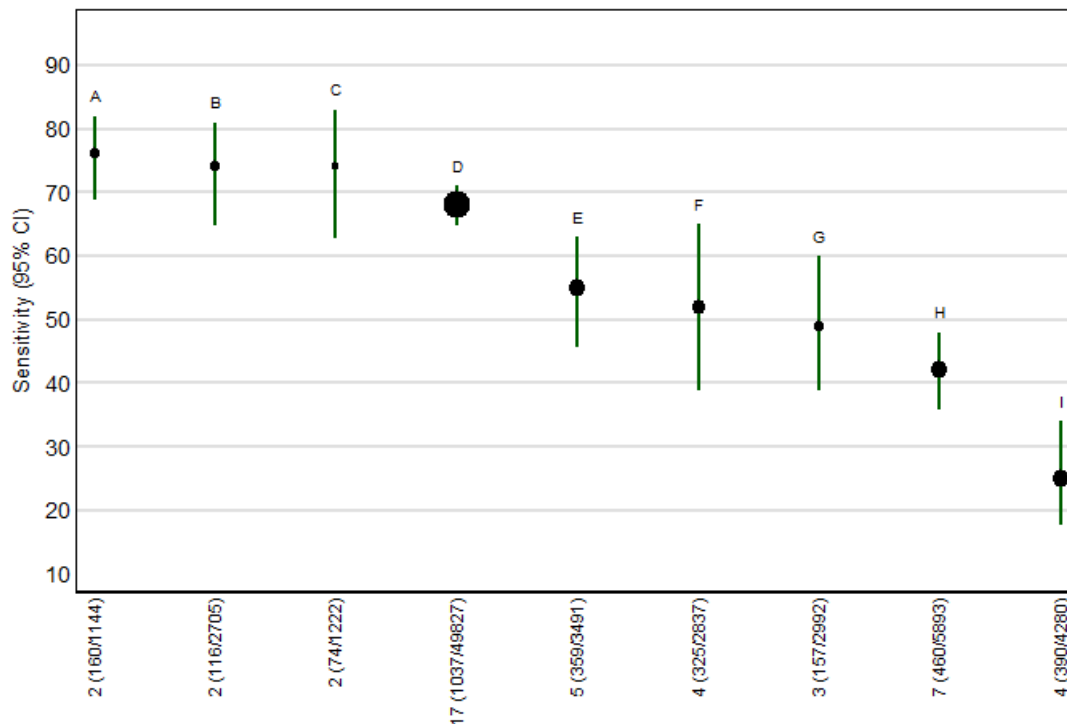


Figure 2.6| Sensitivity (detection rate) at a 5% false positive rate for the 9 selected test strategies

Sensitivity is presented as percentages. Each circle represents the summary sensitivity for a test strategy and the size of each circle is proportional to the number of Down's cases. The estimates are shown with 95% confidence intervals. The test strategies are ordered on the plot according to decreasing detection rate. The number of studies, cases and women included for each test strategy are shown on the horizontal axis. A=Age, PIGF, PAPP-A and free β hCG; B=Age, PAPP-A, free β hCG and AFP; C=Age, ADAM 12, PAPP-A and free β hCG; D=Age, PAPP-A and free β hCG; E=Age, PAPP-A; F=PAPP-A; G=Age, free β hCG and AFP; H=Age, free β hCG; I=Free β hCG.

(Adapted from Alldred et al 2015⁹⁷)

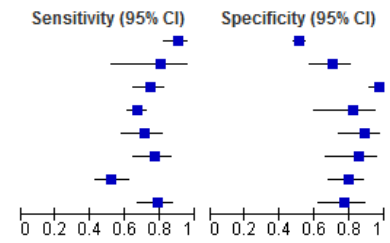
2.3.2.3 Summary curves and points

Example: Screening tests for bipolar disorder– detection of any type of bipolar disorder in mental health centre settings

The total score range from 0–25 points for the BSDS, 0–15 points for the MDQ and 0–32 points for the HCL-32. The cut-off recommended by the developers of each of the screening instruments is 7 for the MDQ,¹⁰⁷ 13 for the BSDS,¹⁰⁸ and 14 for the HCL-32.¹⁰⁹ However, studies used different cut-offs to define a positive screen for each instrument (Figure 2.7).

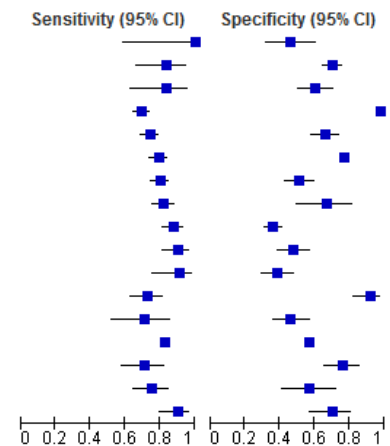
Bipolar spectrum diagnostic scale

Study	TP	FP	FN	TN	Cut-off	Sensitivity (95% CI)	Specificity (95% CI)
Zimmerman 2010	81	427	9	444	8	0.90 [0.82, 0.95]	0.51 [0.48, 0.54]
Nagata 2013	12	19	3	44	11	0.80 [0.52, 0.96]	0.70 [0.57, 0.81]
Chu 2010	74	3	26	97	12	0.74 [0.64, 0.82]	0.97 [0.91, 0.99]
Zarategui 2011	228	4	113	18	13	0.67 [0.62, 0.72]	0.82 [0.60, 0.95]
Vazquez 2010	46	4	19	32	13	0.71 [0.58, 0.81]	0.89 [0.74, 0.97]
Ghaemi 2005	52	4	16	23	13	0.76 [0.65, 0.86]	0.85 [0.66, 0.96]
Shabani 2009	59	14	54	54	14	0.52 [0.43, 0.62]	0.79 [0.68, 0.88]
Castelo 2010a	55	10	15	34	16	0.79 [0.67, 0.87]	0.77 [0.62, 0.89]



Hypomania checklist (HCL-32)

Study	TP	FP	FN	TN	Cut-off	Sensitivity (95% CI)	Specificity (95% CI)
Chou 2012	7	28	0	24	8	1.00 [0.59, 1.00]	0.46 [0.32, 0.61]
Poon 2012	26	82	5	192	11	0.84 [0.66, 0.95]	0.70 [0.64, 0.75]
Carta 2006	20	39	4	60	12	0.83 [0.63, 0.95]	0.61 [0.50, 0.70]
Huang 2013	288	12	129	581	14	0.69 [0.64, 0.73]	0.98 [0.96, 0.99]
Yang 2011	222	53	78	103	14	0.74 [0.69, 0.79]	0.66 [0.58, 0.73]
Yang 2012	244	271	65	907	14	0.79 [0.74, 0.83]	0.77 [0.74, 0.79]
Angst 2005	213	78	53	82	14	0.80 [0.75, 0.85]	0.51 [0.43, 0.59]
Wu 2008	131	13	29	26	14	0.82 [0.75, 0.88]	0.67 [0.50, 0.81]
Meyer 2011	123	223	17	125	14	0.88 [0.81, 0.93]	0.36 [0.31, 0.41]
Leao 2012	66	66	7	61	14	0.90 [0.81, 0.96]	0.48 [0.39, 0.57]
Nallet 2013	30	73	3	46	14	0.91 [0.76, 0.98]	0.39 [0.30, 0.48]
Haghighi 2011	74	5	28	56	14.5	0.73 [0.63, 0.81]	0.92 [0.82, 0.97]
Garcia-Castillo 2012	22	52	9	45	15	0.71 [0.52, 0.86]	0.46 [0.36, 0.57]
Gamma 2013	749	2022	154	2681	15	0.83 [0.80, 0.85]	0.57 [0.56, 0.58]
Forty 2010	42	18	17	58	18	0.71 [0.58, 0.82]	0.76 [0.65, 0.85]
Soares 2010	61	18	20	24	18	0.75 [0.64, 0.84]	0.57 [0.41, 0.72]
Bech 2011	53	19	6	44	18	0.90 [0.79, 0.96]	0.70 [0.57, 0.81]



Mood disorder questionnaire

Study	TP	FP	FN	TN	Cut-off	Sensitivity (95% CI)	Specificity (95% CI)
Wang 2009	5	22	1	25	2	0.83 [0.36, 1.00]	0.53 [0.38, 0.68]
Hu 2012	93	94	216	1084	3	0.30 [0.25, 0.36]	0.92 [0.90, 0.94]
Poon 2012	20	63	11	211	4	0.65 [0.45, 0.81]	0.77 [0.72, 0.82]
Gan 2012	45	16	18	43	4	0.71 [0.59, 0.82]	0.73 [0.60, 0.84]
Zarategui 2011	184	0	157	13	5	0.54 [0.49, 0.59]	1.00 [0.75, 1.00]
Shabani 2009	71	20	42	48	5	0.63 [0.53, 0.72]	0.71 [0.58, 0.81]
Nagata 2013	10	20	5	43	5	0.67 [0.38, 0.88]	0.68 [0.55, 0.79]
Carta 2006	18	14	6	85	6	0.75 [0.53, 0.90]	0.86 [0.77, 0.92]
Hardoy 2005	35	19	11	89	6	0.76 [0.61, 0.87]	0.82 [0.74, 0.89]
Lin 2011	81	10	14	65	6	0.85 [0.77, 0.92]	0.87 [0.77, 0.93]
Chung 2009	0	25	6	83	7	0.00 [0.00, 0.46]	0.77 [0.68, 0.84]
Kim 2008	17	12	42	40	7	0.29 [0.18, 0.42]	0.77 [0.63, 0.87]
van Zaane 2012	15	58	20	77	7	0.43 [0.26, 0.61]	0.57 [0.48, 0.66]
Chung 2008	28	2	34	38	7	0.45 [0.32, 0.58]	0.95 [0.83, 0.99]
Miller 2004	21	12	16	24	7	0.57 [0.39, 0.73]	0.67 [0.49, 0.81]
Zimmerman 2009	33	64	19	364	7	0.63 [0.49, 0.76]	0.85 [0.81, 0.88]
Gervasoni 2009	28	17	16	85	7	0.64 [0.48, 0.78]	0.83 [0.75, 0.90]
Konuk 2007	23	63	13	210	7	0.64 [0.46, 0.79]	0.77 [0.71, 0.82]
Nallet 2013	22	27	11	92	7	0.67 [0.48, 0.82]	0.77 [0.69, 0.84]
Leao 2012	50	47	23	80	7	0.68 [0.57, 0.79]	0.63 [0.54, 0.71]
Soares 2010	57	18	24	24	7	0.70 [0.59, 0.80]	0.57 [0.41, 0.72]
de Sousa Gurgel 2012	36	6	15	96	7	0.71 [0.56, 0.83]	0.94 [0.88, 0.98]
Chou 2012	5	12	2	40	7	0.71 [0.29, 0.96]	0.77 [0.63, 0.87]
de Dios 2008	8	13	3	63	7	0.73 [0.39, 0.94]	0.83 [0.73, 0.91]
Hirschfeld 2000	80	9	29	80	7	0.73 [0.64, 0.81]	0.90 [0.82, 0.95]
Rouget 2005	40	4	14	38	7	0.74 [0.60, 0.85]	0.90 [0.77, 0.97]
Meyer 2011	112	125	28	223	7	0.80 [0.72, 0.86]	0.64 [0.59, 0.69]
Isometsa 2003	17	9	3	8	7	0.85 [0.62, 0.97]	0.47 [0.23, 0.72]
Bech 2011	52	7	7	56	7	0.88 [0.77, 0.95]	0.89 [0.78, 0.95]
Castelo 2010b	63	13	6	32	8	0.91 [0.82, 0.97]	0.71 [0.56, 0.84]

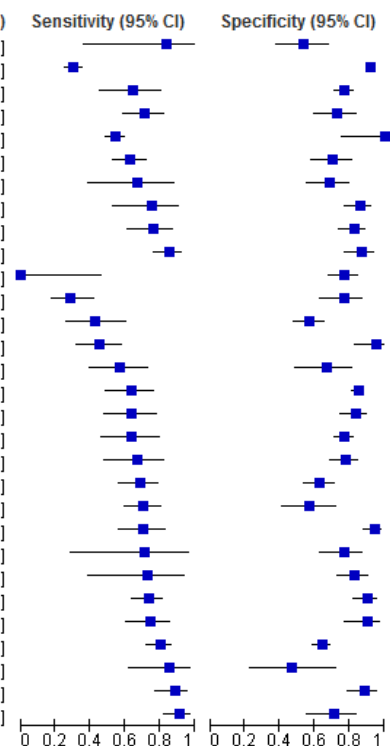


Figure 2.7| Forest plot of screening tests for detection of any type of bipolar disorder (BD type I, BD type II or BD NOS) in mental health centre settings

FN = false negative; FP = false positive; TN = true negative; TP = true positive.

The plot shows study specific estimates of sensitivity and specificity (with 95% confidence intervals) at a specific cut-off. The studies are ordered according to cut-off and sensitivity.

(Adapted from Carvalho et al 2015⁵⁶)

The diagnostic accuracy of the MDQ, the BSDS and the HCL-32 was compared using a HSROC meta-regression model to assess the effect of test type on accuracy, threshold and/or shape parameters of the model (see equations 1.17 and 1.18 in section 1.5.4.2).⁵⁶ The indirect comparison included 44 studies (Figure 2.7). Based on the relationship between the HSROC and bivariate meta-regression models (section 1.5.4.3), summary points were also estimated by applying the HSROC model to only studies that used the recommended cut-off for each instrument. The summary estimates are shown in Table 2.5.

Table 2.5| Accuracy of the BSDS, HCL-32 and MDQ for detection of any type of bipolar disorder in mental health centre settings

Instrument	Cut-off	N	Cases	Patients	Sensitivity (95% CI)	Specificity (95% CI)
BSDS	13	3	474	559	68.8 (63.3–73.7)	85.9 (73.9–92.9)
HCL-32	14	9	1,845	4,807	81.2 (76.7–85.0)	66.7 (46.7–81.9)
MDQ	7	19	969	3,220	65.0 (56.8–72.4)	78.8 (72.5–84.0)

Sensitivity and specificity are presented as percentages. Summary sensitivity and specificity are shown for each instrument at the recommended cut-off.

(Adapted from Carvalho et al 2015⁵⁶)

This example clearly shows that test comparisons may be based on a comparison of summary points and SROC curves in the same review. The feasibility of both types of analyses will depend on the data available and whether common cut-offs are used in practice. In the bipolar disorder review, the choice of a common cut-off was based on recommended cut-offs but may be data driven in other scenarios. Although summary points can be estimated for each test at each threshold for which data are available, ranking of the sensitivities and/or specificities of the tests will not be consistent across thresholds if accuracy depends on threshold. In such situations, a comparison of SROC curves is more appropriate if the curve for each test is allowed to have its own shape (equation 1.17), thus enabling accuracy to depend on threshold

and the crossing of curves. This is evident in Figure 2.4 (panel A) which shows that the SROC curves for the three tests cross, indicating no test is consistently more accurate than any of the others and relative accuracy depends on threshold. This analysis is discussed further in section 2.3.3.1 while Chapter 4 focuses on identifying synthesis methods and test comparison approaches that have been used in recent systematic reviews.

2.3.3 Should a common shape be assumed for SROC curves across tests?

The indirect comparisons performed for detection of any type of bipolar disorder and detection of bipolar disorder type II are considered here.

2.3.3.1 Different asymmetric SROC curves

Example: Screening tests for bipolar disorder – detection of any type of bipolar disorder in mental health centre settings

Using the HSROC model, preliminary assessment of each test separately indicated that the curve for each test may have a different shape. A significant association between test accuracy and threshold, indicated by the shape parameter (β), was found for the HCL-32 ($P < 0.001$) but not for the BSDS ($P = 0.24$) and MDQ ($P = 0.75$). Although the Moses SROC regression framework suffers from methodological problems as noted earlier in section 1.4.3, it is a useful way to graphically explore the relationship between accuracy and threshold. Thus, D was plotted against S for each test to visually characterize how test accuracy, measured by the diagnostic log odds ratio (D), varies with S , a proxy of the positivity threshold across studies (see section 1.4.3 for computation of D and S which are the outcome and explanatory variables of the Moses SROC model). For a study with a zero cell, 0.5 was added to each cell of the 2x2 table for the study.

Figure 2.8 shows the unweighted regression lines of D on S —each test has a different slope, i.e., test accuracy depends on threshold. For the MDQ, D increased as S increased while for the HCL-32 and the BSDS, D decreased as S increased. Since the slope of the MDQ appears to be different to that of the BSDS and HCL-32 and the lines cross, the shape of the SROC curve will not be the same for the three tests. The HCL-32 study highlighted in yellow is potentially an influential study. This study had the highest DOR of 108 for the HCL-32; the DOR for the remaining 16 studies ranged between 2 and 30. There may also be an influential BSDS study (highlighted in blue). The study had a DOR of 92 while the DORs for the remaining seven BSDS studies were between 4 and 19.

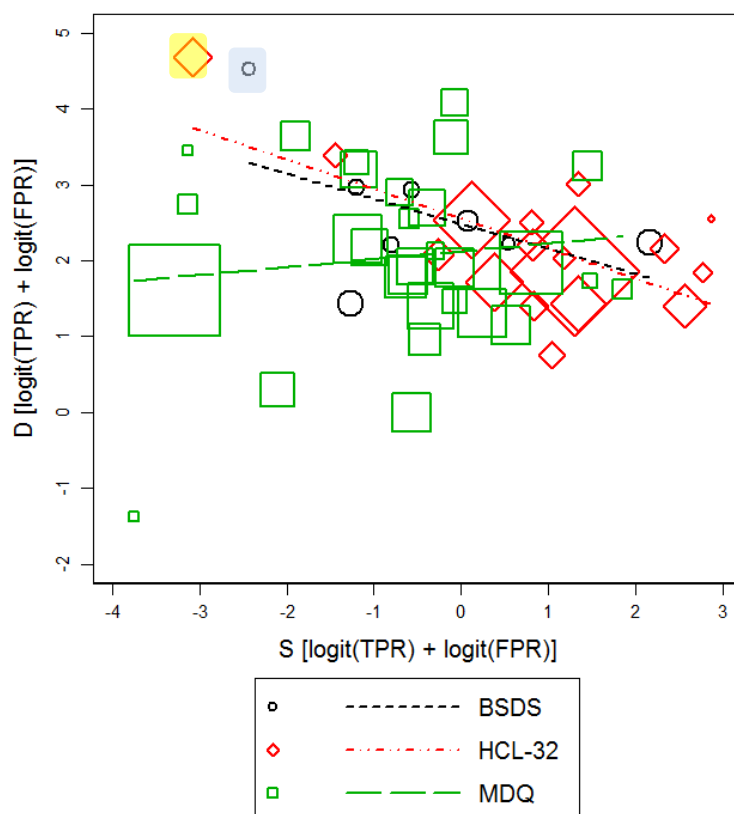


Figure 2.8| Scatterplot of D (log odds ratio) against S (implicit threshold) for the BSDS, HCL-32 and MDQ

An unweighted regression line was fitted for each instrument. The studies highlighted in yellow and blue are potentially influential studies for HCL-32 and BSDS respectively.

When the three tests were compared in a HSROC model, the -2Log likelihood for the model that included covariate terms for shape, accuracy and threshold was 814.8 while that of the model that included covariate terms for only accuracy and threshold was 827.2. The difference in -2Log likelihood between the two models was 12.4 (2 degrees of freedom, $P = 0.002$). Therefore, there was evidence that the shape of the SROC curves differed. Figure 2.4 (panel A) shows that the SROC curves for the three instruments cross, a pattern that is consistent with the regression lines shown in Figure 2.8. Since the curves are asymmetric and do not have a common shape, the relative accuracy of the instruments varies with cut-off and the rDOR cannot be used to quantify relative accuracy. The BSDS curve is consistently above the MDQ curve in the region containing most of the observed data. The HCL-32 curve is above the MDQ and BSDS curve at higher values of specificity, but the curve then crosses both the MDQ and the BSDS curves and accuracy is lower at lower values of specificity. This is also evident in Table 2.6, which shows the sensitivities estimated from the curves at quartiles of the observed specificities in the included studies.

Table 2.6| Comparison of the accuracy of BSDS, HCL-32 and MDQ for detection of any type of bipolar disorder in mental health centre settings

Fixed value of specificity	Instrument	Estimated sensitivity (95% CI)
61	BSDS	86 (74–93)
	HCL-32	82 (78–85)
	MDQ	83 (76–89)
77	BSDS	78 (69–85)
	HCL-32	78 (73–82)
	MDQ	70 (64–77)
85	BSDS	71 (62–79)
	HCL-32	74 (68–80)
	MDQ	58 (50–66)

Sensitivity and specificity are presented as percentages. The sensitivities were estimated from the SROC curves at quartiles of the observed specificity in the included studies.

(Adapted from Carvalho et al 2015⁵⁶)

In a sensitivity analysis, the two outliers highlighted in Figure 2.8 were excluded from the analysis to assess their impact on conclusions about the shape of the SROC curves. To begin, an analysis was performed separately for the BSDS and HCL-32 to assess the impact on each test individually. For the BSDS, the estimated β (standard error) was 0.47 (0.40) when all eight studies were included in the analysis, but it was -0.093 (0.31) when the outlier was excluded (Table 2.7). There was a notable change in the magnitude of the variance parameters, especially the variance of the random effects for accuracy. Similar to the main analysis, there was no statistical evidence of asymmetry in the SROC curve based on the change in -2Log likelihood of models with and without β ($P = 1.0$). Although there was a difference in the magnitude of the estimated shape parameter, the shape of the SROC curve for the BSDS does not appear to be dependent on presence or absence of the outlier.

Table 2.7| Parameter estimates for asymmetric and symmetric HSROC models for the BSDS with and without an outlier

Parameter	BSDS all studies		BSDS outlier excluded	
	Asymmetric	Symmetric	Asymmetric	Symmetric
Accuracy	2.52	2.54	2.19	2.20
Threshold	0.083	-0.178	-0.075	-0.038
Shape	0.47	0	-0.093	0
Variance (accuracy)	0.44	0.47	0.058	0.072
Variance (threshold)	0.34	0.37	0.31	0.30

BSDS = bipolar spectrum disorder scale

For a symmetric SROC curve model, the shape parameter was excluded, i.e., assumed to be zero. The variance parameters are for the random effects for accuracy and threshold.

For the HCL-32, β (standard error) was estimated as 1.33 (0.28) when all 17 studies were included in the analysis, but it was 0.85 (0.01) when the outlier was excluded. Similar to the main analysis, statistical evidence of asymmetry in the SROC curve was observed in the sensitivity analysis ($P = 0.01$). The effect of assuming symmetry of the SROC curve by removing the shape parameter is clearly shown in Table 2.8. When the two outliers were excluded and the meta-regression models comparing the three instruments were refitted, the difference in -2Log likelihood (777.7–771.0) of models with and without covariate terms for β was 6.0 (2 degrees of freedom, $P = 0.05$). There was still evidence that the shape of the SROC curves differed. Thus, the analysis that included all studies was considered robust.

Table 2.8| Parameter estimates for asymmetric and symmetric HSROC models for the HCL-32 with and without an outlier

Parameter	HCL-32 all studies		HCL-32 outlier excluded	
	Asymmetric	Symmetric	Asymmetric	Symmetric
Accuracy	3.10	2.17	2.54	1.96
Threshold	1.21	0.44	0.98	0.55
Shape	1.33	0	0.85	0
Variance (accuracy)	0.071	0.65	0.049	0.21
Variance (threshold)	0.26	0.48	0.16	0.23

HCL-32 = hypomania checklist

For a symmetric SROC curve model, the shape parameter was excluded, i.e., assumed to be zero. The variance parameters are for the random effects for accuracy and threshold.

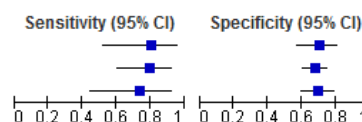
2.3.3.2 Common asymmetric SROC curves

Example: Screening tests for bipolar disorder – detection of bipolar disorder type II in mental health centre settings

Seventeen studies evaluated the BSDS (three studies), HCL-32 (five studies) and MDQ (11 studies) for detection of BD type II (Figure 2.9). Due to limited data for estimation of the shape of each SROC curve, the SROC curves of the three tests were assumed to have the same shape and rDORs were calculated as a summary of the relative accuracy of two screening instruments.

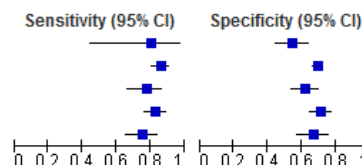
BD-II: Bipolar spectrum diagnostic scale

Study	TP	FP	FN	TN	Cut-off	Sensitivity (95% CI)	Specificity (95% CI)
Nagata 2013	12	19	3	44	11	0.80 [0.52, 0.96]	0.70 [0.57, 0.81]
Chu 2010	23	55	6	116	12	0.79 [0.60, 0.92]	0.68 [0.60, 0.75]
Castelo 2010a	11	30	4	69	16	0.73 [0.45, 0.92]	0.70 [0.60, 0.79]



BD-II: Hypomania checklist (HCL-32)

Study	TP	FP	FN	TN	Cut-off	Sensitivity (95% CI)	Specificity (95% CI)
Carta 2006	8	52	2	61	12	0.80 [0.44, 0.97]	0.54 [0.44, 0.63]
Yang 2012	164	402	27	894	12	0.86 [0.80, 0.90]	0.69 [0.66, 0.71]
Yang 2011	59	59	17	97	13	0.78 [0.67, 0.86]	0.62 [0.54, 0.70]
Mosolov 2014	122	70	25	172	14	0.83 [0.76, 0.89]	0.71 [0.65, 0.77]
Wu 2008	71	35	23	70	14	0.76 [0.66, 0.84]	0.67 [0.57, 0.76]



BD-II: Mood disorder questionnaire

Study	TP	FP	FN	TN	Cut-off	Sensitivity (95% CI)	Specificity (95% CI)
Hu 2012	42	104	149	1192	3	0.22 [0.16, 0.29]	0.92 [0.90, 0.93]
Carta 2006	8	62	2	51	4	0.80 [0.44, 0.97]	0.45 [0.36, 0.55]
Gan 2012	32	21	13	56	4	0.71 [0.56, 0.84]	0.73 [0.61, 0.82]
Hardoy 2005	16	74	4	60	4	0.80 [0.56, 0.94]	0.45 [0.36, 0.54]
Lin 2011	20	57	5	88	6	0.80 [0.59, 0.93]	0.61 [0.52, 0.69]
de Sousa Gurgel 2012	7	30	5	111	7	0.58 [0.28, 0.85]	0.79 [0.71, 0.85]
Gonzalez 2009	9	74	0	116	7	1.00 [0.66, 1.00]	0.61 [0.54, 0.68]
Kim 2008	9	18	24	60	7	0.27 [0.13, 0.46]	0.77 [0.66, 0.86]
Nagata 2013	10	20	5	43	7	0.67 [0.38, 0.88]	0.68 [0.55, 0.79]
Castelo 2010b	11	30	3	70	8	0.79 [0.49, 0.95]	0.70 [0.60, 0.79]
Rouget 2005	11	4	10	38	8	0.52 [0.30, 0.74]	0.90 [0.77, 0.97]

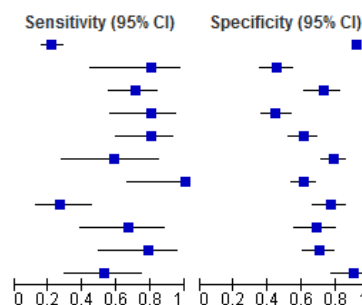


Figure 2.9| Forest plot of BSDS, HCL-32 and MDQ for detection of bipolar disorder type II in mental health centre settings

FN = false negative; FP = false positive; TN = true negative; TP = true positive.

The studies are ordered according to cut-off and study name.

(Adapted from Carvalho et al 2015⁵⁶)

Figure 2.10 presents the SROC curves for the three instruments. The BSDS was not significantly more accurate than the MDQ with an rDOR (95% CI) of 1.7 (0.8 to 3.8, P =

0.19). However, there was evidence that the accuracy of the HCL-32 was superior to that of the MDQ with an rDOR of 2.0 (1.1 to 3.4, $P = 0.018$). It should be noted that although the three BSDS studies are close together in ROC space, the studies used different cut-offs. A curve plotted within the limited range of the specificities of the three studies (68% to 70%) was barely visible and so the range was extended slightly beyond the data (66% to 73%) to make the curve more visible. An alternative approach could be to exclude the BSDS from the test comparison due to limited data.

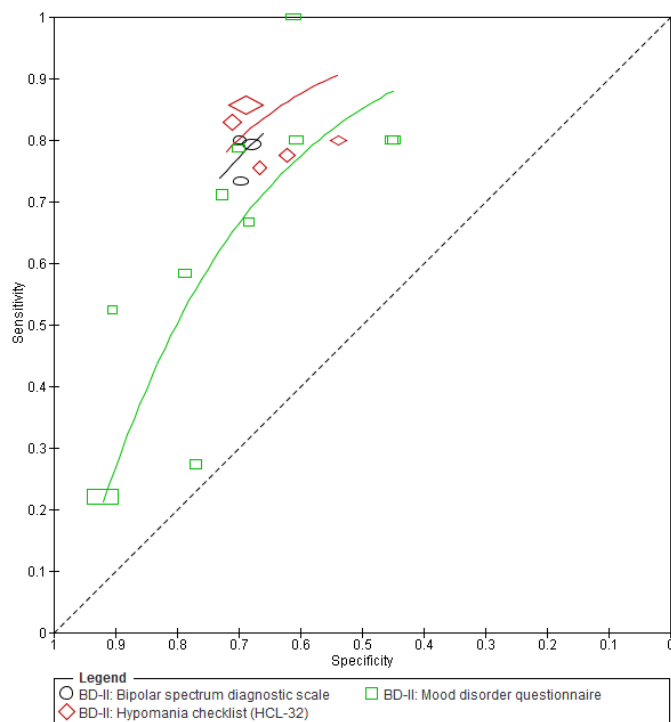


Figure 2.10| Summary ROC plot of the BSDS, HCL-32 and MDQ for detection of bipolar disorder type II in mental health centre settings

For each test, each symbol represents the pair of sensitivity and specificity from a study. The size of the symbols is scaled according to the sample size of the study. Plotted curves are restricted to the range of specificity for each instrument.

(Adapted from Carvalho *et al* 2015⁵⁶)

Models that allow for different assumptions about the shape of SROC curves will be empirically assessed in Chapter 7 to assess the effect of assuming a common underlying shape for the SROC curves of different tests on relative test performance.

2.3.4 Is the assumption of equal variances across tests appropriate?

Hierarchical meta-analytic models include study-specific random effects to account for between-study heterogeneity (see section 1.4.4 and 1.5.4). In a comparative meta-analysis, equality of variance parameters is often assumed for different tests whilst allowing other model parameters to depend on each test (see equations 1.14 and 1.15 for the bivariate model, and equations 1.17 and 1.18 for the HSROC model).²³ This assumption is not always justified. The objective of this example is to demonstrate discrepancy in summary accuracy measures when differences in heterogeneity exist between studies of different tests and equal variances are assumed across tests. In addition, the impact of the approach used to deal with inclusion of comparative studies in the meta-analysis is shown.

2.3.4.1 Exploring equality of variance parameters across tests

Example: Rapid diagnostic tests for uncomplicated *P. falciparum* malaria

Figure 2.11 shows a SROC plot and a plot of D versus S . The scatter of points on both plots suggests that the 65 study cohorts that evaluated Type 1 RDTs are more heterogeneous than the 16 that assessed Type 4 RDTs. A separate meta-analysis of each test also indicated variances may differ between both tests.

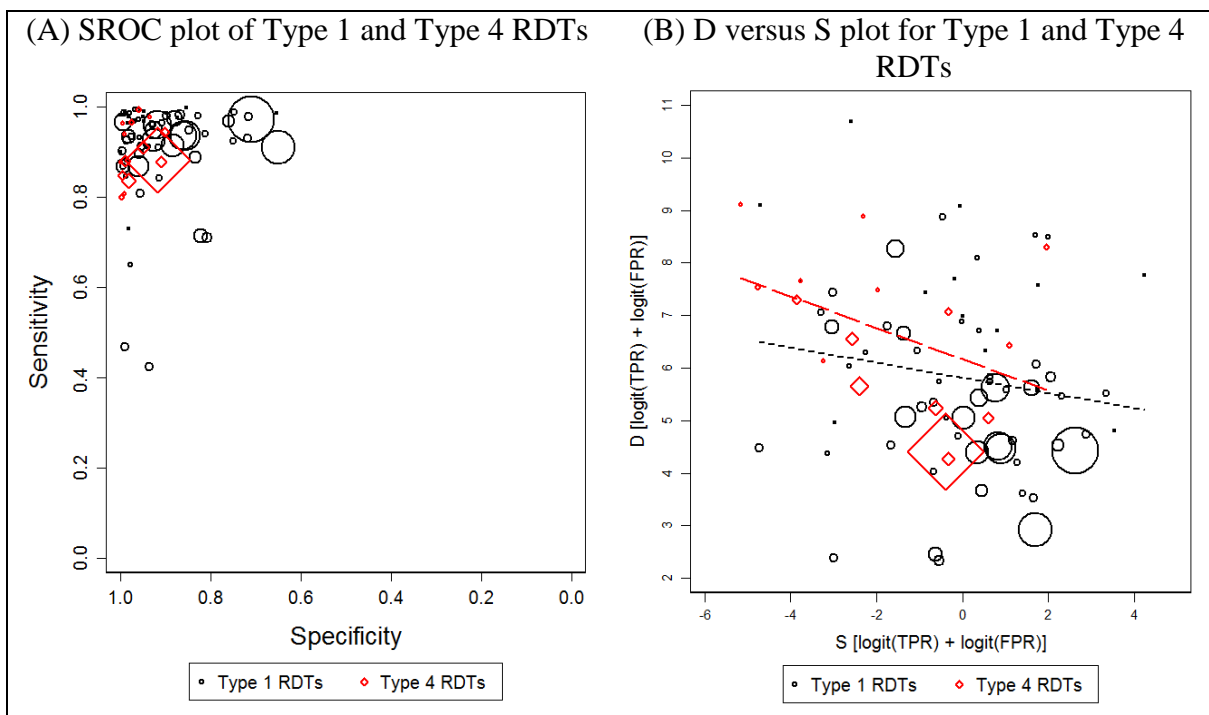


Figure 2.11| Comparison of heterogeneity in test performance for Type 1 and Type 4 rapid diagnostic tests

Type 1 and Type 4 RDTs were compared by including parameters in the HSROC model to allow each RDT type to have a different threshold, accuracy and SROC curve shape.⁹⁵ The impact on the variability of random effects of accuracy and threshold was also investigated. For Model 1, variances of the random effects were assumed equal for the two test types, i.e., no dependence on test type (equation 1.17 in section 1.5.4.2). For the alternative model, Model 2, the variances were allowed to depend on test type (equation 1.20). The two models were compared using likelihood ratio tests. Summary sensitivities and specificities were derived from the models.

The regression equations, distribution of the random effects (equations 1.17, 1.18, 1.20 and 1.21 in section 1.5.4.2), and excerpts of the syntax of the analyses done using SAS Proc NLMIXED are shown in Box 2.1 for the two models. Only the relevant SAS statements where the models differ are shown. See Appendix A.1 for the full SAS program.

Box 2.1| SAS Proc NLMIXED code for each model**Model 1***Regression equation: same shape for the SROC curves of both tests*

$$\text{logit}(\pi_{ij}) = \left((\theta_i + \gamma t_i) + (\alpha_i + \xi t_i) \text{dis}_{ij} \right) \exp(-\beta \text{dis}_{ij})$$

Distribution of the random effects for threshold and accuracy: equal variances for both tests

$$\theta_i \sim N(\Theta, \sigma_\theta^2) \text{ and } \alpha_i \sim N(\Lambda, \sigma_\alpha^2)$$

SAS syntax

```
parms alpha=5 theta=1 beta=1 s2ua=2 s2ut=1 alpha_t4=1 theta_t4=0 covt=0 cova=0;
logitp= ((theta+ut)+(theta_t4)*t4+(alpha+ua)+(alpha_t4)*t4)*dis *exp(-
(beta)*dis);
random ut ua ~ normal([0,0],[s2ut,0,s2ua]) subject=study_id out=randeffects;
```

Model 2*Regression equation: same shape for the SROC curves of both tests*

$$\text{logit}(\pi_{ij}) = \left((\theta_i + \gamma_i t_i) + (\alpha_i + \xi_i t_i) \text{dis}_{ij} \right) \exp(-\beta \text{dis}_{ij})$$

Distribution of the random effects for threshold and accuracy: separate variances for each test

$$\begin{pmatrix} \theta_i \\ \gamma_i \end{pmatrix} \sim N \left(\begin{pmatrix} \Theta \\ \Gamma \end{pmatrix}, \begin{pmatrix} \sigma_\theta^2 & \sigma_\theta \sigma_\gamma \\ \sigma_\theta \sigma_\gamma & \sigma_\gamma^2 \end{pmatrix} \right) \text{ and } \begin{pmatrix} \alpha_i \\ \xi_i \end{pmatrix} \sim N \left(\begin{pmatrix} \Lambda \\ \Xi \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \sigma_\alpha \sigma_\xi \\ \sigma_\alpha \sigma_\xi & \sigma_\xi^2 \end{pmatrix} \right)$$

SAS syntax

```
parms alpha=5 theta=1 beta=1 s2ua=2 s2ut=1 alpha_t4=1 theta_t4=0 s2ua4=1 s2ut4=1
covt=0 cova=0;
logitp= ((theta+ut)+(theta_t4+ut4)*t4 + ((alpha+ua)+(alpha_t4+ua4)*t4)*dis)*exp(-
(beta)*dis);
random ut ua ut4 ua4 ~
normal([0,0,0,0],[s2ut,0,s2ua,covt,0,s2ut4,0,cova,0,s2ua4]) subject=study_id
out=randeffects;
```

Only the relevant statements where the models differ are shown.

In the NLMIXED code above, the *parms* statement defines the model parameters with their starting values, *logitp* specifies the regression equation, and *ut* and *ua* are the random effects for threshold and accuracy respectively. The variable *dis* is the disease indicator which takes the value 0.5 if diseased and -0.5 if non-diseased, and *t4* is a dummy variable which takes the

value 1 for Type 4 and 0 for Type 1. The *random* statement specifies the random effects which are assumed to be independent and normally distributed. The variance parameters for the random effects for accuracy and threshold are $s2ua$ and $s2ut$, respectively. The covariance parameters for the random effects are *covt* and *cova* for accuracy and threshold. NLMIXED assumes that a new realization of the random effects occurs whenever the value of the variable identifying each study (*subject=study_id*) in the dataset changes from the previous one. Therefore, the dataset was sorted by *study_id* and the variable for test type before running NLMIXED. The code for Model 2 includes two additional random effects $ua4$ and $ut4$ in the model equation and random statement, with variance parameters $s2ua4$ and $s2ut4$.

Model 2 had a better fit (-2Log likelihood = 1092.4) than Model 1 (-2Log likelihood = 1152.4), with strong statistical evidence (chi-square = 60.0, 2 degrees of freedom, $P < 0.0001$) of a difference in variances of the random effects between the two test types. The sensitivity of Type 4 RDTs derived from the model with equal variances (Model 1) did not reflect the data, and differed from that of the model with separate variances (Model 2) by 5.5% (Table 2.9). The difference in sensitivity between Type 1 and Type 4 RDTs was statistically significant ($P < 0.001$) in Model 1 but not in Model 2 ($P = 0.20$). Between models, little or no difference in specificities for either test type was observed. This case study shows that assuming common variances across tests may give biased estimates, invalid precision and misleading conclusions. In this case, more complex and appropriate assumptions about the variance parameters led to more conservative differences. Using one of the cohorts of reviews identified in Chapter 3, other examples will be examined in Chapter 7 to determine the feasibility of such complex analyses, and the validity of the common practice of assuming equal variances across tests.

Table 2.9| Comparison of estimates from models based on different comparative approaches and assumptions

	Within-study comparative approach			Between-study comparative approach		
	Model 1 Common shape and equal variances	Model 2 Common shape and separate variances	Model 3 Common shape and equal variances	Model 1 Common shape and equal variances	Model 2 Common shape and separate variances	Model 3 Common shape and equal variances
Variance parameter estimates						
	Accuracy	Threshold	Accuracy	Threshold	Accuracy	Threshold
Type 1 RDTs	2.73	0.88	2.53	0.84	2.51	0.87
Type 4 RDTs	2.73	0.88	1.69	0.67	2.51	0.87
Summary estimates						
	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Type 1 RDTs	95.5 (93.9–96.7)	95.0 (93.1–96.4)	94.8 (93.1–96.1)	95.2 (93.2–96.7)	94.7 (92.9–96.1)	95.1 (93.1–96.6)
Type 4 RDTs	86.0 (81.3–89.7)	98.8 (98.3–99.2)	91.5 (84.7–95.3)	98.7 (96.9–99.5)	92.3 (86.9–95.6)	98.7 (97.1–99.4)
Ratio (95% CI); P-value	0.90 (0.87–0.94); P <0.001	1.04 (1.03–1.05); P <0.001	0.96 (0.91–1.02); P = 0.20	1.04 (1.02–1.06); P <0.001	0.97 (0.93–1.02); P = 0.29	1.04 (1.02–1.06); P <0.001
RDTs = Rapid diagnostic tests						
Sensitivity and specificity are presented as percentages.						

2.3.4.2 Approaches for dealing with comparative studies in a comparative meta-analysis

The approach for handling studies that report the accuracy of more than one test in a comparative meta-analysis may affect the estimation of variance parameters and standard errors. In one approach, coined here as the *between-study comparative approach*, comparative studies are regarded as different studies in the analysis and so test results are analysed between-study at level two of the hierarchical model. This can be written as

$$\begin{pmatrix} \mu_{Aik} \\ \mu_{Bik} \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_A + v_A t_{ik} \\ \mu_B + v_B t_{ik} \end{pmatrix}, \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix} \right) \quad (2.2)$$

where μ_{Aik} and μ_{Bik} are the logit sensitivity and logit specificity for the k th test in the i th study; t_{ik} is the k th test in the i th study; μ_A and μ_B estimate the expected logit sensitivity and logit specificity for the index test used as the reference category, $\mu_A + v_A t_{ik}$ and $\mu_B + v_B t_{ik}$ estimate the expected logit sensitivity and logit specificity of the k th test. The variances are σ_A^2 and σ_B^2 for the logit sensitivities and logit specificities, and σ_{AB}^2 is the covariance between the logits across studies. Each test result from a comparative study is treated as if obtained from a different study and so this approach is not recommended if the number of comparative studies is large because it can lead to inappropriate standard errors.²³

A second approach, coined here as the *within-study comparative approach*, takes each comparative study into account by analysing test results within the study at level two of the hierarchical model. This is the model expressed in equation (1.15), written as

$$\begin{pmatrix} \mu_{Aik} \\ \mu_{Bik} \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_A + v_A t_k \\ \mu_B + v_B t_k \end{pmatrix}, \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix} \right). \quad (2.3)$$

where μ_{Ai} and μ_{Bi} are the logit sensitivity and logit specificity for each test within the i th study and t_k represents the study level covariate for test type. This approach was used for all

the case studies in this thesis. The approach does not account for correlation between test results in studies that used a paired design but rather implies independence or a randomized design. This is a conservative approach. Nevertheless, as previously noted in section 1.5.2.1, studies do not often present the paired results of tests cross classified within the diseased and non-diseased groups.

The effect of a between-study approach on the variance parameters and summary sensitivities and specificities of Type 1 and Type 4 RDTs was investigated. Of the 74 study cohorts (65 for Type 1 and 16 for Type 4) in the meta-analysis, seven were comparative studies. For the within-study approach (Model 1 and Model 2 in the previous section), studies were sorted first by study identifier and then by test type but for the between-study approach, studies were sorted first by test type and then by study identifier. The meta-analyses performed for Model 1 was subsequently repeated to obtain the results shown for Model 3 in Table 2.9. The results from Model 3 were similar to those from Model 1. This finding is likely due to the small number of comparative studies in the meta-analysis. The extent to which results will vary with number of comparative studies is unknown. The impact of adopting a between-study or within-study approach will be investigated further in Chapter 7.

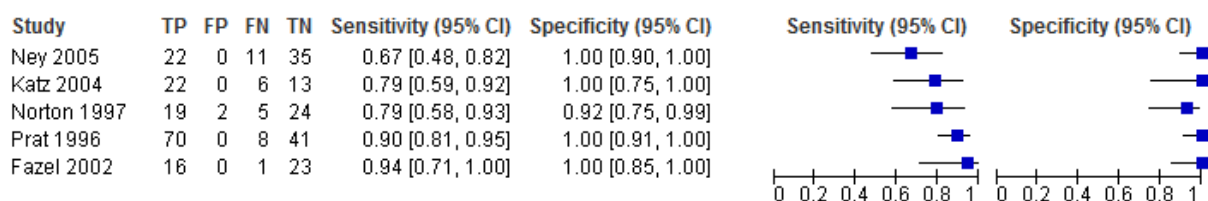
2.3.5 Is comparative meta-analysis feasible with few studies or sparse data?

Model estimation problems, such as unreliable parameter estimates or lack of model convergence, can occur in hierarchical meta-analytic models with small number of studies or sparse data. Sparse data typically arise in meta-analysis of very accurate tests where most of the studies have little or no test errors in the diseased (false negatives) and/or non-diseased (false positives) groups. This may be likened to multivariate meta-analysis of correlated rare

events in intervention reviews. In sparse data situations where a large proportion of studies have 100% sensitivity and/or specificity as shown in Figure 2.12, or when studies are few, the variance-covariance parameters are often on the boundary of the parameter space.^{110,111} The maximum likelihood estimate on the boundary will have at least one of the variances in an HSROC model equal to zero or the correlation parameter in a bivariate model equal to +1 or -1. Model estimation problems will be discussed in more detail later on in the thesis when meta-analysis of few studies and sparse data is investigated in a simulation study in Chapter 8.

It is apparent from all the examples in this chapter that comparative studies tend to be few. Also, one test may have been evaluated less often than others. As such a direct comparison is likely to include few studies, or an indirect comparison may include few studies of one or more of the tests. Regardless of the type of test comparison, there is no recommended minimum number of studies for each test in a comparative meta-analysis. Given the potential number of parameters that can be included in such an analysis, careful consideration should be given to the feasibility of an analysis as well as the degree of model complexity.

Endoscopic retrograde cholangiopancreatography



Intraoperative cholangiography

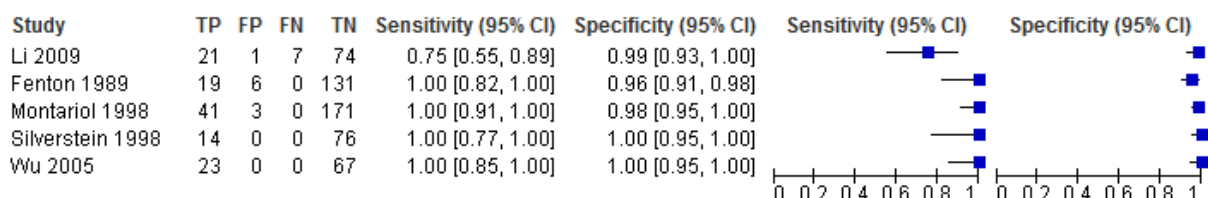


Figure 2.12| Forest plot of endoscopic retrograde cholangiopancreatography (ERCP) and intraoperative cholangiography (IOC) for diagnosis of common bile duct stones
 Studies are ordered by sensitivity and study identifier.
 (Adapted from Gurusamy et al 2015⁷⁰)

2.3.5.1 Simplifying variance-covariance matrices

Example: ERCP versus IOC for common bile duct stones

Five studies assessed ERCP and five assessed IOC. None of the studies were comparative. In preliminary analyses, a separate bivariate model was fitted for each test using the Stata *xtnlogit* command. When the covariance matrix for the random effects for logit sensitivity and logit specificity was unstructured (Model A)—i.e. no constraints imposed so that the variances and covariance were uniquely estimated as in equation 1.8—the variances were poorly estimated, especially the variance parameter for the logit sensitivity of IOC. The forest plot (Figure 2.12) shows that most of the IOC studies had a sensitivity of 100% while most of the ERCP studies had a specificity of 100%. It is unsurprising that the variances were poorly estimated given the small number of studies and sparse data.

Alternative models were investigated as shown in Table 2.10. In Model B, the exchangeable covariance structure estimated a common variance and a covariance for the random effects of the logit sensitivity and logit specificity of each test. The independent covariance structure used in Model C estimated distinct variances for the random effects of the logit sensitivities and logit specificities of both tests but the covariances were assumed to be zero. Model D included only one variance parameter per test; one for the random effects of the logit sensitivity of ERCP and one for the random effects of the logit specificity of IOC. Therefore, no covariances were estimated in Model D.

Table 2.10| Parameter and summary estimates for ERCP and IOC from models with different variance-covariance structure

Test	Logit sensitivity (SE)	Logit specificity (SE)	Variance of random effects for logit sensitivity (SE)	Variance of random effects for logit specificity (SE)	Correlation of the logits (SE)	Sensitivity (95% CI)	Specificity (95% CI)
Model A: Unstructured variance-covariance structure							
ERCP	1.55 (0.30)	5.35 (2.25)	0.22 (0.27)	2.95 (6.34)	0.41 (1.05)	82.5 (72.3–89.5)	99.5 (71.8–100.0)
IOC	7.06 (4.53)	4.15 (0.52)	16.7 (26.9)	0.25 (0.54)	-0.73 (0.98)	99.9 (14.1–100.0)	98.5 (95.8–99.4)
Model B: Exchangeable variance-covariance structure							
ERCP	1.56 (0.33)	4.33 (0.77)	0.26 (0.32)	0.26 (0.32)	0.55 (1.62)	82.6 (71.9–89.8)	98.7 (94.4–99.7)
IOC	4.43 (1.45)	4.80 (1.00)	2.85 (2.71)	2.85 (2.71)	0.03 (0.76)	98.8 (83.1–99.9)	99.2 (94.5–99.9)
Model C: Independent variance-covariance structure							
ERCP	1.56 (0.30)	5.32 (2.19)	0.22 (0.26)	2.83 (5.87)	0	82.6 (72.6–89.5)	99.5 (73.8–100.0)
IOC	6.12 (3.28)	4.19 (0.57)	9.74 (12.7)	0.34 (0.67)	0	99.8 (42.3–100.0)	98.5 (95.6–99.5)
Model D: Fixed effect for sensitivity of IOC and for specificity of ERCP							
ERCP	1.56 (0.30)	4.22 (0.71)	0.22 (0.26)	0	0	82.6 (72.6–89.5)	98.6 (94.4–99.6)
IOC	2.82 (0.39)	4.19 (0.57)	0	0.34 (0.67)	0	94.4 (88.7–97.3)	98.5 (95.6–99.5)

ERCP = endoscopic retrograde cholangiopancreatography; IOC = intraoperative cholangiography; SE = standard error. Sensitivity and specificity are presented as percentages.

An indirect comparison of ERCP and IOC was subsequently performed using a bivariate meta-regression model. Based on results of the preliminary analyses shown in Table 2.10, an exchangeable covariance structure was assumed for the variances of the random effects and the variance-covariance matrix was allowed to depend on test type. This implies that the variance of the random effects for logit sensitivity is the same as that of the logit specificity of each test. Likelihood ratio tests were used to compare the fit of different models. To check the robustness of the assumptions about the variances of the random effects, the estimates of sensitivity and specificity were also compared between models. The Stata program for fitting all the models is included in Appendix A.2.

The analytical approach adopted here was based on the reasoning that it is inappropriate to overfit models by estimating too many parameters from few studies, and to simplify models when parameter estimates cannot be reliably estimated. The importance of choosing an appropriate model will be demonstrated through a simulation study in Chapter 8. The simulation study addresses meta-analysis of a single test. However, the results will be generalised to test comparisons, informed by results in Chapter 7 from empirical evaluations of test comparisons with few studies or sparse data due to highly accurate tests.

2.4 Summary

The seven reviews highlighted the complexity of systematic reviews and meta-analyses of test comparisons. Review complexity increases with increasing number of tests, target conditions, and/or target populations within a single review. Therefore a strategy is needed for structuring the analyses and presenting the review to enable clarity for readers. Although the reviews

addressed different target conditions and test types, a common issue was the scarcity of well-designed comparative studies. Therefore an indirect comparison was the main analysis in the six reviews where a comparative meta-analysis was performed. In the remaining review comparing ultrasound and LFTs, only a meta-analysis of ultrasound was possible. Using the cohorts of reviews identified in the next chapter, these and other issues will be investigated further in chapters 4 and 5.

Only hierarchical (bivariate or HSROC) meta-regression models were used for meta-analyses in all examples. The models that were applied varied in complexity from models that assumed a simpler variance-covariance structure to a more complex unstructured one that was also allowed to depend on test type. The importance of allowing for asymmetry in the SROC curves, and the need for methods that enable estimation of summary points and/or curves was also demonstrated. It is clear that the available data is likely to drive the choice of a meta-analytic method. This makes pre-specification of a method in a systematic review protocol challenging if little is known about the clinical question and literature. The nuances of this meta-regression approach will be examined in Chapter 7.

The use of a within-study or between-study approach for including comparative studies in a meta-analysis was also highlighted. This is a potentially important but little known issue. Where applicable, the impact of the approaches will be evaluated in Chapter 7 using meta-analytic models that do not directly account for comparative data, and the findings will be compared with those of models that explicitly account for comparative data if such models are identified from the searches reported in Chapter 3. The performance of all the methods identified will be empirically assessed using a cohort of reviews.

Sparse data due to frequent zero cells in 2x2 tables and small numbers of studies are not limited to meta-analysis of a single test as shown by the review of ERCP versus IOC.

Simplifying hierarchical models seemed to be a reasonable approach to avoid estimating too many model parameters from very little data. Although the simulation study in Chapter 8 focuses on meta-analyses of a single test, the findings may be applicable to comparative meta-analyses in certain situations and will be discussed in the chapter.

The challenging issues discussed in this chapter were discovered while doing the reviews, and solutions were developed based on theoretical and pragmatic reasoning due to lack of evidence based guidance. The analytical approaches used in completing the reviews were based on what were thought to be statistical best practice. Therefore, the rest of the thesis will seek to understand whether the issues are common and to identify approaches other researchers have used (Chapter 4) and available comparative meta-analysis methods (Chapter 6), and through empirical studies (Chapters 5 and 7) and simulation (Chapter 8) contribute to the evidence base to support reviewers and meta-analysts tackling these and similar problems in the future.

3 IDENTIFYING SYSTEMATIC REVIEWS AND META-ANALYTIC METHODS FOR TEST COMPARISONS

3.1 Introduction

The aims of this chapter are to identify systematic reviews that evaluated at least two index tests and articles which describe meta-analytic methods for comparing test accuracy. The reviews are used in Chapter 4 to provide an overview of data synthesis methods and reporting of test comparisons in published reviews, and in Chapter 5 to determine if discrepancies exist between meta-analyses of direct and indirect comparisons. The meta-analytic methods identified in this chapter are described in Chapter 6, and their performance is empirically assessed in Chapter 7 using a subset of the reviews.

In section 3.2, terminology for different types of test accuracy reviews is defined, and the search strategy and selection process for identification of the cohort of reviews for the thesis is described. Section 3.3 details the methods for identification of meta-analytic methods for test comparisons. The results of both searches are presented in sections 3.4 and 3.5 for the cohort of reviews and meta-analytic methods, respectively. The final section, section 3.6, summarises the results and the strengths and limitations of the searches.

3.2 Identification of systematic reviews

To provide empirical data and to find suitable motivating datasets for Chapters 4, 5 and 7 of this thesis, it was necessary to search for systematic reviews that assessed the accuracy of two or more tests. Comparative studies included in these reviews also need to be identified in order to determine their availability and thus gain an appreciation of the evidence base for test

comparisons. The availability of comparative studies is considered in Chapter 5 and so the identification of these studies from a review cohort will be deferred to that chapter.

3.2.1 Terminology

There is no standard terminology for different types of diagnostic test accuracy systematic reviews. The term comparative accuracy review is used in this thesis to describe a review that met at least one of the following four criteria: (1) clear objective to compare the accuracy of at least two tests; (2) selected only comparative studies; (3) performed statistical analyses comparing the accuracy of all or at least a pair of tests; or (4) performed a direct (head-to-head) comparison of two tests. Reviews that assessed two or more tests but did not meet any of the four criteria were termed a multiple test review. Multiple test reviews have a wide focus, aiming to summarise the accuracy of different tests for the same target condition. Such reviews assess each test individually without making formal comparisons between tests and often involve a large number of tests such as signs and symptoms from clinical examination. An example is a review on tests that can be performed or are accessible in primary care for diagnosis of inflammatory bowel disease.¹¹² The review included 24 studies on 50 tests. The tests included (1) signs and symptoms (including alarm symptoms), individual or in combination (including symptom-based classification systems); (2) blood and faecal tests; and (3) abdominal ultrasonography. Multiple test reviews may be likened to reviews of interventions that include multiple pairwise comparisons without a network meta-analysis to formally compare and rank the effectiveness of the interventions. In Chapter 4, characteristics will be presented separately for comparative and multiple test reviews.

3.2.2 Data sources and searches

Systematic reviews of test accuracy were identified in the Database of Abstracts of Reviews of Effects (DARE) and the Cochrane Database of Systematic Reviews (CDSR issue 11, 2012). DARE is regarded as the most comprehensive source of systematic reviews and was produced by the Centre for Reviews and Dissemination (CRD) between 1994 and March 2015. DARE is based on extensive searches of a wide range of databases and grey literature, and contains over 13,000 critically appraised abstracts.¹¹³ Reviews undergo quality appraisal before inclusion in DARE, and they must meet the first 3 criteria and at least 4 criteria in total from the following list:

1. Were inclusion/exclusion criteria reported?
2. Was the search adequate?
3. Were the included studies synthesised?
4. Was the quality of the included studies assessed?
5. Are sufficient details about the individual included studies presented?

Reviews published between 1994 and 2002 were identified from previous projects.^{94,114} The CRD's in-house content management system (CMS) was used to identify reviews published between January 2003 and October 2012 that had a structured abstract. A CRD database production manager performed the search in the CMS by searching the fields for record type (diagnostic review or not), publication status (fully published with a structured abstract or provisionally published without an abstract), and record date (1st January 2003 to 31st October 2012). The search results were provided in a tagged text file for import into an EndNote (reference management software) library. An EndNote filter for DARE records was also provided by the CRD to enable the import.

Although DARE contains details of all Cochrane reviews, structured abstracts were not prepared for Cochrane reviews by the CRD hence the need to search the CDSR separately. The CDSR is a limited resource because the Cochrane Library only began publishing Cochrane DTA reviews in 2008.³² At the time of the search in October 2012, the CDSR contained eight Cochrane DTA reviews and so the records were easily identified. The records obtained from the CDSR were exported as a text file and imported into the same EndNote library containing the DARE results. All searches were performed without restrictions on language of publication, test type, purpose of the test (for example, screening, staging, diagnosis, etc.), setting, or disease area.

3.2.3 Selection of reviews

Abstracts were screened to identify potentially relevant reviews before retrieval of full text articles. Reviews were selected for inclusion in stages according to the objectives of the thesis. For the first stage, the screening form in Appendix B.1 was used to select reviews for inclusion in the overall cohort for the thesis. The following criteria were applied:

1. Evaluated the diagnostic accuracy of at least two tests
2. Included at least one meta-analysis
3. Full text article could be retrieved
4. Data available on studies included for the evaluation of each test to enable assessment of study design

In the next stage, additional criteria were applied to select reviews for the review of reviews presented in Chapter 4 and for each of the two empirical projects considered in chapters 5 and 7. The eligibility criteria for each cohort are reported separately below. For each cohort, a

random subset of half the reviews judged to be eligible was double checked by a second researcher to confirm eligibility.

3.2.3.1 Eligibility criteria for review cohort for chapter 4

The focus of Chapter 4 is to summarise the characteristics of statistical methods and their reporting, including the presentation of test accuracy estimates from the included studies and pooled estimates from meta-analyses. Key advances in methodology for DTA reviews (including literature searching, quality assessment and meta-analysis) were published between 1993 and 2005.³² For this reason, and to make allowance for dissemination of methods, selection was limited to a five-year period from January 2008 to October 2012. Thus systematic reviews that were likely to reflect current practice were considered in the review of reviews in Chapter 4.

3.2.3.2 Eligibility criteria for review cohort for chapter 5

Chapter 5 focuses on the assessment of the availability of comparative studies and investigation of differences between meta-analyses of direct and indirect comparisons. Where there were multiple reviews of the same tests for the same target condition in the same population, only the most recent review was included in the cohort of reviews. This was done to avoid double counting comparative studies and to limit overlap of similar pairs of meta-analyses in the comparison of meta-analyses of non-comparative studies with those of comparative studies. Where more than two tests were evaluated in a review, each possible paired comparison of tests was considered separately.

3.2.3.3 Eligibility criteria for review cohort for chapter 7

The focus of Chapter 7 is on the empirical assessment of meta-analytic methods for test comparisons. Only reviews that evaluated two tests were considered. Pairwise comparisons can also be made using reviews with more than two tests but the restriction was imposed to make the project manageable. In addition to the number of tests, the reviews had to provide 2x2 data for each test evaluated in the included studies or data that enabled their derivation.

3.3 Identification of methods for meta-analysis of comparative accuracy

3.3.1 Data sources and searches

Diagnostic research literature is poorly indexed and difficult to locate. Therefore searches of electronic bibliographic databases such as MEDLINE and EMBASE were considered an inefficient way to identify papers on methods for comparative meta-analysis. Instead, two systematic review methodology databases—the Cochrane Methodology Register (CMR) and the US Agency for Healthcare Research and Quality (AHRQ) Effective Healthcare Program’s Scientific Resource Center (SRC) Methods Library—were searched.

Both the CMR and SRC databases contain published and unpublished (conference abstracts) literature collated from systematic literature searches using electronic sources and hand searching. In addition, the CMR includes book chapters and reports of ongoing methodological research.¹¹⁵ The CMR is one of the databases available in The Cochrane Library. However, the database is no longer being updated; the last submission was in July 2012 with data from 1985 to March 2012.¹¹⁶ The CMR is a comprehensive database that predates the SRC Methods Library, and has an archive of 15,764 records.¹¹⁷ The rationale for considering unpublished literature was so that authors could be contacted if an abstract was

found to be potentially relevant. The CMR contains only articles related to systematic review methodology while in addition the SRC Methods Library contains articles related to comparative effectiveness research methodology. Searching the SRC Methods Library as an adjunct to the CMR enabled identification of articles published since the last update of the CMR and articles missed in the CMR.

The CMR was searched on 24 July 2014 via the Cochrane Library 2014 (Issue 7). Following advice from an information specialist who contributed to the production of the CMR, the 'Advanced Search' functionality on the Cochrane Library was used with the search limited to 'Keywords'. The combination of the keywords "diagnostic test accuracy" AND "meta-analysis" was used to identify records that had been coded specifically as methods papers on diagnostic test accuracy meta-analysis. The list of results from the CMR was obtained by selecting the 'Methods Studies' option. Using the terms "diagnostic test" and "meta-analysis", the SRC Methods Library was searched in-house by an SRC research librarian on 31 July 2014. The search results were provided in a .ris file for upload into an EndNote library bibliographic database.

To augment the database searches, methodological experts and research groups known to have an interest in meta-analytic methods for test accuracy were contacted regarding ongoing work or papers in press. A similar approach had been used successfully to identify additional literature for a methodology review.¹¹⁴ A database of published and unpublished methodological studies was thus assembled from a variety of sources. Additionally, the methods section of systematic reviews identified in section 3.5 for the cohort of reviews

assessed in Chapter 4 will be examined for novel methods or modifications of existing methods.

3.3.2 Selection of methodology studies

After removal of duplicates, titles and abstracts identified by the search were screened to identify potentially relevant studies. Papers were initially selected for inclusion if they proposed a novel method for meta-analysis of diagnostic accuracy. Papers that extended, evaluated or explained the properties of an existing method were also considered. Following this initial assessment, full text reports were retrieved for further assessment of eligibility. In this second and final stage, only papers that considered a method for comparative meta-analysis were included. The selection process was not verified by a second researcher because it was obvious whether or not a paper described a meta-analytic method for test comparisons.

3.4 Search results for systematic reviews

The flow of reviews through the screening process and reasons for exclusion are shown in Figure 3.1. The searches identified 1023 reviews, of which 914 evaluated test accuracy and 466 (51%) included results for two or more tests. Of these, 286 met the main inclusion criteria. Due to the search period for the reviews, only one of the reviews⁹⁵ described in Chapter 2 was part of this cohort. For the Chapter 4 cohort, 130 reviews published between 2008 and 2012 were identified. For 38 of the 286 reviews, there was a more recent review on the same review question and so the cohort for Chapter 5 included 248 reviews. In 101 of the 286 reviews, only two tests were evaluated. These reviews formed the cohort for Chapter 7. Additional selection criteria and the characteristics of the final cohort used for each assessment are described within the relevant chapter.

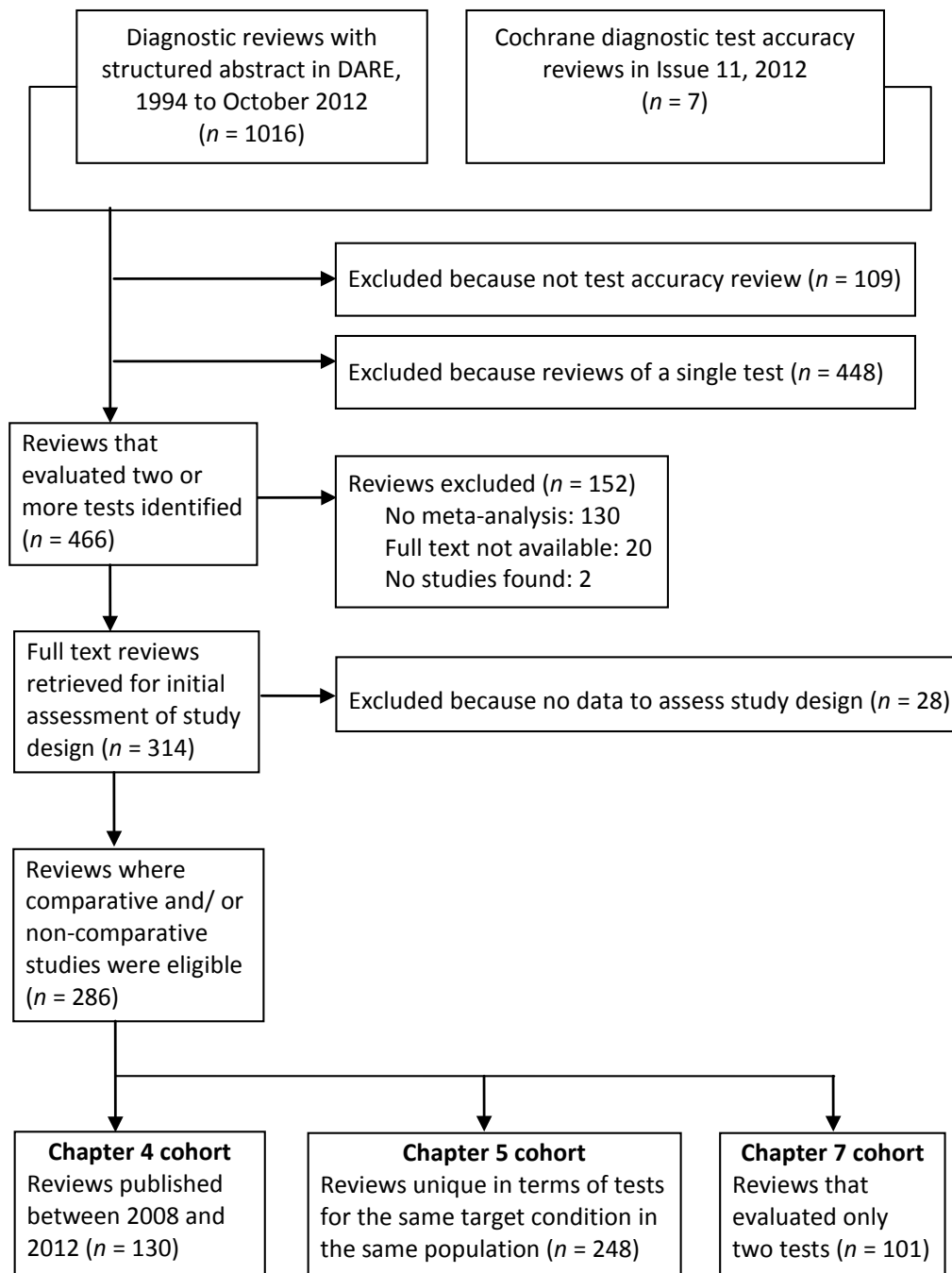


Figure 3.1| Flowchart of selection of systematic reviews

Several reviews were eligible for inclusion in more than one cohort and so the three cohorts do not sum up to 286.

(Adapted from Takwoingi et al 2013⁷¹)

3.5 Search results for meta-analytic methods

The flow of articles through the screening process and reasons for exclusion are shown in Figure 3.2. The searches identified 395 unique titles and abstracts, of which 48 appeared to be about meta-analytic methods for test accuracy. Of the 48 papers, 11 presented a method for comparative meta-analysis. Three of the 11 papers explained or adapted an existing method while the remaining eight were original papers. Contact with methodological experts at the Universities of Amsterdam, Sydney, Düsseldorf, Minnesota, and Brown University yielded two conference papers, a PhD thesis and one manuscript under peer review. The manuscript was excluded because it was shared confidentially, and was yet to be published as at 22 July 2015 when data extraction began. Two relevant papers were identified from the thesis. Both papers were systematic reviews with a detailed account of the novel approach that was used for comparative meta-analysis. Altogether, 13 papers and two presentations describing 13 methods were included.

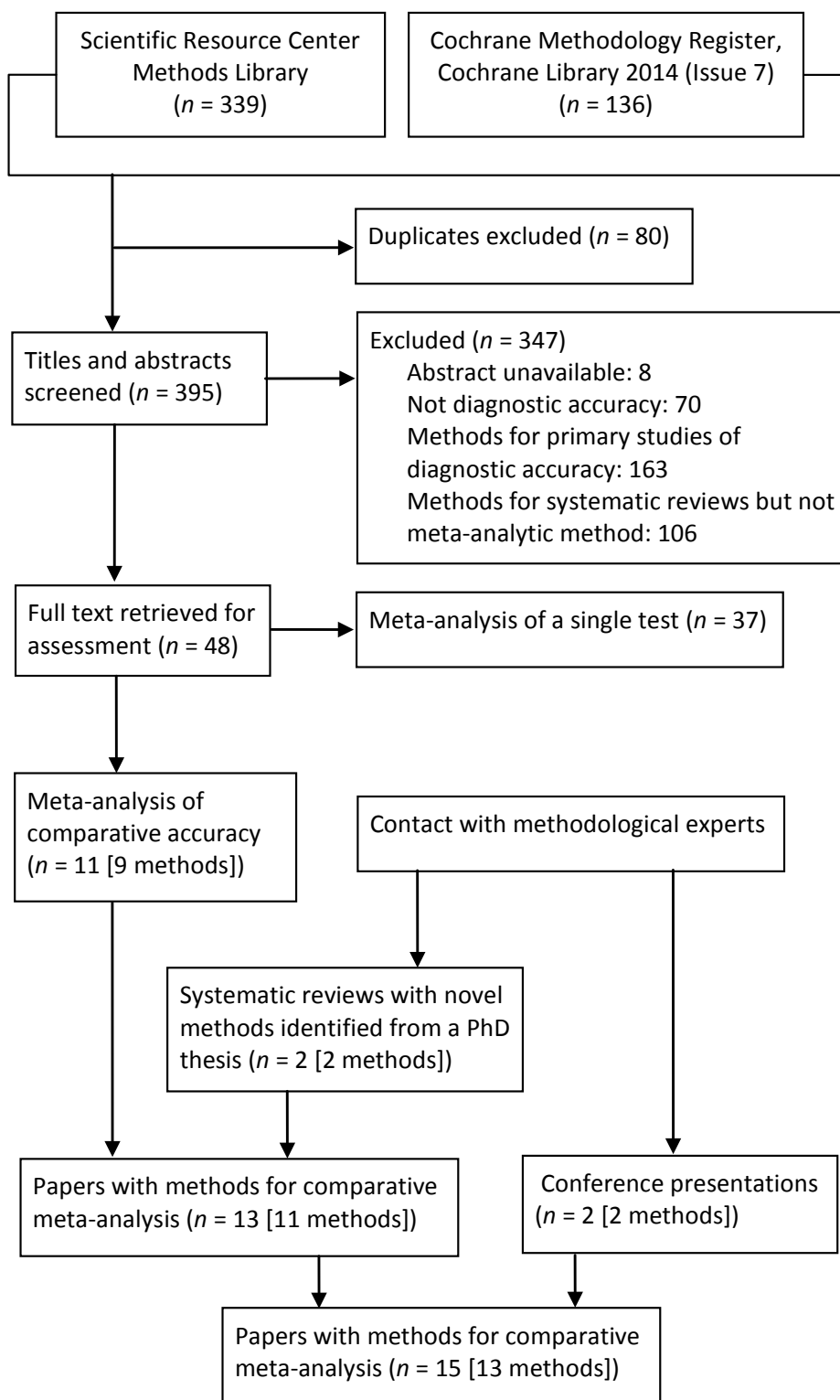


Figure 3.2| Flowchart of selection of reports of meta-analytic methods

3.6 Discussion

3.6.1 Summary of findings

The search identified 466 systematic reviews that evaluated at least two tests. Only two reviews were empty, i.e., no studies were included and a meta-analysis was not included in many reviews (130/466, 28%). Altogether, 269 of the 286 reviews that met the main inclusion criteria were included across the three cohorts for Chapters 4, 5 and 7. The reviews will be used to provide an overview of current practice in Chapter 4, and for empirical evaluations in chapters 5 and 7. The characteristics of the reviews in each cohort will be presented in each chapter.

Nine meta-analytic methods for test comparisons were identified from the 395 titles and abstracts screened during the search for methods papers. These methods as well as the four identified from other sources will be examined in detail in Chapter 6.

3.6.2 Strengths and limitations

Using DARE facilitated the identification of a large number of DTA reviews across a wide range of clinical topics and test types. Some reviews may have been missed but the number of reviews identified is likely to ensure qualitative saturation of the various issues and current standards that will be explored in later chapters. Identifying method papers is not trivial because of the lack of standard terminology. Using the CMR and the SRC Methods Library enabled a focused and resource-efficient search. Contacting methodological experts proved useful in identifying ongoing work as well as any relevant published work that may have been missed, thus providing reassurance that the search was comprehensive and no relevant paper was missed. Unlike the systematic reviews, the method papers were not double checked

because it was obvious that excluded papers described methods for meta-analysis of a single test.

3.6.3 Conclusions

The large number of included reviews will facilitate a detailed examination of the issues addressed in the thesis, including the empirical evaluation of the performance of the meta-analytic methods.

4 REVIEW OF PUBLISHED SYSTEMATIC REVIEWS OF COMPARATIVE TEST ACCURACY

4.1 Introduction

Test evaluation is often limited to the assessment of test accuracy. Therefore, it is vital that in the rapidly expanding evidence base, systematic reviews and meta-analyses that compare the accuracy of two or more tests are conducted appropriately to avoid misleading conclusions and recommendations. Furthermore, it is essential that the reviews are well reported to facilitate transparency and credibility. To the author's knowledge, a comprehensive overview of data synthesis methods and presentation of results in reviews of comparative accuracy has not been undertaken.

Using reviews identified in Chapter 3, the overarching aims of this chapter is to assess current usage of data synthesis methods in reviews that evaluated at least two tests; to examine reporting characteristics of the methods and findings in order to highlight deficiencies and good practice; and to propose recommendations for improving the reporting of future reviews. Specifically, this survey of recent reviews aimed to address the following research questions:

1. What summary measures are used in DTA systematic reviews? How often are alternatives to sensitivity/specificity used, and what are the alternatives?
2. Do DTA reviews plan to use comparative studies when they are available? What strategies are used to support best use of available evidence?
3. Do reviews use statistical methods which account for correlated bivariate data, and both within- and between-study heterogeneity? If not, what methods are used and what limitations do they have?

4. Do reviews use statistical methods that estimate differences between tests and evaluate statistical significance? If not what methods are used?
5. Do reviews investigate heterogeneity?
6. Do reviews report:
 - a. Objectives for a comparative question;
 - b. Types of studies used for test comparisons ;
 - c. Strategy used for comparing tests;
 - d. Statistical methods;
 - e. Study characteristics;
 - f. Study findings;
 - g. Limitations of using indirect comparisons?

Several guidelines exist for the conduct of systematic reviews of test accuracy including the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy.¹¹⁸ In contrast, there is no specific guideline for *reporting* DTA reviews and meta-analyses. While the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist designed for systematic reviews of interventions¹¹⁹ can be used, it is inadequate for DTA reviews, especially those that compare test accuracy given the complexities illustrated in Chapter 2. Extensions to the PRISMA statement have been developed or suggested for other types of reviews and meta-analyses such as individual patient data (PRISMA-IPD),¹²⁰ adverse events (PRISMA harms),^{121,122} overviews,¹²³ and network meta-analysis.^{93,124} Similarly, an extension for DTA reviews (PRISMA-DTA) needs to be developed.

The chapter is structured as follows. In section 4.2, the methods for data extraction from the cohort of reviews identified in section 3.4 are described. The section also describes the methods for data analysis. The methodological and reporting characteristics of the reviews are summarised in section 4.3. In section 4.4 the findings are discussed and recommendations for reporting the statistical methods and results of comparative accuracy reviews are suggested.

4.2 Methods

4.2.1 Review selection and data extraction

Reviews published between 2008 and 2012 were selected from the overall cohort identified in section 3.4 (see also Figure 3.1). Full text reports and supplementary files of eligible reviews were read in full. Test comparisons were considered irrespective of whether or not a meta-analytic model was applied to the comparison to estimate differences in test performance. A comparative meta-analysis was considered to have been feasible in a review if there were at least five studies for at least two of the tests evaluated in the review. This choice is informed by the fact that there are five parameters in the hierarchical models recommended for meta-analysis of a single test (see section 1.4.4)⁴⁴ and convergence problems can occur when few studies are available.

To address the questions posed in the introduction (section 4.1), data were extracted from each review in the cohort using the form in Appendix B.2. Broadly, information on general characteristics, statistical methods and reporting were collected as outlined in Table 4.1. Data extraction of a random subset of half of the reviews was done by a second assessor for data checking.

Table 4.1| Summary of information extracted from review cohort

Item	Description
<i>General characteristics*</i>	
Target condition	The disease or condition of interest. A review may include multiple target conditions or different sub types.
Tests evaluated	Number and type of tests evaluated
Publication details	Type of publication and year of publication
<i>Statistical methods</i>	
Summary statistics	Measures of test performance such as sensitivity and specificity, likelihood ratios, DOR, AUC, etc. Several summary statistics can be used to quantify test accuracy as previously summarised in section 1.3.2. These statistics may be single or paired measures. For the purpose of this study, a pair such as sensitivity and specificity was considered a single measure if one or both were reported.
Test comparison strategy	Test comparisons (direct or indirect) were examined irrespective of whether or not a meta-analytic model was applied to the comparison.
Meta-analytical methods	Methods used for meta-analysis. This includes methods that may have been used for the analysis of each test separately as well as methods used to compare test accuracy (comparative meta-analysis methods). Also includes methods for investigating variation in test performance between studies.
<i>Reporting</i>	
Role of test(s)	Proposed role of the test(s) within a diagnostic pathway as replacement, triage or add on (see section 1.5.1).
Study characteristics	Reported in text or in tables
Study type (comparative or non-comparative)	Reported in text or flow diagrams. The design of the studies in a review is important for assessing the strength of the evidence
Limitations of indirect comparisons	Only applicable for reviews that include indirect comparisons. Caveats about the quality and strength of the evidence may be given in the discussion and conclusions of reviews.

*Since these reviews are mainly a subset of a larger cohort that will be described in Chapter 5, general characteristics will not be presented in this chapter to avoid duplication.

4.2.2 Data analysis

Since this is a qualitative review of current analysis and reporting characteristics, descriptive statistics were computed. Categorical variables were summarised using frequencies and percentages, and continuous variables were summarised using the median, range and

interquartile range. Reviews were broadly categorised into two types based on the definition described in section 3.2.1—comparative (a comparative objective was explicitly stated or inferred because only comparative studies were included or direct comparisons were done) and multiple test reviews. Multiple test reviews lack a comparative objective and assess each test individually without formally comparing tests. Comparative reviews were subdivided into two groups—those with and without statistical analyses for comparing test accuracy. As such all the reviews were categorised into three groups. All data analyses and graphs of the distribution of continuous variable were done using STATA SE version 13.0 (Stata-Corp, College Station, Texas, USA).

4.3 Results

Of the 286 reviews in the overall cohort for the thesis, 130 were published between 2008 and 2012 (Figure 3.1). It became apparent during review of the full text and data extraction that three reviews were ineligible and so they were excluded. One of the three reviews compared populations rather than tests, and the other two reviews combined data from both tests together as a single test in the meta-analysis. Therefore, 127 reviews were included in this study. Based on the terminology defined in section 3.2.1, there were 80 comparative reviews and 47 multiple test reviews (Figure 4.1). Of the 80 comparative reviews, 53 (66%) formally compared test accuracy.

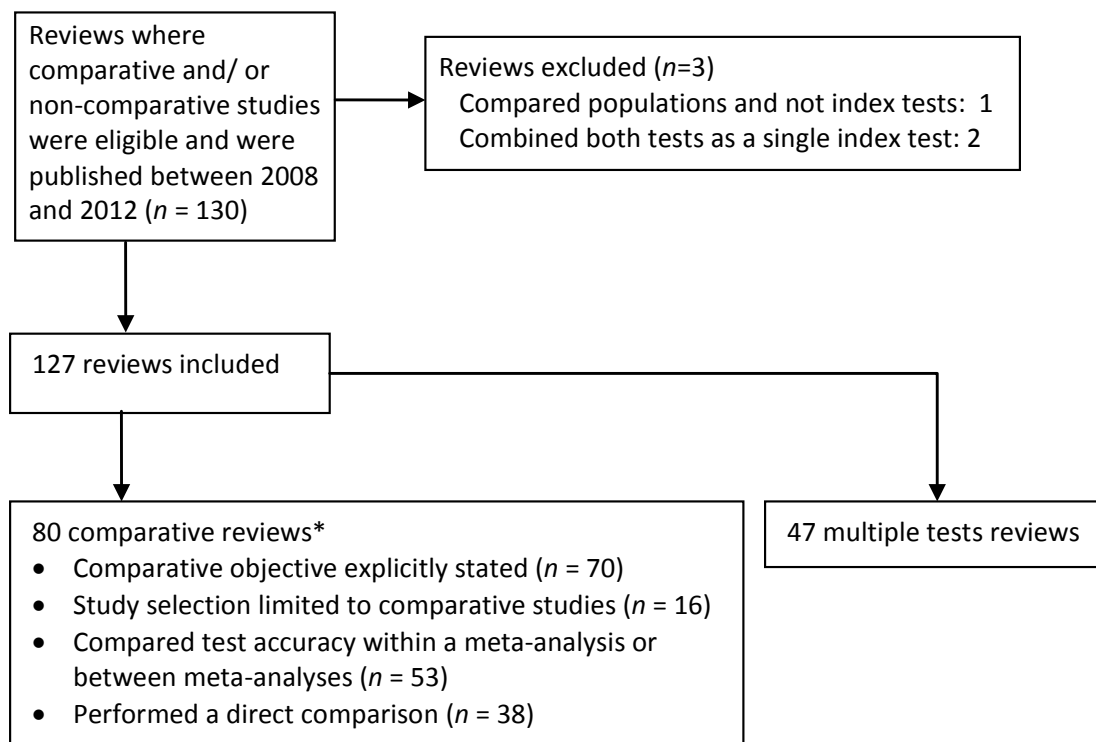


Figure 4.1| Cohort of reviews included in the review of reviews

*The 80 comparative accuracy reviews met at least one of the following four criteria: (1) clear objective to compare the accuracy of at least two tests; (2) selected only comparative studies; (3) performed statistical analyses comparing the accuracy of all or at least a pair of tests; or (4) performed a direct (head-to-head) comparison of two tests.

4.3.1 General characteristics

The 127 reviews were published in 74 different journals. Most of the reviews (93/127, 73%) were in specialist medical journals while there were 23 (18%) in general medical journals, six (5%) technology assessment reports and five (4%) Cochrane reviews. Characteristics of the 127 reviews are briefly summarised in Table 4.2. Over half of the reviews (54%) assessed the accuracy of two or three tests.

Table 4.2| Descriptive characteristics of 127 reviews of comparative accuracy and multiple tests

Characteristic	Comparative reviews		Multiple test reviews	Total
	Statistical analyses to compare test accuracy			
	Yes	No or unclear ¹		
Number of reviews ²	53 (42)	29 (23)	45 (35)	127 (100)
Year of publication				
2008	14 (26)	11 (38)	13 (29)	38 (30)
2009	6 (11)	10 (34)	8 (18)	24 (19)
2010	16 (30)	4 (14)	11 (24)	31 (24)
2011	13 (25)	3 (10)	7 (16)	23 (18)
2012 ³	4 (8)	1 (3)	6 (13)	11 (9)
Number of tests evaluated				
2	20 (38)	14 (48)	12 (27)	46 (36)
3	12 (23)	6 (21)	4 (9)	22 (17)
4	8 (15)	3 (10)	4 (9)	15 (12)
≥5	13 (25)	6 (21)	25 (56)	44 (35)
Clinical topic (according to ICD-10 Version: 2015)				
Circulatory system	9 (17)	5 (17)	5 (11)	19 (15)
Digestive system	3 (6)	1 (3)	8 (18)	12 (9)
Infectious and parasitic diseases	3 (6)	4 (14)	9 (20)	16 (13)
Injury, poisoning and certain other consequences of external causes	2 (4)	1 (3)	2 (4)	5 (4)
Mental and behavioural disorders	2 (4)	1 (3)	3 (7)	6 (5)
Musculoskeletal system and connective tissue	1 (2)	1 (3)	4 (9)	6 (5)
Neoplasms	28 (53)	12 (41)	7 (16)	47 (37)
Other ICD-10 codes ⁴	5 (9)	4 (14)	7 (16)	16 (13)
Type of tests evaluated				
Biopsy	0	1 (3)	0	1 (1)
Clinical and physical examination	5 (9)	3 (10)	15 (33)	23 (18)
Device	1 (2)	0	0	1 (1)
Imaging	32 (60)	13 (45)	9 (20)	54 (43)
Laboratory	8 (15)	8 (28)	12 (27)	28 (22)
RDT or POCT	1 (2)	0	4 (9)	5 (4)
Self-administered questionnaire	1 (2)	1 (3)	0	2 (2)
Combinations of any of the above ⁵	5 (9)	3 (10)	5 (11)	13 (10)

ICD-10 = International Classification of Diseases, Tenth Revision; RDT = Rapid diagnostic test; POCT = Point of care test.

¹In 3 reviews, it was unclear whether a statistical comparison of test accuracy was done.

²Numbers in parentheses are row percentages.

³Includes only studies published up to October 2012.

⁴Includes 8 ICD-10 codes with fewer than 5 reviews across the three groups.

⁵Tests evaluated in a review were not of the same type.

Numbers in parentheses are column percentages unless otherwise stated. Percentages may not add up to 100% because of rounding.

The reviews covered a broad array of target conditions (categorised under 15 different ICD-10 codes, see Table 4.2) and test types (including clinical signs and symptoms). Neoplasms were the most often assessed target condition (37%), and imaging tests were the most frequently assessed type of test (43%). Primary studies included in a review often evaluated only one of the tests of interest for that review (i.e. non-comparative). The median (interquartile range) number of comparative and non-comparative studies included per review were 6 (3 to 11) and 14 (3 to 24), respectively.

4.3.2 Statistical characteristics

4.3.2.1 Summary measures of test performance

Table 4.3 shows that most reviews (75%) used two or more different measures with sensitivity and specificity being the most commonly used metric (93%). Other measures such as the DOR (43%) and likelihood ratios (52%) are frequently used unlike predictive values (11%). The area under the curve or Q* were used to quantify the SROC curve in 34 (27%) and 19 (15%) reviews respectively; both measures were used in 12 reviews. Amongst the 53 reviews that formally compared test accuracy, 18 (34%) reviews used relative measures to summarise differences in accuracy.

4.3.2.2 Test comparison strategy and use of comparative studies

Sixteen (13%) reviews restricted study selection and test comparisons to comparative studies while the other 111 (87%) reviews included any study type (Table 4.3). A qualitative or quantitative test comparison was not done in 40 (32%) reviews; these were all multiple test reviews. In 22 reviews (17%), both direct and indirect comparisons were performed with the direct comparisons performed as secondary analyses using pairs of tests for which data were

available. Direct comparisons were not performed in 49 (39%) reviews even though comparative studies were available in 40 of the reviews and qualitative or quantitative syntheses would have been possible.

Table 4.3| Outcome measures and test comparison strategy in the reviews

Characteristic	Comparative reviews		Multiple test reviews	Total
	Statistical analyses to compare test accuracy			
	Yes	No or unclear		
Number of reviews*	53 (42)	29 (23)	45 (35)	127 (100)
Number of summary measures used				
1	12 (23)	9 (31)	11 (24)	32 (25)
2	15 (28)	4 (14)	20 (44)	39 (31)
3	17 (32)	6 (21)	9 (20)	32 (25)
≥4	9 (17)	10 (34)	5 (11)	24 (19)
Summary measures used†				
Area under the curve	11 (21)	13 (45)	10 (22)	34 (27)
Diagnostic odds ratio	29 (55)	14 (48)	12 (27)	55 (43)
Likelihood ratios	19 (36)	17 (59)	30 (67)	66 (52)
Predictive values	6 (11)	3 (10)	5 (11)	14 (11)
Q* statistic	12 (23)	5 (17)	2 (4)	19 (15)
Sensitivity and specificity	52 (98)	25 (86)	41 (91)	118 (93)
Other	2 (4)	2 (7)	1 (2)	5 (4)
Relative measures used to summarise differences in test accuracy	18 (34)	0	0	18 (14)
Study type				
Comparative only	8 (15)	8 (28)	0	16 (13)
Any study type	45 (85)	21 (72)	45 (100)	111 (87)
Test comparison strategy				
Both direct and indirect comparison	17 (32)	5 (17)	0	22 (17)
Direct comparison only	8 (15)	8 (28)	0	16 (13)
Indirect comparison only – comparative studies available	26 (49)	10 (34)	4 (9)	40 (32)
Indirect comparison only – no comparative studies available	2 (4)	6 (21)	1 (2)	9 (7)
None	0	0	40 (89)	40 (32)

*Numbers in parentheses are row percentages.

†For each summary statistic, the number in parentheses is the percentage that reported the statistic out of the total number of reviews in the group. Most reviews (75%) reported multiple statistics and so the total does not add up to 100%.

Paired measures such as sensitivity and specificity were considered a single measure if one or both were reported. Numbers in parentheses are column percentages. Percentages may not add up to 100% because of rounding.

4.3.2.3 *Methods for meta-analysis of a single test*

Almost all of the reviews (124/127, 98%) specified the method used for meta-analysis (Table 4.4). Hierarchical models were used for meta-analysis of individual tests in 46 (36%) reviews, either alone (26 reviews used the bivariate model and 11 used the HSROC model), in combination with each other (one review) or other meta-analytic methods (eight reviews). The Moses SROC regression approach was frequently used (42/127, 33%); it was used alone in five (4%) reviews, in combination with bivariate and univariate models in one review (1%) or with univariate models in 36 (28%) reviews in order to obtain summary sensitivity and specificity, likelihood ratios and/or diagnostic odds ratios. Univariate random effects or fixed effect models were also commonly used alone; three (2%) reviews used univariate logistic regression but the majority, 29 (19%) reviews, used traditional univariate models. Other methods were used in the remaining eight (6%) reviews. Of the eight reviews, two used a simple pooling method by summing up the true positives, false positives, false negatives and true negatives across studies, and then computed sensitivity and specificity using the totals.

Table 4.4| Meta-analysis methods used in the reviews

Characteristic	Comparative reviews		Multiple test reviews	Total
	Statistical analyses to compare test accuracy			
	Yes	No or unclear		
Number of reviews *	53 (42)	29 (23)	45 (35)	127 (100)
Hierarchical meta-analytic model used				
Yes	25 (47)	5 (17)	16 (36)	46 (36)
No	28 (53)	22 (76)	28 (62)	78 (61)
Method not specified	0	2 (7)	1 (2)	3 (2)
Meta-analytic method				
Bivariate model	11 (21)	5 (17)	10 (22)	26 (20)
HSROC model	10 (19)	0	1 (2)	11 (9)
Bivariate and HSROC models	1 (2)	0	0	1 (1)
Bivariate model, Moses SROC regression and univariate random effects model	0	0	1 (2)	1 (1)
Bivariate model and univariate methods	2 (4)	0	3 (7)	5 (4)
HSROC model and univariate methods	1 (2)	0	1 (2)	2 (2)
Moses SROC regression	1 (2)	2 (7)	2 (4)	5 (4)
Moses SROC regression and univariate fixed effect or random effects model	16 (30)	11 (38)	9 (20)	36 (28)
Univariate fixed effect or random effects logistic regression	1 (2)	1 (3)	1 (2)	3 (2)
Univariate fixed effect model	1 (2)	3 (10)	3 (7)	7 (6)
Univariate random effects model	2 (4)	2 (7)	12 (27)	16 (13)
Univariate fixed effect and random effects models	3 (6)	3 (10)	0	6 (5)
ANCOVA	2 (4)	0	0	2 (2)
Simple pooling	1 (2)	0	1 (2)	2 (2)
Weighted mean difference	1 (2)	0	0	1 (1)
Method not specified	0	2 (7)	1 (2)	3 (2)

*Numbers in parentheses are row percentages.

Numbers in parentheses are column percentages unless otherwise stated. Percentages may not add up to 100% because of rounding.

4.3.2.4 Methods for comparative meta-analysis

Comparative meta-analyses were considered feasible (see criteria in section 4.2.1) in 102

(80%) reviews but such analyses were performed in 50 (49%) of these reviews (Table 4.5).

Novel methods or modifications of existing methods were not found. Broadly classified, three

methods were used in the 53 comparative reviews (listed in Appendix B.3) that statistically

compared test accuracy: (1) naïve comparison (19/53, 36%) which refers to a comparison

where a statistical test, e.g. a Z-test, was used to compare summary estimates from separate

meta-analysis of one test with summary estimates from the meta-analysis of another test; (2) univariate pooling of differences in sensitivity and specificity, or pooling of differences in the diagnostic odds ratio (6/53, 11%); and (3) meta-regression by adding test type as a covariate to a meta-analytic model (23/53, 44%). For the remaining 5 (9%) reviews, the method used was unclear. Naïve comparisons were done using the Z-test in 15 of the 19 reviews (79%), and the remaining four reviews performed one of the following: paired t-test, unpaired t-test, chi-squared test, or comparison of Q* statistics and their standard errors.

Of the 23 reviews that used a meta-analytic meta-regression approach, 18 (78%) used a bivariate or HSROC model. The remaining five reviews used analysis of covariance (ANCOVA), Moses SROC or logistic regression (Table 4.5). Of the 29 comparative reviews that did not formally compare tests, three (10%) determined the statistical significance of differences in test accuracy based on whether or not confidence intervals overlapped, nine (31%) narratively compared tests, 14 (48%) did not perform a comparison and three (10%) were unclear. Forty-two reviews (33%) included studies that reported test accuracy at different thresholds (the number and set of thresholds reported differed between studies). Of these, 13 reviews formally compared tests and 6 (46%) of them accounted for multiple thresholds in the analysis.

Table 4.5| Test comparison strategy and comparative meta-analysis methods

Characteristic	Comparative reviews		Multiple test reviews	Total
	Statistical analyses to compare test accuracy			
	Yes	No or unclear		
Number of reviews *	53 (42)	29 (23)	45 (35)	127 (100)
Test comparison feasible	50 (94)	25 (86)	27 (60)	102 (80)
Test comparison method				
Meta-regression – hierarchical model	18 (34)	0	0	18 (14)
Meta-regression – SROC regression	2 (4)	0	0	2 (2)
Meta-regression – ANCOVA	2 (4)	0	0	2 (2)
Meta-regression – logistic regression	1 (2)	0	0	1 (1)
Univariate pooling of difference in sensitivity and specificity or DORs	6 (11)	0	0	6 (5)
Naïve (comparison of pooled estimates from separate meta-analyses)		0	0	
Z-test	15 (28)	0	0	15 (12)
Paired t-test	1 (2)	0	0	1 (1)
Unpaired t-test	1 (2)	0	0	1 (1)
Chi-squared test	1 (2)	0	0	1 (1)
Comparison of Q* and their SEs	1 (2)	0	0	1 (1)
Overlapping confidence intervals	0	3 (10)	0	3 (2)
Narrative	0	9 (31)	4 (9)	13 (10)
None	0	14 (48)	40 (89)	54 (43)
Unclear	5 (9)	3 (10)	1 (2)	9 (7)
Multiple thresholds included	13 (25)	12 (41)	17 (38)	42 (33)
If multiple thresholds included, were they accounted for in the comparative meta-analysis (meta-analysis at each threshold or fitted appropriate model)				
Yes	6 (46)	0	0	6 (46)
No	4 (31)	0	0	4 (31)
Unclear	3 (23)	0	0	3 (23)

*Numbers in parentheses are row percentages.

Numbers in parentheses are column percentages unless otherwise stated. Percentages may not add up to 100% because of rounding.

4.3.2.5 Investigations of heterogeneity

Investigations of heterogeneity were performed in 67 (53%) reviews, of which 24 (36%) used meta-regression, 35 (52%) used subgroup analyses, and 8 (12%) used both methods (Table 4.6). These analyses were performed for individual tests. Amongst the 53 comparative reviews that statistically compared test accuracy, 33 (62%) investigated heterogeneity. Of these 33 reviews, five (15%) assessed the effect of potential confounders on relative accuracy.

Four of the five reviews used subgroup analyses where test comparisons were made within each subgroup, while the remaining review used a Bayesian bivariate meta- regression. One of the 33 reviews planned an investigation of heterogeneity on relative accuracy but there was insufficient data for the analyses.

Table 4.6| Investigations of heterogeneity in the reviews

Characteristic	Comparative reviews		Multiple test reviews	Total
	Statistical analyses to compare test accuracy			
	Yes	No or unclear		
Number of reviews *	53 (42)	29 (23)	45 (35)	127 (100)
Formal investigation performed				
Yes – meta-regression and subgroup analyses	5 (9)	1 (3)	2 (4)	8 (6)
Yes – meta-regression	15 (28)	5 (17)	4 (9)	24 (19)
Yes – subgroup analyses	13 (25)	8 (28)	14 (31)	35 (28)
No – limited data	8 (15)	2 (7)	1 (2)	11 (9)
No – only tested for heterogeneity	3 (6)	8 (28)	16 (36)	27 (21)
No – nothing reported	7 (13)	5 (17)	8 (18)	20 (16)
Unclear	2 (4)	0	0	2 (2)
If yes above, was effect on relative accuracy also investigated?				
Yes	5 (15)	0	0	5 (15)
No	21 (64)	0	0	21 (64)
Planned but no data	1 (3)	0	0	1 (3)
Unclear	6 (18)	0	0	6 (18)

*Numbers in parentheses are row percentages.

Numbers in parentheses are column percentages unless otherwise stated. Percentages may not add up to 100% because of rounding.

4.3.3 Presentation and reporting

Although there is no dedicated guideline for reporting DTA reviews and meta-analyses, 13 reviews (10%) used a reporting guideline (Table 4.7). Five reviews used PRISMA; four used QUOROM (Quality of Reporting of Meta-analyses), the precursor to PRISMA; one used both QUOROM and PRISMA; one used both STARD (Standards for the Reporting of Diagnostic accuracy) and MOOSE (Meta-analysis of Observational Studies in Epidemiology); and the remaining two stated they followed recommendations of the Cochrane DTA working Group.

4.3.3.1 Summary of reporting quality

Figure 4.2 shows the grid of results for 10 characteristics (derived from Table 4.7) reported in each of the 127 reviews, sorted within each of the three review groups by year of publication and the total number of missing (unreported or unclear) items in a review. The figure clearly shows that the reporting of several items—in particular the role of the index tests, test comparison strategy and limitations of indirect comparisons—was deficient in many reviews. All multiple test reviews did not state a clear comparative objective (this was one of the four criteria used for classifying the reviews as outlined in section 3.2.1).

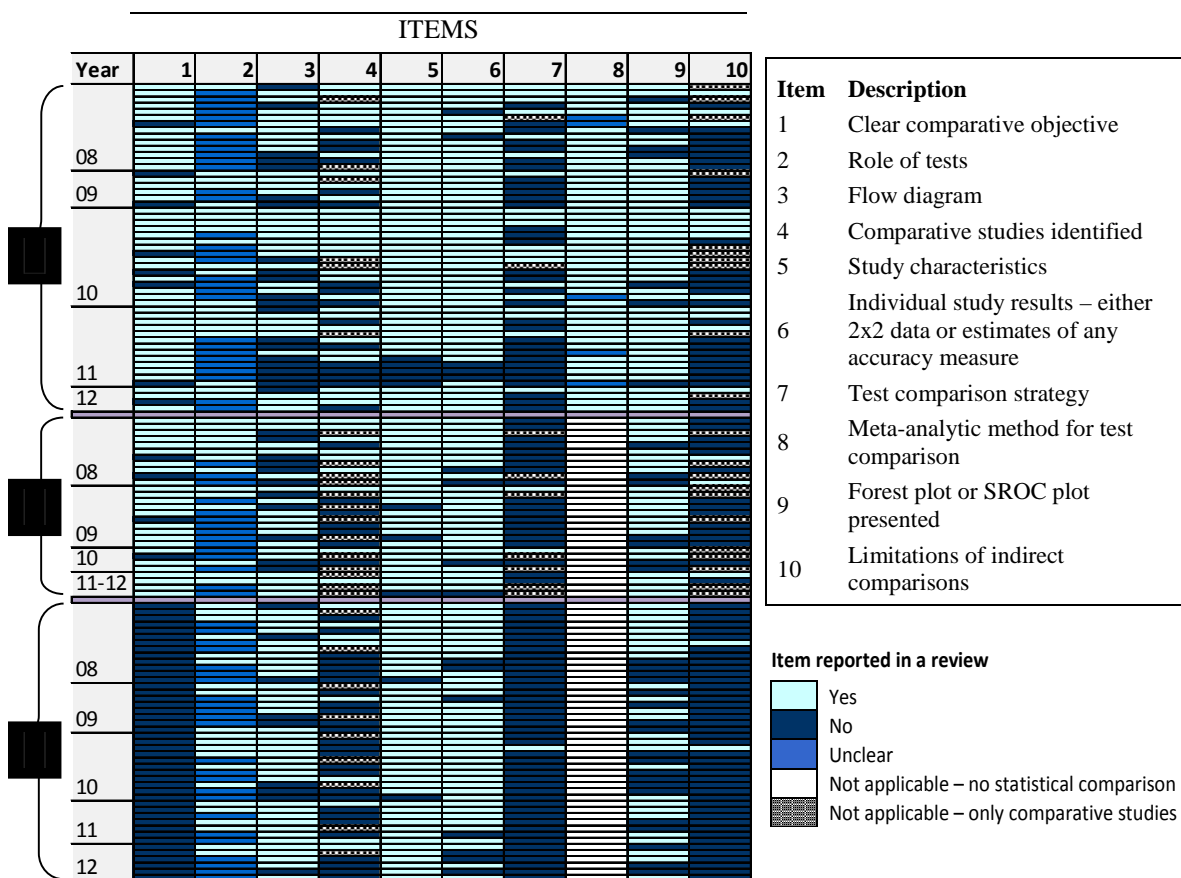


Figure 4.2| Reporting characteristics of 127 reviews

A– Comparative reviews with statistical analyses performed to compare accuracy; B – Comparative reviews without statistical analyses to compare accuracy; C – Multiple test reviews. The coloured cells in each row illustrate the reporting of the 10 items in each review. The box to the right of the figure gives the description of the reporting items. Reviews were ordered by year of publication and the number of missing items within each of the three review categories A to C. Each review category is separated by a purple strip.

4.3.3.2 Review objectives and role of the index tests

A comparative objective was explicitly stated in 70 (55%) reviews (Table 4.7). It was possible to deduce the role of the tests in 57 (45%) reviews as add on, triage and/or replacement for an existing test. For 28 of the 57 (49%) reviews, the role was explicitly stated while for the remaining 29 (51%) reviews, the role had to be deduced using implicit information in the background and discussion sections of the reviews.

4.3.3.3 Study identification and characteristics

A flow diagram illustrating the selection of studies was not presented in 41 (32%) reviews (Table 4.7). In 61 (48%) reviews, a flow diagram was presented without the number of studies per test while 25 (20%) reviews presented a comprehensive flow diagram with the number of studies per test. Of these 25 reviews, the flow diagrams in five reviews¹²⁵⁻¹²⁹ were good examples. These flow diagrams clearly showed the number of studies included in the analysis of each test, and also indicated the number of comparative studies available. Of the 99 reviews that had at least one comparative study, 50 (51%) reviews did not identify the comparative studies. Most of the reviews (92%) reported study characteristics though the detail reported varied.

Table 4.7| Reporting and presentation characteristics of the reviews

Characteristic	Comparative reviews		Multiple test reviews	Total
	Statistical analyses to compare test accuracy			
	Yes	No or unclear		
Number of reviews *	53 (42)	29 (23)	45 (35)	127 (100)
Reporting guideline used	2 (4)	5 (17)	6 (13)	13 (10)
Clear comparative objective stated	45 (85)	25 (86)	0	70 (55)
Role of the tests				
Add-on	6 (11)	3 (10)	2 (4)	11 (9)
Replacement	8 (15)	6 (21)	6 (13)	20 (16)
Triage	4 (8)	1 (3)	11 (24)	16 (13)
Any two of the above	4 (8)	4 (14)	2 (4)	10 (8)
Unclear	31 (58)	15 (52)	24 (53)	70 (55)
Flow diagram presented				
Yes – included number of studies per test	11 (21)	6 (21)	8 (18)	25 (20)
Yes – excluded number of studies per test	21 (40)	12 (41)	28 (62)	61 (48)
No	21 (40)	11 (38)	9 (20)	41 (32)
Comparative studies identified				
Yes	31 (58)	9 (31)	9 (20)	49 (39)
No	16 (30)	7 (24)	27 (60)	50 (39)
No comparative studies in review	6 (11)	13 (45)	9 (20)	28 (22)
Study characteristics presented	48 (91)	26 (90)	43 (96)	117 (92)
Test comparison strategy				
Yes	19 (36)	2 (7)	1 (2)	22 (17)
No	32 (60)	20 (69)	44 (98)	96 (76)
No but only comparative studies included	2 (4)	7 (24)	0	9 (7)
Meta-analytic method for test comparison				
Yes	48 (91)	NA	NA	48 (91)
Unclear	5 (9)	NA	NA	5 (9)
2x2 data for each study	30 (57)	10 (34)	14 (31)	54 (43)
Individual study estimates of test accuracy	46 (87)	25 (86)	36 (80)	107 (84)
Forest plot(s)	30 (57)	19 (66)	16 (36)	65 (51)
SROC plot				
SROC plot comparing summary points or curves for 2 or more tests	19 (36)	7 (26)	2 (4)	28 (22)
Separate SROC plot per test	17 (32)	11 (38)	19 (42)	47 (37)
No SROC plot	17 (32)	11 (38)	24 (53)	52 (41)
Limitations of indirect comparison acknowledged				
Yes	13 (25)	3 (10)	2 (4)	18 (14)
No	30 (57)	15 (52)	43 (96)	88 (69)
No but only comparative studies included	10 (19)	11 (38)	0	21 (17)

NA = not applicable.

*Numbers in parentheses are row percentages.

Numbers in parentheses are column percentages unless otherwise stated. Percentages may not add up to 100% because of rounding.

4.3.3.4 Strategy for comparing test accuracy

Seventy three comparative reviews included both comparative and non-comparative studies and 21 (29%) reviews stated their strategy for comparing tests, i.e., direct and/or indirect comparisons. Of the 21 reviews, 19 (90%) formally compared test accuracy (Table 4.7).

4.3.3.5 Review findings and limitations

Results from individual studies and meta-analyses

Four (3%) reviews reported 2x2 data only, 57 (45%) reviews reported test accuracy estimates only, and 50 (39%) reviews reported both. The accuracy estimates or 2x2 data were reported in tables (43 reviews) or on forest plots (64 reviews). One review presented a forest plot but no study results were shown. Therefore, a total of 65 (51%) reviews presented forest plots (Table 4.7). Forest plots were more commonly presented in both types of comparative reviews (57% and 66%) than in multiple test reviews (36%). A SROC plot showing results for two or more tests was presented in 28 (22%) reviews, 47 (37%) reviews presented each test on a separate SROC plot, and the remaining 52 (41%) reviews did not present a SROC plot. Two multiple test reviews and seven comparative reviews without a formal test comparison presented a SROC plot showing a test comparison.

Limitations of indirect comparisons

Twenty one (17%) reviews restricted inclusion to comparative studies (Table 4.7). Of the remaining 106 reviews that included any study type, 18 (17%) acknowledged the limitations of indirect comparisons. Furthermore, 9 of these 18 reviews also recommended that future primary studies should directly compare the performance of tests within the same patient population.

4.4 Discussion

4.4.1 Principal findings

The findings of this “review of reviews” show variation in methods and reporting of systematic reviews of comparative accuracy and multiple tests. The findings are discussed in relation to the questions posed in the introduction (section 4.1).

4.4.1.1 *What summary measures are used?*

Several outcome measures can be used to quantify test accuracy. Many reviews (75%) used two or more different measures with sensitivity and specificity being the most commonly used and predictive values the least used metric. The probability that the test will give the correct diagnosis, i.e. predictive values, is generally of clinical interest and may be more intuitive to decision makers.¹³⁰ However, the poor usage of predictive values may be due to their dependence on the prevalence of the target condition which may differ substantially between studies and so raises issues about transferability. There is increasing recognition that the effect of prevalence is not limited to predictive values.^{130,131} Due to clinical variability (such as spectrum effects, referral filters or reader expectation) and other mechanisms, prevalence can affect measures such as sensitivity and specificity.¹³² The Q* statistic—the point on the SROC curve where sensitivity is equal to specificity—is not recommended for use in DTA reviews because it can be misleading (see section 1.4.3), yet it was sometimes used to quantify the Moses SROC curve. Relative measures of test performance are not frequently used to express differences in test performance. Amongst reviews that formally compared test accuracy, about a third used relative measures.

4.4.1.2 What strategies are used to compare tests?

For reasons highlighted in section 1.5, comparative studies are ideal for comparing test accuracy. In brief, such designs ensure validity by comparing like-with-like thus avoiding confounding. Although some reviews restricted study selection to comparative studies (direct comparisons), the majority (87%) included any study type (indirect comparisons). This is likely due to scarcity of comparative studies. The availability of comparative studies will be assessed in the next chapter. It is worth noting that most (81%) of the 16 reviews with only direct comparisons evaluated two tests. Direct comparisons were also performed in some reviews that performed indirect comparisons as the main analysis, thus making use of all the available evidence as well as comparing the consistency of indirect and direct evidence.

4.4.1.3 Do reviews use robust meta-analysis methods?

As explained in section 1.4.4.2, hierarchical models which account for between-study correlation in sensitivity and specificity while also allowing for variability within and between studies are recommended for meta-analysis of test accuracy studies.^{23,43} The use of these methods was limited. The most commonly used approach was a combination of traditional univariate methods (see section 1.4.2) and the Moses SROC regression. Both approaches require the use of a continuity correction if there are zeros in the 2x2 table of any included study thus leading to a downward bias in test accuracy. Traditional univariate methods also have the disadvantage that they ignore potential correlation between sensitivity and specificity. While the Moses SROC regression approach estimates the underlying SROC curve by modelling the trade-off between sensitivity and specificity,^{37,133} it has methodological limitations which lead to inaccurate standard errors and invalid statistical inference.^{39,40} The Moses method does not give summary estimates of sensitivity and

specificity which may explain why it was seldom used alone but was used in conjunction with traditional univariate methods to obtain summary sensitivity and specificity.

4.4.1.4 Do reviews use robust comparative meta-analysis methods?

Formal test comparisons were not always done even when comparative meta-analyses were deemed feasible; such reviews relied on informal inferences. The reason may be lack of familiarity with methods and/or the capability to perform such analyses. Comparative meta-analyses using hierarchical methods may be based on a comparison of summary points and/or SROC curves. The estimation of SROC curves allows data from studies that have used different thresholds to define test positivity. Of the reviews that formally compared tests using studies that reported test accuracy at different thresholds, only about half of them accounted for multiple thresholds in the analysis. The other reviews presented summary sensitivities and specificities which do not have clinically meaningful interpretation because the analyses were not based on data at a given threshold.

4.4.1.5 Do reviews appropriately investigate heterogeneity?

Heterogeneity is often observed in test accuracy reviews and differences between tests may be confounded by differences in study characteristics. Reviews commonly investigated heterogeneity using meta-regression or subgroup analyses. The analyses were often done separately for each test with sufficient data rather than examining the effect jointly on all tests in a comparison; the latter was rarely possible due to limited data. Given the preponderance of indirect comparisons, adjusting for potential confounders is unlikely to be feasible in most situations. In a case study using individual patient data, differences between direct and

indirect comparisons persisted, even after adjusting for differences in threshold, reference standard and patient characteristics.¹³⁴

4.4.1.6 Are reviews reported appropriately?

Complete and transparent reporting is essential for understanding and judging the validity of a review. Clear review objectives aid study selection, provide a template for devising an analysis plan, and facilitate appropriate interpretation of the review findings but a comparative objective was not explicitly stated in many (45%) reviews. Additionally, the role of the index tests in the diagnostic pathway was ambiguous in many reviews. Although it was expected that this would be poorly reported, the extent was not anticipated.

Distinguishing between the different types of studies that contribute to different analyses in a review enhances clarity and facilitates judgements about the source of the evidence but half of the reviews failed to identify which of the studies were comparative. To aid understanding of the review methods and interpretation of findings, the strategy (direct comparisons, indirect comparisons or both) adopted for test comparisons should be clearly specified in addition to the methods for meta-analysis. However, many reviews did not specify test comparison strategies. Except for three reviews, the methods used for meta-analysis of individual tests were specified. Most reviews also indicated the methods used for comparative meta-analysis although the clarity of reporting varied. Adequate reporting is essential to enable appraisal of the validity of the methods used.

The applicability of the findings of a review to the review question depends on characteristics of the included studies¹³⁵ but reviews did not always report study characteristics. Furthermore,

the level of detail provided by reviews that reported study characteristics varied considerably between reviews. Reviews usually reported study specific estimates of test performance either in tables or in forest plots. In addition to forest plots, reviews may include SROC plots which show SROC curves or summary points with corresponding confidence and/or prediction regions. Ideally, results from a test comparison should be shown on a single SROC plot (for examples, see Figure 2.3 and Figure 2.4) instead of showing the results for each test on a separate SROC plot. For reviews that included SROC plots, it was more common to present each test on a separate SROC plot rather than compare tests on one plot. Interestingly, two multiple test reviews and seven comparative reviews without a formal test comparison presented a SROC plot showing a test comparison. This justifies the inclusion of such reviews in this methodological review instead of including only reviews that formally compared tests or clearly defined a comparative objective.

There are several potential sources of bias and variation in test accuracy studies,¹³⁶⁻¹³⁸ and review findings must be interpreted in the context of the quality and the strength of the evidence. It is possible that results of direct comparisons may not be consistent with those of indirect comparisons but reviews seldom acknowledged the limitations of indirect comparisons. In Chapter 5, meta-analyses of direct and indirect comparisons will be investigated to determine the existence and magnitude of differences in estimates of test performance between both types of test comparisons.

4.4.2 Comparison with other studies

Published methodological reviews have focused primarily on systematic reviews of a single test,^{27,139} specific clinical area^{140,141} or specific methodological issue.^{94,142,143} Mallett et al

assessed reviews of diagnostic tests in cancer published between 1990 and 2003.¹⁴⁰ They included 89 reviews, 25 of which were assessed in detail. The authors concluded that the reliability and relevance of the reviews was compromised by poor reporting and review methods. Cruciani et al assessed 55 meta-analyses in infectious diseases and also reported problems in the conduct and reporting of DTA reviews.¹⁴¹

Dahabreh et al conducted a comprehensive overview of DTA reviews published between 1987 and 2009.²⁷ Of the 760 included reviews, 396 (52%) evaluated a single test and 364 (48%) evaluated two or more tests. Statistical comparative analyses were reported in 131/364 reviews (36%), with no change over time in the proportion of reviews reporting comparative analyses of at least two index tests. Thirty three of the 131 (25%) reviews performed direct comparisons while the remaining performed indirect comparisons. Similarly, in the present study, 42% of reviews included statistical comparative analyses while 15% of these restricted analyses to direct comparisons and 32% performed both direct and indirect comparisons.

4.4.3 Implications for research and practice

Long-term RCTs of test-plus-treatment strategies are advocated for evaluating the benefits of a new test relative to current best practice but such RCTs are not always feasible, available or necessary.^{20,21} Since these RCTs are rare,²² comparative accuracy reviews are a useful surrogate for guiding test selection and decision making. Complete and unambiguous reporting facilitates critical appraisal of the evidence which is vital since comparative reviews commonly rely on indirect comparisons. While several items of the PRISMA checklist are relevant for DTA reviews, the challenges of a DTA review require a dedicated reporting guideline or extension to the PRISMA statement; extensions have been developed for other

review types (see section 4.1).^{93,120,121,123,124} As such PRISMA-DTA is now being developed¹⁴⁴ and the results of this study will be used to inform special guidance provided for comparative DTA reviews.

In the interim, based on findings from the reviews and methodological recommendations in the Cochrane Handbook¹¹⁸ and published literature,^{32,43,60} seven reporting criteria were devised. They are (1) role of the tests in the diagnostic pathway; (2) identification of studies for each test; (3) test comparison strategy; (4) meta-analytic method; (5) study characteristics; (6) presentation of study estimates of test performance and graphical summaries; and (7) limitations of indirect comparisons (if such comparisons were performed). The rationale and explanation for the criteria are summarised in Box 4.1. There is some repetition from contents of this and earlier chapters because the set of criteria was designed for use outwith the thesis. These criteria can be used by authors for reporting comparative DTA reviews, and by peer reviewers and journal editors to appraise the reviews.

Box 4.1| Criteria for reporting test comparisons in systematic reviews of test accuracy

Item	Description	Rationale and explanation
1	Role of tests in diagnostic pathway	Test evaluation requires a clear objective and definition of the intended use and role of a test within the context of a clinical pathway for a specific population with the target condition. The intended role of a test guides formulation of the review question and provides a framework for assessing test accuracy, including the choice of a comparator(s) and selection of studies. The role of a test is therefore important for understanding the context in which the tests will be used and the interpretation of the meta-analytic findings. The existing diagnostic pathway and the current or proposed role of the index test(s) in the pathway should be described. A new test may replace an existing one (replacement), be used before the existing test (triage) or after the existing test (add-on). ⁶⁰
2	Identification of included studies for each test	Review complexity increases with increasing number of tests, target conditions, uses and/or target populations within a single review. Therefore, distinguishing between the different groups of studies that contribute to different analyses in the review enhances clarity. The PRISMA flow diagram can be extended to show the number of included studies for each test or group of tests if inclusion is not limited to comparative studies. The detail shown—individual tests or groups of tests, settings and populations—will depend on the volume of information and the ability of the review team to neatly summarise the information. If such a comprehensive flow diagram is not feasible, the studies contributing to the assessment of each test should be clearly identified in the manuscript in some other way. The source of the evidence should be declared by stating types of included studies and studies contributing direct evidence should also be clearly identified in the review.
3	Test comparison strategy	Comparative studies are ideal but they are scarce. ⁷¹ An indirect between-study (uncontrolled) test comparison uses a different set of studies for each test and so does not ensure like-with-like comparisons; the difference in accuracy is prone to confounding due to differences in patient groups and study methods. Although direct comparisons based on only comparative studies are likely to ensure an unbiased comparison and enhance validity, such analyses may not always be feasible due to limited availability of comparative studies. Conversely, an indirect comparison uses all eligible studies that have evaluated at least one of the tests of interest thus maximising use of the available data. If study selection is not limited to comparative studies and comparative studies are available, a direct comparison should be considered in addition to an indirect comparison. The direct comparison may be narrative or quantitative depending on the availability of comparative studies.
4	Meta-analytic methods	Hierarchical models which account for between-study correlation in sensitivity and specificity while also allowing for variability within and between studies are recommended for meta-analysis of test accuracy studies. ^{23,43} The two main hierarchical models are the bivariate and the hierarchical summary receiver characteristic operating (HSROC) models which focus on the estimation of summary points (summary sensitivities and specificities) and SROC curves respectively. ^{41,42} For the summary point of a test to have a clinically meaningful interpretation, the analysis

Box 4.1 continued...

Item	Description	Rationale and explanation
5	Study characteristics	<p>should be based on data at a given threshold. For the estimation of a SROC curve, data from all studies, regardless of threshold, can be included. As such test comparisons may be based on a comparison of summary points and/or SROC curves. For the estimation of a SROC curve, one threshold per study is selected for inclusion in the analysis. If multiple cut-offs were considered, the description of methods should include how the cut-offs were selected and handled in the analyses. Methods have been proposed which allow inclusion of data from multiple thresholds for each study but these are rarely used in practice due to their complexity and limitations. Also, the methods have not been applied to test comparisons.</p> <p>Relevant characteristics for each included study should be provided. This may be summarised in a table and should include elements of study design if eligibility was not restricted to specific design features. Heterogeneity is often observed in test accuracy reviews and differences between tests may be confounded by differences in study characteristics. Confounders can potentially be adjusted for in indirect test comparisons, though this is likely to be unachievable due to small number of studies and/or incomplete information on confounders. The effect of factors that may explain variation in test performance is typically assessed separately for each test.</p>
6	Study estimates of test performance and graphical summaries (forest plot or SROC plot)	<p>It is desirable to report 2x2 data (number of true positives, false positives, false negatives and true negatives) and summary statistics of test performance from each included study. This may be done graphically (e.g. forest plots) or in tables. Such summaries of the data will inform the reader about the degree to which study specific estimates deviate from the overall summaries, as well as the size and precision of each study. It is plausible that study results for one test may be more consistent or precise than those of another test in an indirect comparison. In addition to forest plots, reviews may include SROC plots. A SROC plot of sensitivity against specificity displays the results of the included studies as points in ROC space. The plot can also show meta-analytic summaries such as SROC curves or summary points (summary sensitivities and specificities) with corresponding confidence and/or prediction regions to illustrate uncertainty and heterogeneity, respectively. Ideally, results from a test comparison should be shown on a single SROC plot instead of showing the results for each test on a separate SROC plot. Furthermore, for pairwise direct comparisons, the pair of points representing the results of the two tests from each study can be identified on the plot by adding a connecting line between the points.</p>
7	Limitations of the evidence from indirect comparisons	<p>This is only applicable for reviews that include indirect comparisons. Be clear about the quality and strength of the evidence when interpreting the results, including limitations of including non-comparative studies in a test comparison. The results of indirect comparisons should be carefully interpreted taking into account the possibility that differences in test performance may be confounded by clinical and/or methodological factors. This is essential because it is seldom feasible to assess the effect of potential confounders on relative accuracy.</p>

It should be emphasised that other aspects relevant to all DTA reviews such the index tests, reference standards and target conditions, should also be clearly described. Space constraint in journals may limit proper reporting but may be a minor issue because many journals accept online supplementary materials. It was noted that 56 (44%) reviews used appendices or online supplementary files to provide additional data and information. Based on recommendations in the Cochrane Handbook,¹¹⁸ five comparative reviews^{103,129,145-147} were judged exemplary in methods used for test comparisons. In addition to the suggested criteria in Box 4.1, these reviews can serve as examples of good practice to guide authors. A summary of the features of the reviews is given in Appendix B.4. Two of the exemplar reviews were among the five Cochrane reviews included.

4.4.4 Strengths and limitations

The work was undertaken without a formal power calculation to determine the number of reviews to assess. Although a method for sample size estimation for meta-epidemiological studies was recently proposed,¹⁴⁸ this applies to trials of interventions and the assumptions of the method may not be tenable for test accuracy studies and meta-analyses. Furthermore, this descriptive survey aimed to provide an understanding of the methodological and reporting characteristics of the reviews and statistical testing was not done—frequencies and percentages were presented to summarise the characteristics of the reviews.

Although the search was limited to DARE, for a review to be included in DARE it must meet certain quality criteria.¹¹³ As such the quality of the literature may be even poorer than has been shown using a large sample of reviews published in a wide range of journals. The reviews assessed a variety of test types for different uses and target conditions. The

classification of reviews was inclusive such that reviews were considered irrespective of whether or not tests were formally compared. Thus a broad perspective of the literature was gained and the generalisability of the findings was increased; the author believes this comprehensive overview adequately reflects practice.

In addition to documenting the characteristics of the reviews, examples of good practice in terms of methods and reporting were highlighted. Review authors can use these reviews as a guide. Only a list of items considered relevant when reporting comparative reviews was developed to supplement items applicable to any DTA review. These generic criteria have not been stated in this chapter and the reader should consult the Cochrane Handbook¹¹⁸ for further information. This study was limited to analysis methods and strategies for test comparisons though various stages of the conduct of a systematic review can affect its results and conclusions. For instance, this review of reviews did not consider quality assessment of comparative studies. There are special issues to consider in these studies which are worthy of investigation but are beyond the scope of this thesis. Therefore, the list of reporting items suggested for comparative reviews in this chapter is a "*starting point*" rather than the "*finished checklist*".

A limitation of this study is the lack of standard terminology for study designs in test accuracy research. To keep the review broad, any review that assessed more than one test was included even if a comparative objective was not stated or tests were not compared in a meta-analysis. Unexpectedly, two multiple test reviews and seven comparative reviews without a formal test comparison presented a SROC plot showing a test comparison. This finding provides support for the decision to include such reviews. Second, assessment of the role of the tests in a

review was sometimes subjective and relied on the judgement of the assessor. Therefore, reporting was judged in terms of whether or not the role was stated rather than the quality of the description provided. Any uncertainty in a judgement was discussed with supervisors before making a final decision.

4.4.5 Conclusions

Sometimes a choice needs to be made between alternative tests that can be used at the same point in the diagnostic pathway, which can be informed by a systematic review of the comparative accuracy of the tests. The complexities inherent in comparative reviews make clear reporting a necessity. This review of reviews has highlighted deficiencies in the methods and reporting of comparative reviews, and identified examples of good practice. Comparative accuracy reviews can inform decisions about test selection but suboptimal conduct and reporting will compromise their validity and relevance, and contribute to research waste. To improve quality and transparency, and to increase confidence in decision making informed by these reviews, a reporting guideline for test accuracy reviews that includes considerations for comparative accuracy is essential. Improved reporting is urgent because of the increasing prevalence of test accuracy reviews and their role in health technology assessment and clinical guideline development.

In the next chapter, meta-analyses based on direct and indirect comparisons will be examined to determine the magnitude of differences between the two types of analyses and the implications of such differences on review conclusions. In Chapter 7, comparative meta-analysis methods identified from the search for methods described in Chapter 3 will be

Chapter 4: Review of systematic reviews of comparative test accuracy

assessed to determine if different methods give different estimates and lead to different conclusions.

5 EMPIRICAL EVIDENCE OF THE IMPORTANCE OF COMPARATIVE STUDIES OF DIAGNOSTIC TEST ACCURACY

A paper based on the content of this chapter has been published.

Citation: **Takwoingi Y**, Leeflang MM, Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. *Annals of Internal Medicine* 2013;158(7):544-54.

5.1 Introduction

Systematic reviews of comparative accuracy may undertake meta-analyses by including all available studies (indirect comparison) or by restricting the analyses to only comparative studies (direct comparison) as previously described in section 1.6 and illustrated in Figure 1.11. In Chapter 2, seven case studies highlighted the potential scarcity of comparative accuracy studies and so indirect comparisons were the main or only analyses performed in the reviews. The findings of the review of reviews in Chapter 4 further showed that such indirect comparisons were indeed the major source of evidence on the comparative accuracy of many tests used in practice.

Chapter 4 also characterised statistical methods and reporting in published comparative or multiple test reviews to provide an understanding of the current state of the literature on which evidence of comparative test accuracy is based. In this chapter an empirical study is presented to build on the review by investigating the robustness of indirect comparisons in regard to potential bias due to confounding. Based on logical reasoning, direct comparisons are less likely to be prone to bias due to confounding because comparative studies compare test accuracy within the same population using either within- or between-subject randomized

designs. Moreover, because heterogeneity is the norm in test accuracy reviews, such comparative accuracy studies should provide the most reliable evidence on relative test performance. Since *true* relative accuracy is unknown, the term bias will be used in this chapter to simply refer to differences or discrepancies between estimates from direct and indirect comparisons.

The aim of this empirical study is to assess the availability of comparative studies of test accuracy and to assess the validity of indirect test comparisons by comparing summary *indirect* estimates of comparative accuracy derived from non-comparative studies with summary *direct* estimates from comparative studies. In particular, it was a concern that indirect comparisons were biased towards favouring the newer or experimental test rather than the older test or current practice. There may be optimism bias in early studies of the accuracy of a single test due to involvement of industry and test developers; use of poor study methods; and/or evaluation in less challenging situations such as in populations which may be unrepresentative of clinical practice.

Unlike adjusted indirect comparisons of interventions, indirect test comparisons do not use a common comparator. These indirect test comparisons are analogous to naïve or unadjusted indirect comparisons of interventions where the results of individual arms of RCTs are compared. For interventions, the simplest form of an adjusted indirect comparison involves three interventions where two interventions of interest (e.g. A and B) have been compared in an RCT to a common comparator (e.g. C which could be a standard intervention or placebo).¹⁴⁹ Randomization is preserved by deriving the comparison A versus B from the A versus C and B versus C comparisons. However, primary studies of test accuracy often assess

the performance of one test at a time and so the simplest form of an indirect test comparison involves two tests (e.g. A and B) without a loop of evidence. An adjusted indirect test comparison is plausible if a common comparator is available.

Therefore, to avoid confusion or ambiguity that may arise from the normal usage of the terms direct and indirect comparisons to refer to different types of meta-analyses (see further explanation in section 1.6), in this chapter meta-analyses of comparative studies are regarded as direct comparisons and meta-analyses restricted to non-comparative studies (i.e. studies that evaluated a single test or evaluated only one of the tests of interest) as indirect comparisons. This distinction also makes it possible to perform one meta-analysis for each test comparison dataset to investigate the effect of study type on test accuracy (i.e. a meta-regression). The terms direct and indirect comparisons will only be used where necessary.

The outline of this chapter is as follows. In section 5.2, the methods for review selection, data extraction and data synthesis are described. Results of the assessment of the availability of comparative studies, comparison of meta-analytic findings obtained from comparative and non-comparative studies, and analysis of the direction of differences between them are reported in section 5.3. In section 5.4 the findings are discussed, including the implications for research and practice.

5.2 Methods

5.2.1 Review selection and data extraction

Reviews were selected for inclusion in this study in two stages from the overall cohort of reviews identified in section 3.4 based on the two aims of this chapter. In the first stage, for

assessment of the availability of comparative primary studies within each review, the eligibility of a review was assessed, first on the basis of the abstract, then confirmed by review of the full publication. Inclusion decisions that remained uncertain were resolved by discussion with supervisors. Possible comparative studies were identified by checking citations in each review. For pragmatic reasons and because the purpose is merely to gauge the availability of comparative studies, a comparative design was presumed if two tests shared the same citation in a review. If citations differed but shared authors and were published within a year, the citations were assessed further by reviewing full text papers to confirm study design.

In the second stage, for the meta-analytic investigation, only reviews identified in the first stage that had both comparative and non-comparative primary studies were considered. Reviews that 1) included at least three comparative studies; 2) included at least seven non-comparative studies (with a minimum of two studies for each test); and 3) provided data to reconstruct 2×2 tables were eligible for inclusion. The aforementioned minimum study numbers were chosen to ensure adequate data to achieve convergence of the meta-analytic models. Unlike in the first stage, the full text of all primary comparative studies included in each review identified in this stage was assessed for confirmation of study design; if uncertain the supervisors were consulted and a final decision reached upon consensus. Some seemingly comparative studies compared the accuracy of two tests between non-randomly allocated groups and thus were judged to be prone to selection bias. For example, some studies based the allocation of the tests on clinical signs and symptoms. Therefore these studies were classified as non-comparative. Only randomized and paired (within person) comparisons were regarded as comparative designs because these studies are more likely to make fair

comparisons. The effect of this study design classification was assessed in a sensitivity analysis.

Data were abstracted from the full text of included reviews. A random subset containing half the reviews was double checked by a second researcher to confirm consistency. The target condition, patient population, tests evaluated, purpose of the tests, and reference standard were recorded for each review. In addition, for the subset of reviews included in the meta-analytic stage, the numbers of true positive, false positive, false negative, and true negative results (or summary statistics that allowed their derivation) and the analytic strategy by which tests were compared were also extracted.

5.2.2 Data synthesis and analysis

The analyses were undertaken in four parts. First, differences between comparative and non-comparative studies in comparisons of overall accuracy (measured as diagnostic odds ratios) were assessed within each meta-analysis by investigating the interaction between accuracy and study type. Second, the magnitude of the difference between summary estimates obtained from comparative and non-comparative studies were considered in terms of absolute differences in sensitivity and specificity using the subset of meta-analyses where primary studies shared common thresholds. Third, differences in estimates from non-comparative studies and comparative studies were pooled across all the meta-analyses to investigate whether differences occurred more often than can be expected by chance. Fourth, an assessment was undertaken to determine whether there was evidence of a common direction to the differences.

5.2.2.1 Estimation of differences in comparative accuracy for each test comparison

HSROC models^{42,57} were used to summarize the accuracy of each test and to compare the accuracy of the pair of tests in each meta-analysis. A separate model was fitted for each of the pairwise comparisons. The HSROC model was chosen to enable meta-analyses of all datasets irrespective of whether studies used different or common thresholds. The HSROC model uses study specific estimates of the true positive rate (sensitivity) and the false positive rate (1-specificity) to estimate a SROC curve. The accuracy parameter estimates the expected log DOR (for further details of the HSROC model, see sections 1.4.4.2). The HSROC analyses were performed using the SAS NLMIXED procedure.

The initial HSROC model fitted to each meta-analytic dataset compared SROC curves of tests by allowing studies using different test thresholds to be included in the same meta-analysis.

The HSROC meta-regression model with an indicator variable for test type, t , that affects the accuracy (α), threshold (θ) and shape (β) parameters can be written as

$$\text{logit}(\pi_{ij}) = \left((\theta_i + \gamma t_i) + (\alpha_i + \xi t_i) \text{dis}_{ij} \right) \exp(-(\beta + \delta t_i) \text{dis}_{ij}). \quad (5.1)$$

In this model, γ assesses whether the underlying threshold differ between tests, ξ assesses whether test accuracy differ between tests, and δ assesses whether the shape of the curves differ by test.

Comparative studies were accounted for in the HSROC model by clustering data for both tests within each study as outlined in sections 2.3.4.2. The HSROC model in (5.1) was extended to assess the effect of study type on comparative accuracy by including an additional covariate to indicate study type (non-comparative or comparative). Thus two covariates, one indicating test type (test A or test B) and the other indicating study type were included in the HSROC

model. In this model, the test type covariate was used to estimate differences in the accuracy and threshold parameters, but the underlying shape of the SROC curve and the variances of the random effects for threshold and accuracy were assumed to be common to both tests. The study type covariate was used to estimate differences in accuracy only. These constraints were imposed to reduce the risk of convergence problems that occur when undertaking test accuracy meta-analyses with small numbers of studies. It also simplifies interpretation as where SROC curves have a common asymmetric shape (as described in section 1.5.4.2 and exemplified in section 2.3.3.2), the relative accuracy of tests is constant at all thresholds. This assumption is commonly made for trials of treatments in a network meta-analysis, where every source of direct evidence is assumed to have the same heterogeneity variance, implying a single heterogeneity variance for the entire network.¹⁵⁰

The difference in test performance between the two tests (relative accuracy) was quantified as the rDOR (see section 1.5.4.2). A term for the interaction between test type and study type was included in the model to assess whether the rDOR was associated with study type. This model can be written as

$$\text{logit}(\pi_{ij}) = \left((\theta_i + \gamma t_i) + (\alpha_i + \xi_1 t_i + \xi_2 z_i + \xi_3 t_i z_i) \text{dis}_{ij} \right) \exp(-\beta \text{dis}_{ij}), \quad (5.2)$$

where z is an indicator variable for study type that takes the value 0 for comparative studies and 1 for non-comparative studies. The exponent of the coefficient ξ_3 of the interaction term $t_i z_i$ gave the difference in comparative accuracy due to difference in study type, the ratio of the rDOR. A ratio of rDORs equal to one indicates no difference in estimates between the two study types; values that differed from one indicate a difference, with the direction dependent on whether the rDORs were greater than or less than one.

Assessment of assumption of equal variances of random effects

For each meta-analysis, study estimates of sensitivity and specificity were plotted in ROC space to observe whether there were marked differences in heterogeneity between studies for the two tests in order to check whether the assumption of equal variances of random effects for the two tests was appropriate. Where possible, alternative models were fitted and model fit was assessed by using likelihood ratio tests to compare nested models, i.e., model with and without a different shape for SROC curves of the two tests, and model with and without distinct variance parameters for the random effects (for further details, see section 1.5.4.4). Estimates of the DORs, rDORs, and ratios of rDORs were also compared between the alternative models to assess whether conclusions were robust to the assumptions.

5.2.2.2 Differences in terms of sensitivity and specificity

Since meaningful estimates of differences in average sensitivity and specificity should be obtained only when studies share a common test threshold²³ as explained in section 1.4.4, a subset of meta-analyses was identified in which common or consistent thresholds for each of the two tests had been used. Differences in sensitivity and specificity between tests were estimated from HSROC models by using the ESTIMATE statement in the NLMIXED procedure. As shown by Harbord et al,⁴⁴ the mean logit sensitivity and mean logit specificity can be obtained from parameters of the HSROC model as follows:

$$\mu_A = \exp\left(-\frac{\beta}{2}\right)\left(\theta + \frac{\Lambda}{2}\right), \mu_B = -\exp\left(\frac{\beta}{2}\right)\left(\theta - \frac{\Lambda}{2}\right). \quad (5.3)$$

Based on (5.3) differences in summary sensitivity and summary specificity between tests were computed using these equations for non-comparative studies

$$\Delta_{sensitivity} = \left[\frac{\exp\left(e^{-0.5\beta}(\theta + \gamma + 0.5(\Lambda + \xi_1 + \xi_2 + \xi_3))\right)}{1 + \exp\left(e^{-0.5\beta}(\theta + \gamma + 0.5(\Lambda + \xi_1 + \xi_2 + \xi_3))\right)} - \frac{\exp\left(e^{-0.5\beta}(\theta + 0.5(\Lambda + \xi_2 + \xi_3))\right)}{1 + \exp\left(e^{-0.5\beta}(\theta + 0.5(\Lambda + \xi_2 + \xi_3))\right)} \right]$$

$$\Delta_{specificity} = \left[\frac{\exp\left(-e^{0.5\beta}(\theta + \gamma - 0.5(\Lambda + \xi_1 + \xi_2 + \xi_3))\right)}{1 + \exp\left(-e^{0.5\beta}(\theta + \gamma - 0.5(\Lambda + \xi_1 + \xi_2 + \xi_3))\right)} - \frac{\exp\left(-e^{0.5\beta}(\theta - 0.5(\Lambda + \xi_2 + \xi_3))\right)}{1 + \exp\left(-e^{0.5\beta}(\theta - 0.5(\Lambda + \xi_2 + \xi_3))\right)} \right]$$

and these equations for comparative studies

$$\Delta_{sensitivity} = \left[\frac{\exp\left(e^{-0.5\beta}(\theta + \gamma + 0.5(\Lambda + \xi_1))\right)}{1 + \exp\left(e^{-0.5\beta}(\theta + \gamma + 0.5(\Lambda + \xi_1))\right)} - \frac{\exp\left(e^{-0.5\beta}(\theta + 0.5\Lambda)\right)}{1 + \exp\left(e^{-0.5\beta}(\theta + 0.5\Lambda)\right)} \right]$$

$$\Delta_{specificity} = \left[\frac{\exp\left(-e^{0.5\beta}(\theta + \gamma - 0.5(\Lambda + \xi_1))\right)}{1 + \exp\left(-e^{0.5\beta}(\theta + \gamma - 0.5(\Lambda + \xi_1))\right)} - \frac{\exp\left(-e^{0.5\beta}(\theta - 0.5\Lambda)\right)}{1 + \exp\left(-e^{0.5\beta}(\theta - 0.5\Lambda)\right)} \right]$$

in the ESTIMATE statement. Estimates derived from non-comparative studies were then compared with those from comparative studies.

5.2.2.3 Assessment of differences in comparative accuracy across meta-analyses

To assess whether differences in estimates from non-comparative studies and comparative studies were greater than expected by chance across all the meta-analyses, the estimated ratios of rDORs were plotted against the inverse of their standard errors in a one-sided contour-enhanced funnel plot, marked with a contour line corresponding to statistical significance at the 5% level.¹⁵¹ In this analysis, the ordering of the two tests was chosen such that all ratios were greater than one. The observed proportion significant at the 5% level was compared with the expected value of 5% by using a binomial test.

5.2.2.4 Assessment of direction of differences between study type

To test the directional hypothesis concerning differences between summary estimates derived from comparative and non-comparative studies favouring newer test technologies, it was necessary to first order the tests in each comparison. The experimental or "newer" test was identified as test A, and current practice or the "older" test was identified as test B. These definitions were based on information about the roles of the test reported in the review,

supplemented by clinical opinion where unclear. When clear categorization could not be achieved, comparisons were excluded from the directional analysis. The average difference in estimates from non-comparative studies versus those from comparative studies across the topics was computed by a second-level meta-analysis (meta-meta-analysis) of ratios of rDORs (on the log scale) by using an inverse variance weighted model with the estimate of heterogeneity obtained via the DerSimonian and Laird method of moments. This meta-analysis assumes independence of the ratios of rDORs. The assumption may have been violated because some reviews contributed multiple pairwise test comparisons and may have had some studies in common. As such, a sensitivity analysis was conducted where the meta-meta-analysis included one test comparison selected at random such that each review contributed only one estimate of the ratio of rDORs. The assessment of differences in comparative accuracy across meta-analyses and the directional analyses were performed using version 11 of the Stata software.

5.3 Results

Of the 286 reviews included in the cohort for the thesis, 248 met the inclusion criteria for the assessment of availability of comparative studies (see section 3.4 and Figure 3.1).

5.3.1 Availability of studies with comparative designs

Of the 248 reviews, 177 (71%) included both comparative and non-comparative studies, 28 (11%) included only comparative studies, and 43 (17%) found no comparative studies.

Generally, characteristics were similar across the three groups of review (Table 5.1). The 248 reviews contained 6915 studies: 2113 (31%) comparative and 4802 (69%) non-comparative. Median (interquartile range) numbers of comparative and non-comparative studies per review

were 6 (2 to 11) and 14 (4 to 28), respectively. The clinical purpose of most tests evaluated in the reviews was diagnosis (87%). Imaging methods were the most common test type (40%). Cancers were the most frequently evaluated target conditions (27%), especially neoplasms of the breast, lung, colon, liver, and skin.

Table 5.1| Characteristics of reviews included in the assessment of availability of comparative studies

Characteristic	Type of included studies			Total
	Comparative	Non-comparative	Both	
Number of reviews	28 (11)	43 (17)	177 (71)	248
Number of test accuracy studies in reviews				
Median (range)	11 (3–32)	23 (7–98)	23 (4–162)	22 (3–162)
Interquartile range	6–15	15–35	13–40	13–35
Number of tests evaluated				
2	19 (68)	19 (44)	45 (25)	83 (33)
3	3 (11)	12 (28)	31 (18)	46 (19)
4	3 (11)	3 (7)	21 (12)	27 (11)
≥5	3 (11)	9 (21)	80 (45)	92 (37)
Clinical topic (according to ICD-10 Version: 2010)				
Circulatory system	2 (7)	11 (26)	32 (18)	45 (18)
Digestive system	2 (7)	2 (5)	18 (10)	22 (9)
External causes of morbidity and mortality (Complications of medical and surgical care)	0	4 (9)	6 (3)	10 (4)
Genitourinary system	3 (11)	1 (2)	6 (3)	10 (4)
Infectious and parasitic diseases	5 (18)	8 (19)	18 (10)	31 (13)
Injury, poisoning and certain other consequences of external causes	2 (7)	2 (5)	6 (3)	10 (4)
Musculoskeletal system and connective tissue	1 (4)	1 (2)	11 (6)	13 (5)
Neoplasms	9 (32)	8 (19)	50 (28)	67 (27)
Other ICD-10 codes*	4 (14)	6 (14)	30 (17)	40 (16)
Type of tests evaluated				
Biopsy	0	1 (2)	2 (1)	3 (1)
Clinical and physical examination	2 (7)	6 (14)	31 (18)	39 (16)
Drug	1 (4)	0	1 (0.6)	2 (0.8)
Imaging	9 (32)	21 (49)	70 (40)	100 (40)
Laboratory	9 (32)	11 (26)	41 (23)	61 (25)
RDT or POCT	0	1 (2)	5 (3)	6 (2)
Self administered†	2 (7)	1 (2)	2 (1)	5 (2)
Combinations of any of the above‡	5 (18)	2 (5)	25 (14)	32 (13)
Clinical purpose of the tests				
Diagnostic	23 (82)	37 (86)	156 (88)	216 (87)
Diagnosis and staging	0	1 (2)	0	1 (0.4)
Monitoring	1 (4)	0	2 (1)	3 (1)
Prognostic/prediction	1 (4)	1 (2)	4 (2)	6 (2)
Screening	3 (11)	2 (5)	6 (3)	11 (4)
Staging	0	2 (5)	9 (5)	11 (4)
Type of publication				
Cochrane review	1 (4)	0	4 (2)	5 (2)
General medical journal	5 (18)	8 (19)	48 (27)	61 (25)
Specialist medical journal	22 (79)	34 (79)	113 (64)	169 (68)
Technology assessment report	0	1 (2)	12 (7)	13 (5)

ICD-10 = International Classification of Diseases, Tenth Revision; RDT = Rapid diagnostic test; POCT = Point of care test.

* Includes 11 ICD-10 codes that had fewer than 10 reviews across the 3 groups.

† Includes questionnaires and home testing kits.

‡ Tests evaluated in a review were not of the same type.

Numbers in parentheses are column percentages unless otherwise stated. Percentages may not add up to 100% because of rounding.

(Adapted from Takwoingi et al 2013⁷¹)

5.3.2 Characteristics of reviews included in comparison of meta-analytic findings

Forty two reviews initially appeared eligible for the meta-analyses comparing summary estimates from different study types (Figure 5.1), but after checking the design of comparative and possibly comparative studies from full-text reports, only 39 were confirmed eligible.

Twelve seemingly comparative studies in the reviews were reclassified as non-comparative because allocation of patients to tests was non-random; after this reclassification, three reviews no longer met the requirement of having at least three comparative studies, which reduced the total number of eligible reviews from 42 to 39.

Evaluation of the full text of 13 possibly comparative studies that shared authors but had different citations found two additional comparative studies (the other 11 evaluated only one of the two tests). In the end, the 39 selected reviews contributed 55 pairwise test comparisons. A meta-analysis was performed separately for each test comparison. The 55 meta-analyses contained 1138 studies. Of these studies, 283 (25%) were comparative; one comparative study used a randomized design while the remaining 282 used a within-subject paired design. Table 5.2 shows the characteristics of the 39 reviews^{95,103,147,152-187} and test comparisons within them.

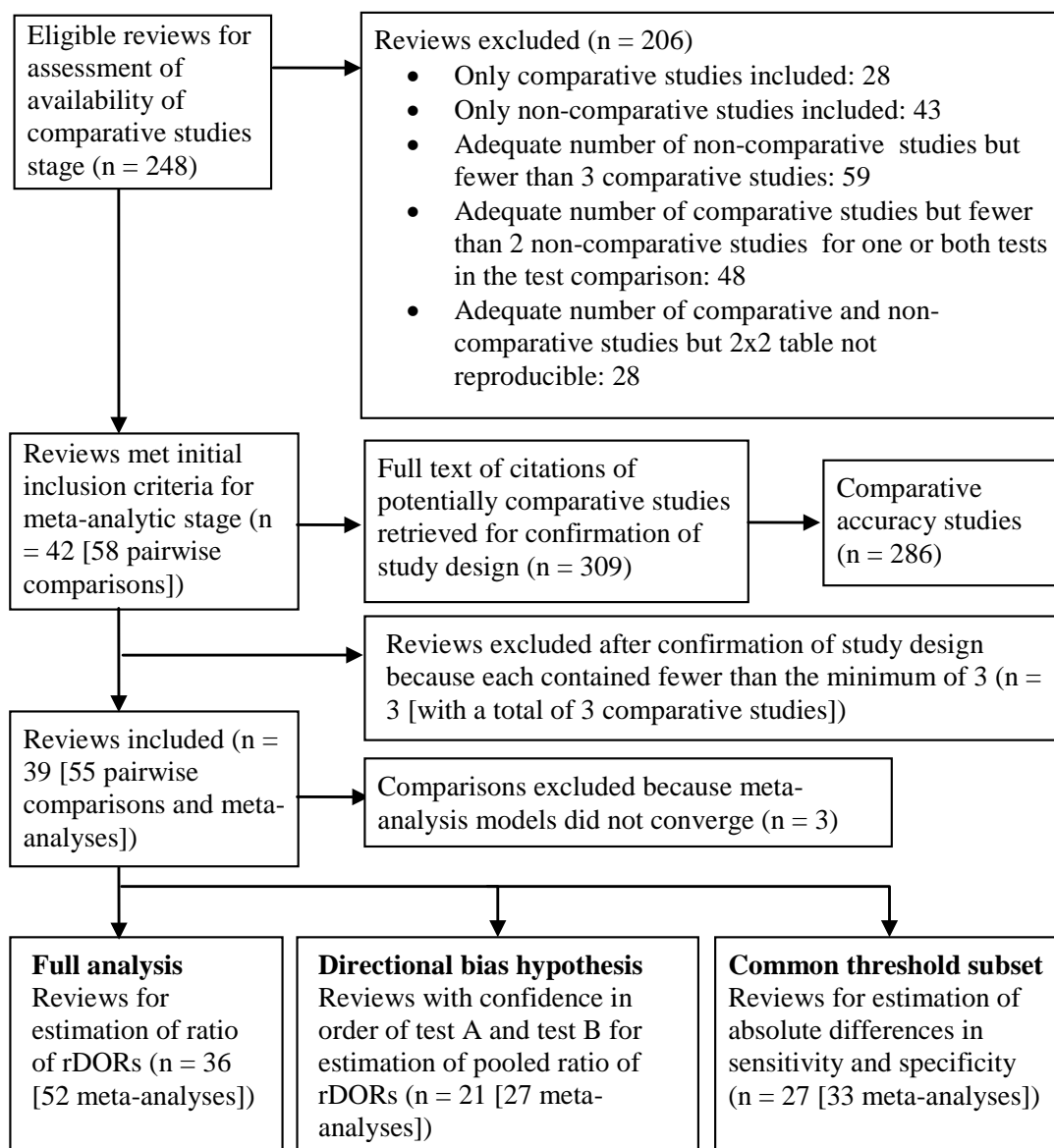


Figure 5.1| Flowchart of review and meta-analysis selection

DARE = Database of Abstracts of Reviews of Effects; rDOR = relative diagnostic odds ratio.

(Adapted from Takwoingi et al 2013⁷¹)

Table 5.2| Characteristics of included reviews and test comparisons in each meta-analysis

Review	Target condition	Test comparison		Number of studies				Number diseased/Total			
		Test A*	Test B†	Comparative	Non-comparative	Comparative	Non-comparative	Test A	Test B	Test A	Test B
Abba et al, 2011 ⁹⁵	<i>Plasmodium falciparum</i> malaria	Type 4 RDT [‡]	Type 1 RDT	7	9	58	3433/9761	3433/9764	841/3249	8533/30298	
Al-Khayal and Al-Omran, 2007 ¹⁵²	Equivocal acute appendicitis	CT	US	3	22	22	279/419	279/419	1291/3506	2462/8702	
Allred et al, 2012 ¹⁰³	Down syndrome	Total hCG, AFP and maternal age [‡]	Free fhCG, AFP and maternal age	4	11	8	176/7231	176/7231	297/126552	165/38366	
		Total hCG, AFP, uE3 and maternal age [‡]	Free fhCG, AFP and maternal age [‡]	5	19	7	199/9603	199/9603	449/79444	142/35994	
		Total hCG, AFP, uE3 and maternal age [‡]	Total hCG, AFP and maternal age [‡]	6	18	9	260/16666	260/16808	388/72381	213/116975	
Balk et al, 2001 ¹⁵³	Acute myocardial infarction	Presentation CK-MB [‡]	Presentation CK	6	13	6	270/1044	270/1050	777/5384	344/2102	
		Presentation CK-MB [‡]	Presentation myoglobin	13	6	5	582/3557	582/3556	465/2872	172/578	
		Presentation CK [‡]	Presentation myoglobin	6	6	12	280/972	281/974	334/2180	473/3160	
		Serial CK-MB [‡]	Presentation CK-MB	8	6	11	459/4044	467/4044	694/7639	580/2385	
Choi et al, 2010 ¹⁵⁴	Cervical cancer	PET	MRI	4	6	16	50/117	50/117	122/370	250/1319	
de Bondt et al, 2007 ¹⁵⁵	Head and neck cancer	CT	US	3	5	4	215/364	216/366	174/380	310/593	
Elamin et al, 2008 ¹⁵⁶	Cushing's syndrome	1 mg overnight DST [‡]	UFC	5	9	9	163/ 370	162/368	84/1695	483/4213	
Fleischmann et al, 1998 ¹⁵⁷		Exercise SPECT [‡]	Exercise ECHO	6	21	17	314/443	322/454	2211/2794	1425/2183	

Table 5.2 continued...

Review	Target condition	Test comparison	Number of studies				Number diseased/Total			
			Test A*	Test B†	Comparative	Non-comparative	Test A	Test B	Comparative	Non-comparative
Gisbert and Abraira, 2006 ¹⁵⁸	<i>Helicobacter pylori</i> infection in patients with upper gastrointestinal bleeding	Serology [‡] RUT	5	2	7	219/366	246/356	251/328	549/759	
Glas et al, 2003 ¹⁵⁹	Bladder cancer	Serology [‡] BTA BTA stat NMP22 [‡] CYFRA 21-1 [‡]	3 4 6 4 8	4 2 2 10 4	5 22 20 4 7	116/177 137/323 232/828 171/659 766/1339	114/178 137/323 229/742 171/653 766/1339	354/517 53/392 235/472 511/1631 124/281	436/540 1087/3121 995/2702 296/647 638/1608	
Gu et al, 2007 ¹⁶⁰	Malignant pleural effusion	PET-CT	5	6	5	128/183	128/183	220/285	111/165	
Gu et al, 2009 ¹⁶¹	Ovarian carcinoma	PET-CT	4	7	11	134/213	110/187	214/255	557/836	
Heim et al, 2004 ¹⁶²	Deep venous thrombosis	Automated rapid ELFA [‡]	4	5	2	199/424	199/425	306/915	294/574	
Mahajan et al, 2010 ¹⁶³	Left main and triple vessel coronary artery disease	SE [‡] MPI	6	8	9	76/508	76/468	178/895	573/1802	
Mant et al, 2009 ¹⁶⁴	Heart failure	ECG [‡] NT-proBNP NT-proBNP CEUS [‡]	6 6 7 5	5 10 9 2	3 13 4 16	1395/3181 437/1293 443/2245 71/236	1269/3181 437/1293 444/2260 71/236	300/1497 866/2936 860/1984 10/49	323/603 1701/3280 1251/2419 368/2374	
Mitchell et al, 2010 ¹⁶⁶	Depression	GDS ₁₅	4	6	3	178/1144	178/1144	295/1868	182/542	
Mowatt et al, 2010 ¹⁶⁷	Bladder cancer	FISH Cytology	5	7	32	428/1119	468/1198	411/1632	3276/13268	
Ngamruengphong et al, 2010 ¹⁶⁸	Esophageal cancer	NMP22 EUS	16 3	12 4	20 12	1220/5436 76/159	1311/5448 67/142	1724/4698 80/162	2346/8812 188/504	
Niekel et al, 2010 ¹⁶⁹	Colorectal liver metastases	FDG-PET [‡] CT	4	2	5	432/628	432/628	42/128	281/768	

Table 5.2 continued...

Review	Target condition	Test comparison		Number of studies				Number diseased/Total			
		Test A*	Test B†	Comparative	Non-comparative	Test A	Test B	Comparative	Non-comparative	Test A	Test B
Nishimura et al, 2007 ¹⁷⁰	Rheumatoid arthritis	anti-CCP antibody	RF	28	9	22	3962/8504	3792/8292	2338/6445	2508/7067	
Noguchi et al, 2005 ¹⁷¹	Coronary artery disease	Dipyridamole SE	Dobutamine SE	12	28	68	742/1028	742/1028	1703/2438	4812/6884	
Safdar et al, 2005 ¹⁷²	Intravascular device related bloodstream infection	Dipyridamole SE	Exercise SE	4	36	40	229/293	229/293	2216/3173	2315/3401	
Scheidler et al, 1997 ¹⁷³	Cervical cancer	MRI	CT	4	6	15	53/300	52/300	103/537	207/742	
Scherer et al, 2005 ¹⁷⁴	Myasthenia gravis	Anticholinesterase tests ^{‡,§}	Ice test	3	4	4	27/54	27/54	129/266	65/135	
Schuetz et al, 2010 ¹⁴⁷	Coronary artery disease	CT	MRI	5	84	14	159/334	142/307	3961/7192	381/671	
Selman et al, 2005 ¹⁷⁵	Vulvar cancer	SNB 99mTc ^{‡,§}	SNB blue dye	3	7	5	16/69	10/56	68/244	28/141	
Shiga et al, 2006 ¹⁷⁶	Thoracic aortic dissection	TEE ^{‡,§}	MRI	3	7	4	91/150	106/185	253/480	110/207	
Smith et al, 2011 ¹⁷⁷	Acetabular labral tears	MR arthrography	MRI	4	11	4	70/84	70/84	430/505	107/136	
Smith-Bindman et al, 2001 ¹⁷⁸	Down syndrome	NFT [‡]	FS	11	15	18	441/14729	430/11505	615/67824	492/9773	
St John et al, 2006 ¹⁷⁹	Urinary tract infection	NFT [‡]	HS	6	20	5	338/7488	323/7423	718/75065	138/4449	
Terasawa et al, 2004 ¹⁸⁰	Acute appendicitis in adults and adolescents	LE or nitrite	LE	9	6	5	941/6896	942/6896	497/2576	163/1068	
		CT	US	4	8	10	118/275	102/253	407/885	635/1240	

Table 5.2 continued...

Review	Target condition	Test comparison		Number of studies				Number diseased/Total			
		Test A*	Test B†	Comparative	Non-comparative	Comparative	Non-comparative	Test A	Test B	Test A	Test B
Tian et al, 2012 ¹⁸¹	<i>Helicobacter pylori</i> infection in patients with partial gastrectomy	UBT [‡]	RUT	4	5	3	114/196	119/204	200/362	82/228	
Tolosa et al, 2003 ¹⁸²	Lung cancer	PET	CT	8	10	12	200/564	196/578	137/481	742/2858	
van der Windt et al, 2010 ¹⁸³	Celiac disease	IgA-tTG [‡]	EmA	3	4	5	210/918	210/918	217/3363	73/861	
van Vliet et al, 2008 ¹⁸⁴	Esophageal cancer	EUS	CT	12	19	5	442/671	478/737	751/1170	105/206	
Wardlaw et al, 2006 ¹⁸⁵	Carotid stenosis	FDG-PET	EUS	5	5	26	161/243	145/222	102/181	1048/1619	
Xu et al, 2011 ¹⁸⁶	Head and neck cancer	MR angiography	DUS	4	8	4	192/475	227/501	79/299	87/215	
Yin et al, 2010 ¹⁸⁷	Scaphoid fracture	PET-CT	PET	3	4	5	11/147	11/147	77/640	121/648	
		MRI	BS	4	6	11	33/223	33/224	76/289	116/878	

AFP = alpha-fetoprotein; anti-CCP = anti-cyclic citrullinated peptide; βhCG = beta human chorionic gonadotrophin; BNP = B-type natriuretic peptide; BS = bone scintigraphy; BTA = bladder tumour antigen; CA 125 = cancer antigen 125; CEA = carcinoembryonic antigen; CEUS = contrast enhanced ultrasound; CXR = chest X-ray; CK = creatinine kinase; CK-MB = creatine kinase-MB; CT = computed tomography; CYFRA 21-1 = Cytokeratin fragment 19; DST = dexamehasone suppression test; DUS = Doppler ultrasound; ECG = electrocardiogram; ECHO = echocardiography; ELFA = enzyme-linked fluorescent assay; ELISA = enzyme-linked immunosorbent assay; EmA = IgA antiendomysial antibodies; EUS = endoscopic ultrasonography; FDG-PET = fluorine18 fluorodeoxyglucose positron emission tomography; FISH = fluorescence in situ hybridisation; FS = femur shortening; GDS15 = geriatric depression scale (15-item questionnaire); GDS30 = geriatric depression scale (30-item questionnaire); hCG = human chorionic gonadotrophin; HS = humerus shortening; IgA-tTG = IgA tissue transglutaminase; LE = leukocyte esterase; MPI = myocardial perfusion imaging; MR = magnetic resonance; MRI = magnetic resonance imaging; NFT = nuchal fold translucency; NMP22 = nuclear matrix protein; NT-proBNP = N-terminal pro-B-type natriuretic peptide; PET = positron emission tomography; PET-CT = positron emission tomography/computed tomography; QCSC = quantitative catheter segment culture; RDT = rapid diagnostic test; RF = rheumatoid factor; RUT = rapid urease test; SE = stress echocardiography; SNB = sentinel node biopsy; SNB

^{99m}Tc = sentinel node biopsy using technetium-99m-labelled nanocolloid; SPECT = single photon emission computed tomography; SQSC = semi-quantitative catheter segment culture; TEE = transesophageal echocardiography; UBT = urea breath test; uE3 = unconjugated oestriol; UFC = urinary free cortisol; US = ultrasound; USS = duplex ultrasound.

*Test A = Experimental or newer test.
†Test B = Current practice or older test.
‡Test comparisons where the ordering of test A and test B was uncertain. These were excluded from the pooled analysis of the ratio of relative diagnostic odds ratios.
§Test comparisons for which the HSROC meta-analytic model failed to converge.
(Adapted from Takwoingi *et al* 2013⁷¹)

Six reviews^{147,152,169,176,180,185} included only prospective studies. Few reviews (23%) used the same reference standard in all included studies to verify disease status. Eleven reviews^{95,103,147,152,157,163,167-170,180} performed a meta-analysis restricted to comparative studies (or summarised the studies in a narrative) in addition to a primary meta-analysis based on all studies. The remaining 28 reviews did not directly compare tests or compared tests mixing together both comparative and non-comparative studies.

5.3.3 Evidence of difference in meta-analyses of comparative and non-comparative studies

The findings were based on 52 meta-analyses (derived from studies in 36 reviews) because analyses of the HSROC model failed to converge for three¹⁷⁴⁻¹⁷⁶ of the 55 meta-analyses. Figure 5.2 shows ratios of rDORs comparing meta-analytic estimates from non-comparative studies with estimates from comparative studies. Ten (19%) meta-analyses showed qualitative changes: the estimates from non-comparative studies ranking the two tests in a comparison in the opposite order of those of comparative studies. Twenty-five (48%) meta-analyses produced more than a two-fold difference in rDOR between results from the two study types, with as much as a 10-fold difference in four meta-analyses.

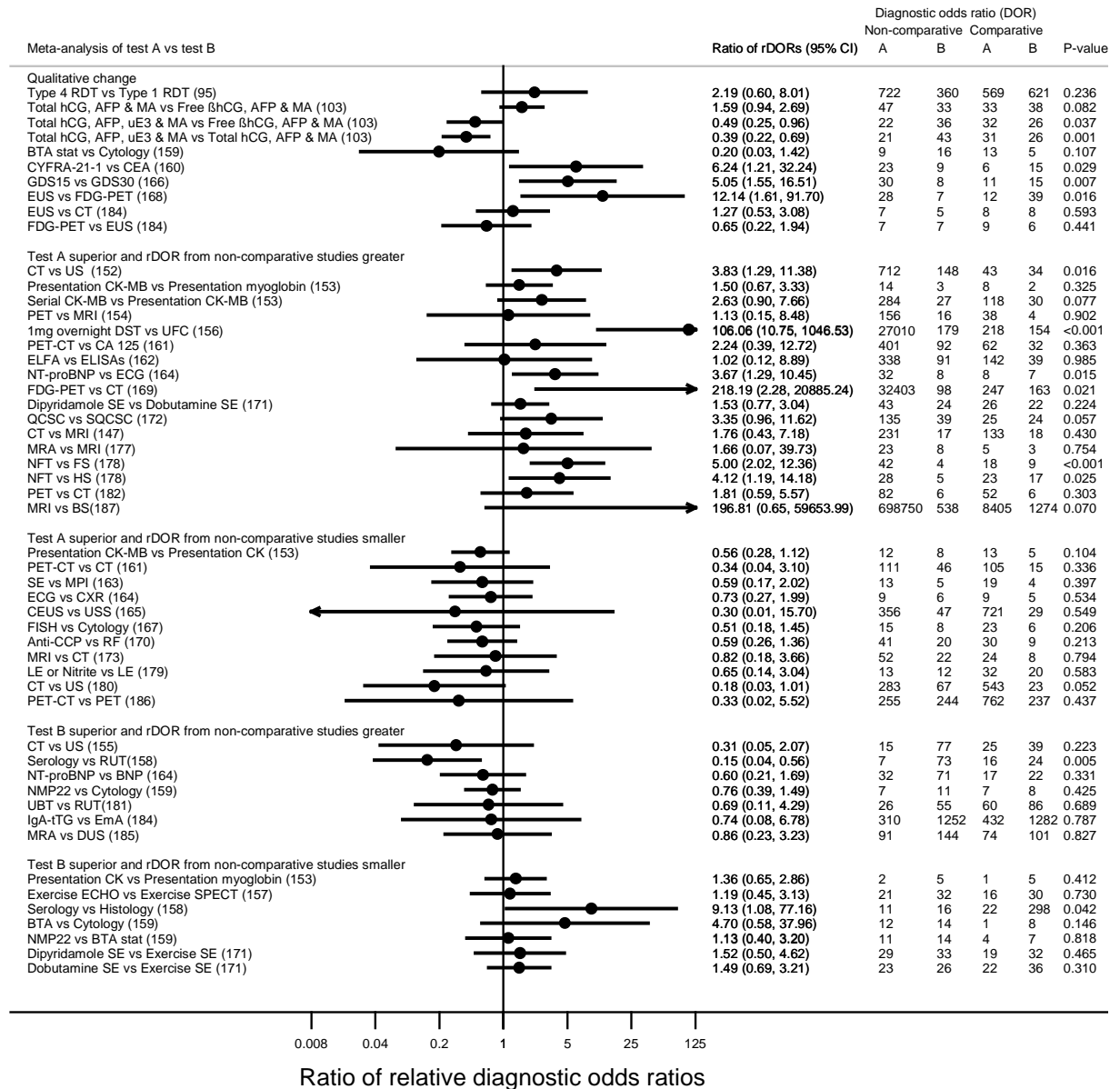


Figure 5.2| Ratio of relative diagnostic odds ratios (with 95% confidence intervals)

A = test A; AFP = alpha-fetoprotein; anti-CCP = anti-cyclic citrullinated peptide; B = test B; ßhCG = beta human chorionic gonadotrophin; BNP = B-type natriuretic peptide; BS = bone scintigraphy; BTA = bladder tumour antigen; CA 125 = cancer antigen 125; CEA = carcinoembryonic antigen; CEUS = contrast enhanced ultrasound; CXR = chest X-ray; CK = creatinine kinase; CK-MB = creatine kinase-MB; CT = computed tomography; CYFRA 21-1 = Cytokeratin fragment 19; DOR = diagnostic odds ratio; DST = dexamethasone suppression test; DUS = Doppler ultrasound; ECG = electrocardiogram; ECHO = echocardiography; ELFA = enzyme-linked fluorescent assay; ELISA = enzyme-linked immunosorbent assay; EmA = IgA antiendomysial antibodies; EUS = endoscopic ultrasonography; FDG-PET = fluorine18 fluorodeoxyglucose positron emission tomography; FISH = fluorescence in situ hybridisation; FS = femur shortening; GDS15 = geriatric depression scale (15-item questionnaire); GDS30 = geriatric depression scale (30-item questionnaire); hCG = human chorionic gonadotrophin; HS = humerus shortening; IgA-tTG = IgA antitissue transglutaminase antibodies; IgM RF = rheumatoid factor (Immunoglobulin M subtype); LE = leukocyte esterase; MA = maternal age; MPI = myocardial perfusion imaging; MR =

magnetic resonance; MRI = magnetic resonance imaging; NFT = nuchal fold translucency; NMP22 = nuclear matrix protein; NT-proBNP = N-terminal pro-B-type natriuretic peptide; PET = positron emission tomography; PET-CT = positron emission tomography/computed tomography; QCSC = quantitative catheter segment culture; RDT = rapid diagnostic test; ; rDOR = relative diagnostic odds ratio; RUT = rapid urease test; SE = stress echocardiography; SPECT = single photon emission computed tomography; SQSC = semi-quantitative catheter segment culture; UBT = urea breath test; uE3 = unconjugated oestriol; UFC = urinary free cortisol; US = ultrasound; USS = duplex ultrasound.

Test comparisons are grouped according to meta-analytic findings—whether rDOR was greater than or less than 1 and ratio of rDOR was greater than or less than 1, or whether non-comparative studies ranked tests in opposite order of the comparative studies. The P value is the probability of a ratio of rDORs at least as extreme as the observed ratio, assuming no difference in comparative accuracy between non-comparative and comparative studies. The numbers in parentheses in the first column are the reference citations for the reviews from which the meta-analysis was obtained.

(Adapted from Takwoingi et al 2013⁷¹)

In the one-sided contour-enhanced funnel plot (Figure 5.3), 13 points were observed to the right of the 5% contour line indicating statistically significant differences between estimates from comparative and non-comparative studies. This observed proportion (13/52, 25%) was higher than the 5% expected by chance ($P < 0.001$). It should be noted that points from the same review, i.e. reviews that contributed multiple pairwise test comparisons, are likely to be correlated. Therefore, in a sensitivity analysis, one point was randomly selected from each review. This analysis included 36 instead of 52 points (see Appendix C.1). Eight of the 36 points (22%) were above the contour line. The observed proportion was also higher than the 5% expected by chance ($P < 0.001$).

It was possible to classify tests as newer or older in 27 of the 52 meta-analyses, allowing investigation of direction of the differences. The pooled estimate of the ratios of rDORs was 1.15 (95% CI 0.81 to 1.64), a difference that was not statistically significant ($P = 0.4$). In sensitivity analysis where only one randomly chosen estimate was included per review, the

pooled estimate of the ratios of rDORs from the 21 meta-analyses was 1.07 (95% CI 0.69 to 1.65; P = 0.8).

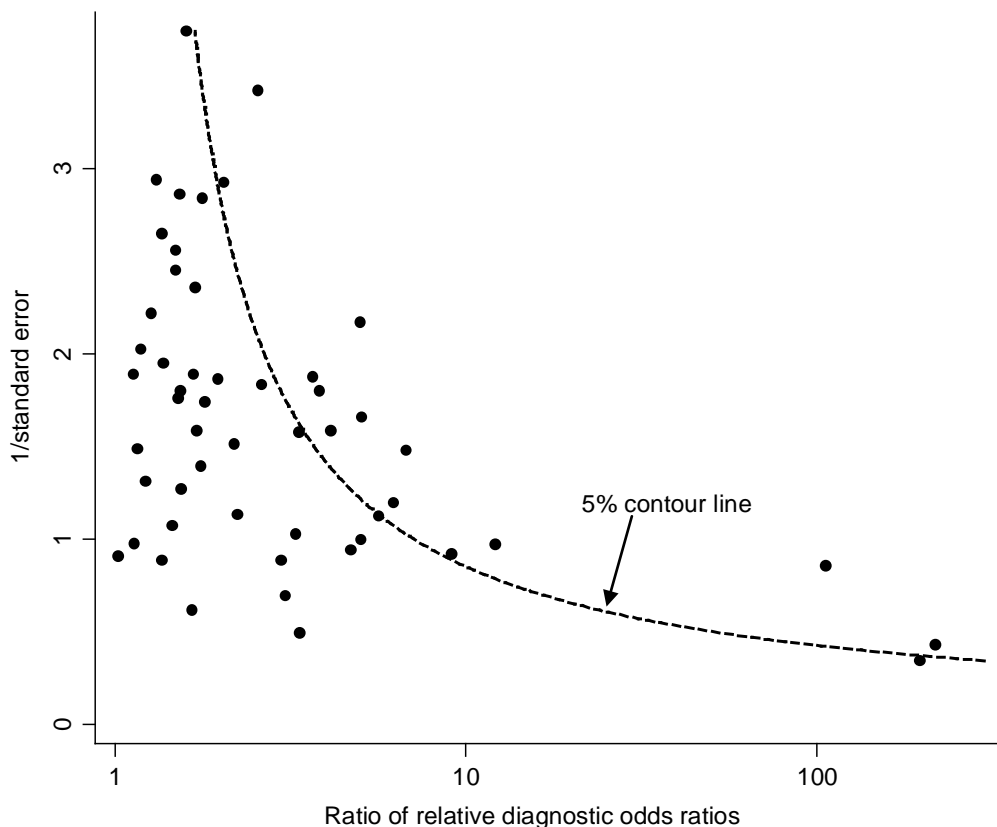


Figure 5.3| One-sided contour-enhanced funnel plot of the ratio of relative diagnostic odds ratio

By chance, assuming comparative and non-comparative studies give the same summary estimates, 5% of the points would be expected to lie above the contour line indicating statistical significance at the 5% level. Order of tests was chosen such that ratios were greater than one.

(Adapted from Takwoingi et al 2013⁷¹)

Thirty-two of the 52 meta-analyses evaluated tests at a common threshold that allowed computation of average sensitivities and specificities. Figure 5.4 shows the magnitude of the differences in sensitivities and specificities between tests from using non-comparative studies compared with comparative studies.

Six meta-analyses showed discrepancies in estimates of differences in sensitivity between comparative and non-comparative studies that were greater than 10%, and four showed discrepancies in differences in specificity greater than 10%. For sensitivity, differences were in opposing directions in seven (22%) meta-analyses and similarly for specificity in five (16%) meta-analyses. For example, in the comparison of endoscopic ultrasonography (EUS) and fluorine-18 fluorodeoxyglucose positron emission tomography (FDG-PET) for esophageal cancer,¹⁶⁸ the non-comparative studies gave a difference in summary sensitivity of 4% (95% CI -11% to 20%), suggesting that EUS was marginally more sensitive than FDG-PET, whereas the comparative studies gave a difference in sensitivity of -14% (95% CI -29% to 1%), suggesting that FDG-PET was much more sensitive than EUS. Similarly, for differences in the summary specificities, the non-comparative studies gave a difference of 17% (95% CI 2% to 32%), suggesting that EUS was much more specific than FDG-PET, whereas the comparative studies gave a difference of -3% (95% CI -15% to 10%), suggesting that FDG-PET was marginally more specific than EUS.

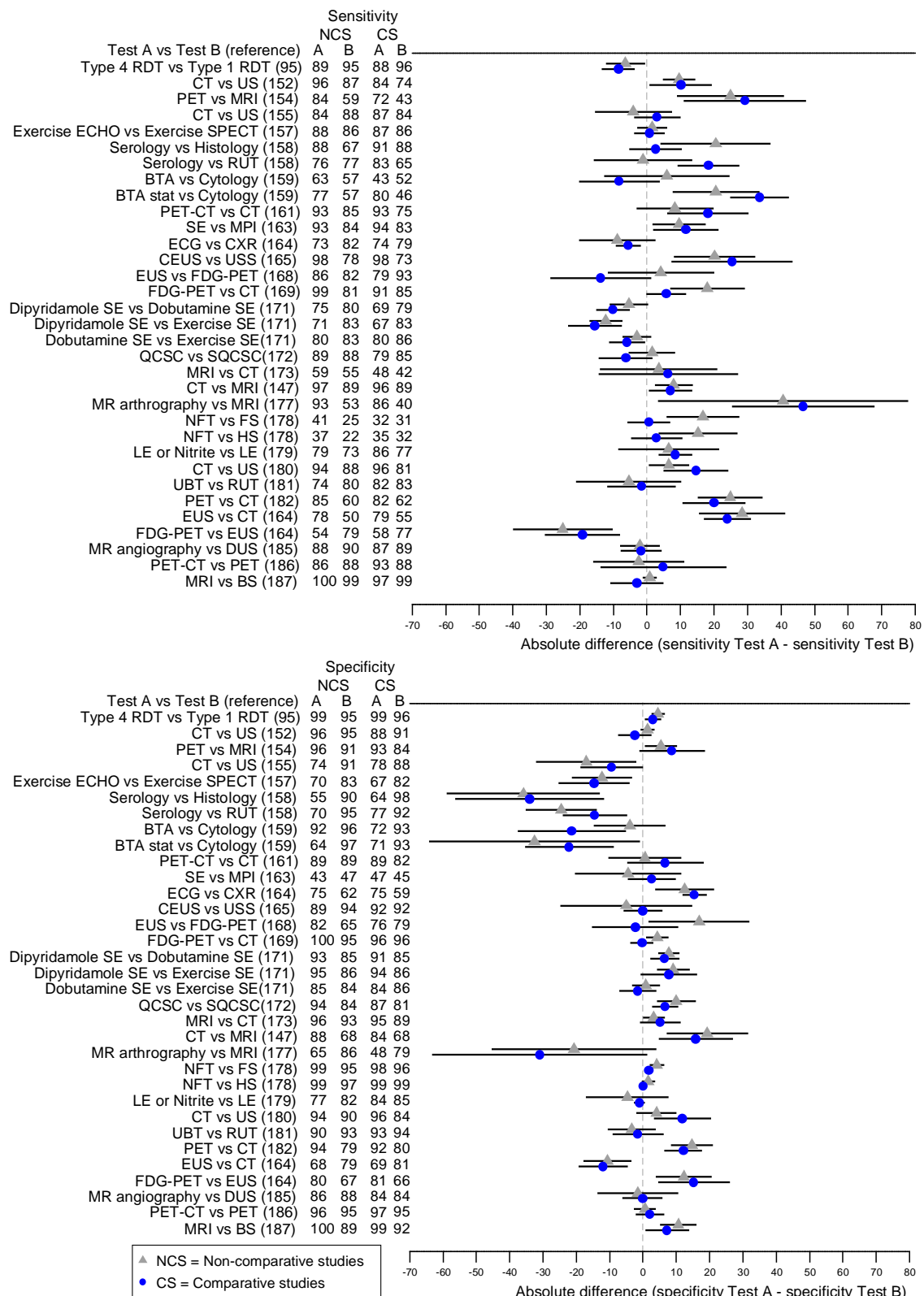


Figure 5.4| Absolute differences in sensitivity and specificity between tests (with 95% confidence intervals)

BS = bone scintigraphy; BTA = bladder tumour antigen; CEUS = contrast enhanced ultrasound; CXR = chest X-ray; CT = computed tomography; DUS = Doppler ultrasound; ECG = electrocardiogram; ECHO = echocardiography; EUS = endoscopic ultrasonography;

FDG-PET = fluorine18 fluorodeoxyglucose positron emission tomography; FS = femur shortening; HS = humerus shortening; LE = leukocyte esterase; MPI = myocardial perfusion imaging; MR = magnetic resonance; MRI = magnetic resonance imaging; NFT = nuchal fold translucency; PET = positron emission tomography; PET-CT = positron emission tomography/computed tomography; QCSC = quantitative catheter segment culture; RDT = rapid diagnostic test; RUT = rapid urease test; SE = stress echocardiography; SPECT = single photon emission computed tomography; SQCSC = semi-quantitative catheter segment culture; UBT = urea breath test; US = ultrasound; USS = duplex ultrasound

The numbers in parentheses in the first column are the reference citations for the reviews from which the meta-analysis was obtained.

(Adapted from Takwoingi et al 2013⁷¹)

5.4 Discussion

This study and the review of reviews (Chapter 4) have shown that it is common practice for systematic reviews that compare the accuracy of two or more tests to include different sets of studies for each test (i.e. non-comparative studies) due to paucity of comparative studies.

There was evidence that meta-analyses of non-comparative studies are likely to yield summary estimates that significantly differ from those of comparative studies that use a within-subject multiple test or between-subject randomized design. There was no evidence of a systematic direction in the differences but several examples were identified in which the difference was of a magnitude likely to lead to an effect on policy and practice.

5.4.1 Possible explanations

Differences in estimates between the two study types could have arisen for two reasons. First, studies of one test may systematically differ from those of the other test, in delivery of the index tests, patient characteristics or study methods, which may lead to bias in indirect estimates of comparative accuracy. For example, in a review comparing multislice computed tomography and magnetic resonance imaging for ruling out coronary artery disease, Schuetz and colleagues¹⁴⁷ reported that although almost all computed tomographic studies evaluated all coronary artery segments, most of the magnetic resonance imaging studies evaluated only

proximal vessels or segments. Through such a mechanism, meta-analysis of non-comparative studies would confound the estimate of comparative accuracy, with effects introduced by differences in the vessels imaged. This was interpreted as evidence of a greater risk of bias in meta-analysis of non-comparative studies, explained through lack of like-with-like comparisons, leading to interpretation of the difference between comparative and non-comparative studies as indicating bias in the non-comparative studies.

Second, because this study itself was observational, it is possible that systematic differences between non-comparative studies and comparative studies may partially explain the observed differences in some reviews. For instance, in a comparison of rheumatoid factor with anti-cyclic citrullinated peptide antibody for diagnosis of rheumatoid arthritis, non-comparative studies were more likely to be published before 2000, and earlier studies tended to report higher sensitivity and specificity for rheumatoid factor than studies published later.¹⁷⁰ It is thus impossible to interpret the evidence shown in this study as being conclusive about bias in non-comparative studies.

Biases may be present in direct and indirect comparisons. In the absence of knowledge about *true* estimates of comparative accuracy for a specific test comparison, one cannot say with certainty whether direct and/or indirect comparisons are biased; other factors may distort the true difference between tests rather than systematic errors in study design and conduct. If there is a lack of comparability between studies of different tests in an indirect comparison, the comparison will suffer from biases due to confounding. Consequently, the validity of an indirect comparison relies on the different sets of studies for each test being similar, on average, in factors that may affect test accuracy. Empirical evidence about potential sources

of bias in comparative accuracy studies is lacking. Nonetheless, one may argue that because comparative studies compare test accuracy within the same study population, the risk of bias would be lower than can be expected when comparisons are made between studies particularly when heterogeneity is present.

Others have noted that studies of new tests reported high sensitivities and specificities significantly more often than did studies of existing tests.¹⁸⁸ but here there was no clear evidence of a direction in the differences observed in the main analysis in this empirical study. This was contradicted by the sensitivity analysis which gave a larger and significant difference, favouring the newer test relative to the older test or current practice. Song and colleagues¹⁸⁹ also observed no directional bias in treatment effects in indirect comparisons of interventions. Given the difference in conclusions between the main and sensitivity analyses reported here, it is difficult to make a firm conclusion about a systematic direction to the effect across studies. Therefore, a future update of this empirical study should explore this issue further.

5.4.2 Comparison with existing evidence

This is the first study to provide empirical evidence across multiple topics of the effect of using non-comparative studies on meta-analytic estimates of comparative test accuracy. Previous empirical research has focused on the effect of features of the design, execution, analysis, and reporting of test accuracy studies on estimates of performance for a single test.^{136-138,190} The potential for bias in estimates from meta-analysis that rely on non-comparative studies has been argued from a theoretical viewpoint only in guidance, such as that from the Cochrane Collaboration's Diagnostic Test Accuracy initiative²³ and the

Australian Medical Services Advisory Committee.¹⁸ This study provides empirical evidence in support.

5.4.3 Implications for research and practice

The findings have important implications for the design of future primary studies of test accuracy, for systematic reviews and meta-analyses that compare the accuracy of tests, and also for clinical practice.

Besides the view that comparative studies are likely to be less prone to bias, comparative studies also allow assessment of consistency in the direction of the difference in accuracy, enable comparison of cases of uninterpretable or indeterminate results between tests, and can be used to explore and develop test combinations for improving diagnostic accuracy. Given the merits of comparative test accuracy studies, they deserve more appreciation by clinical investigators, researchers, and grant-awarding organizations funding test research.

Although there were some good examples in which review authors were explicit about the type of studies they used to provide evidence for the superior accuracy of one or more tests over other tests, it was not possible to assess the type of studies included in 28 reviews because only summary results were presented for each test and authors did not report on the type of studies included. When all eligible studies that have evaluated at least one of the tests of interest are included in a comparative meta-analysis, identification of the comparative studies in the review and investigation of the effect of study type on results is recommended. Journal editors can facilitate the implementation of appropriate reporting (as suggested in section 4.4) by requiring review authors to present a clear analytic strategy for comparing

tests and to identify the type of evidence included. As the demand for evidence-based diagnosis grows, test accuracy reviews will increasingly be commissioned by health technology agencies and guideline developers. The number of reviews being published has increased tremendously, from fewer than 10 per year in the early 1990s to almost 100 per year in recent years.²⁷

For clinicians, the results warrant caution in interpreting the results of meta-analyses comparing tests. The highest level of evidence in such a case would be a well conducted meta-analysis that contains only high quality studies that evaluated both (or all) tests against the same reference standard, providing that these studies include studies done in a setting and patient spectrum similar to one's own. In addition, meta-analyses that contain both comparative and non-comparative studies may be valuable, especially if they show that the estimates from the non-comparative studies do not differ from the estimates in the comparative studies. If such analyses are not provided, or if a meta-analyses only contains non-comparative studies, then the results need to be used with caution.

5.4.4 Strengths and limitations

For pragmatic reasons, study design was not confirmed for all primary studies included in the assessment of the availability of comparative studies. This limitation may have under- or overestimated the number of comparative studies. The margin of error is unknown but unlikely to overturn conclusions given the aim was to assess the availability of comparative studies and not to provide a definitive estimate of the number in the literature. Related to this is the classification of studies as non-comparative. The term non-comparative was carefully chosen to reflect the notion that a primary accuracy study may have assessed one or more

tests but if only one of the tests is relevant to a comparative review question, then it is not a (head-to-head) comparative study. Therefore, the term also captures studies of a single test. This is contrary to an intervention review where non-comparative RCTs can be included in an indirect comparison or mixed treatment comparison if there is a common comparator.

This study was based on an extensive database of systematic reviews of diagnostic tests but due to the low availability of comparative studies, the number of test comparisons for the meta-analytic stage was limited. In making best use of the available data, multiple comparisons from the same reviews were selected, which may have been correlated in their findings. Addition of further evidence as more comparisons appear over time will increase the robustness of the findings and should be undertaken.

Although comparative studies are advocated, study design alone will not lead to valid estimates of comparative accuracy because other aspects of study methods and conduct can also affect test performance. It is worth noting that comparative studies can also be prone to bias. Studies that created groups by using non-random allocation are at risk for confounding. Only 10 studies of this design were found, and results were not sensitive to the categorization of these studies. Other threats to the validity of multiple test comparative studies exist. In clinical care, patients whose test results are uninformative often undergo further testing—if a multiple test study is created by identifying such patients, its results are unlikely to be generalisable. A further requirement is that the tests are used in a standard way and are independent. If, for example, one test alters the condition of the patient (for example, through injecting a contrast dye), or if the result of one imaging test is interpreted with knowledge of

the other, then the results of the test comparison may be biased. The possibility of these limitations was not assessed in this study.

The existence of bias due to study type cannot be proven in a meta-epidemiologic study of this nature, and it is acknowledged that findings could be attributable to other differences in study or patient characteristics that were not investigated because of data limitations and the widely recognized issue of poor reporting of test accuracy studies. However, this further strengthens the argument for comparing tests within the same population and study.

5.4.5 Conclusions

Between meta-analysis comparison of tests based on non-comparative studies may currently be the major source of evidence available to guide decision making for many tests, but the results should be interpreted with caution because variation in patient groups, study design, reference standard, and other sources can confound differences in test performance.

Comparative accuracy evidence obtained from robustly designed comparative studies should be regarded as representing a higher level of evidence. When alternative tests exist, important comparative questions that reflect the clinical context in which the tests will be used should be addressed in designing future test accuracy studies. Where possible, analyses limited to comparative studies should also be conducted and reported along with the analysis of all studies in systematic reviews.

6 METHODOLOGICAL REVIEW OF STATISTICAL METHODS FOR COMPARATIVE META-ANALYSIS

6.1 Introduction

Hierarchical meta-regression models (see section 1.5.4) are recommended by Cochrane for comparative meta-analysis.^{23,43} The HSROC model⁴² was published in 2001 and the bivariate model adapted for test accuracy meta-analysis by Reitsma et al⁴¹ was published in 2005. The current version of the statistical analysis chapter of the Cochrane Handbook for Systematic Reviews of Diagnostic Accuracy²³ was published in 2010. Scientific literature is not static and there have been methodological advances since publication of the two models and the handbook. The literature on methods for meta-analysis of a single test has been prominent,^{36,40,49,191-199} and little is known about available methods for comparative meta-analysis.

The issues discussed thus far in this thesis would suggest that different methods for comparative meta-analyses may be needed depending on the type of test comparison (direct or indirect), type of data (e.g. common or mixed thresholds), and availability of studies. Therefore, the aims of this chapter are to identify methods for comparative meta-analysis and to characterise the methods, including hierarchical meta-regression models and any extensions of the models. The specification (including notation and any underlying assumptions), strengths and weaknesses, and software requirements of each approach will be examined in order to provide an overview of the available methods. In Chapter 7, relevant methods will then be empirically assessed in order to provide practical guidance for meta-analysts and evidence based recommendations in an update of the Cochrane handbook mentioned above.

The outline of this chapter is as follows. In section 6.2, methods for selection and data extraction of the meta-analytic methods identified from the search results reported in section 3.5 are outlined. In section 6.3, the meta-analytic methods are explained and only illustrated with an example if required for highlighting the impact of modelling choices within a particular method. Section 6.4 concludes the chapter by summarising the characteristics of the methods, including their strengths and limitations.

6.2 Identification of comparative meta-analysis methods

Published and unpublished papers and presentations describing meta-analytic methods for comparing test accuracy were identified as outlined in section 3.3. Based on knowledge of previous and current methodological developments in comparative meta-analysis, methods that specifically address simultaneous comparisons of several tests, as illustrated in the network plot shown in Figure 2.2, were likely to be non-existent or developed within a Bayesian framework. Although the thesis emphasises methods implemented within a frequentist framework, Bayesian methods that are identified will be summarised in order to provide a comprehensive overview of all available methods. However, Bayesian methods will not be empirically evaluated as already explained in section 1.4. The following data were extracted from each methods paper or presentation: publication details, underlying principles and assumptions, and model specification. To gauge the popularity of the methods, the number of citations per year was obtained for each paper from Scopus®, the largest abstract and citation database of peer-reviewed literature.²⁰⁰

6.3 Statistical methods for comparative meta-analysis of test accuracy

Thirteen papers and two presentations describing 13 methods were identified from the searches (see section 3.5 for full results of the search) and are summarised in Table 6.1. Two of the 13 papers were systematic reviews. No additional novel methods were identified in the 53 reviews that statistically compared test accuracy (see Appendix B.3) that were described in section 4.3.2.4. Sensitivity and specificity or the DOR were the outcome measures used in most (nine out of 13) of the methods.

Table 6.1| Meta-analytic methods for comparing test accuracy

	Reference	Citations in Scopus	Method	Test accuracy measure
1	Moses et al 1993 ³⁷	775	Comparison of Q*	Q*
	Littenberg and Moses 1993 ¹³³	283		
2	Hasselblad and Hedges 1995 ²⁰¹	260	Standardized distance between the means of two populations	Effectiveness measure (<i>d</i>) proportional to log DOR
3	Rutter and Gatsonis 2001 ⁴²	289	HSROC meta-regression	DOR
4	Kowalski et al 2001 ^{202*}	7	Generalized estimating equation	Sensitivity and specificity
5	Lijmer et al 2002 ²⁰³	232	Moses SROC meta-regression	DOR
6	Worster et al 2002 ²⁰⁴	56	General linear mixed model	Likelihood ratios
7	Suzuki et al 2004 ²⁰⁵	38	Conditional relative odds ratio	DOR
8	Siadaty and Shu 2004 ²⁰⁶	12	Proportional odds ratio	DOR
9	Siadaty et al 2004 ²⁰⁷	14	Repeated measures modelling	DOR
10	Reitsma et al 2005 ⁴¹	634	Bivariate meta-regression	Sensitivity and specificity
	Hamza et al 2009 ²⁰⁸	1		
11	Trikalinos et al 2014 ²⁰⁹	0	Bivariate analysis of paired data	Sensitivity and specificity
12	Cheng et al 2013 ^{210†}	N/A	Network meta-analysis	Sensitivity and specificity
13	Verde 2013 [†]	N/A	Bivariate analysis of paired data	Sensitivity and specificity

N/A = not applicable because they are unpublished.

*This systematic review was also identified as part of the cohort of reviews for the empirical evaluation in this chapter.

†Conference presentation

Thirteen papers and two conference presentations describing 13 methods are listed according to year of publication. Citations were not applicable for presentations.

The number of citations for each paper was obtained from Scopus on 8th January 2016. The Moses SROC model,³⁷ first published in 1993, is the earliest of the methods while extensions of hierarchical models are the most recent. The number of citations per year from 1993 to

2015 is shown in Figure 6.1 for the three most popular models—the Moses SROC model, bivariate model and HSROC model—and are based solely on the original papers that proposed each method. The citations are not specific to systematic reviews (comparative or single test reviews) but also include methodology papers, and other article types. Nonetheless, growing popularity of hierarchical models can be inferred from Figure 6.1. In particular, citations of the Reitsma et al⁴¹ paper, the first paper to adapt the bivariate model to test accuracy meta-analysis, has grown rapidly since its publication in 2005 even though the specification of the binomial likelihood for modelling within-study variability that was later suggested by Chu and colleagues^{53,54} is preferred (see section 1.4.4.1).

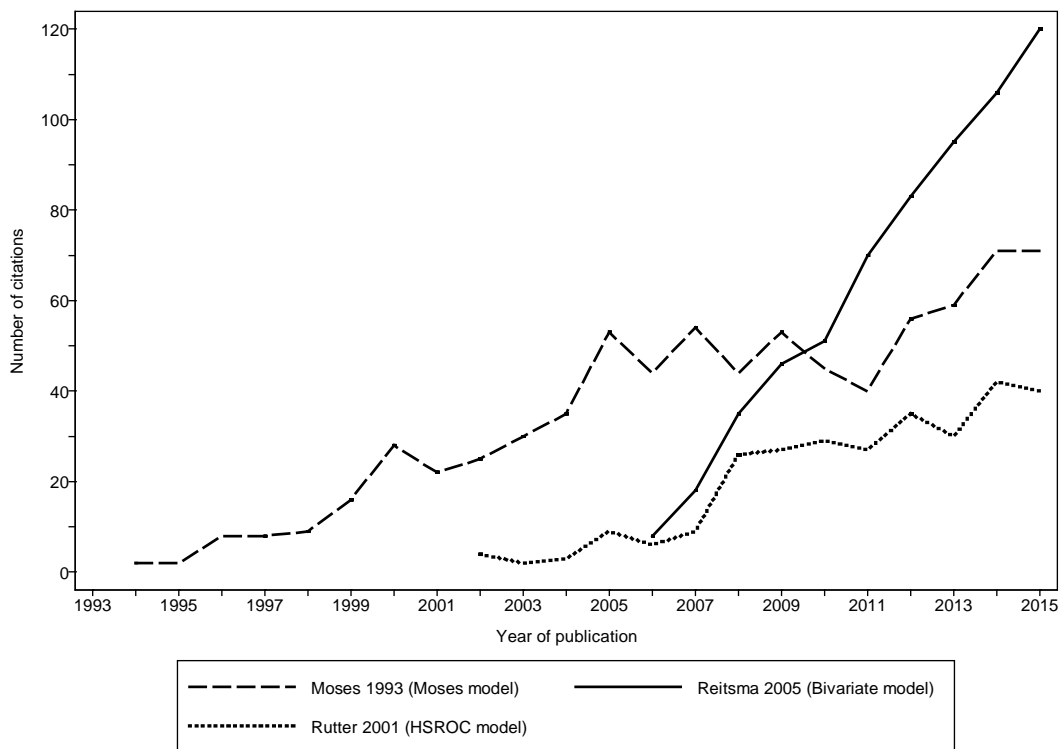


Figure 6.1| Number of citations per year for original publications of the Moses model, bivariate model and HSROC model for test accuracy meta-analysis
 Number of citations for each paper was obtained from Scopus on 08/01/2016.

The 13 comparative meta-analysis methods are discussed below. Methods that do not account for correlated data arising from paired test results are the focus of sections 6.3.1 to 6.3.3, while in section 6.3.4, methods that specifically address correlated data are considered.

6.3.1 Methods for comparing summary points

Meta-analysis of treatment effects is based on pooling statistics (e.g. risk ratio) which contrast outcomes between two groups— typically experimental and control. On the contrary, DTA meta-analysis pool statistics (e.g. sensitivity and specificity) which summarise outcomes in a single group—those receiving the index test—without contrasting that value with a comparator. In this section, methods that focus on estimating summary points such as summary sensitivities, specificities and likelihood ratios from a comparative meta-analysis are described.

6.3.1.1 General linear mixed model for comparing likelihood ratios

A linear mixed model approach for pooling likelihood ratios was described in a systematic review comparing the accuracy of non-contrast helical CT (NHCT) and intravenous pyelography (IVP) for diagnosis of acute urolithiasis.²⁰⁴ In the review, Worster et al first computed likelihood ratios for each study and then estimated 95% CIs using the method of Simel et al.²¹¹ The LR+ and LR– for a single study can be calculated as

$$\text{LR+} = \frac{\text{sensitivity}}{1-\text{specificity}},$$

and

$$\text{LR-} = \frac{1-\text{sensitivity}}{\text{specificity}}.$$

Simel et al,²¹¹ showed that likelihood ratios are algebraically identical to relative risk ratios and so variances of log likelihood ratios can be approximated as follows:

$$\text{var}(\log\text{LR}+) = \frac{1-\text{sensitivity}}{\text{TP}} + \frac{\text{specificity}}{\text{FP}}$$
$$\text{var}(\log\text{LR}-) = \frac{\text{sensitivity}}{\text{FN}} + \frac{1-\text{specificity}}{\text{TN}},$$

where TP, FP, FN and TN are the number of true positives, false positives, false negatives and true negatives. From the equations given above, it is apparent that computations of likelihood ratios and their variances require a continuity correction if the cells of any of the 2x2 tables contains zero. Following computation of likelihood ratios and their variances for each individual study, the likelihood ratios were then log-transformed and a general linear model with test type (NHCT or IVP) as a fixed effect and study as a random factor was applied to pool LR+ and LR- separately. The dependent variable (log LR+ or log LR-) was weighted by the inverse of its variance as suggested by Deeks.²¹² The software package used was not stated.

Zwinderman and Bossuyt⁵⁵ have shown that univariate (fixed or random effects) meta-analysis or bivariate meta-analysis of likelihood ratios using a bivariate normal distribution, can give nonsensical results due to a mis-specified statistical model. In an illustration, they show that a pooled LR+ and 1/LR- of 1.34 and 0.94 correspond to a sensitivity of -0.31 and specificity of +1.23. Therefore, simple univariate or bivariate meta-analysis of log transformed likelihood ratios is discouraged. Consequently, comparative meta-analysis of likelihood ratios will not be discussed further in this thesis. As already noted in section 1.4.4.1, if likelihood ratios are of interest, they should be derived from the parameters of a bivariate model.

6.3.1.2 Bivariate meta-regression of sensitivity and specificity

The bivariate model focuses on estimation of summary sensitivities and specificities as previously detailed in sections 1.4.4.1 and 1.5.4.1. Bivariate meta-regression models were explained in section 1.5.4.1, applied to examples in Chapter 2, and used in seven (10%) of the reviews identified in Chapter 4 (see list in Appendix B.3).

The within-study likelihood for a bivariate model with a covariate for test type, t (indexed by k), can be written as

$$y_{Aik} \sim \text{Binomial}(n_{Aik}, g^{-1}(\mu_{Aik})), y_{Bik} \sim \text{Binomial}(n_{Bik}, g^{-1}(\mu_{Bik})), \quad (6.1)$$

where y_{Aik} and y_{Bik} represent the number of true positives and true negatives, n_{Aik} and n_{Bik} the number of diseased and non-diseased subjects, and $g^{-1}(\mu_{Aik})$ and $g^{-1}(\mu_{Bik})$ the sensitivity and specificity for the k th test from the i th study. The logit link $g(\cdot)$ is commonly used.⁴⁰ A bivariate normal distribution is used for modelling between-study variation.

Assuming common variances between tests, this is written as

$$\begin{pmatrix} \mu_{Aik} \\ \mu_{Bik} \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_A + v_A t_k \\ \mu_B + v_B t_k \end{pmatrix}, \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix} \right), \quad (6.2)$$

where μ_A estimates the expected logit sensitivity for the referent test (note not the reference standard but a comparator) and $\mu_A + v_A t_k$ estimates the expected logit sensitivity for the k th test. Thus, $\exp(v_A)$ gives the estimated odds ratio for the sensitivity of the k th test relative to that of the referent test. The same applies to specificity where μ_B is the expected logit specificity for the referent test and $\mu_B + v_B t_k$ estimates the expected logit specificity for the k th test. The variances are σ_A^2 and σ_B^2 for the logit sensitivities and logit specificities, and σ_{AB}

is the covariance between the logits across studies. The model that allows for separate variances for each test can be written as

$$\begin{pmatrix} \mu_{Aik} \\ \mu_{Bik} \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_A + v_A t_k \\ \mu_B + v_B t_k \end{pmatrix}, \begin{pmatrix} \sigma_{Ak}^2 & \sigma_{ABk} \\ \sigma_{ABk} & \sigma_{Bk}^2 \end{pmatrix} \right). \quad (6.3)$$

where σ_{Ak}^2 and σ_{Bk}^2 are the variances for the logit sensitivities and logit specificities for the k th test, and σ_{ABk} is the covariance between the logits across studies evaluating the test.

Based on the author's experience of reviewing the literature, the variance-covariance structure in equation 6.3 is typically modelled assuming independence between tests because most comparative meta-analyses are based on indirect comparisons with few or no comparative studies. For two tests, this can be expressed as

$$\begin{bmatrix} \mu_{Ai1} \\ \mu_{Ai2} \\ \mu_{Bi1} \\ \mu_{Bi2} \end{bmatrix} \sim N \left(\begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}, \Sigma \right) \text{ with } \Sigma = \begin{bmatrix} \sigma_{A1}^2 & 0 & \sigma_{A1B1} & 0 \\ & \sigma_{A2}^2 & 0 & \sigma_{A2B2} \\ & & \sigma_{B1}^2 & 0 \\ & & & \sigma_{B2}^2 \end{bmatrix} \quad (6.4)$$

The means $\mu_A = \begin{pmatrix} \mu_{A1} \\ \mu_{A2} \end{pmatrix}$ and $\mu_B = \begin{pmatrix} \mu_{B1} \\ \mu_{B2} \end{pmatrix}$ are column vectors of the means of logit sensitivities and logit specificities for the two tests. For direct comparisons, the bivariate model can allow for correlation in test performance between tests by estimating all between-study and between test variability using the following unstructured variance-covariance matrix:

$$\begin{bmatrix} \mu_{Ai1} \\ \mu_{Ai2} \\ \mu_{Bi1} \\ \mu_{Bi2} \end{bmatrix} \sim N \left(\begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}, \Sigma \right) \text{ with } \Sigma = \begin{bmatrix} \sigma_{A1}^2 & \sigma_{A1A2} & \sigma_{A1B1} & \sigma_{A1B2} \\ & \sigma_{A2}^2 & \sigma_{A2B1} & \sigma_{A2B2} \\ & & \sigma_{B1}^2 & \sigma_{B1B2} \\ & & & \sigma_{B2}^2 \end{bmatrix}, \quad (6.5)$$

Note that potential within-study correlation between tests is not taken into account—requires individual patient data or aggregate data in the form shown in Table 1.3 which is seldom

reported in primary studies. This is also a common issue for multivariate meta-analysis of multiple outcomes where the within-study correlations needed to fit the multivariate model are unavailable from primary studies.²¹³ According to Riley et al, the within-study correlation is most influential in a multivariate meta-analysis when the within-study variation is large relative to the between-study variation in the underlying true study values, and the converse is true for the between-study correlation.²¹⁴ An example of a network meta-analysis of multiple outcomes showed changes in the ranking of treatments due to accounting for correlation between multiple outcomes.²¹⁵ The impact of ignoring within-study correlation on joint inferences about differences in test performance is yet to be shown in empirical or simulation studies for test comparisons. There may be biological/clinical justification for other variants of (6.5), such as assuming no correlation between the sensitivity of one test and the specificity of another test (i.e. $\rho_{A_1B_2} = 0$ and $\rho_{A_2B_1} = 0$), but is beyond the scope of this thesis. Comparative meta-analysis methods that explicitly account for paired data are described in section 6.3.4.

Since equations 6.2 and 6.4 were used to model between-study variability in comparative meta-analyses presented so far in this thesis, it is worth examining bivariate models with different parameterisations of the covariance matrix as expressed in equations 6.2, 6.4 and 6.5. The example used below was chosen because it included a good number (13) of paired accuracy studies for illustrating the potential effect of the three covariance structures on findings. The review by Kittler et al²¹⁶ is a direct comparison of dermoscopy versus inspection by the unaided eye (i.e. clinical inspection without dermoscopy) for detection of melanoma. Figure 6.2 shows the estimates of sensitivity and specificity for each pair of tests from each of the 13 included studies.

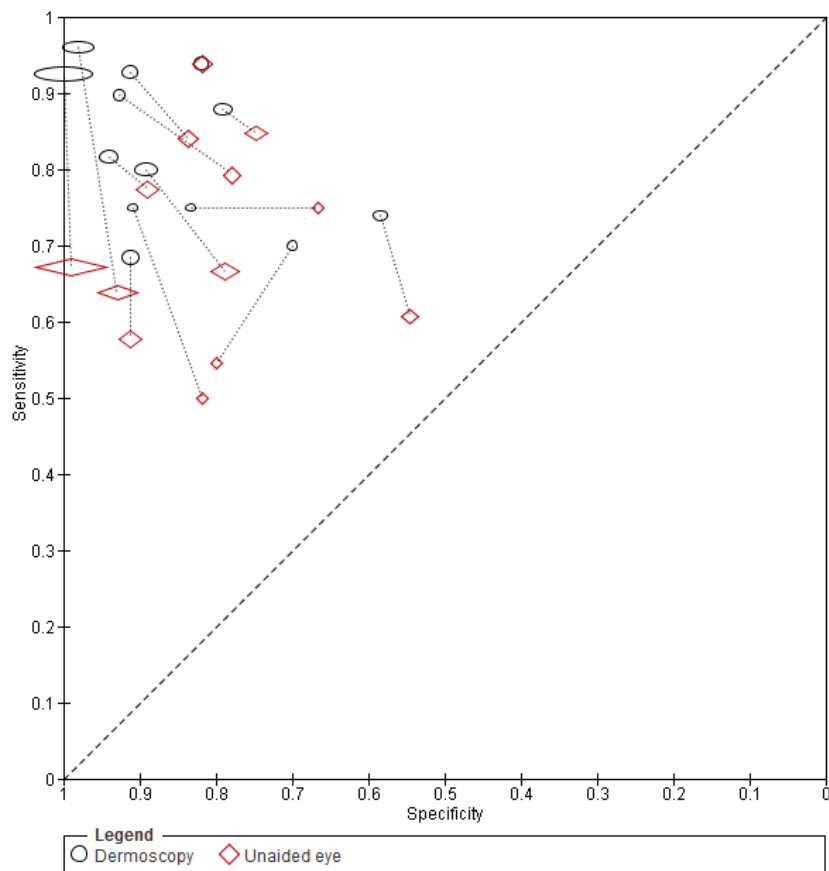


Figure 6.2| SROC plot of dermoscopy and unaided eye for diagnosis of melanoma

Point estimates for dermoscopy and unaided eye from the same study are connected by a dotted line. Each study point was scaled by the sample sizes for diseased and non-diseased groups to reflect precision of the sensitivity and specificity estimates.

Bivariate models that assume equal variances across tests (Model 1, equation 6.2), and unequal variances without dependence (Model 2, equation 6.4), and with dependence (Model 3, equation 6.5) between tests were fitted to the data using maximum likelihood estimation as described in section 2.2.5. Based on the likelihood ratio test comparing Model 2 and Model 3, there was statistical evidence of a difference in model fit ($P < 0.0001$). Parameter and summary estimates from the three models are presented in Table 6.2. As the magnitude of correlation is more readily interpreted by readers, correlations are presented instead of covariances.

Table 6.2| Parameter and summary estimates from bivariate models with increasing complexity of the variance-covariance structure

Estimate	Model		
	(1) Equal variances across tests	(2) Unequal variances without dependence between tests	(3) Unequal variances with dependence between tests
Model parameter			
σ_A^2	0.312		
σ_B^2	1.393		
ρ_{AB}	0.053		
σ_{A1}^2		0.339	0.387
σ_{A2}^2		0.308	0.319
σ_{B1}^2		3.366	2.873
σ_{B2}^2		1.075	1.015
ρ_{A1B1}		0.530	0.456
ρ_{A2B2}		-0.244	-0.175
ρ_{A1A2}			0.775
ρ_{A1B2}			0.446
ρ_{A2B1}			-0.187
ρ_{B1B2}			0.994
Summary estimate (95% CI)			
Sensitivity of dermoscopy	0.86 (0.80, 0.90)	0.86 (0.80, 0.90)	0.85 (0.79, 0.90)
Specificity of dermoscopy	0.92 (0.85, 0.96)	0.93 (0.82, 0.97)	0.92 (0.82, 0.97)
Sensitivity of unaided eye	0.73 (0.65, 0.80)	0.74 (0.66, 0.81)	0.74 (0.66, 0.81)
Specificity of unaided eye	0.85 (0.74, 0.92)	0.86 (0.77, 0.92)	0.85 (0.76, 0.91)
Relative sensitivity	1.16 (1.08, 1.26)	1.16 (1.03, 1.30)	1.15 (1.05, 1.25)
Relative specificity	1.08 (1.03, 1.14)	1.08 (0.96, 1.21)	1.08 (1.05, 1.11)

σ_A^2 = variance of logit sensitivity across both tests; σ_{A1}^2 = variance of logit sensitivity for dermoscopy; σ_{A2}^2 = variance of logit sensitivity for unaided eye; σ_B^2 = variance of logit specificity across both tests; σ_{B1}^2 = variance of logit specificity for dermoscopy; σ_{B2}^2 = variance of logit specificity for unaided eye; ρ_{AB} = correlation of logit sensitivity and logit specificity across both tests; ρ_{A1B1} = correlation of logit sensitivity and logit specificity for dermoscopy; ρ_{A1A2} = correlation of logit sensitivities for dermoscopy and unaided eye; ρ_{A1B2} = correlation of logit sensitivity for dermoscopy and logit specificity for unaided eye; ρ_{A2B1} = correlation of logit sensitivity for unaided eye and logit specificity for dermoscopy; ρ_{A2B2} = correlation of logit sensitivity and logit specificity for unaided eye; ρ_{B1B2} = correlation of logit specificities for dermoscopy and unaided eye.

Comparing Models 2 and 3 to Model 1, it is apparent that variances of the random effects for the logit sensitivities and the logit specificities, as well as the correlation between the logits, differed between dermoscopy (test 1) and unaided eye (test 2). It is not unreasonable to expect

positive dependence between the logit sensitivities and between the logit specificities of two tests while negative dependence can be expected between the logit sensitivities and logit specificities due to threshold effects. However, heterogeneity can distort the relationships. Also, as mentioned earlier, one can speculate that between-study correlations may be misspecified because within-study correlations of test results in paired studies are ignored. This cannot be verified without individual patient data or within-study correlations from primary studies.

The summary estimates and their 95% CIs from the three models were similar except for relative specificity. In contrast to the other two models, the difference in specificity derived from Model 2 was not statistically significant. The example illustrates that different models may give similar estimates of summary sensitivities and specificities yet lead to different conclusions about relative accuracy. This finding is similar to those of the case study presented in section 2.3.4.1. Other examples will be investigated in the empirical evaluation of comparative meta-analyses methods in Chapter 7 to determine if such findings are common.

6.3.2 Methods for comparing SROC curves

The two methods that compare SROC curves are the Moses SROC regression and the HSROC model. Both models were introduced in section 1.4 for analysis of a single test.

6.3.2.1 Moses SROC approach

The DOR, AUC and Q^* have been used to compare test accuracy in the Moses SROC approach as observed in Chapter 4. The AUC or Q^* of two tests can be compared using a z -

test.³⁷ The AUC will not be discussed further in this thesis due to the limitations outlined in section 1.4.3. The Q^* statistic and other test comparison approaches that use a z -test or t -test to compare a test accuracy measure will be examined briefly in section 6.3.3. As observed in Chapter 4 (section 4.3.2.3), the Moses SROC approach is still a commonly used meta-analytic method, either alone or in conjunction with other methods, even though it has important methodological limitations noted in section 1.4.3. Regardless of its popularity for meta-analysis of a single test, only two of the 53 reviews listed in Appendix B.3 used a Moses SROC meta-regression model to compare tests. Given the prevalence of the Moses SROC approach in published reviews and the lack of empirical or simulation studies of the performance of the Moses SROC meta-regression approach for comparing tests, the meta-regression approach is discussed extensively below.

Moses SROC meta-regression

The meta-regression approach proposed by Moses et al³⁷ for investigating heterogeneity was adapted for test comparisons by using a covariate for test type to investigate effect of test type on diagnostic accuracy (log DOR).^{203,217,218} The regression equation in (1.2) can be extended to include an indicator variable for test type (k). The model becomes

$$D_i = a + b_0 S_i + b_1 k_i + e_i \quad (6.6)$$

with

$$e_i \sim N(0, \sigma^2).$$

The intercept is a , b_0 and b_1 are the regression coefficients for S and k , and e_i is the random error. Recall D (difference in the logits) is the natural log of the DOR and S (sum of the logits) is a proxy for threshold computed as follows for the i th study:

$$D_i = \ln\left(\frac{TPR_i}{1-TPR_i}\right) - \ln\left(\frac{FPR_i}{1-FPR_i}\right)$$

$$S_i = \ln\left(\frac{TPR_i}{1-TPR_i}\right) + \ln\left(\frac{FPR_i}{1-FPR_i}\right).$$

The meta-regression model specified in (6.6) assumes the shape of the SROC curve does not differ between tests and represents parallel lines with common slope b_0 in the (S, D) space.

Thus b_1 , the coefficient of k estimates the difference in the log DOR between two tests (vertical distance between the two regression lines) and the exponent of b_1 is the relative DOR. Moses et al suggested that if $b = 0$ in separate models (equation 1.2) for each test, then a t -test can be applied to compare values of D for two tests.³⁷ This eliminates $b_1 S$ in (6.6) and equates to

$$D_i = a + b_1 k_i + e_i, \quad (6.7)$$

for an unpaired t -test.

The model in (6.6) can be extended to include an interaction term (kS) that allows the shape of the SROC curve to differ between tests as follows:

$$D_i = a + b_0 S_i + b_1 k_i + b_2 k_i S_i + e_i. \quad (6.8)$$

In this interaction model, the regression lines are no longer parallel and relative accuracy depends on threshold (S).

The Moses model is generally implemented using simple (weighted or unweighted) linear regression and so the model can be fitted using any statistical package. The weight for each study i in an inverse variance weighted SROC regression, w_i , is computed as $1/V_i$ where

$$V_i = \frac{1}{TP_i} + \frac{1}{FP_i} + \frac{1}{FN_i} + \frac{1}{TN_i}.$$

The variance (similar to computation of D and S) is undefined if any of the four cells of the 2x2 table is zero.

To illustrate the impact of weighting the regression analyses as well as assumptions about the shape of the SROC curves (equations 6.6 and 6.8) on findings, a review that compared the performance of rapid enzyme-linked immunosorbent assay (ELISA) and standard radioimmunosorbent assay (RIA) tests for diagnosis of congestive heart failure²¹⁹ was chosen. The forest plot of the data in Figure 6.3 shows that four studies had a zero cell and so a continuity correction of 0.5 was added to the cells of the 2x2 table for each study included in the meta-analysis as is typically done in practice.

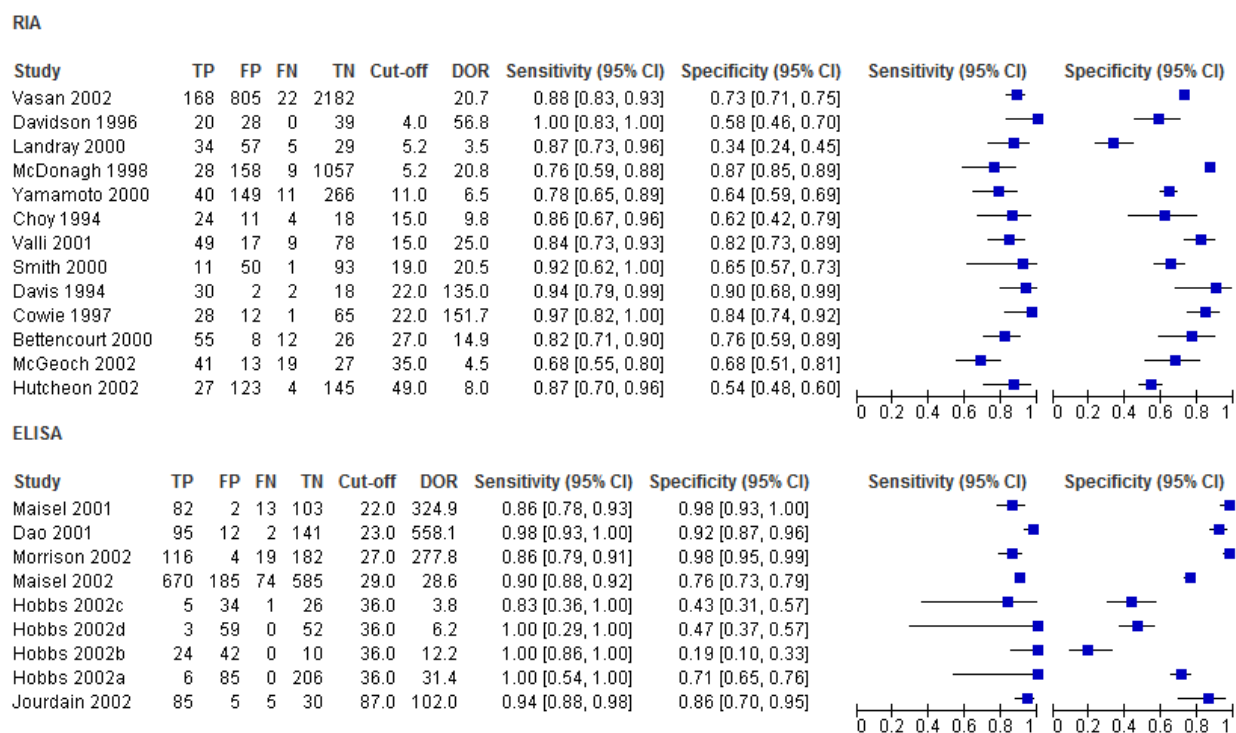


Figure 6.3| Forest plot of RIA and ELISA for diagnosis of congestive heart failure

DOR = diagnostic odds ratio; ELISA = enzyme-linked immunosorbent assay; RIA = radioimmunosorbent assay.

Studies were ordered on the plot according to cut-off and DOR. The Hobbs 2002 study included four different non-overlapping population cohorts and results were given separately for each cohort hence the study names with suffixes a to d. Data were extracted from Battaglia et al.²¹⁹

Assuming the same shape for the SROC curves of both tests (equation 6.6), Figure 6.4 shows the unweighted regression lines of D on S for both tests in panel A and the corresponding SROC curves in panel B. The SROC curves were produced by applying equation 1.5 (see section 1.4.3) to the range of specificities for each test.

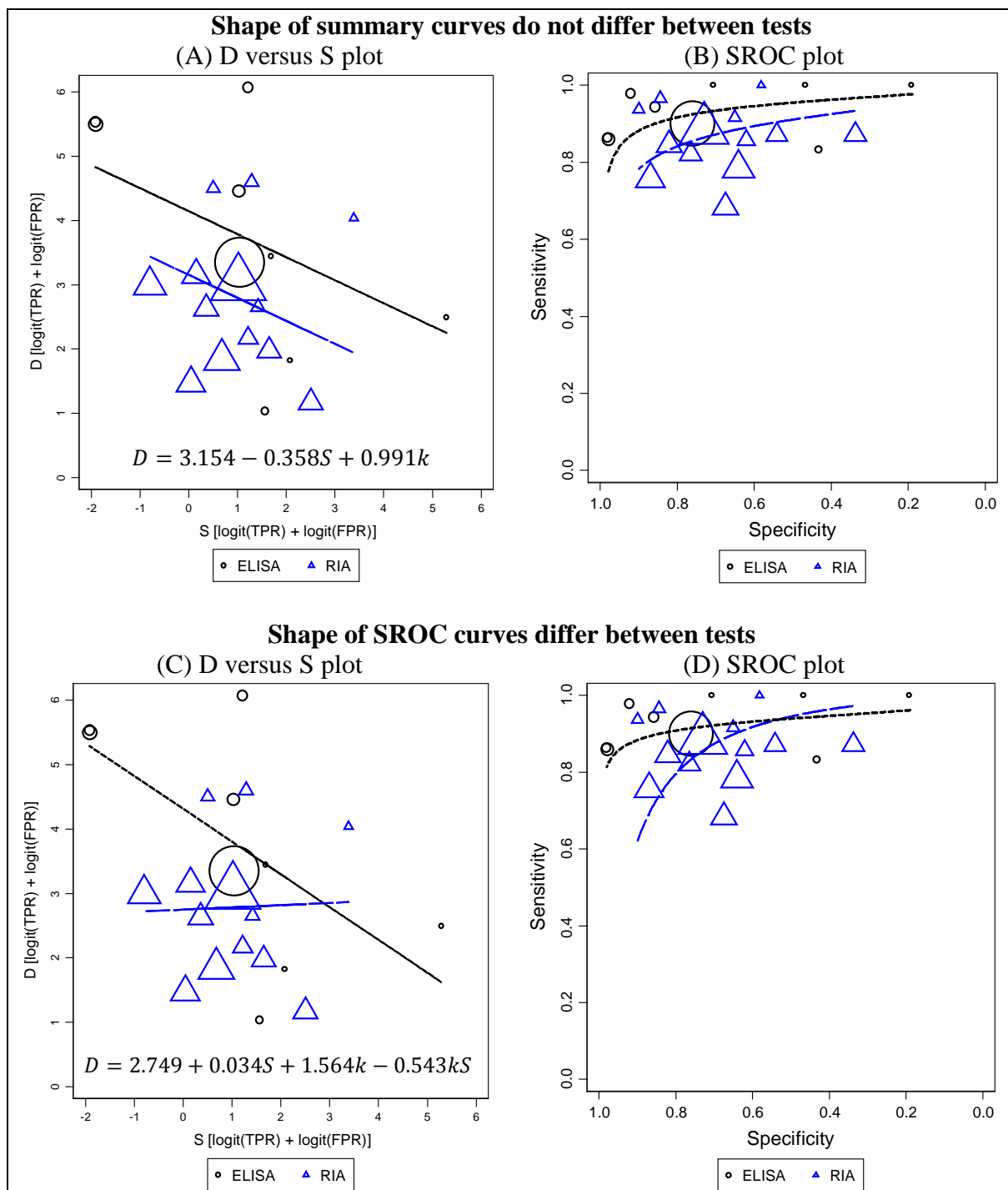


Figure 6.4| Comparison of summary curves from Moses SROC meta-regression models
 ELISA = enzyme-linked immunosorbent assay (blue triangles and line with long dashes); RIA = radioimmunosorbent assay (black circles and line with short dashes).

The models were fitted using unweighted linear regression and the regression equation is shown on the D versus S plot for each of the two models. The regression equation for the plots in panels A and B does not include the interaction term (kS), i.e. assumes same shape across tests, while the regression equation for the plots in panels C and D include the interaction term thus allowing shape to differ by test.

The unweighted regression gave an rDOR of 2.69 (95% CI 0.82 to 8.87; $P = 0.098$) meaning the DOR of the ELISA test was 2.69 times higher than that of the RIA test. For the weighted analysis (Table 6.3), the rDOR was 3.03 (1.50 to 6.12; $P = 0.004$). In this example, the difference in diagnostic accuracy from unweighted and weighted regression differed in magnitude and statistical significance.

A limitation of the unweighted analysis is that equal weight is given to all studies thus ignoring the precision of each included study. If there is much greater between-study variability relative to within-study variability, the unweighted analysis approximates a random effects model because the weights of the studies in the latter will be dominated by the between-study variance and so studies are weighted relatively more equally.^{37,57,217} In contrast, an analysis weighted by the inverse of the variance of the DOR accounts for the precision of each study but may be biased.²¹⁷ It is known that for high DORs, such as is frequently encountered in test accuracy meta-analysis where the DOR is much larger than 1, the standard error of the DOR is correlated with the DOR.²²⁰ For example, in Figure 6.3, the DORs (standard error) for Hobbs 2002a and Hutcheon 2002 are 31.4 (1.47) and 8.0 (0.52). Hutcheon 2002 had lower test performance and greater weight (3.66) in the meta-analysis compared to the weight (0.46) for Hobbs 2002a even though both studies have similar sample sizes (299 and 297 respectively). Greater weight is given to studies that report poorer test accuracy (because cells have similar counts) compared to those reporting higher test accuracy (i.e. lower false negative and/or false positive counts) even if studies have similar sample sizes.

Table 6.3| Comparison of unweighted and weighted Moses SROC meta-regression models

Parameter	Unweighted regression Estimate (95% CI); P value	Weighted regression Estimate (95% CI); P value
<i>Summary curves do not differ between tests ($D = a + b_0S + b_1k$)</i>		
a	3.154 (2.295, 4.013)	2.889 (2.324, 3.455)
b_0	-0.358 (-0.741, 0.025); P = 0.065	-0.441 (-0.832, -0.050); P = 0.029
b_1	0.991 (-0.202, 2.183); P = 0.098	1.108 (0.405, 1.811); P = 0.004
<i>Summary curves differ between tests ($D = a + b_0S + b_1k + b_2kS$)</i>		
a	2.749 (1.699, 3.799)	2.713 (2.061, 3.365)
b_0	0.034 (-0.679, 0.748); P = 0.921	-0.202 (-0.792, 0.388); P = 0.481
b_1	1.564 (0.095, 3.033); P = 0.038	1.437 (0.508, 2.367); P = 0.004
b_2	-0.543 (-1.383, 0.296); P = 0.191	-0.424 (-1.209, 0.362); P = 0.272

a is the intercept, b_0 and b_1 are the regression coefficients for S and k , and b_2 is the coefficient of the interaction term (kS) in the models.

For the model that allowed the shape of the SROC curves to differ by test (equation 6.8), the plots are shown in panels C and D of Figure 6.4. The parameter estimates from the unweighted and weighted analyses are presented in Table 6.3. From both analyses, it can be seen that the interaction between the shape of the curve and test type (b_2) was not statistically significant and so the same shape can be assumed for both tests. A weighted interaction model was used by de Vries et al²¹⁸ to compare the accuracy of duplex and colour-guided duplex ultrasonography for peripheral arterial disease. Similar to the results above, they found that the slopes of the two regression lines were nearly identical and so the interaction term was removed from their final model. Issues regarding weighting and assumptions about the shape of SROC curves will be investigated further using an empirical cohort in Chapter 7.

6.3.2.2 HSROC meta-regression

The HSROC model also defines test accuracy in terms of the DOR. HSROC meta-regression models were explained in section 1.5.4.2 and extensively illustrated in Chapter 2 using the bipolar disorder example described in section 2.2.2. Of the 53 reviews listed in Appendix B.3, 10 (19%) used a HSROC meta-regression model. HSROC meta-regression models were also used in Chapter 5 to compare summary estimates from meta-analyses of direct and indirect comparisons. In the HSROC model, the number of test positives from the i th study, y_{ij} , is assumed to follow a binomial distribution

$$y_{ij} \sim \text{Binomial}(n_{ij}, \pi_{ij}).$$

For the non-diseased group $j = 0$ and for the diseased group $j = 1$, and n_{ij} is the number in group j . The model incorporating an indicator variable for test type, t , can be written as

$$\text{logit}(\pi_{ij}) = \left((\theta_i + \gamma t_i) + (\alpha_i + \xi t_i) \text{dis}_{ij} \right) \exp(-(\beta + \delta t_i) \text{dis}_{ij}), \quad (6.9)$$

where π_{ij} is the probability of a positive result, and dis_{ij} is coded -0.5 for $j = 0$ and 0.5 for $j = 1$. The model includes β the shape parameter. θ is the threshold parameter and α is the accuracy parameter and both parameters are modelled as random effects with independent normal distributions as follows:

$$\theta_i \sim N(\Theta, \sigma_\theta^2) \text{ and } \alpha_i \sim N(\Lambda, \sigma_\alpha^2). \quad (6.10)$$

In this model, γ assesses the difference in the underlying threshold between tests, ξ assesses the difference in test accuracy, and δ assesses the difference in shape of the curves. If a common shape is assumed (see example in section 2.3.3), equation 6.9 reduces to

$$\text{logit}(\pi_{ij}) = \left((\theta_i + \gamma t_i) + (\alpha_i + \xi t_i) \text{dis}_{ij} \right) \exp(-\beta \text{dis}_{ij}), \quad (6.11)$$

Models that make different assumptions about the shape of the SROC curves between tests will be empirically assessed and compared with the Moses SROC meta-regression models described in section 6.3.2.1.

If differences in accuracy and threshold are modelled as random effects, the model specified in (6.9) and (6.10) becomes

$$\text{logit}(\pi_{ij}) = ((\theta_i + \gamma_i t_i) + (\alpha_i + \xi_i t_i) \text{dis}_{ij}) \exp(-(\beta + \delta t_i) \text{dis}_{ij}), \quad (6.12)$$

and

$$\theta_i \sim N(\Theta, \sigma_\theta^2), \gamma_i \sim N(\Gamma, \sigma_\gamma^2), \alpha_i \sim N(\Lambda, \sigma_\alpha^2), \text{ and } \xi_i \sim N(\Xi, \sigma_\xi^2). \quad (6.13)$$

In the HSROC framework, the random effects for accuracy and threshold are modelled using independent normal distributions. For two tests, if all the variances in equation 6.13 are assumed to be independent., this can be written as

$$\begin{bmatrix} \theta_i \\ \gamma_i \\ \alpha_i \\ \xi_i \end{bmatrix} \sim N \left(\begin{pmatrix} \Theta \\ \Gamma \\ \Lambda \\ \Xi \end{pmatrix}, \Sigma \right) \text{ with } \Sigma = \begin{bmatrix} \sigma_\theta^2 & 0 & 0 & 0 \\ & \sigma_\gamma^2 & 0 & 0 \\ & & \sigma_\alpha^2 & 0 \\ & & & \sigma_\xi^2 \end{bmatrix}. \quad (6.14)$$

However, assuming that the covariances between σ_θ^2 and σ_γ^2 , and between σ_α^2 and σ_ξ^2 are zero forces the variances for the index test to be larger than those of the comparator test. Therefore, covariance terms are needed between the random effects for accuracy (i.e. α_i and ξ_i) and between the random effects for threshold (i.e. θ_i and γ_i) such that for two tests

$$\begin{bmatrix} \theta_i \\ \gamma_i \\ \alpha_i \\ \xi_i \end{bmatrix} \sim N \left(\begin{pmatrix} \Theta \\ \Gamma \\ \Lambda \\ \Xi \end{pmatrix}, \Sigma \right) \text{ with } \Sigma = \begin{bmatrix} \sigma_\theta^2 & \sigma_\theta \sigma_\gamma & 0 & 0 \\ & \sigma_\gamma^2 & 0 & 0 \\ & & \sigma_\alpha^2 & \sigma_\alpha \sigma_\xi \\ & & & \sigma_\xi^2 \end{bmatrix}. \quad (6.15)$$

To illustrate the different covariance structures represented by equations 6.10 (Model 1), 6.14 (Model 2) and 6.15 (Model 3), HSROC models were fitted to the Kittler et al²¹⁶ example used in section 6.3.1.2 to illustrate different parameterisations of the covariance matrix for the bivariate model. However, the algorithm used for fitting the HSROC model failed to converge for the more complex parameterisation expressed in equation 6.15. Since estimation of σ_{γ}^2 was truncated at zero, Model 3 was then simplified so that both tests had the same variance for threshold but unequal variances were allowed for accuracy. The results are given in Table 6.4 for the three models. Overall, there were small differences between point estimates and summary estimates from the models. The confidence intervals from Model 1 were the narrowest. In all models, there was a significant difference in the sensitivity of the two tests, though the difference was 5% higher in Model 2 and Model 3 compared to Model 1.

Table 6.4| Parameter and summary estimates from HSROC models with increasing complexity of the variance-covariance structure

Estimate	Model		
	(1) Equal variances across tests	(2) Unequal variances without dependence between tests	(3) Unequal variances with dependence between tests (accuracy only)
Model parameter (95% CI)*			
Alpha for dermoscopy	4.80 (3.44, 6.16)	4.58 (3.13, 6.03)	4.27 (3.05, 5.50)
Theta for dermoscopy	1.19 (0.10, 2.28)	0.94 (-0.46, 2.34)	0.48 (-0.58, 1.54)
Beta for dermoscopy	1.43 (0.61, 2.25)	1.21 (0.07, 2.36)	0.80 (-0.10, 1.70)
Alpha for unaided eye	2.68 (1.89, 3.46)	2.70 (1.88, 3.51)	2.72 (1.97, 3.47)
Theta for unaided eye	-0.03 (-0.68, 0.62)	-0.10 (-0.80, 0.60)	-0.17 (-0.82, 0.48)
Beta for unaided eye	0.47 (-0.28, 1.23)	0.38 (-0.43, 1.19)	0.28 (-0.45, 1.02)
σ_{α}^2	1.26	1.28	1.09
σ_{θ}^2	0.33	0.34	0.35
σ_{ξ}^2	–	0.18	0.50
σ_{γ}^2	–	0 [†]	–
$\sigma_{\theta}\sigma_{\gamma}$	–	–	–
$\sigma_{\alpha}\sigma_{\xi}$	–	–	0.52
Summary estimate (95% CI)			
Sensitivity of dermoscopy [‡]	0.86 (0.77, 0.92)	0.84 (0.76, 0.90)	0.83 (0.70, 0.92)
Sensitivity of unaided eye [‡]	0.76 (0.63, 0.86)	0.72 (0.56, 0.83)	0.72 (0.57, 0.83)
Relative sensitivity	1.12 (1.00, 1.26)	1.17 (1.02, 1.36)	1.17 (1.04, 1.31)

σ_{α}^2 = variance of accuracy; σ_{θ}^2 = variance of threshold; σ_{ξ}^2 = variance of difference in accuracy; σ_{γ}^2 = variance of difference in threshold; $\sigma_{\alpha}\sigma_{\xi}$ = covariance for accuracy; $\sigma_{\theta}\sigma_{\gamma}$ = covariance for threshold.

*Confidence intervals are not given for the variance parameters because they may be unreliable due to the assumption of normality.

[†]Estimation truncated at the lower boundary of zero.

[‡]Estimated at a fixed specificity using the median specificity of 0.94 for dermoscopy and 0.87 for unaided eye obtained from the included studies.

HSROC models are non-linear generalized mixed models and the range of software for fitting such models is limited; HSROC models are usually fitted using WinBUGS or SAS Proc NLMIXED. Throughout this thesis, HSROC models were fitted using SAS.

6.3.3 Methods comparing pooled estimates between meta-analyses

Methods that use a statistical test—such as a z -test or t -test—to compare summary estimates of any test accuracy measure obtained from separate meta-analysis of each test are discussed in this section.

6.3.3.1 Comparison of Q^*

Assuming two tests are independent, Moses et al proposed a ‘global’ comparison of diagnostic accuracy by comparing values of Q^* using a z -test.³⁷ Recall from section 1.4.3 that Q^* is the point on the SROC curve where sensitivity = specificity. Thus, Q^* represents the diagnostic threshold at which the probability of a correct diagnosis is constant for all subjects and is a point of ‘indifference’ between false positive and false negative test errors.³⁸ The Q^* statistic is a function of the intercept of the regression line and does not depend on the slope as given by

$$Q^* = \frac{1}{1+e^{-a/2}} . \quad (6.16)$$

As shown by Moses et al³⁷, the standard error (SE) of Q^* is given by

$$SE(Q^*) = \frac{SE(a)}{8(\cosh(a/4))^2} .$$

Equation 6.16 implies that all SROC curves with a certain value of a pass through the same Q^* point irrespective of their value of b . Therefore, Q^* is of limited use because SROC curves with different shapes can have the same Q^* value. In addition, as highlighted earlier in section 1.4.3, Q^* may not be located within the region of the observed data. To compare the accuracy of two tests using Q^* , an approximate z statistic is given by

$$Z = \frac{Q_1^* - Q_2^*}{\sqrt{[SE_1^2(Q_1^*) + SE_2^2(Q_2^*)]}} . \quad (6.17)$$

Moses et al suggested that if the number of studies is large, e.g. at least 10 studies, a normal distribution can be used as an approximation to the sampling distribution of this statistic in order to obtain a P value.

6.3.3.2 Comparison of effectiveness measure

Hasselblad and Hedges proposed an effectiveness measure, d , that is proportional to the log DOR.²⁰¹ Their approach makes two key assumptions: (1) that underlying distributions of continuous measurements are logistic and (2) have equal variances (see section 1.3.2.2).

Using the counts of the 2x2 table, d_i can be estimated for the i th study as follows:

$$d_i = \frac{\sqrt{3}}{\pi} [\ln(TP_i) + \ln(FP_i) - \ln(FN_i) - \ln(TN_i)] \quad (6.18)$$

The variance of d_i is given by

$$\text{var}(d_i) = 3 \left(\frac{1}{TP_i} + \frac{1}{FP_i} + \frac{1}{FN_i} + \frac{1}{TN_i} \right) / \pi^2. \quad (6.19)$$

From equations (6.18) and (6.19), it is clear that a continuity correction will be needed if any of the cells of the 2x2 table is zero. In a meta-analysis, d can be pooled using an inverse variance weighted fixed effect model or the DerSimonian and Laird random effects approach. This can be implemented in any statistical package.

To compare two tests with pooled effectiveness estimate \hat{d}_1 and variance $\text{var}(\hat{d}_1)$ for one test, and \hat{d}_2 and $\text{var}(\hat{d}_2)$ for the other test, the z statistic is

$$z = \frac{\hat{d}_1 - \hat{d}_2}{\sqrt{\text{var}(\hat{d}_1) + \text{var}(\hat{d}_2)}}. \quad (6.20)$$

The difference in average effectiveness and its 95% confidence interval is given by

$$\hat{d}_1 - \hat{d}_2 \pm 1.96 \sqrt{\text{var}(\hat{d}_1) + \text{var}(\hat{d}_2)}.$$

6.3.4 Methods for comparative meta-analysis of correlated (paired) data

Contrary to the standard comparative meta-analysis methods that assume independence between test results from the same individuals (paired data), the comparative meta-analysis methods discussed in this section explicitly account for correlated multivariate data structures. Methods based on odds ratios or directly on sensitivity and specificity have been proposed as described below.

6.3.4.1 Conditional relative odds ratio method

Suzuki et al proposed the conditional relative odds ratio (CROR) as a measure of relative test performance conditioned on counts of discordant test results.²⁰⁵ The CROR approach requires tables of the joint classification of the results of two tests against those of the reference standard (see Table 1.3). The CROR only requires the number of discordant test results in the diseased group and in the non-diseased group. This is illustrated in Table 6.5 using similar notation as in Table 1.3.

Table 6.5| Discordant test results in the diseased and non-diseased groups for study i

	Reference standard positive		Reference standard negative	
	Test 1 positive	Test 1 negative	Test 1 positive	Test 1 negative
Test 2 positive		$y_{i,01}^D$		$y_{i,01}^{\bar{D}}$
Test 2 negative	$y_{i,10}^D$		$y_{i,10}^{\bar{D}}$	

The odds ratio in study i comparing test A and B in the diseased group is given by

$$OR_{i(\text{diseased})} = y_{i,10}^D / y_{i,01}^D,$$

and in the non-diseased group by

$$OR_{i(\text{non-diseased})} = y_{i,10}^{\bar{D}}/y_{i,01}^{\bar{D}}.$$

Note that $OR_{i(\text{diseased})}$ and $OR_{i(\text{non-diseased})}$ are measures similar to odds ratios from a matched case control or cohort study. The ratio of these two ORs is the CROR which represents the relative accuracy of test A and test B conditioned on discordant test results.

This can be written as

$$CROR_i = \frac{OR_{i(\text{diseased})}}{OR_{i(\text{non-diseased})}} = \frac{y_{i,10}^D/y_{i,01}^D}{y_{i,10}^{\bar{D}}/y_{i,01}^{\bar{D}}}, \quad (6.21)$$

and the variance as

$$\text{var}(CROR) = \frac{1}{y_{i,10}^D} + \frac{1}{y_{i,01}^D} + \frac{1}{y_{i,10}^{\bar{D}}} + \frac{1}{y_{i,01}^{\bar{D}}}.$$

After obtaining the CROR and its variance from each study, the estimates are then pooled using traditional univariate methods.

Since the CROR approach only requires discordant results, the true diagnosis of concordant results in a primary study is not needed. However, the use of only discordant results has limitations. The number of individuals with discordant test results tends to be small and so the standard error of the CROR may be large. Also, it may not be possible to construct Table 6.5 for several studies, therefore limiting the number of studies included in the meta-analysis.

6.3.4.2 Generalized estimating equation modelling of sensitivity and specificity

Kowalski et al used generalized estimating equations (GEEs) to compare the accuracy of enzyme immunosorbent assays (EIAs) used in HIV antibody testing.²⁰² The primary aim of this systematic review was to estimate the relative test performance of test kits from different manufacturers applied to the same serum samples in each included study. The marginal (population average) model, models the averaged result for each study, and allows for missing

data if all studies do not evaluate all tests. The approach assumes a fixed threshold across studies and so sensitivity and specificity were analysed separately. However, test threshold will often vary across studies thus inducing correlation between sensitivity and specificity in addition to heterogeneity. Although this approach accounts for pairing of test results, it fails to account for between-study heterogeneity and potential correlation between sensitivity and specificity across studies. This approach will not be discussed further because some of the other more sophisticated models described later in this section were implemented using GEEs.

6.3.4.3 Repeated measures modelling of diagnostic odds ratios

An extension to the Moses SROC model to account for dependencies between tests was proposed by Siadaty et al.²⁰⁷ The results of tests applied to the same participants in a study were regarded as a cluster of repeated measurements that are potentially correlated and so statistical methods for repeated measurements were applied. The proposed marginal model is written as

$$\text{logit}(\pi_{ik}) = \beta_0 + \beta_1 D_i + \beta_2 T_{ik} + \beta_3 D_i T_{ik} + e_{ik}, \quad (6.22)$$

with

$$e_{ik} \sim N(0, \sigma^2).$$

π_{ik} is the probability of a positive test result for the k th test from the i th study and e_{ik} is the random error for each test within each study. Disease status, D_i , is an indicator variable which is coded 0 for the non-diseased and 1 for the diseased group. T is a categorical variable for test type and is represented by indicator variables (T_{ik}) in the model. Therefore, regression

coefficients β_2 and β_3 are vectors of coefficients. For the comparison of two tests, β_3 is the log of the ratio of the two DORs from the two tests.

The method was developed for data grouped at the study level but can be modified to expand each 2x2 table to the original sample size such that the units of analysis are individuals and not studies. Whilst this approach enables a simple transition from aggregated data to individual patient data, it does not account for between-study variability. The marginal logistic regression was fitted using SAS Proc GENMOD.

6.3.4.4 Proportional odds ratio model

The proportional odds ratio (POR) regression model proposed by Siadaty and Shu allows each test in a test comparison to have its own trend of odds ratios across studies and the trends of two tests are assumed to be *proportional* to each other, the "proportional odds ratio" assumption.²⁰⁶ The model assumes a binomial distribution, accounts for correlated test results and allows for missing data if all studies do not evaluate all tests. Using the odds ratio as the measure of test performance, the proportional odds ratio model can be written as

$$\text{logit}(\pi_{ik}) = \beta_0 + \beta_1 D_i + \beta_2 P_{ik} + \beta_3 D_i P_{ik} + \beta_4 T_{ik} + \beta_5 D_i T_{ik} + e_{ik}, \quad (6.23)$$

with

$$e_{ik} \sim N(0, \sigma^2).$$

π_{ik} is the probability of a positive test result for the k th test from the i th study and e_{ik} is the random error for each test within each study. Disease status, D_i , is coded 0 for the non-diseased and 1 for the diseased group. T and P are categorical variables for test type and study and are represented by indicator variables, P_{ik} and T_{ik} , in the model. Therefore,

regression coefficients β_2 , β_3 , β_4 and β_5 are all vectors of coefficients when there are more than two tests and more than two studies. The average log DOR is given by β_1 and components of β_5 estimate the deviation of the log DOR of each test from the average log DOR. Similar to the previous model, the POR model was also fitted using Proc GENMOD.

The POR model (equation 6.23) assumes that the SROC curves for the tests have the same shape with the difference between the curves being their position in ROC space. This assumption can be relaxed by using a nonlinear mixed effects model in which study is modelled as random effects and random interaction effects are included for the interaction of study with disease, and interaction of study with disease and test. However, the authors note that analyses of such models will have difficulty converging, especially for datasets where there are many studies that evaluated only one or a small number of the tests.

6.3.4.5 Bayesian hierarchical models for joint meta-analysis of paired data

Trikalinos et al extended the bivariate model (equation 6.4) to enable joint meta-analysis of studies comparing multiple index tests using within-subject study designs.²⁰⁹ The model, implemented within a Bayesian framework, accounts for the relationship between sensitivities and specificities across two or more tests, and also captures information on the concordance between two tests in those with and without disease by incorporating the ‘joint true positive rate’ (JTPR), and ‘joint false positive rate’ (JFPR). The JTPR is the probability of positive test results for both tests in the diseased group while the JFPR is the probability of positive test results for both tests in the non-diseased group. From Table 6.6, the JTPR = $\pi_{i,11}^D$ and the JFPR = $\pi_{i,11}^{\bar{D}}$. Thus, the approach considers the joint probability of two tests in addition to the marginal probabilities of each test.

Table 6.6| Probability of each combination of test results for two tests in study i

	Reference standard positive		Reference standard negative	
	Test 1 positive	Test 1 negative	Test 1 positive	Test 1 negative
Test 2 positive	$\pi_{i,11}^D$	$\pi_{i,01}^D$	$\pi_{i,11}^{\bar{D}}$	$\pi_{i,01}^{\bar{D}}$
Test 2 negative	$\pi_{i,10}^D$	$\pi_{i,00}^D$	$\pi_{i,10}^{\bar{D}}$	$\pi_{i,00}^{\bar{D}}$

π indicates probability of positive and negative test results in those with (D) and without disease (\bar{D}). 0 denotes negative test results and 1 denotes positive test results.

Analogous to the binomial likelihood used in the bivariate model (see equation 1.11), the multinomial distribution can be used to model the cross-classification of the results of two or more tests within each study.²⁰⁹ Table 6.7 shows the cross-classified results for two tests in study i . These are the observed counts of concordant ($y_{i,00}^D$ and $y_{i,11}^D$) and discordant ($y_{i,01}^D$ and $y_{i,10}^D$) test results in the diseased group, and the observed counts of concordant ($y_{i,00}^{\bar{D}}$ and $y_{i,11}^{\bar{D}}$) and discordant ($y_{i,01}^{\bar{D}}$ and $y_{i,10}^{\bar{D}}$) test results in the non-diseased group. Trikalinos et al²⁰⁹ modelled the column vector of counts

$$\mathbf{y}^D = \begin{pmatrix} y_{i,00}^D \\ y_{i,01}^D \\ y_{i,10}^D \\ y_{i,11}^D \end{pmatrix} \text{ and } \mathbf{y}^{\bar{D}} = \begin{pmatrix} y_{i,00}^{\bar{D}} \\ y_{i,01}^{\bar{D}} \\ y_{i,10}^{\bar{D}} \\ y_{i,11}^{\bar{D}} \end{pmatrix}$$

in the diseased and non-diseased groups using conditionally independent multinomial distributions as follows:

$$\mathbf{y}^D \sim M(N_i^D, \boldsymbol{\pi}_i^D) \tag{6.24}$$

$$\mathbf{y}^{\bar{D}} \sim M(N_i^{\bar{D}}, \boldsymbol{\pi}_i^{\bar{D}}) \tag{6.25}$$

where N_i^D and $N_i^{\bar{D}}$ are the total number in the diseased and non-diseased groups, and $\pi_i^D = (\pi_{i,00}^D, \pi_{i,01}^D, \pi_{i,10}^D, \pi_{i,11}^D)'$ and $\pi_i^{\bar{D}} = (\pi_{i,00}^{\bar{D}}, \pi_{i,01}^{\bar{D}}, \pi_{i,10}^{\bar{D}}, \pi_{i,11}^{\bar{D}})'$ are vectors of the probabilities in Table 6.6.

Table 6.7| Observed counts of test results cross-classified for two tests in study i

	Reference standard positive		Reference standard negative	
	Test 1 positive	Test 1 negative	Test 1 positive	Test 1 negative
Test 2 positive	$y_{i,11}^D$	$y_{i,01}^D$	$y_{i,11}^{\bar{D}}$	$y_{i,01}^{\bar{D}}$
Test 2 negative	$y_{i,10}^D$	$y_{i,00}^D$	$y_{i,10}^{\bar{D}}$	$y_{i,00}^{\bar{D}}$

0 = negative test result; 1 = positive test result; D = diseased; \bar{D} = non-diseased.

The joint distribution of the random effects for the logit transformed TPRs, FPRs, JTPR, and JFPR is modelled using a six-dimensional normal distribution analogous to equation 6.4 if an unstructured variance-covariance matrix is used. The authors expressed the model in terms of logit TPR and logit FPR instead of logit sensitivity (μ_A) and logit specificity (μ_B) used for expressing bivariate models in this thesis. For consistency with the notation in the thesis, the distribution is modelled in terms of μ_A and μ_B , and written as

$$\begin{bmatrix} \mu_{Ai1} \\ \mu_{Ai2} \\ \mu_{Ai*} \\ \mu_{Bi1} \\ \mu_{Bi2} \\ \mu_{Bi*} \end{bmatrix} \sim N \left(\begin{pmatrix} \boldsymbol{\mu}_A \\ \boldsymbol{\mu}_B \end{pmatrix}, \boldsymbol{\Sigma} \right) \text{ with } = \begin{bmatrix} \sigma_{A1}^2 & \sigma_{A1A2} & \sigma_{A1A*} & \sigma_{A1B1} & \sigma_{A1B2} & \sigma_{A1B*} \\ & \sigma_{A2}^2 & \sigma_{A2A*} & \sigma_{A2B1} & \sigma_{A2B2} & \sigma_{A2B*} \\ & & \sigma_{A*}^2 & \sigma_{A*B1} & \sigma_{A*B2} & \sigma_{A*B*} \\ & & & \sigma_{B1}^2 & \sigma_{B1B2} & \sigma_{B1B*} \\ & & & & \sigma_{B2}^2 & \sigma_{B2B*} \\ & & & & & \sigma_{B*}^2 \end{bmatrix}, \quad (6.26)$$

where μ_{Ai1} , μ_{Ai2} , μ_{Bi1} and μ_{Bi2} are the logit sensitivities and logit specificities for test 1 and test 2, and μ_{Ai*} and μ_{Bi*} are the logit JTPR and logit JTNR (joint true negative rate) in the i th study. The means $\boldsymbol{\mu}_A = \begin{pmatrix} \mu_{A1} \\ \mu_{A2} \\ \mu_{A*} \end{pmatrix}$ and $\boldsymbol{\mu}_B = \begin{pmatrix} \mu_{B1} \\ \mu_{B2} \\ \mu_{B*} \end{pmatrix}$ are column vectors of the overall means of the

logit sensitivities (μ_{A1} and μ_{A2}), logit specificities (μ_{B1} and μ_{B2}), logit JTPR (μ_{A*}) and logit

FPR (μ_{B^*}) for the two tests. Σ is an unstructured between-study covariance matrix with 21 parameters. The six diagonal elements (σ_{A1}^2 , σ_{A2}^2 , $\sigma_{A^*}^2$, σ_{B1}^2 , σ_{B2}^2 and $\sigma_{B^*}^2$) are the variances while the 15 off diagonal elements are the covariances. Structure can be imposed by setting equal variances and covariances between tests and so reduce the number of parameters to 12. The structured variance-covariance matrix can be written as

$$\Sigma_s = \begin{bmatrix} \sigma_A^2 & \sigma_{AA} & \sigma_{AA^*} & \sigma_{AB} & \sigma_{AB} & \sigma_{AB^*} \\ & \sigma_A^2 & \sigma_{AA^*} & \sigma_{AB} & \sigma_{AB} & \sigma_{AB^*} \\ & & \sigma_{A^*}^2 & \sigma_{A^*B} & \sigma_{A^*B} & \sigma_{A^*B^*} \\ & & & \sigma_B^2 & \sigma_{BB} & \sigma_{BB^*} \\ & & & & \sigma_B^2 & \sigma_{BB^*} \\ & & & & & \sigma_{B^*}^2 \end{bmatrix}. \quad (6.27)$$

The Trikalinos et al model can be extended to more than two tests. However, the number of model parameters grows rapidly with each additional test. For k tests, the model requires a total of $2^{k+1}-2^k-1$ parameters assuming the unstructured covariance matrix in (6.26). For example, for three, four and five tests, this equates to 119, 495 and 2015 parameters.

Therefore, due to the large number of parameters, numerical difficulties will be encountered in both unstructured and structured parameterizations.^{209,221} While this model may in theory offer better statistical properties than bivariate and HSROC models that ignore within-study correlation, it clearly comes at a price; simplifications such as using a multivariate normal approximation to model within-study variation instead of the correct multinomial distribution are likely to be needed.

The model is applicable when a substantial number of studies report data on the cross classification of results from several tests, and for meta-analyses in which estimation of summary points is appropriate due to common thresholds. There is a common perception

among diagnostic test researchers that most comparative accuracy studies that use a within-subject design do not typically report the cross classification of test results as in Table 1.3 and Table 6.7. If this is true, then the value of this approach is limited. In their application of the model to an example—accuracy of two second trimester ultrasound markers (shortened femur and shortened humerus) for Down syndrome screening—the payoff of multivariate analyses was modest. Differences in the summary estimates and credible intervals from separate meta-analyses for each test and joint meta-analyses were very small (Table 6.8). On the other hand, larger differences in estimates of comparative accuracy were observed due to the joint meta-analysis using all the information in the cross tabulation of test results, thus giving smaller standard deviations than from separate meta-analyses.²⁰⁹

Verde 2013 proposed a new Bayesian hierarchical model for meta-analysis of paired data in which the observed rates are modelled as the marginal results of unobserved rates, thus enabling direct comparison between tests.²²² Since the model does not rely on the availability of joint classification of test results, it appears to overcome the limitation of the Trikalinos et al²⁰⁹ approach. However, the Verde model is yet to be published and further details are unavailable. Cheng et al have developed a Bayesian network meta-analysis approach for pooling data from direct and indirect test comparisons,²¹⁰ but their work is also yet to be published.

Table 6.8 | Summary estimates and 95% credible intervals from alternative meta-analysis models for comparing accuracy of short femur and short humerus for Down syndrome screening

Model	Summary TPR (95% credible interval) %	Summary TPR (95% credible interval) %	Summary JTPR (95% credible interval) %	Summary FPR (95% credible interval) %	Summary FPR (95% credible interval) %	Summary JFPR (95% credible interval) %	Difference in TPR	Difference in FPR
	Short humerus	Short femur		Short humerus	Short femur			
Separate meta-analyses of the two tests using a bivariate model with a binomial within-study likelihood	37.9 (27.7, 50.3)	35.4 (23.1, 49.5)	-	4.8 (3.0, 7.4)	7.4 (5.0, 10.7)	-	2.6 (-14.7, 19.8)	-2.5 (-6.3, 1.1)
Joint meta-analysis of the two tests using a multinomial within-study likelihood (uses within-study correlations)	35.3 (26.9, 41.8)	35.0 (22.4, 46.2)	26.1 (16.6, 34.0)	4.9 (2.8, 7.5)	7.3 (4.6, 10.5)	2.7 (1.6, 4.2)	0.0 (-8.9, 9.5)	-2.5 (-5.4, 0.3)
Joint meta-analysis of the two tests using a binomial within-study likelihood (ignores within-study correlations)	34.6 (20.3, 44.2)	35.9 (20.5, 50.4)	26.8 (11.3, 39.2)	4.8 (2.9, 7.7)	7.3 (4.6, 11.5)	2.8 (1.7, 4.4)	-1.4 (-10.2, 8.8)	-2.5 (-6.4, 0.5)

FPR = false positive rate (1 – specificity); JFPR = joint false positive rate; JTPR = joint true positive rate; TPR = true positive rate (sensitivity).

For the joint meta-analyses comparing accuracy of short humerus and short femur, Trikalinos et al²⁰⁹ used the unstructured variant of the between-study covariance matrix (see equation 6.26).

(Adapted from Trikalinos et al.²⁰⁹)

6.4 Summary of comparative meta-analysis methods

A range of methods, varying in complexity and methodological rigour, are available for comparative meta-analysis. Key characteristics of the methods are summarised in Table 6.9. Some methods synthesise studies in one step in a model while others use a two stage approach, first estimating new measures or variables before the actual meta-analysis. Of all the methods, only hierarchical models (including Bayesian extensions for modelling correlated data) explicitly account for both within and between-study variability, while also accounting for the bivariate and logistic (no continuity correction required) nature of the data. Frequentist based methods which directly model correlated data, such as the CROR or POR approaches, fail to address between-study variability. In contrast, a Bayesian bivariate model extension for modelling correlated data proposed by Trikalinos et al²⁰⁹ is promising but data availability and modelling assumptions/complexities are likely to compromise its use.

For bivariate and HSROC models, different parameterisations of the covariance matrix are possible. In practice, it is often assumed that the variances of the respective model parameters are identical across tests, i.e. differences in parameter estimates are modelled as fixed effect. A similar assumption is commonly made about between-study variances for treatment effects in multiple treatment comparisons (network meta-analysis). While this assumption has the advantage of simplifying estimation of the models and may be appropriate for treatment effects, it is unlikely to be generalisable for test comparisons. The assumption will be evaluated and discussed further in Chapter 7.

Table 6.9| Summary of comparative meta-analysis methods

Method	Characteristic							
	Models paired data	One stage analysis	Regression model	Avoids continuity correction	Credibly accounts for within-study variability ¹	Accounts for between-study variability through inclusion of random effects	Accounts for bivariate data ²	Extends to more than two tests
Comparison of Q*	No	No	No	No	No	No	Yes	No
Comparison of effectiveness measure	No	No	No	No	No	No	Yes	No
Linear mixed model for LRs	No	No	Yes	No	No	Yes	No	Yes
Moses SROC regression ³	No	No	Yes	No	No	No	Yes	Yes
Bivariate meta-regression	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
HSROC meta-regression	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Conditional relative odds ratio approach	Yes	No	No	No	No	No	Yes	No
GEE modelling of sensitivity and specificity	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
Repeated measures	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
Proportional odds ratio model	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
Bayesian models for joint meta-analysis of paired data ⁴	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

¹Explicitly models within-study variability using a binomial likelihood to model proportions or approximates within-study variability using a normal distribution.

²Methods that either considered sensitivity and specificity jointly or used the diagnostic odds ratio or a variant of it as the outcome measure were deemed to account for the bivariate nature of test accuracy data.

³The meta-regression approach for comparing tests was not proposed in the paper by Moses et al 1993 but the meta-regression approach suggested in the paper for investigating heterogeneity was later adopted for test comparisons.

⁴Refers to the three Bayesian models by Trikalinos et al 2014,²⁰⁹ Cheng et al 2013²¹⁰ and Verde 2013²²².

Almost all the methods can be extended to include more than two tests in a meta-analysis. The comparisons between different tests constitute a network (see Figure 2.2) that can enable inferences about the relative merits of tests that have not been compared directly. Appropriate methodology has been developed for combining the data from a network of RCTs, allowing the incorporation of evidence from all direct and indirect comparisons toward estimating summary treatment effects.^{91,223,224} The basic network meta-analysis approach cannot be directly applied to test accuracy studies because test accuracy is not normally reported using a single measure of effect but rather two correlated measures such as sensitivity and specificity to quantify test performance in the diseased and non-diseased groups respectively. Also, unlike RCTs where the relative treatment effects of comparator interventions is summarised in meta-analysis using statistics such as relative risks and odds ratios, the sensitivity and specificity of each test are usually the measures meta-analysed and not relative sensitivity and relative specificity. Cheng et al²¹⁰ have generalised the Trikalinos et al²⁰⁹ approach to network meta-analysis but the work is yet to be published.

Given the number of published papers and those in progress, it seems the development of comparative meta-analysis methods is an area of active research. Recent advances have been in Bayesian methods but given the poor uptake of Bayesian methods for meta-analysis of a single test, it remains to be seen if Bayesian comparative meta-analysis methods will be used in practice. Methods that use a frequentist approach can be fitted in standard statistical packages and accessibility to appropriate software is likely to drive their use. In the next chapter, the performance of robust or frequently used frequentist based methods described in this chapter will be explored in detail.

6.5 Conclusions

Various comparative meta-analysis methods are available but few of them take into account the potential for correlation between sensitivity and specificity, in addition to explicit modelling of within- and between-study variability. In particular, none of the frequentist approaches designed to fully account for paired data meet all of these requirements.

Consequently, use of such methods should be discouraged. Rigorous and more integrated approaches to comparative meta-analysis are evolving, especially within the Bayesian framework, but bivariate and HSROC meta-regression models are currently the most sophisticated and theoretically sound approach in use.

Though the limitations of the Moses SROC method for basic meta-analysis are well documented in the literature, the simplicity of the approach probably explains its continued popularity. The approach is likely to give erroneous standard errors and misleading conclusions, and so should not be used for making formal inferences about relative test performance. The extent to which results from Moses SROC meta-regression deviate from those of hierarchical models and the impact on conclusions will be thoroughly investigated in the next chapter.

7 EMPIRICAL ASSESSMENT OF COMPARATIVE META-ANALYSIS METHODS

7.1 Introduction

The case studies in Chapter 2 raised the prospect that assumptions made in fitting hierarchical meta-regression models, such as assuming a common shape for SROC curves (section 2.3.3) or common variances across tests (section 2.3.4) may threaten the validity of review conclusions. The complexity of models and/or the number of test comparisons in relation to the number of studies available (sections 2.3.1 and 2.3.5) also pose a challenge. However, there is a lack of evidence based guidance on meta-analysis of test comparisons. Therefore, the aim of this chapter is to empirically investigate the validity of assumptions commonly made when comparing test accuracy in hierarchical meta-regression models, and to examine the impact of alternative comparative meta-analytic models on findings. Given the scope of the challenges illustrated in Chapter 2, understanding the properties of common and novel methods when assessing two tests is an essential precursor to undertaking more complex evaluations involving more than two tests. **Therefore, this empirical assessment is limited to meta-analyses comparing the diagnostic accuracy of two tests.**

To help the reader navigate this chapter, the chapter is divided into three parts. Part I (sections 7.2 and 7.3) describes the methods and dataset used for the empirical study; Part II (sections 7.4 and 7.5) gives the results of the investigation of modelling assumptions used in hierarchical models; and Part III (sections 7.6 to 7.8) gives the results of the assessment of the impact of different modelling complexity and comparative meta-analysis models on findings, and concludes the chapter with a discussion. A more detailed outline for this chapter is as

follows. In section 7.2, methods for selection of systematic reviews identified from the search results in section 3.4 and data extraction are outlined. Data analysis methods for the empirical evaluation of the most commonly used or promising comparative meta-analytic methods described in Chapter 6 are also detailed in this section. Characteristics of the review cohort are described in section 7.3. In section 7.4, results of the investigation of modelling assumptions based on separate meta-analyses of tests in each test comparison are presented, and key findings are summarised in section 7.5. Section 7.6 builds on the previous two sections by addressing the impact of different modelling complexity on relative test performance. In section 7.7, different comparative meta-analysis methods are compared. Finally, section 7.8 concludes the chapter with a discussion of the findings, and gives recommendations for appropriate use of comparative meta-analysis methods.

To avoid repetition detailed research questions will be specified at the beginning of Part II and Part III.

PART I: METHODS AND DESCRIPTION OF EMPIRICAL DATA

The first part of this chapter (sections 7.2 and 7.3) describes the methods for selecting systematic reviews from the cohort identified in Chapter 3; methods for extracting data from the reviews; data analysis methods used for investigating modelling assumptions and for comparing the findings of different comparative meta-analysis methods; and characteristics of the reviews selected. Therefore, Part I is a description of the methodology and dataset used for the empirical study.

7.2 Methods

7.2.1 Selection of systematic reviews and data extraction

From the cohort of 101 systematic reviews identified in section 3.4, reviews were selected if they provided sufficient data to enable derivation of the number of true positives, false positives, false negatives and true negatives from each study included a meta-analysis. For reviews with data for multiple units of analysis, e.g. lesion-based and patient-based data, patient level data was selected. If a review stratified analyses by population, setting, target condition, etc., only the dataset for the main analysis was selected. Where it was not possible to make this judgement, the stratum containing the largest number of studies was selected. As such only one pairwise test comparison was selected per review.

For each review, primary studies that only reported sensitivity or specificity (data for one half of the 2×2 table), or had no cases were excluded from the meta-analysis. For studies with discrepancies in the 2×2 data, full text papers were retrieved for verification of the data. Information about target condition, index tests, unit of analysis, and thresholds were extracted from the reviews. The introduction/background and discussion sections of each paper were

examined in order to identify the index ("newer") test and comparator (current practice or the "older" test). If it was not possible to make this distinction, either explicitly or by inference from the available information, an arbitrary choice was made. The test comparison in each meta-analysis was classified as direct or indirect depending on the type of primary studies included. Data extraction was done only by the author. To check for errors in the 2x2 data extracted, forest plots containing the 2x2 data as well as the corresponding sensitivities and specificities were produced using Review Manager version 5 (The Nordic Cochrane Centre, The Cochrane Collaboration, 2014) and compared with the sensitivities and specificities reported in the original review.

7.2.2 Selection of comparative meta-analysis models

Of the comparative meta-analysis methods described in section 6.3, only those that were: (1) commonly used methods for synthesising studies in a regression model; or (2) theoretically rigorous approaches that model the bivariate nature of test accuracy data while also accounting for within- and between-study variability were considered. With the exception of Bayesian models for joint meta-analysis of paired data, only three models—Moses SROC, bivariate and HSROC meta-regression models—meet one or both requirements. Empirical evaluations of meta-analysis of a single test have suggested that results are often similar between univariate and bivariate meta-analyses.^{36,225} Since univariate random effects logistic regression models for sensitivity and specificity represent a special case obtained by simplifying the covariance structure of bivariate models (i.e. zero covariance model, see equation 1.12), univariate models were also included in the empirical evaluation. The Moses model in (6.7) simply averages the log DORs, and so only Moses SROC meta-regression

models without interaction (equation 6.6) and with interaction (equation 6.8) terms, using unweighted and weighted analyses, were assessed.

7.2.3 Data analysis

All test comparisons were meta-analysed using methods for comparing points and curves, irrespective of whether common or mixed thresholds were used in the included studies. In an empirical investigation such as this, the aim is not to estimate summary estimates that will be interpreted and translated into clinical practice, but merely for highlighting differences between different models. Therefore, to enable the use of all the available datasets, the distinction between common and mixed thresholds was ignored.

A two-stage approach was used to assess modelling assumptions commonly used in hierarchical models; first by preliminary assessments of test comparisons using separate meta-analysis of each test in a test comparison and second by investigation in comparative meta-analyses. This was appropriate because commonly used comparative meta-analysis methods assume independence between test results from the same individuals in comparative accuracy studies with a paired design.

7.2.3.1 Assessment of modelling assumptions using separate meta-analyses of two tests in test comparisons

Preliminary analyses were undertaken using bivariate and HSROC models for meta-analyses of each test in a test comparison. The two models were fitted for each test in order to explore estimation issues specific to a particular model, e.g. shape in the HSROC model and correlation in the bivariate model, as well as exploring the relationship between the two models in situations where one of the models may be unstable but the other is stable.

Thorough investigation and meta-analysis of each test was undertaken prior to comparative meta-analyses in order to (1) gain insight into model stability and potential model fitting problems that may be encountered with increased model complexity when test comparisons are made under various modelling assumptions; (2) to assess heterogeneity in the performance of each test; and (3) to investigate the shape of each SROC curve. The methods used are described below.

Assessment of model stability

Estimation of the variances of the random effects can be problematic when the estimates are zero or close to zero. A small number of studies may lead to unreliable estimation of the correlation (ρ) between sensitivity and specificity across studies and the variances of the random effects in a bivariate model. Based on the examples in Chapter 2, the cohort of reviews in Chapter 5 and a systematic overview,²⁷ it is common for meta-analyses of test accuracy to only include a small number of studies. Sparse data also poses another challenge as highlighted in section 2.3.5.1. For similar reasons, the shape parameter (β) and variances in a HSROC model may be poorly estimated.

To explore the robustness of ρ in a bivariate model, univariate random effects logistic regression models were used to pool sensitivity and specificity by specifying an independent variance-covariance structure (i.e. $\rho = 0$) for a bivariate model (equation 1.12). For brevity, throughout the rest of this chapter, univariate random effects logistic regression models will be referred to simply as univariate models. The effect of removing ρ on estimates of variances of the random effects and standard errors of mean logit sensitivities and mean logit specificities is of interest. To investigate estimation of the shape parameter (β), estimates of β

from HSROC models were compared with those derived using functions of the bivariate model parameters. According to Harbord et al,⁴⁴ β is determined solely by the ratio of the variances of logit sensitivity and logit specificity in the bivariate model as follows:

$$\beta = \log(\sigma_B/\sigma_A) .$$

Given this relationship, it can be argued that if estimates of β from the two models disagree, then poor estimation of the variances in a bivariate model or poor estimation of β in an HSROC model can be inferred, depending on which of the two models is considered to be valid based on criteria outlined in section 7.2.3.3.

Assessment of heterogeneity

Graphical plots are a useful tool for visual exploration of heterogeneity, and so for each test comparison, study specific estimates of sensitivity and specificity were plotted in ROC space to illustrate the spread of study results for the tests. The extent of heterogeneity in a random effects meta-analysis is quantified by the variances of the random effects. It is plausible that the extent of heterogeneity may differ between tests, thus, the assumption of common variances (i.e. variances of the random effects do not depend on test type) in hierarchical meta-regression models may be untenable as illustrated earlier in section 2.3.4.1. In that example, initial meta-analysis of each test indicated variances may differ between both tests. Therefore, variances of the random effects obtained from meta-analyses of individual tests were compared for each pair of test in a test comparison to determine their similarity. This was examined graphically across test comparisons by plotting the variance estimates for an index test against those of the corresponding comparator test.

Assessment of shape of SROC curves

Estimation of SROC curves using the HSROC model allowed investigation of the shape of SROC curves for individual tests prior to exploring assumptions of a common asymmetric shape for SROC curves in test comparisons. For each pair of tests in a test comparison, estimates of β (shape parameter) from separate meta-analyses of the tests were graphically compared to assess their similarity. For each test, the 95% confidence interval of β was examined to explore uncertainty in the estimation of β . Since the number of studies for each test in a test comparison may differ, uncertainty in the estimation of β was also investigated by using scatter plots to examine the effect of the number of studies included in a meta-analysis on the magnitude of β and its standard error. A second meta-analysis in which symmetry of the SROC curve was assumed by removing β from the HSROC model, i.e. constraining β to zero, was performed for each test. Likelihood ratio tests were used to assess whether the observed difference in the fit of HSROC models with and without β was statistically significant.

7.2.3.2 Assessment of impact of different modelling assumptions on relative test performance

Comparative meta-analyses were performed for each test comparison by using bivariate and HSROC meta-regression models. Issues that were illustrated in Chapter 2 were investigated as follows.

Dealing with comparative studies in a comparative meta-analysis

Two approaches for handling comparative studies in a comparative meta-analysis were discussed in section 2.3.4.2. One approach—the between-study comparative approach—ignores the clustering of test results within comparative studies while the other approach—the

within-study comparative approach—takes each comparative study into account. The effect of a between-study approach was investigated in bivariate models (using equations 6.1 and 6.2) by giving each study and test combination a unique identifier. This variable was then used to determine the clusters for estimation of the random effects instead of the study identifier used for the within-study approach. These models were compared with bivariate models that used a within-study approach. The within-study approach was used for all other comparative meta-analyses performed using bivariate or HSROC models in the rest of the thesis.

Testing for differences in variances of random effects

Assumptions about the variances of the random effects were investigated in bivariate models by fitting the models specified in equations 6.2 (Model 1) and 6.4 (Model 2). The statistical significance of different covariance structures was assessed using likelihood ratio tests to compare models that assumed common variances across tests and those that allowed variances to vary by test. Univariate models were also used to examine assumptions about the variances of the random effects. The covariance structure in HSROC models can also be investigated but such analyses were considered unnecessary given the close relationship between bivariate and HSROC models, and because the bivariate model directly models sensitivity and specificity which are often the quantities of interest to meta-analysts and clinicians instead of DORs.

Additional analyses restricted to direct comparisons were conducted if paired data were available for at least 10 studies in a test comparison. Three bivariate models were fitted to each direct comparison using the different covariance structures represented by equations 6.2 (Model 1), 6.4 (Model 2) and 6.5 (Model 3). The minimum number of studies was selected

based on the number of parameters to be estimated and to facilitate convergence. A larger number of studies should be preferred because there are 14 parameters in (6.5) but the limit of 10 was chosen to allow investigation of any numerical issues.

Testing for differences in shape of SROC curves

For methods that compare SROC curves, assumptions about the shape of the SROC curves were explored by fitting models that assumed a common shape and those that allowed for differences in shape between tests (i.e. equations 6.6 and 6.8 for Moses models, and equations 6.9 to 6.11 for HSROC models). Association between shape and test type was statistically assessed using likelihood ratio tests to compare models with and without covariate terms for shape.

7.2.3.3 Comparison of meta-analytic models

Bivariate and HSROC meta-regression models were considered the benchmark against which to compare the performance of univariate and Moses SROC meta-regression models respectively. This is because hierarchical models are statistically rigorous—argued from a theoretical viewpoint^{23,35,44} and demonstrated in simulation studies of their application in the meta-analysis of single tests.^{53,54,226} To illustrate differences between models, univariate models were compared against bivariate models; weighted Moses SROC models against unweighted models; and Moses SROC models against HSROC models.

7.2.3.4 Model fitting and estimation of summary estimates

Hierarchical models were fitted in SAS and Stata using maximum likelihood estimation, via adaptive Gaussian quadrature as described in section 2.2.5. Bivariate models were fitted using

the *xtmelogit* command in Stata. If the analysis of a model failed to converge, the number of quadrature points was increased from five to 10. If the analysis still failed to converge, the user-written program *gllamm*²²⁷ was used. Prior to version 10 of Stata, *gllamm* was the only option for fitting generalized linear mixed models. Based on the author's experience, *gllamm* appears to be better at obtaining feasible starting values for the likelihood estimation than *xtmelogit*. However, *xtmelogit* was preferred to *gllamm* for several reasons. First, because *gllamm* is a user-written program in a Stata ado file, it is slower to run than the in-built *xtmelogit* command. Second, *gllamm* has no option for displaying the gradient vector in the iteration log. Third, it is not possible to fit bivariate models with the covariance structure in (6.4) using *gllamm*. For bivariate and univariate models, differences in diagnostic accuracy between tests were presented as relative sensitivity and relative specificity.

HSROC models were fitted using the NLMIXED procedure in SAS. A negative variance component is an underestimate of a small or zero variance component.²²⁸ To prevent estimation of such negative variances, boundary constraints ($\sigma^2 \geq 0$) were specified for the variances of the random effects in the HSROC model (σ_α^2 and σ_θ^2). A bivariate or HSROC model may satisfy a convergence criterion but may be unstable or have missing standard errors due to issues with model identifiability. Therefore, for an analysis to be deemed valid, the convergence criterion had to be met, and the analysis had to give standard errors along with gradient values close to zero for all model parameters. If boundary constraints were triggered for variance parameters of the HSROC model, there was no requirement for small gradient values or standard errors since estimation of these parameters would be truncated.

If a common shape was assumed for SROC curves, the rDOR was used to quantify differences in DORs by taking the exponent of the difference in accuracy (ξ in equation 6.11). For HSROC models that allowed β to differ by test (equation 6.9), relative sensitivity was computed by using the ESTIMATE statement described earlier in section 2.2.5. Relative sensitivity was derived from estimates of sensitivities at the median specificity values obtained from the included studies for each test. The log of relative sensitivity was computed by taking the difference between the estimated summary sensitivities on the log scale [$\log(\text{sensitivity of index test}) - \log(\text{sensitivity of comparator})$] to ensure appropriate estimation of standard errors using the delta method. The sensitivity of each test was computed using equation 1.19.

Weighted and unweighted Moses SROC regression models were fitted using the *regress* command in Stata. If a common shape was assumed for SROC curves, the rDOR was used to quantify differences in DORs by taking the exponent of the regression coefficient b_1 in equation 6.6. For the Moses model that included the kS interaction term (equation 6.8), relative sensitivity was estimated using equation 1.5 and the *nlcom* command in Stata post-estimation of the regression model. The *nlcom* command uses nonlinear combinations of the parameter estimates to compute point estimates and standard errors are also computed using the delta method.

7.2.3.5 Assessment of performance of different models

The performance of different models in Part III (same comparative meta-analysis method but under different modelling assumptions or different comparative meta-analysis methods) was assessed by examining estimates of measures of relative test performance (rDOR, relative

sensitivity and relative specificity) and their standard errors. Performance was assessed using the following four criteria.

1. Difference in magnitude of relative test performance computed as ratio of rDORs, or ratio of relative sensitivities and ratio of relative specificities.
2. Difference in precision of measures of relative test performance computed as a ratio of standard errors.
3. Change in statistical significance at the 5% level determined by assessing whether or not the confidence intervals for rDORs, or relative sensitivities and relative specificities include 1.
4. Change in direction of effect, i.e., qualitative change where the ranking of a pair of tests in terms of superior sensitivity or specificity was inconsistent between two different models.

Across the cohort of meta-analyses, descriptive statistics were computed to summarise how often differences occurred. Where applicable, the statistical significance of differences in the fit of two models was assessed using likelihood ratio tests.

7.3 Description of cohort of systematic reviews

Characteristics of the 57 reviews and test comparisons included in this cohort are summarised in Table 7.1. Further details are provided in Appendix D.1. Complete 2x2 data were not available for 30 primary studies in nine of the 57 reviews. These studies were excluded and so the number of studies presented for each test in Table 7.1 is the number of studies with complete 2x2 tables.

Table 7.1| Characteristics of empirical dataset

ID	Reference	Type of test comparison	N _{IC} ¹	N _I ²	N _C ³	N _T ⁴	Unit of analysis	Thresholds ⁵
1	Alkhayal 2007 ¹⁵²	Indirect	3	25	25	47	Patient	No
2	Arbyn 2004 ²²⁹	Direct only	4	4	4	4	Patient	No
3	Bafounta 2001 ²³⁰	Direct only	8	8	8	8	Lesion	Yes
4	Basaran 2009 ²³¹	Indirect	0	4	3	7	Patient	No
5	Battaglia 2006 ²¹⁹	Indirect	0	6	13	19	Patient	Yes
6	Birim 2005 ²³²	Direct only	17	17	17	17	Patient	No
7	Brazzelli 2009 ²³³	Direct only	7	7	7	7	Patient	No
8	Carlson 1994 ²³⁴	Indirect	0	7	3	10	Patient	No
9	Cavallazzi 2008 ²³⁵	Indirect	1	3	5	7	Patient	Yes
10	de Vries 1996 ²¹⁸	Indirect	0	6	8	14	Segment	Yes
11	Deville 2000 ²³⁶	Indirect	6	6	11	11	Patient	Yes
12	Dong 2008 ²³⁷	Indirect	1	8	21	28	Patient	No
13	Dong 2009 ²³⁸	Indirect	0	4	14	18	Patient	No
14	Doria 2006 ²³⁹	Indirect	5	8	23	26	Patient	No
15	Ewald 2008 ¹²⁵	Direct only	7	7	7	7	Patient	Yes
16	Fleischmann 1998 ¹⁵⁷	Indirect	6	27	23	44	Patient	Yes
17	Gisbert 2006 ²⁴⁰	Indirect	9	16	9	16	Patient	No
18	Gould 2003 ²⁴¹	Indirect	24	33	24	33	Patient	No
19	Granader 2008 ²⁴²	Direct only	4	4	4	4	Lesion	Yes
20	Gu 2007 ¹⁶⁰	Indirect	8	12	15	19	Patient	Yes
21	Hamon 2007 ²⁴³	Indirect	0	12	17	29	Patient	No
22	Hayashino 2005 ²⁴⁴	Indirect	2	9	5	12	Patient	Yes
23	Hodgkinson 2011 ²⁴⁵	Indirect	1	7	3	9	Patient	No
24	Hovels 2008 ²⁴⁶	Indirect	3	10	17	24	Patient	No
25	Karger 2007 ²⁴⁷	Indirect	5	6	5	6	Patient	No
26	Kearon 1998 ²⁴⁸	Indirect	1	18	6	23	Patient	No
27	Kittler 2002 ²¹⁶	Direct only	13	13	13	13	Image /Patient	Yes
28	Koumans 1998 ²⁴⁹	Indirect	0	10	3	13	Patient	No
29	Kriston 2008 ²⁵⁰	Direct only	5	5	5	5	Patient	Yes
30	Ledro-Cano 2007 ²⁵¹	Direct only	7	7	7	7	Patient	No
31	Lewis 2006 ²⁵²	Direct only	34	34	34	34	Patient	No
32	Lewis 2010 ²⁵³	Direct only	11	11	11	11	Patient	No
33	Mahajan 2010 ¹⁶³	Indirect	6	14	15	23	Patient	No
34	Mirza 2010 ¹⁶⁵	Indirect	5	7	21	23	Patient	No
35	Mitchell 2010 ¹⁶⁶	Indirect	4	10	7	13	Patient	Yes
36	Ngamruengphong 2010 ¹⁶⁸	Indirect	3	7	15	19	Patient	No
37	Nishimura 2007 ¹⁷⁰	Indirect	28	37	50	59	Patient	Yes
38	Olatidoye 1998 ²⁵⁴	Indirect	2	12	9	19	Patient	Yes
39	Roos 2007 ²⁵⁵	Direct only	27	27	27	27	Patient	Yes
40	Schuetz 2010 ¹⁴⁷	Indirect	5	89	19	103	Patient	No
41	Schuijf 2006 ²⁵⁶	Indirect	1	24	28	51	Segment	No

Table 7.1 continued...

ID	Reference	Type of test comparison	N _{IC} ¹	N _I ²	N _C ³	N _T ⁴	Unit of analysis	Thresholds ⁵
42	Shie 2008 ²⁵⁷	Direct only	4	4	4	4	Lesion	No
43	Smith 2011 ¹⁷⁷	Indirect	4	15	8	19	Hips	No
44	Sun 2011 ²⁵⁸	Direct only	14	14	14	14	Biopsy	No
45	Tan 2002 ²⁵⁹	Indirect	2	12	15	25	Artery	No
46	Terasawa 2004 ¹⁸⁰	Indirect	4	12	14	22	Patient	No
47	van Randen 2008 ²⁶⁰	Direct only	6	6	6	6	Patient	No
48	Verma 2006 ²⁶¹	Direct only	5	5	5	5	Patient	No
49	Vestergaard 2008 ²⁶²	Direct only	9	9	9	9	Lesion	Yes
50	Visser 2000 ²⁶³	Indirect	0	10	21	31	Segment	No
51	Wang 2005 ²⁶⁴	Direct only	6	6	6	6	Patient	No
52	Wiese 2000 ²⁶⁵	Indirect	7	7	30	30	Patient	No
53	Wijnberger 2001 ²⁶⁶	Direct only	6	6	6	6	Patient	Yes
54	Worster 2002 ²⁰⁴	Direct only	4	4	4	4	Patient	No
55	Xu 2011 ¹⁸⁶	Indirect	3	7	8	12	Patient	No
56	Yang 2009 ²⁶⁷	Direct only	12	12	12	12	Patient	Yes
57	Zhu 2010 ²⁶⁸	Direct only	7	7	7	7	Patient	No

¹N_{IC} = Number of studies comparing index test and comparator (i.e. comparative studies).

²N_I = Index test – experimental or newer test.

³N_C = Comparator – current practice or older test.

⁴N_T = Total number of studies

⁵Whether or not thresholds were used to determine test positivity.

ID uniquely identifies each test comparison. The table was sorted according to ID.

The SROC plots for the 57 test comparisons were grouped into three figures (Figure 7.1, Figure 7.2 and Figure 7.3) for ease of viewing each plot. The number on each plot is a unique identifier for each test comparison and corresponds to the ID column in Table 7.1. The test comparisons will be referred to using these IDs in the rest of the chapter. When reference is made to the meta-analysis of one of the tests in a test comparison, the index test or comparator will be identified along with the test comparison ID. Figure 7.1 shows test comparisons that included studies with different thresholds while Figure 7.2 and Figure 7.3 show test comparisons involving studies that used common thresholds. The latter were presented on two figures because of the number of datasets.

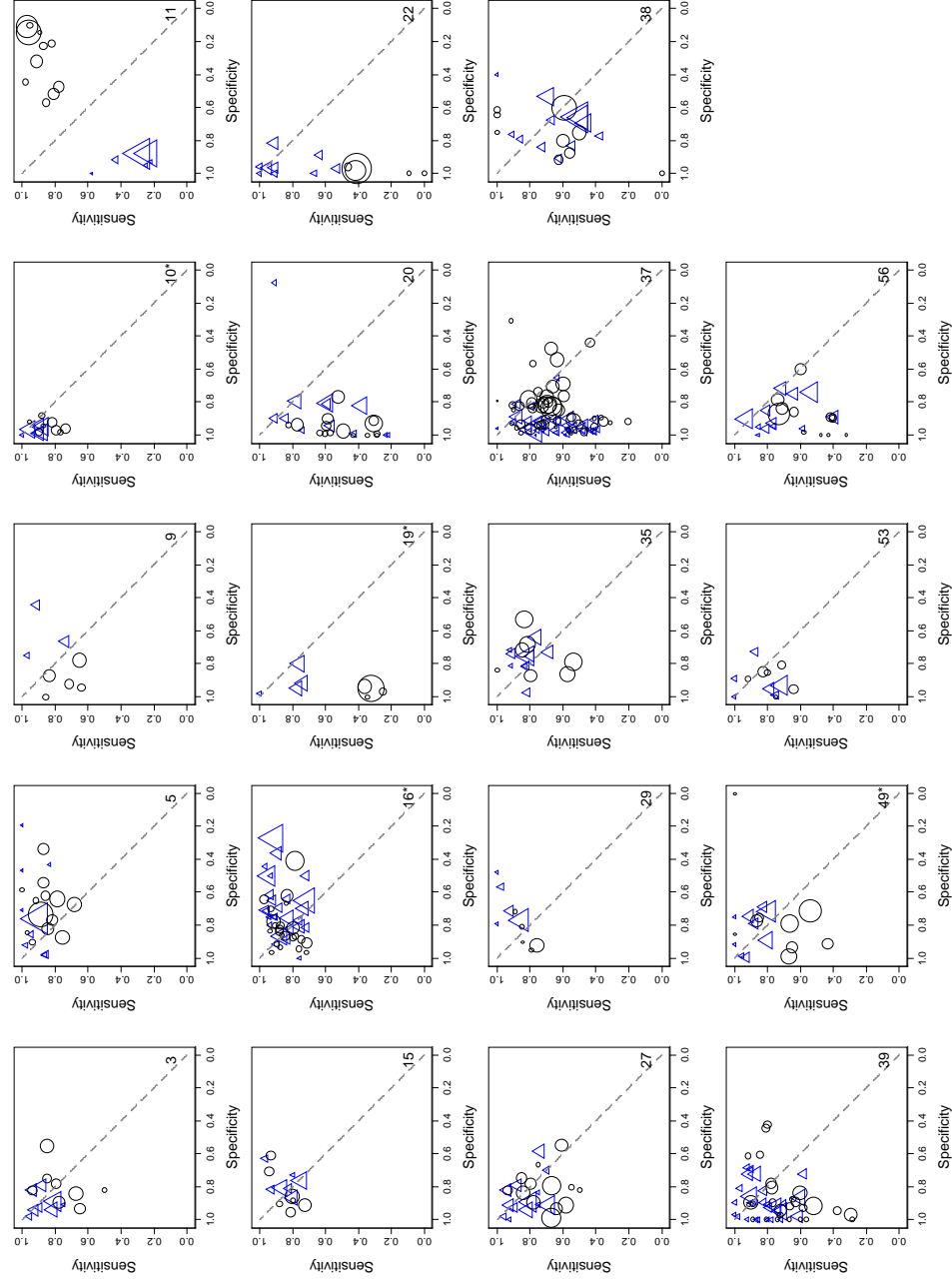


Figure 7.1 | SROC plots of meta-analytic datasets with mixed thresholds

*Studies in these meta-analyses used non-numeric thresholds. In each SROC plot, the index test is represented by the blue triangles and the comparator by the black hollow circles. The number on each SROC plot is the ID which uniquely identifies each test comparison and corresponds to the ID in Appendix D.1 and Table 7.1.

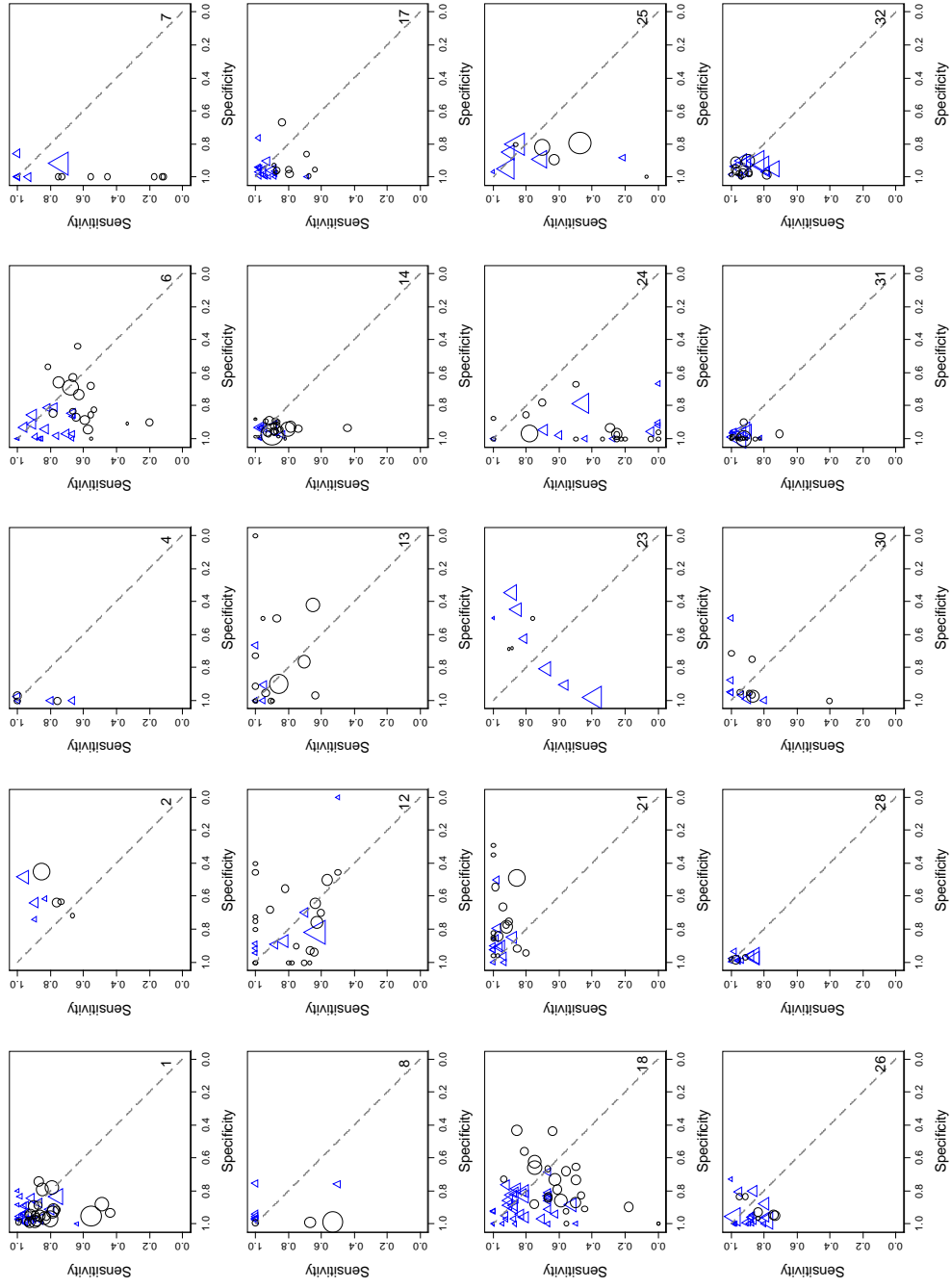


Figure 7.2| SROC plots of meta-analytic datasets with common thresholds (1)

In each SROC plot, the index test is represented by the black hollow circles and the comparator by the blue hollow triangles. The number on each SROC plot is the ID which uniquely identifies each test comparison and corresponds to the ID in Appendix D.1 and Table 7.1.

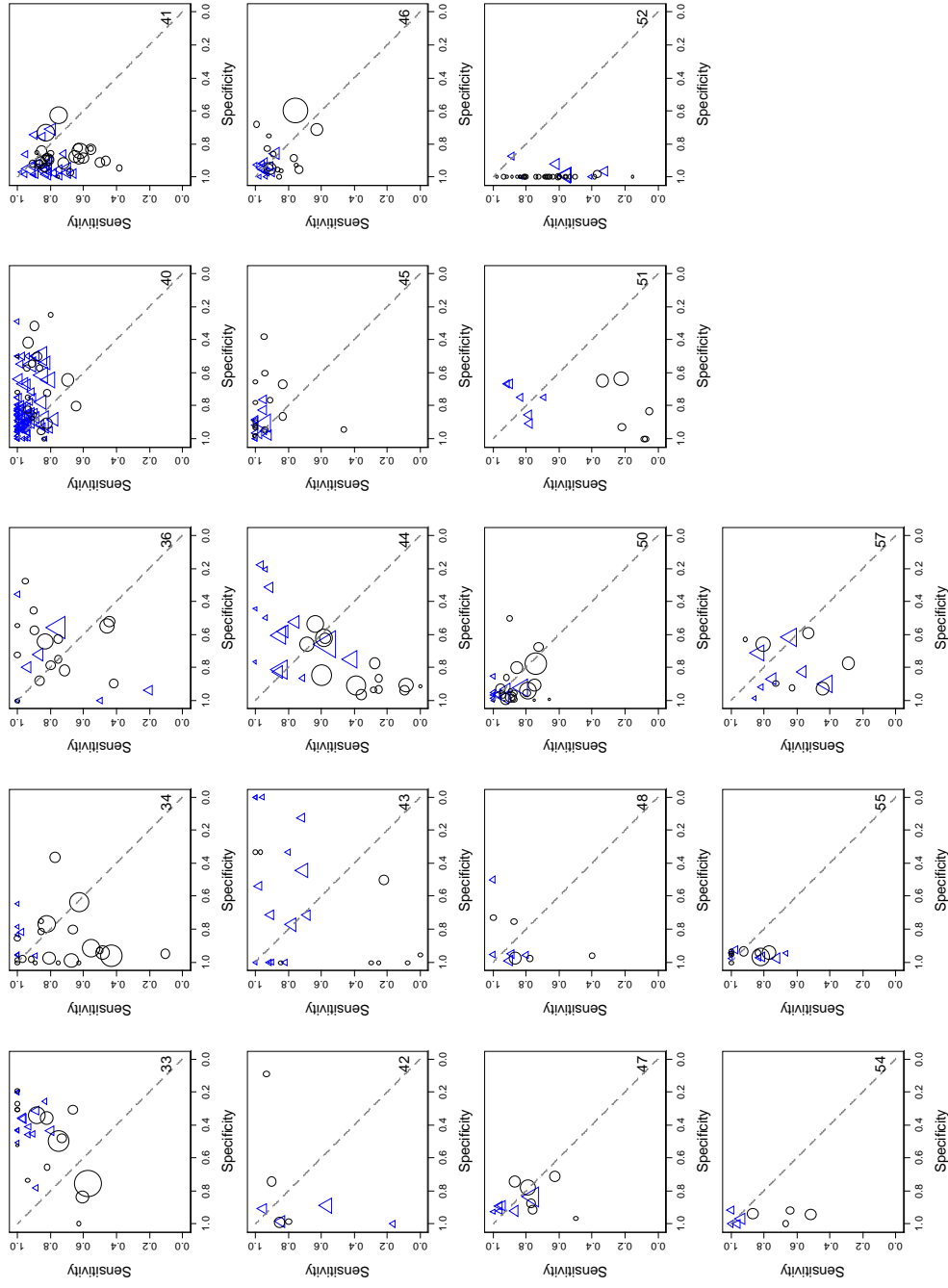


Figure 7.3| SROC plots of meta-analytic datasets with common thresholds (2)

In each SROC plot, the index test is represented by the black hollow circles and the comparator by the blue hollow triangles. The number on each SROC plot is the ID which uniquely identifies each test comparison and corresponds to the ID in Appendix D.1 and Table 7.1.

7.3.1 Types of test comparisons

Twenty two (39%) reviews restricted test comparison to direct comparisons while the remaining 35 (61%) reviews performed indirect comparisons (Table 7.1). Of the 35 indirect comparisons, eight (23%) had no comparative studies and 27 (77%) included between one and 28 comparative studies. The total number of studies in the test comparisons ranged between six and 103. The distribution of the number of studies for the index test and comparator in the test comparisons is shown in Figure 7.4. The median (interquartile range) number of studies for the index and comparator tests were 8 (6 to 13) and 10 (6 to 17). Nineteen (33%) test comparisons included studies with different numeric thresholds or different non-numeric criteria to determine test positivity.

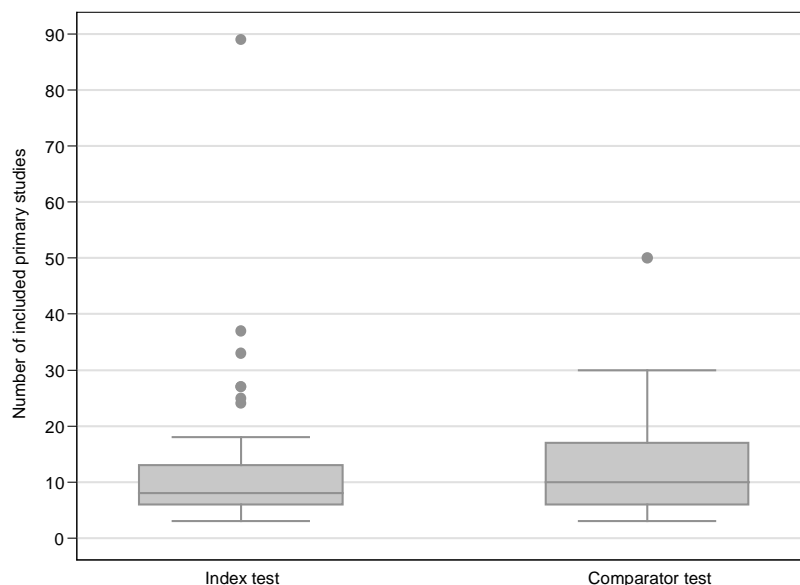


Figure 7.4| Distribution of number of primary studies for the index and comparator tests in the test comparisons

7.3.2 Frequency of zero cells

The distribution of the number of studies with zeros in any of the cells of the 2x2 tables in the 57 comparative meta-analyses is shown in Figure 7.5. The median (interquartile range)

number of studies with zero cells was 4 (2 to 7). In four test comparisons, there were no zero cells in any of the included studies while there were 39 studies with zero cells in one test comparison. This means that for the empirical evaluation of Moses SROC meta-regression, zero cell corrections will be applied in 53 of the 57 (93%) comparative meta-analyses.

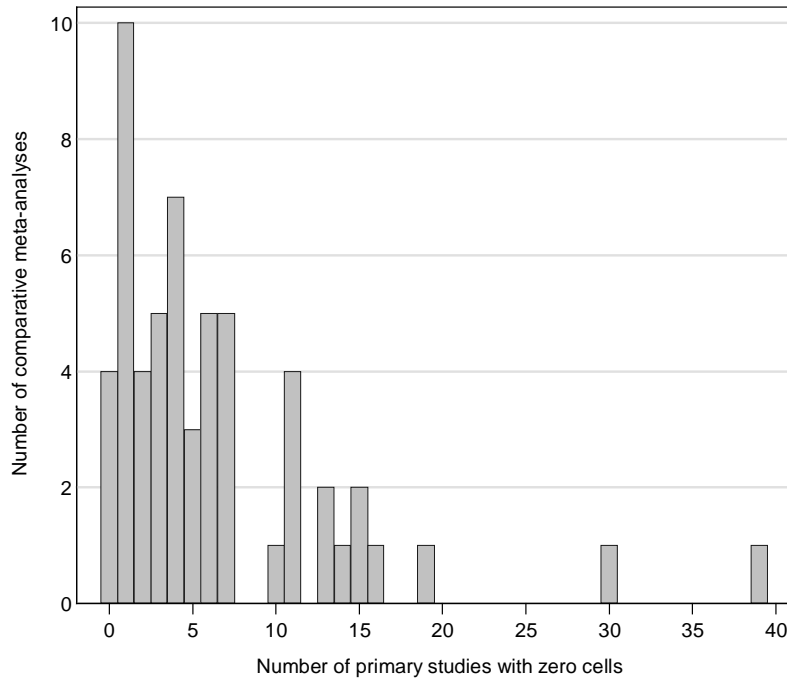


Figure 7.5| Number of studies with zero cells in each test comparison

PART II: ASSESSMENT OF MODELLING ASSUMPTIONS IN HIERARCHICAL MODELS

In the second part of this chapter, results of the data analysis methods described in section 7.2.3 for investigating modelling assumptions commonly made in bivariate and HSROC models are described. In section 7.4, results are presented for the analyses based on separate meta-analyses of tests in each test comparison. The analyses explored the reliability of the models, as well as the potential for similarity of variances of random effects and shape of SROC curves between tests in test comparisons. These preliminary analyses addressed the following questions that were relevant to methodological issues raised in Chapter 2:

1. How valid is the estimation of parameters in hierarchical models?
 - a. Did analyses of hierarchical models converge?
 - b. Is the correlation parameter in the bivariate model reliably estimated?
 - c. If estimated reliably, is the correlation parameter important for valid estimation of variances and standard errors of the estimates of the mean logit sensitivity and logit specificity?
 - d. Is the shape parameter in the HSROC model reliably estimated?
2. Is heterogeneity in test performance similar between tests in test comparisons?
3. Is the shape of SROC curves similar between tests in test comparisons?

Part II concludes in section 7.5 with a summary of the findings.

7.4 Assessment of modelling assumptions based on separate meta-analyses of tests in test comparisons

The number of studies in the 114 meta-analyses (separate meta-analyses of two tests from 57 test comparisons) ranged between 3 and 89 (Figure 7.6). More than half of the meta-analyses

(60/114, 53%) had less than 10 studies, 36 (32%) had between 10 and 20 studies, and the remaining 18 (16%) had more than 20 studies.

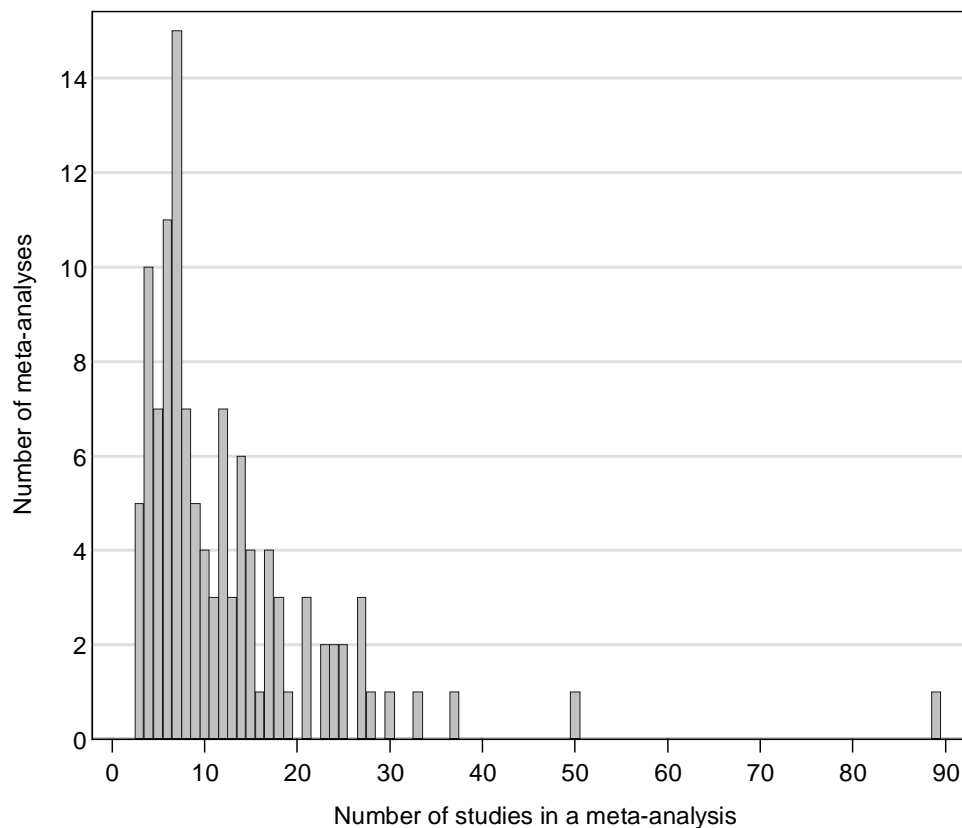


Figure 7.6| Distribution of number of studies in each meta-analysis of a single test

7.4.1 Are hierarchical models stable in meta-analyses of individual tests?

7.4.1.1 Did bivariate and HSROC models converge?

Using the bivariate model, the 114 meta-analyses converged though two meta-analyses required the use of *gllamm* instead of *xtmelogit*. The estimates appeared to be valid based on the criteria defined in section 7.2.3.3. For the HSROC model, optimization failed to complete for one meta-analysis and so 113 meta-analyses satisfied the NLMIXED convergence criteria. However, in five of the 113 meta-analyses, standard errors were missing for all parameters or were extremely large. The parameter estimates for the five meta-analyses are shown in Table

7.2. The table also shows that gradient values for the parameters were close to zero except for variance parameters where boundary constraints were activated. These five meta-analyses were different to the two that were difficult to fit using *xtnlogit*.

Table 7.2| Parameter estimates from unstable HSROC models

Parameter	Estimate	Standard error	95% confidence interval	Gradient
<i>Basaran 2009 (ID 4) – index test (4 MRI studies)</i>				
Λ	6.25	122	-233 to 246	3.58E-8
Θ	0.99	193	-378 to 380	-3.71E-8
β	1.57	124	-241 to 244	2.26E-8
σ_{α}^2	0	–	–	0.468849
σ_{θ}^2	0	–	–	0.876607
<i>Brazzelli 2009 (ID 7) – index test (7 CT studies)</i>				
Λ	10.5	823	-1605 to 1626	0.000016
Θ	-6.05	409	-809 to 797	0.000033
β	1.21	41.1	-79.4 to 81.8	0.000608
σ_{α}^2	0.21	–	–	-0.000030
σ_{θ}^2	4.56	–	–	-0.000130
<i>Hayashino 2005 (ID 22) – comparator (5 ventilation perfusion scanning studies)</i>				
Λ	0.76	–	–	0.000031
Θ	-1.28	–	–	-0.000010
β	1.57	–	–	0.000042
σ_{α}^2	0	–	–	11.60103
σ_{θ}^2	0	–	–	44.21522
<i>Koumans 1998 (ID 28) – comparator (3 nuclei acid amplification (LCR) studies)</i>				
Λ	7.57	28.0	-47.4 to 62.6	-2.77E-7
Θ	0.97	54.8	-107 to 108	8.54E-7
β	0.68	28.9	-56.1 to 57.5	-1.35E-6
σ_{α}^2	0	–	–	0.173067
σ_{θ}^2	0	–	–	6.763934
<i>Worster 2002 (ID 54) – index test (4 non-contrast helical CT studies)</i>				
Λ	7.23	–	–	3.11E-7
Θ	-0.59	–	–	8.94E-8
β	-0.19	–	–	2.22E-8
σ_{α}^2	0	–	–	0.217482
σ_{θ}^2	0	–	–	0.834333

– indicates no estimate available.

Each ID uniquely identifies a test comparison dataset. See Figure 7.1, Figure 7.2 and Figure 7.3 for SROC plots that correspond to the IDs. Λ , Θ , β , σ_{α}^2 and σ_{θ}^2 are the five parameters of the HSROC model representing accuracy, threshold, shape of the SROC curve, and variances of the random effects for accuracy and threshold. Where σ_{α}^2 and σ_{θ}^2 are zero, boundary constraints ($\sigma_{\alpha}^2 \geq 0$ and $\sigma_{\theta}^2 \geq 0$) were activated. Forest plots of study specific estimates of sensitivity and specificity are presented in Appendix D.2 for the five tests.

Forest plots of the data for each of the five meta-analyses and the one that did not converge are shown in Appendix D.2. A common observation across the six datasets was that the number of studies was small and/or data were sparse due to studies with 100% sensitivity and/or specificity. In one meta-analysis for ID 7 (also see Figure 7.2), all studies of the comparator (CT) had a specificity of 100% with sensitivities between 11% and 75%.

7.4.1.2 Is the correlation parameter in the bivariate model reliably estimated?

For eight of the 57 pairs of tests, ρ was estimated as +1 or -1 for both tests. Figure 7.7 shows that estimates of ρ were rarely similar for a pair of tests, with very few points lying close to the diagonal. Furthermore, negative estimates of correlation occurred more frequently for the index tests (43/57, 75%) than for the comparator tests (29/57, 51%). In most (80/114, 70%) meta-analyses, ρ was estimated within the boundary of the parameter space (above -1 and below +1). For the remaining 34 (30%) meta-analyses, ρ was estimated as +1 or -1; estimates of -1 occurred more frequently (22/34, 65%).

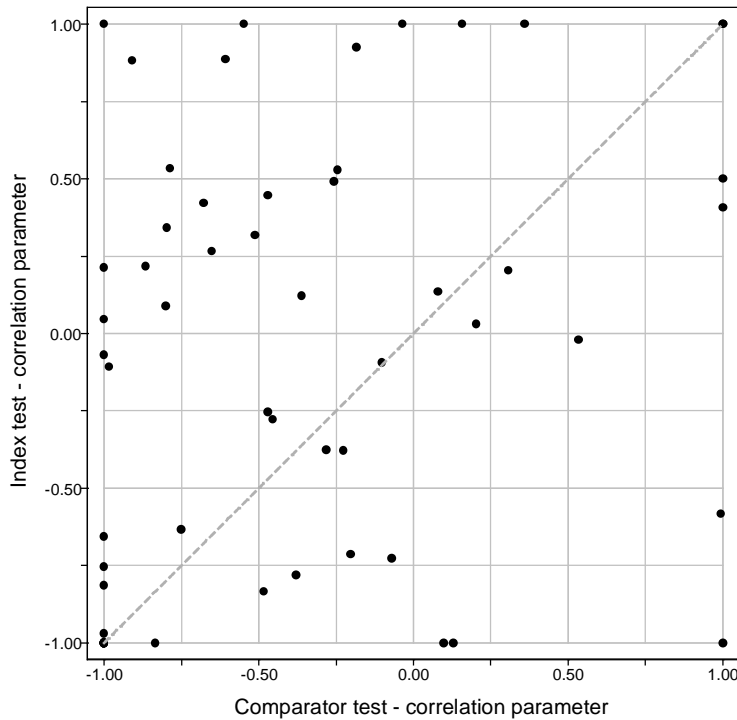


Figure 7.7| Estimates of the correlation parameter from meta-analyses of individual tests in 57 test comparisons

If the correlation between the logits of sensitivity and specificity for the index and comparator tests in the test comparisons were similar, then the points should lie along or close to the dashed diagonal line.

7.4.1.3 *Is the correlation parameter needed for valid estimation of variances and standard errors?*

The variance of logit specificity from one univariate meta-analysis was extremely high, 60.6, while the corresponding variance from the bivariate meta-analysis was 1.05. Since this was the meta-analysis of CT, mentioned earlier, where all studies had 100% specificity, this result is clearly invalid and so was not used in the comparison of the two models to avoid skewing the results. Point estimates of the variances of the random effects of the logits from univariate and bivariate models were thus compared for 113 meta-analyses (Figure 7.8). Variance estimates from univariate and bivariate models were fairly similar (especially for variances of logit specificities). Using univariate models as the reference category, the median (interquartile range) of differences between the variance estimates from both models were

0.012 (-0.005 to 0.073) and 0.004 (-0.003 to 0.035) for logit sensitivity and logit specificity respectively. For the subset of 34 meta-analyses where the correlation was +1 or -1, the median (interquartile range) of the differences were 0.070 (0.016 to 0.142) and 0.021 (0.003 to 0.063) for logit sensitivity and logit specificity respectively. The higher estimates observed in this subset is likely due to poor estimation of correlation as +1 or -1 in the bivariate models.

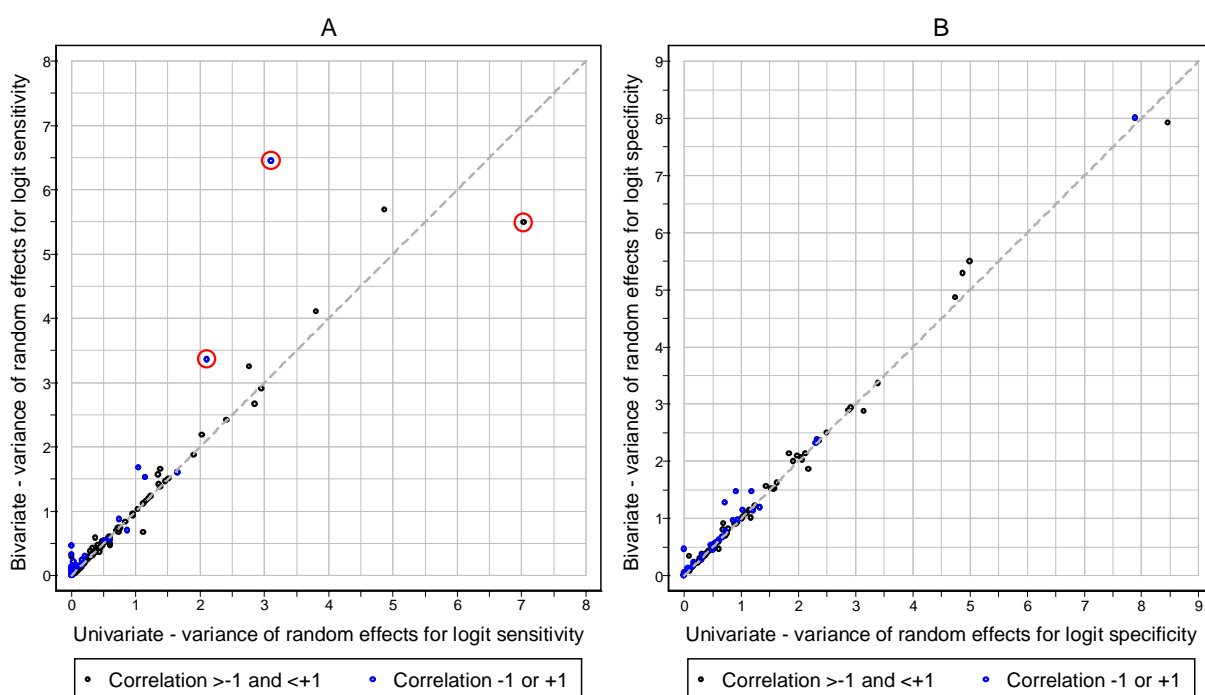


Figure 7.8| Comparison of estimates of variances from bivariate and univariate meta-analyses

If estimates from both meta-analyses were similar, the black dots would lie along or close to the dashed diagonal line. In panel A, the three points circled in red have the largest difference between both models.

The three points circled in red in Figure 7.8 (panel A) have the largest absolute difference in variance estimates for logit sensitivity. These three meta-analyses had less than 10 studies. Of the three meta-analyses, variances of logit sensitivity were higher in bivariate compared to univariate models for two meta-analyses and the correlation from the bivariate models was

+1. For the third meta-analysis where the variance of logit sensitivity was lower in the bivariate compared to the univariate model, the correlation was -0.66 . The meta-analysis with the largest difference (3.34) had only 20 diseased patients (Figure 7.9) and the sensitivity of abdominal US (index test) for screening for ovarian cancer was 100% in six of the seven studies (see also SROC plot 8 in Figure 7.2).

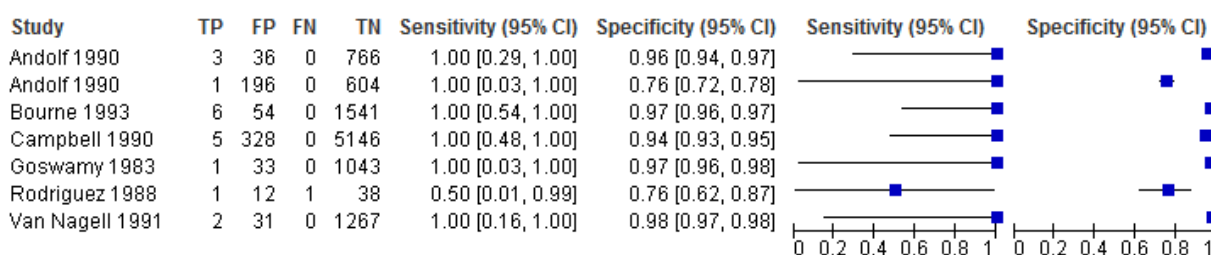


Figure 7.9| Forest plot of sensitivity and specificity for the dataset with the largest difference between univariate and bivariate meta-analyses
Data taken from Carlson et al.²³⁴

Estimates of the standard errors of mean logit sensitivities and specificities were also compared as shown in Figure 7.10. Estimates of the standard errors of both logit sensitivity and logit specificity from univariate models were often (76/113, 67%) smaller than those from bivariate models though the absolute differences tended to be small. The median (interquartile range) of the differences between the estimates were 0.001 (-0.001 to 0.024) and 0.001 (-0.001 to 0.009) for logit sensitivity and logit specificity respectively. For the subset of 34 meta-analyses where the correlation was +1 or -1 , the median (interquartile range) of the differences were 0.020 (0.002 to 0.059) and 0.005 (0.000 to 0.026) for logit sensitivity and logit specificity respectively. In panel A, the point circled in red was the largest difference (0.497) between both models and is from meta-analysis of the index test in test comparison 8 (see SROC plot in Figure 7.2). In panel B, the point circled in red was the largest difference

(-1.43) on the plot and is from meta-analysis of the comparator for ID 52 (see SROC plot in Figure 7.3).

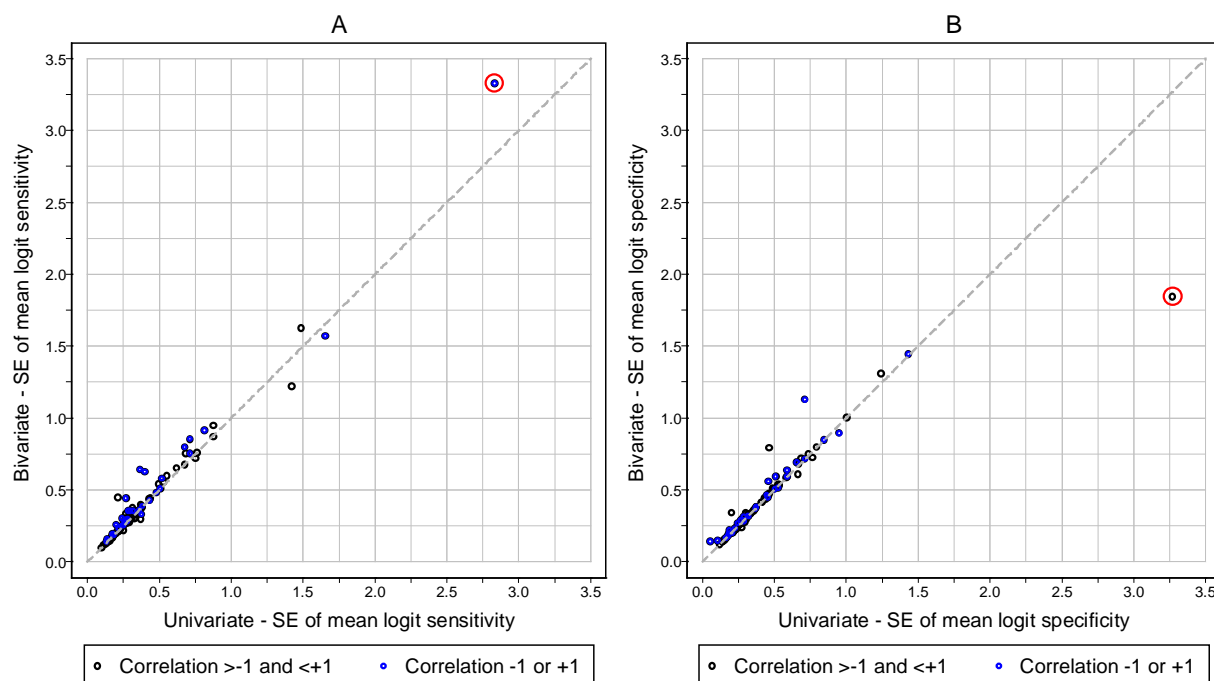


Figure 7.10| Comparison of estimates of standard errors of mean logit sensitivities and mean logit specificities from univariate and bivariate meta-analyses

SE = standard error.

If estimates from both meta-analyses were similar, the black dots would lie along or close to the dashed diagonal line. In panels A and B, the points circled in red show the largest absolute difference between both models on the plots.

7.4.1.4 Is the shape parameter in the HSROC model reliably estimated?

Estimates of β from HSROC models were compared with those derived using functions of the bivariate model parameters as shown in Figure 7.11. The scatterplot shows that most of the estimates (108/113, 96%) lie on or close to the dashed diagonal line. This indicates that estimates of β from HSROC models and those derived from functions of bivariate model parameters were similar. The five points circled in red are the five pairs of estimates that differ substantially between both models. The five estimates from HSROC models were from

the models that converged but gave unreliable estimates (see Table 7.2); the corresponding bivariate models appeared to give reliable estimates.

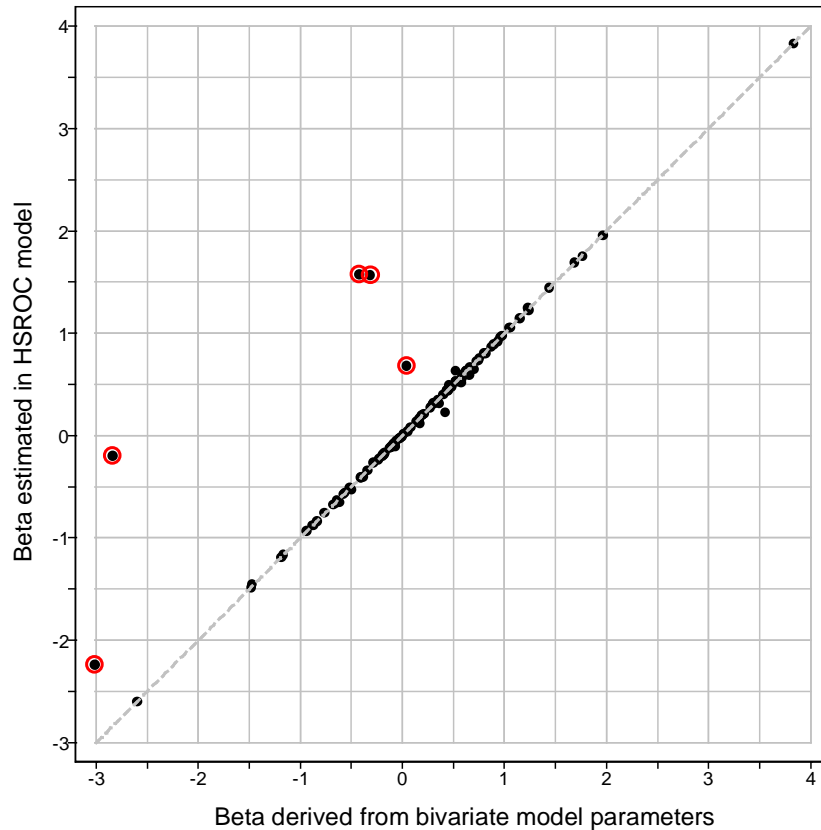


Figure 7.11| Comparison of estimates of beta from HSROC models with estimates derived from bivariate models

The five pairs of estimates that do not agree (circled in red) correspond to the five meta-analyses where the HSROC model converged but gave unreliable estimates. For these five meta-analyses, the bivariate model appeared to give reliable estimates.

The effect of the number of studies in a meta-analysis on the estimation of β was explored using scatterplots of estimates of β and its standard error against the number of studies in each meta-analysis as illustrated in Figure 7.12. The two plots included the 108 meta-analyses where the HSROC models were considered valid. The magnitude of β appears to decrease as the number of studies increased (panel A). The relationship between the standard error of β and the number of studies is evident with increased precision as the number of studies

increased (panel B). As shown by the blue circles in each plot, there were 16 meta-analyses where there was statistical evidence that β differed from zero.

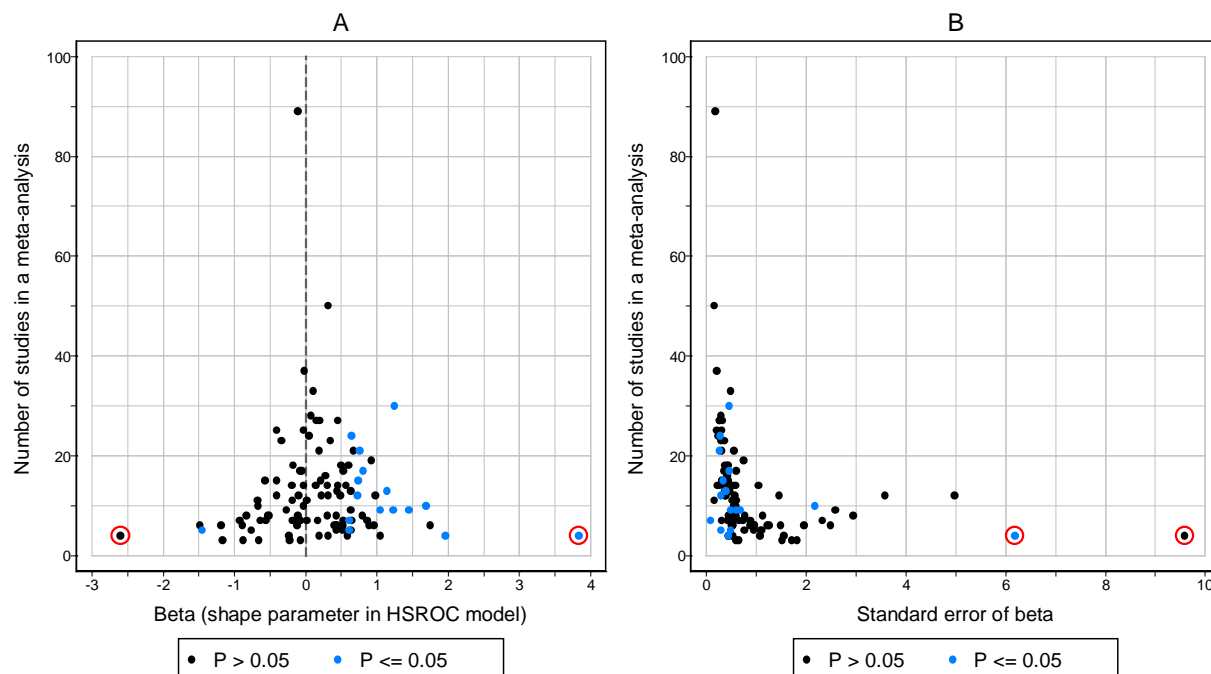


Figure 7.12| Estimates of beta and its standard error against number of studies in each meta-analysis

The plots show only the 108 meta-analyses where the HSROC model converged and gave reliable estimates. The blue solid circles represent meta-analyses where there was statistical evidence (at the 5% significance level) that the shape of the SROC curve was asymmetric (i.e. accuracy depends on threshold); the black circles represent meta-analyses where there was no statistical evidence of asymmetry in the SROC curve. On each plot, the two points surrounded by a red circle have extremely large values for beta and their standard errors.

Estimates of β were also plotted with their 95% confidence intervals for each of the 108 meta-analyses as shown in Figure 7.13. The three vertical dashed lines delineate four sections according to number of included studies; sections A, B, C and D include meta-analyses with ≤ 5 studies, 6 to 10 studies, 11 to 20 studies, and >20 studies respectively. As the number of studies increased, confidence intervals became narrower but there were three notable exceptions in section C. These three meta-analyses (index tests in test comparisons 21, 45 and

46) each included 12 studies. Uncertainty in the estimation of beta may be partly due to little or no heterogeneity in threshold (Figure 7.14).

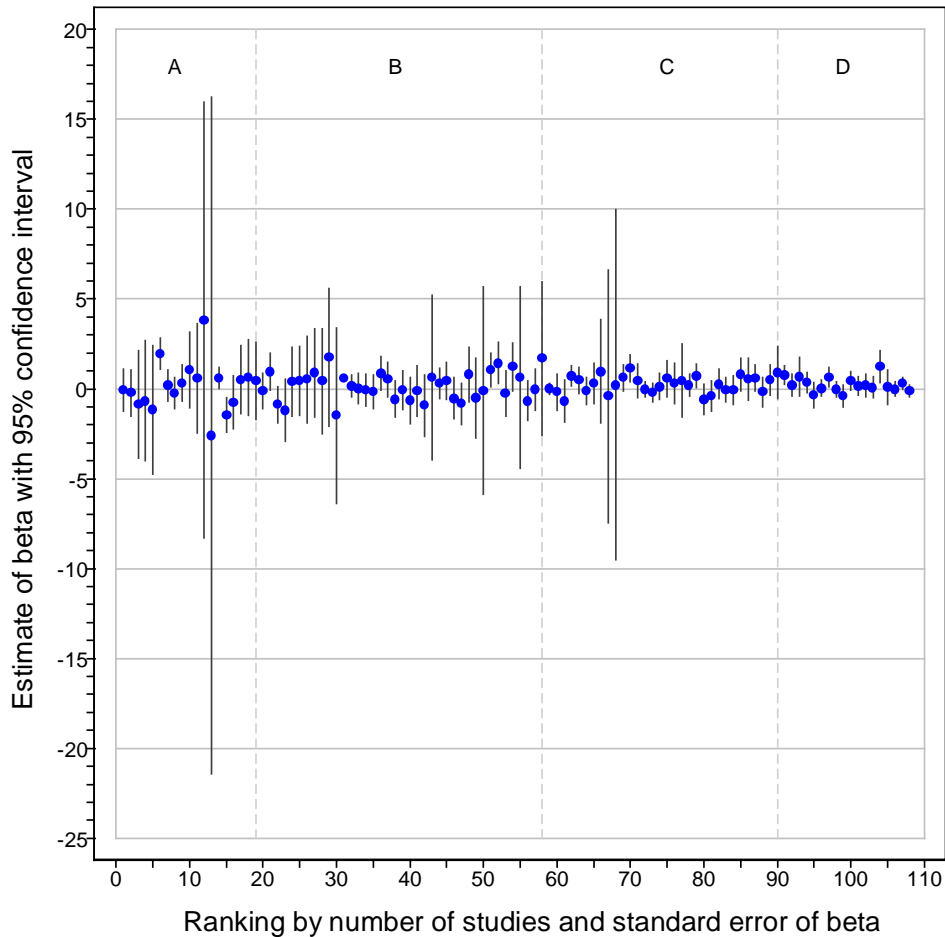


Figure 7.13| Estimates of beta and their 95% confidence intervals from HSROC models Each blue dot represents the point estimate of beta and is drawn with its 95% confidence interval. Estimates of beta from the 108 meta-analyses are ranked on the plot according to the number of included studies and the standard error of beta. The three vertical dashed lines delineate four sections based on the number of included studies in the meta-analyses; sections A, B, C and D include meta-analyses with ≤ 5 studies, 6 to 10 studies, 11 to 20 studies, and >20 studies respectively.

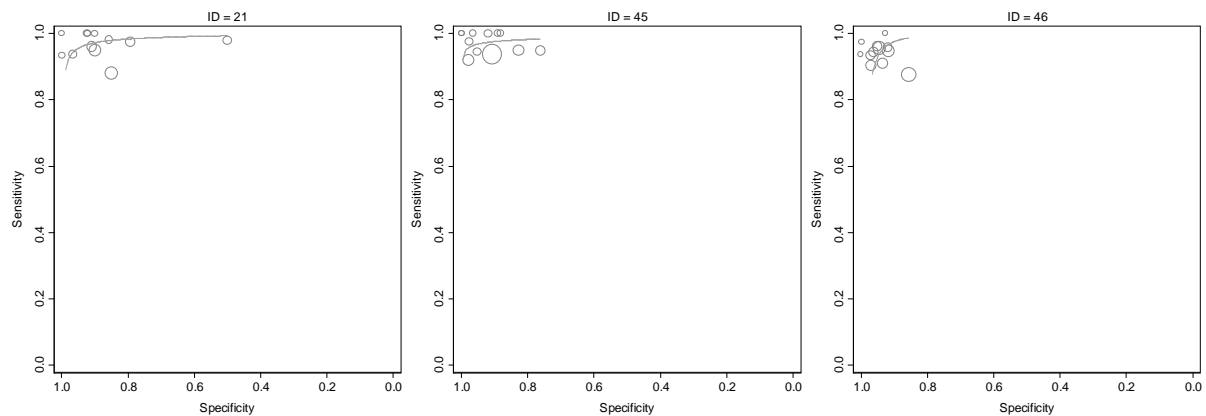


Figure 7.14| SROC plots of three meta-analyses with 12 included studies and substantial uncertainty in estimation of beta

7.4.2 Is heterogeneity in test performance similar between tests in test comparisons?

Scatterplots of the 57 test comparisons shown in Figure 7.1, Figure 7.2 and Figure 7.3 often indicated heterogeneity in test performance. Sometimes, strong heterogeneity in sensitivity but homogeneity in specificity (e.g. ID 55 in Figure 7.3), or vice versa was observed (e.g. ID 21 in Figure 7.2). Differences in patterns of heterogeneity between tests in a comparison were also observed. For example, ID 46 in Figure 7.3 shows homogeneity in sensitivities and specificities of the index test but heterogeneity in both measures for the comparator. For each pair of tests in a test comparison, estimates of the variances of the random effects for logit sensitivities and logit specificities were obtained from a separate bivariate model for each test and compared as shown in Figure 7.15 (panels A and B). The plots indicate that the variances for both tests in a test comparison sometimes differed substantially, even in meta-analyses of tests from a direct comparison. The differences may be due to chance. Therefore, the magnitude and importance of these differences will be explored in comparative meta-analyses in section 7.6.2 by using bivariate meta-regression models.

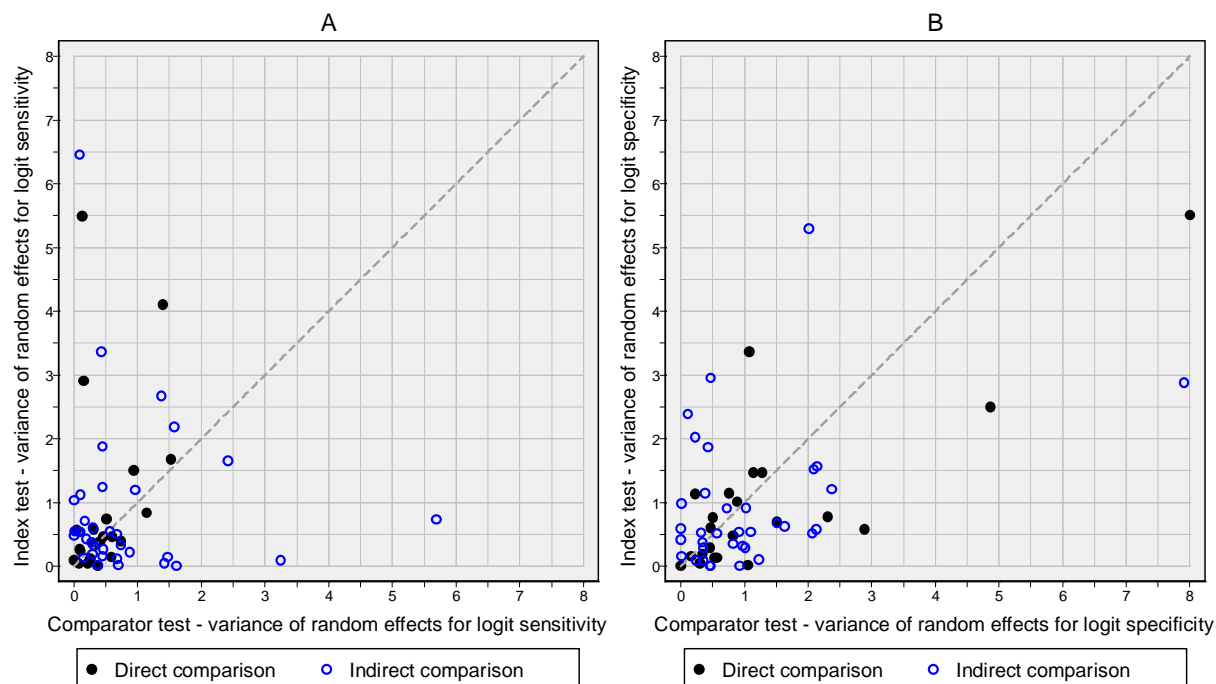


Figure 7.15| Variance estimates for random effects for logit sensitivity and logit specificity from meta-analysis of each test in the test comparisons

A scatterplot comparing the variances of the random effects for logit sensitivity from a bivariate model fitted to each test in a test comparison is shown in (A). A similar plot is shown for variances of logit specificity in (B). Each plot shows results from 114 meta-analyses for 57 pairs of tests. The black solid circles represent estimates from meta-analyses of direct comparisons (i.e. all studies evaluated both tests) while the blue hollow circles represent indirect comparisons.

Estimates of the variances for accuracy and threshold from the HSROC model were also examined. In 36 of the 108 (33%) meta-analyses, boundary constraints were activated for variances of the accuracy or threshold parameters, i.e., estimation truncated at zero. Similar to the findings for the bivariate model, the plots indicate that the variances for both tests in a test comparison sometimes differ (Figure 7.16).

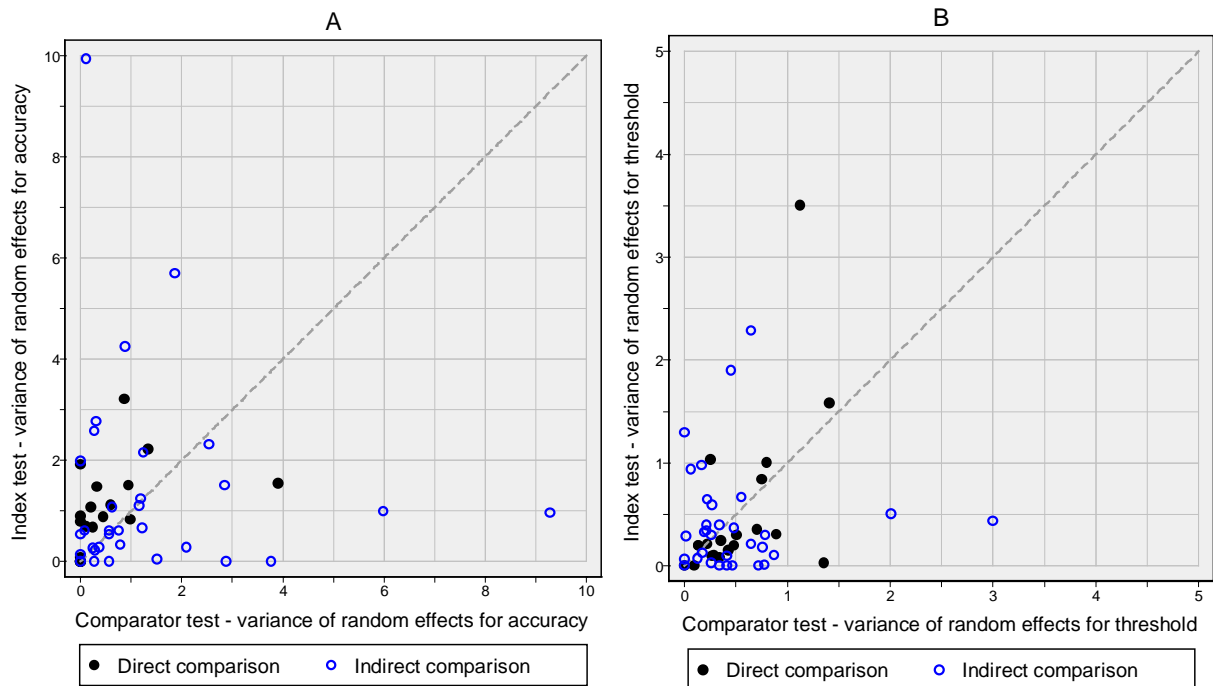


Figure 7.16| Variance estimates for random effects for accuracy and threshold from meta-analysis of each test in the test comparisons

A scatterplot comparing the variances of the random effects for accuracy from an HSROC model fitted to each test in a test comparison is shown in (A). A similar plot is shown for variances of threshold in (B). The black solid circles represent estimates from meta-analyses of direct comparisons (i.e. all studies evaluated both tests) while the blue hollow circles represent indirect comparisons. Of the 108 meta-analyses with no convergence issues, results were available for 52 pairs of tests and so the plot shows results from 104 meta-analyses.

7.4.3 Is the shape of SROC curves similar between tests in test comparisons?

For each test comparison, the estimate of β for the index test was compared with that of the comparator test. Figure 7.17 shows that β often differs between tests but it cannot be inferred from the plot whether these are likely to be important differences or not. Section 7.6.3 will address this further where the shape parameter will be investigated in comparative meta-analyses by using HSROC meta-regression models.

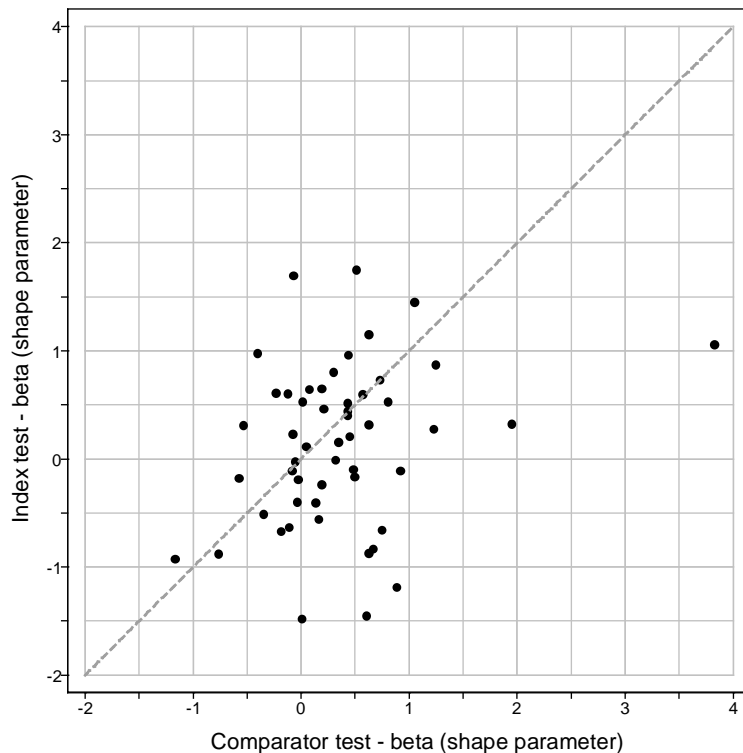


Figure 7.17| Estimates of beta (shape parameter) from separate meta-analyses of pairs of tests in 52 test comparisons

The plot shows 52 pairs of tests because the HSROC model did not converge or estimates were unreliable for one or both tests in five test comparisons. If the SROC curves of both tests in the test comparisons have similar shape, then the black dots should lie along or close to the dashed diagonal line.

7.5 Summary of assessment of modelling assumptions in hierarchical models

Problematic datasets which gave unreliable estimates either had a small number of studies and/or sparse data due to studies with 100% sensitivity and/or specificity. In one particular meta-analysis, the seven studies all had a specificity of 100% with sensitivities between 11% and 75%. In the original review, the authors pooled sensitivities and specificities separately using a random effects model for sensitivity and a fixed effect model for specificity.²³³

Methods for meta-analysis with sparse data (where non-convergence is most likely to occur) will be fully investigated in a simulation study in Chapter 8.

Estimates of the correlation and variance parameters for pairs of tests were seldom similar which raised doubts about the assumption of equal variances in comparative meta-analyses. This will be investigated in section 7.6. In about a third of the meta-analyses, the correlation of the logits was estimated on the boundary of the parameter space as +1 or -1. In the remaining meta-analyses, estimates of the correlation were predominantly negative. A negative correlation can be expected across studies due to trade-off in sensitivity and specificity as threshold varies across studies. However, due to other sources of heterogeneity besides threshold, such as the factors stated in section 1.4.1, positive estimates can occur. There was limited effect on estimates of the variances or standard errors of both logit sensitivity and logit specificity when the bivariate model was reduced to two separate univariate models for sensitivity and specificity. This implies minimal impact of using a bivariate structure to model the data which may provide a solution to non-convergence when data are sparse. The impact of univariate and bivariate meta-regression models on findings will be examined in section 7.7.

There was substantial uncertainty in estimates of the shape parameter when there were few studies or little or no heterogeneity in threshold. The shape of SROC curves often differed between tests with the magnitude of the shape parameter decreasing as the number of studies increased. This implies that strong asymmetry in meta-analyses with few studies is likely to be a chance finding. In a few (14%) meta-analyses, there was statistical evidence of asymmetry of the SROC curve. Therefore, assuming symmetry of SROC curves by eliminating the shape parameter may be appropriate especially when there are few studies, and substantial uncertainty in its estimation and/or variances of the random effects. This will also be investigated further in the simulation study in Chapter 8.

Having examined different modelling assumptions, it is clear that findings may differ depending on the assumptions used. Therefore, the validity of assumptions should be investigated if data permits. Assuming the same asymmetry for SROC curves in HSROC meta-regression models may have less of an impact on relative test performance compared to assumptions about covariance structures in bivariate models. This hypothesis will be explored in Part III.

PART III: IMPACT OF DIFFERENT MODELS ON FINDINGS

The third and final part of this chapter builds on the findings in Part II by exploring the extent to which different hierarchical meta-regression models agree or disagree under the two key modelling assumptions—same variances and same shape across tests—that were extensively explored earlier in Part II. The impact of clustering test results within comparative studies and the impact of using a bivariate structure (by comparing univariate and bivariate meta-regression models) will also be investigated. Part III also considers Moses SROC meta-regression by comparing findings from unweighted and weighted analyses that assume a common shape or allow shape of SROC curves to differ by test. In addition, Moses and HSROC meta-regression models will be compared. As such Part III details the impact of different models on findings. Specifically, the six questions addressed were:

- I. Is there a difference between the findings from meta-analyses in which test results are clustered within comparative studies in bivariate meta-regression models and findings from meta-analyses that ignore such clustering?
- II. Are there important differences between findings from bivariate meta-regression models that assume common variances across tests and those which allow variances to differ by test?
- III. Should shape of SROC curves differ in HSROC meta-regression models or can a common shape be assumed without impact on relative test performance?
- IV. Are the findings from univariate meta-regression models similar to those from bivariate meta-regression models?
- V. How comparable are findings from unweighted and weighted Moses SROC meta-regression models if shape is assumed to be common or allowed to differ by test?

- VI. Do results differ between HSROC and Moses SROC meta-regression models if shape is assumed to be common or allowed to differ by test?

The impact of different modelling complexity on findings from hierarchical meta-regression models are presented in section 7.6 (questions 1 to 3) while the performance of different comparative meta-analysis models are presented in section 7.7 (questions 4 to 6).

7.6 Impact of different modelling complexity on findings

7.6.1 Do test results need to be clustered within comparative studies in bivariate meta-regression models?

The within-study and between-study approaches are the same when there are no comparative studies in a test comparison. Therefore, the seven test comparisons that contained only non-comparative studies (IDs 4, 5, 10, 13, 21, 28 and 50) were excluded from these analyses. For IDs 8 and 51, the models failed to converge. Of the remaining 48 test comparisons, one (ID 7) did not converge for the within-study approach while 9 (IDs 11, 16, 19, 23, 25, 32, 48, 53 and 55) did not converge for the between-study approach (Figure 7.19). Therefore, the two models were compared using estimates from 38 test comparisons of which 16 (42%) were direct comparisons.

Using the between-study approach as the reference category, Figure 7.18 shows ratios of the relative sensitivities and relative specificities (panel A) and ratios of their standard errors from the two models (panel B). For eight (21%) test comparisons, there was more than a 10% difference in relative sensitivities (IDs 20, 24, 36, 42 and 52) or relative specificities (IDs 29, 33 and 44). On average, the magnitude of differences in the point estimates (panel A) were negligible with a median (interquartile range) of 1.00 (0.98 to 1.02) for ratio of relative

sensitivities and 1.00 (0.99 to 1.02) for ratio of relative specificities. However, estimates from the within-study approach were on average more precise than those from the between-study approach (panel B); median (interquartile range) ratios of the standard errors of log relative sensitivities and log relative specificities were 0.74 (0.54 to 0.90) and 0.69 (0.43 to 0.78).

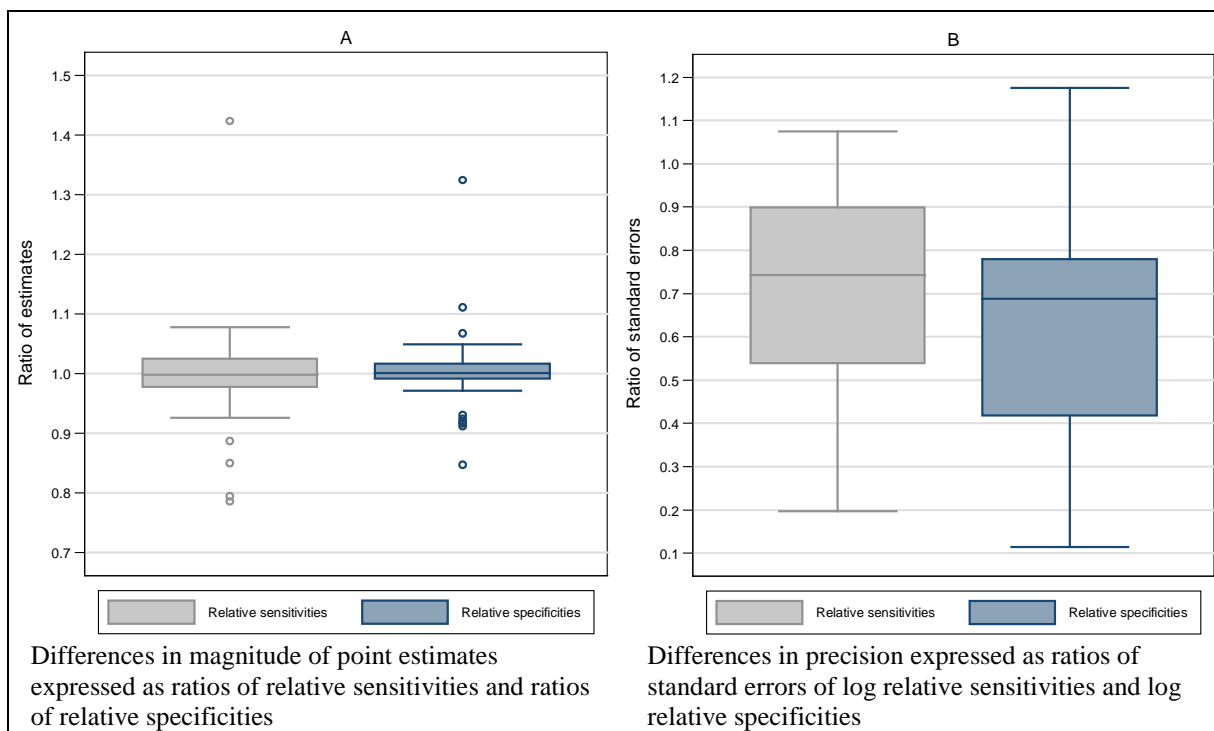


Figure 7.18| Differences in estimates from bivariate meta-regression models with and without clustering of test results in comparative studies

Panel A shows the ratios of the relative sensitivities and relative specificities between the within-study and between-study approaches. The between-study approach was used as the reference category. Relative sensitivities and relative specificities were estimated on the log scale and so Panel B shows the ratios of the standard errors of the log relative sensitivities and log relative specificities.

Figure 7.19 shows changes in the statistical significance of 16 out of 38 (42%) test comparisons (IDs 29, 31, 42, 43, 52 and 57 for relative sensitivity and IDs 2, 3, 15, 27, 35, 38, 41, 45, 47, 49 and 57 for relative specificity). Of the 16 test comparisons, 10 (63%) were direct comparisons. A few qualitative differences were observed in relative sensitivity (IDs 20, 24 and 38) or in relative specificity (IDs 33 and 41).

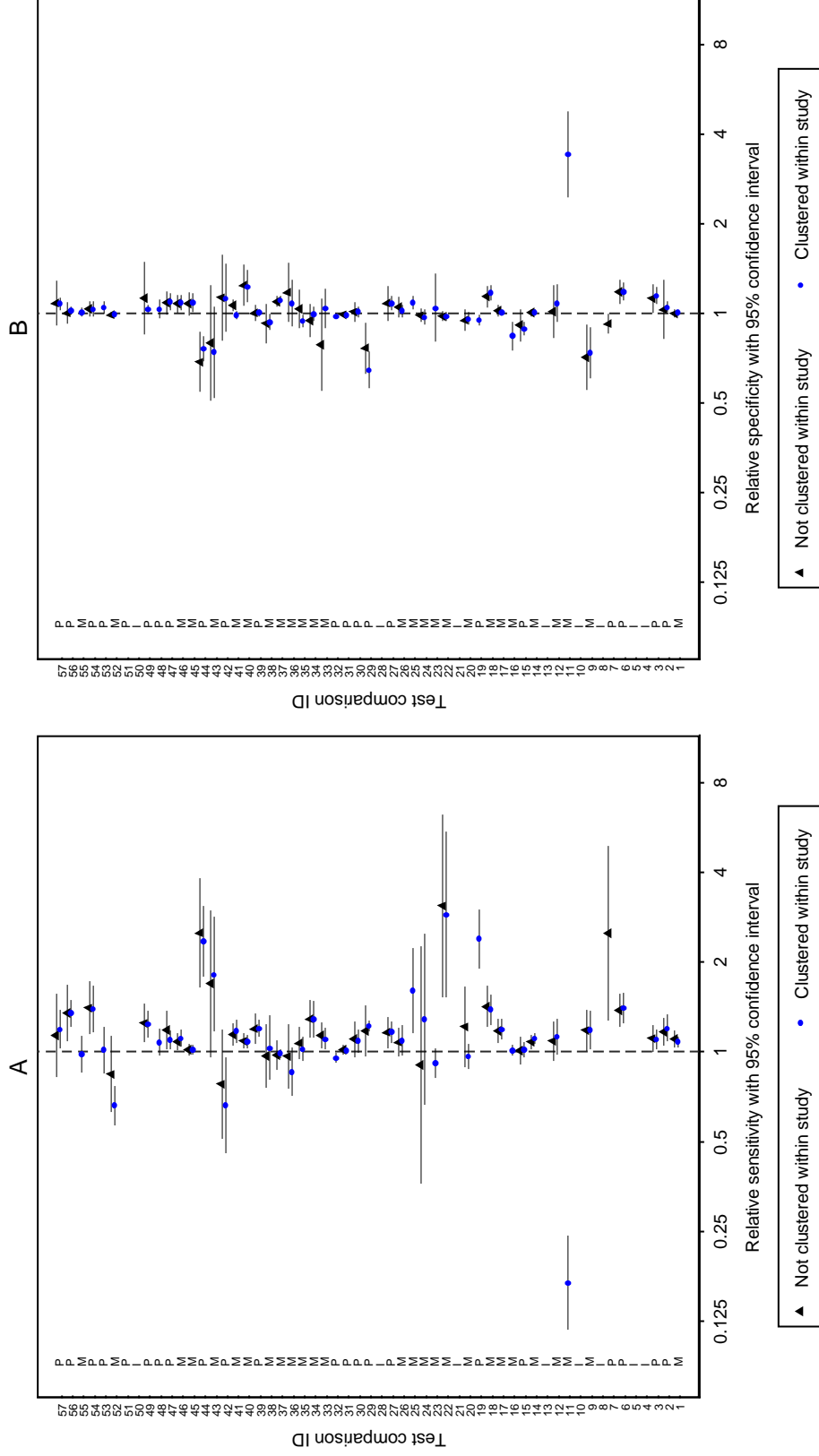


Figure 7.19| Comparison of relative sensitivity and relative specificity from bivariate meta-regression models with and without clustering of test results in comparative studies

The plots show estimates for each test comparison. The letters I, P, and M on the plots indicate the type of test comparison as follows: I = indirect only with no comparative studies; P = paired data only; and M = mixture of both comparative and non-comparative studies. Since the within-study approach does not apply to test comparisons without comparative studies, these test comparisons were not included in the analyses. Models for both approaches did not converge for two test comparisons (IDs 8 and 51). The dashed line on plots A and B is the line of no difference in sensitivity and no difference in specificity between the index and comparator tests in a test comparison.

7.6.2 Is it important to allow variances to differ by test in bivariate meta-regression models?

7.6.2.1 All types of test comparisons

For the comparison of bivariate models that assumed common variances for the random effects of the logits (Model 1), and bivariate models that allowed for unequal variances with independence between tests (Model 2), only one of the two models converged for six test comparisons (IDs 2, 11, 19, 23, 42 and 51). For two test comparison (IDs 7 and 8) neither of the two models converged (Figure 7.21). Thus, results from both models were compared for 49 test comparisons using the common variance model as the reference category.

Although there were 11 test comparisons with more than a 10% difference in relative sensitivity (IDs 20, 22, 24, 36 and 52) and/or relative specificity (IDs 29, 33, 35, 36, 43, 44 and 49), differences between both models were often small (Figure 7.20 panel A). Across all 49 test comparisons, the median (interquartile range) ratios of relative sensitivities and relative specificities were 1.00 (0.99 to 1.01) and 1.00 (0.98 to 1.01). In contrast, there were marked differences in precision of the estimates (Figure 7.20 panel B). One test comparison (ID 49) was excluded from the plot in panel B because the ratio was too large (10.3) and made the plot visually unhelpful. The standard errors were on average higher for models with unequal variances compared to models with equal variances with median (interquartile range) of 1.37 (1.09 to 1.77) and 1.39 (1.15 to 2.05) for ratios of standard errors of log relative sensitivities and log relative specificities. See Appendix D.3 for full results, including estimates of the variances.

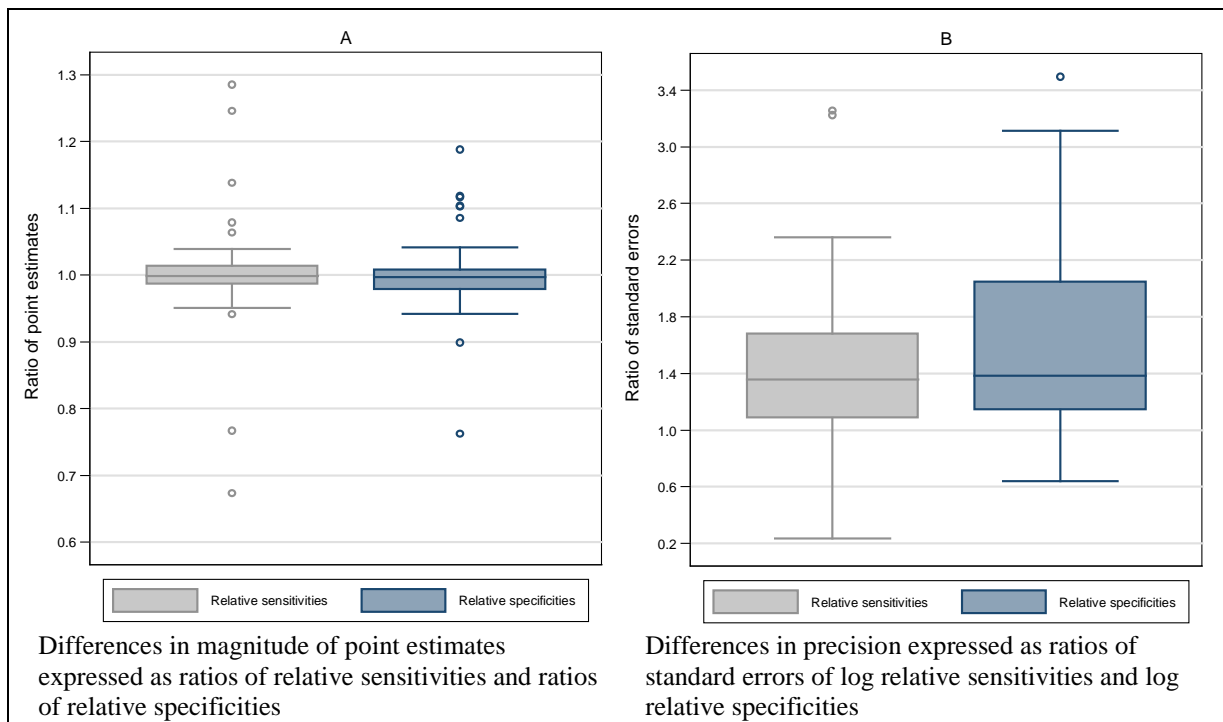


Figure 7.20 | Differences in estimates from bivariate meta-regression models with equal and unequal variances

Panel A shows the ratios of the relative sensitivities and relative specificities between models with unequal variances and those with equal variances (reference category). Relative sensitivities and relative specificities were estimated on the log scale and so Panel B shows the ratios of the standard errors of the log relative sensitivities and log relative specificities.

For 21 (43%) of the 49 test comparisons, likelihood ratio tests indicated statistically significant ($P \leq 0.05$) differences in model fit (see red asterisks on Figure 7.21). Fifteen (31%) test comparisons had a change in the statistical significance of relative sensitivity (IDs 3, 13, 31, 43, 49 and 52) or relative specificity (IDs 15, 26, 27, 35, 38, 41, 45, 47 and 53) while four (8%) had a change in both measures (IDs 25, 32, 46 and 57). Qualitative differences were observed for 11 (22%) test comparisons (IDs 20, 21, 24, 45, 55 for relative sensitivity and IDs 1, 33, 35, 39, 41 and 56 for relative specificity).

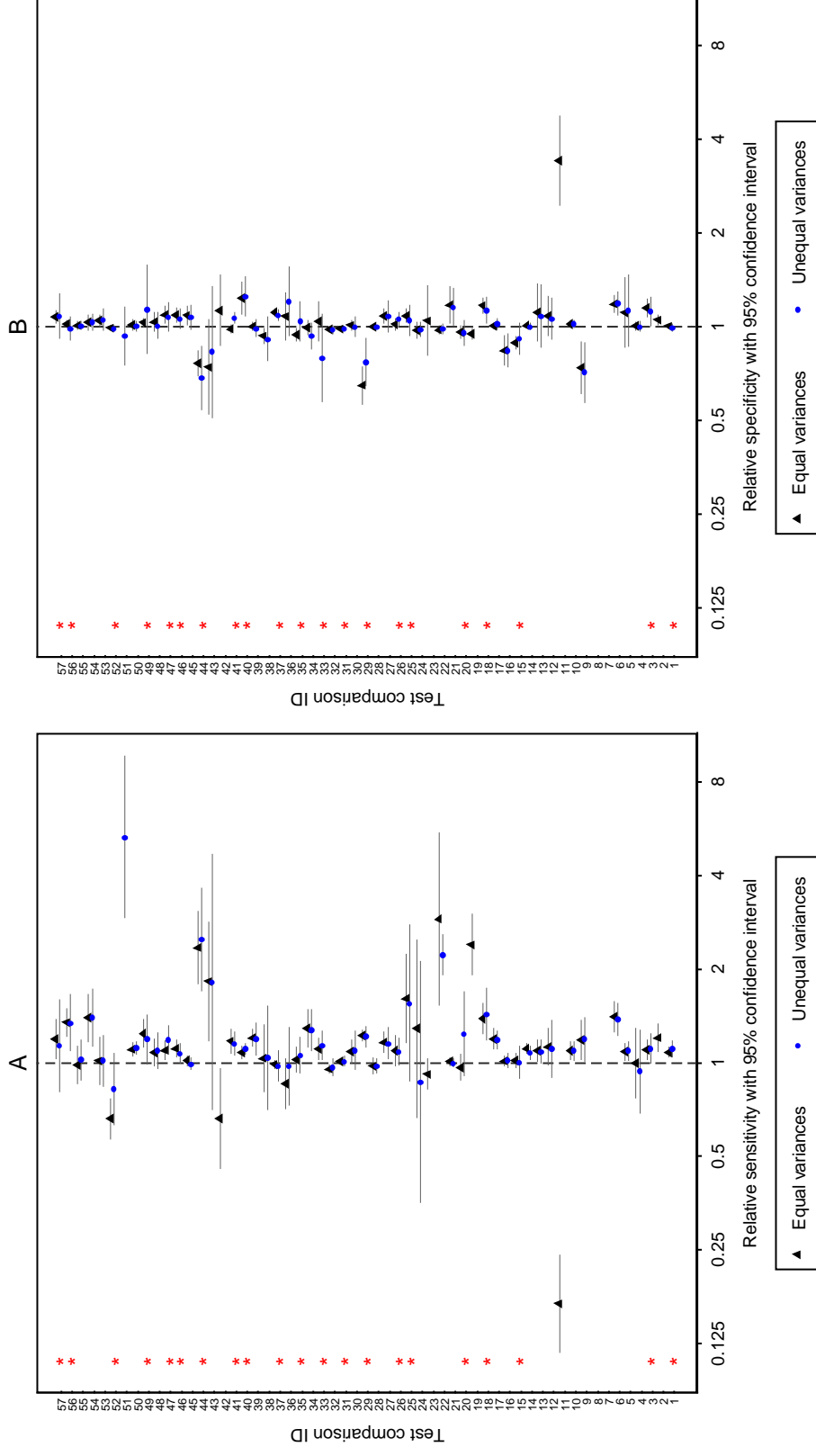


Figure 7.21| Comparison of relative sensitivity and relative specificity from bivariate meta-regression models with equal and unequal variances

The plots show estimates from bivariate models that assumed variances of the random effects for logit sensitivities and logit specificities were the same for both tests (black triangles), and bivariate models that allow for unequal variances (blue circles) for each test comparison. The red asterisks identify test comparisons where likelihood ratio tests indicated the differences between both models were statistically significant ($P \leq 0.05$). For eight test comparisons, one or both models did not converge. The dashed line on plots A and B is the line of no difference in sensitivity and no difference in specificity between the index and comparator tests in a test comparison.

7.6.2.2 Direct comparisons only

Nine of the 57 (16%) test comparisons each had at least 10 paired comparative studies and so direct comparisons were also performed. A model with unequal variances that allowed for correlation between tests (Model 3) was compared to the other two models investigated in section 7.6.2.1. For seven of the eight test comparisons where all three models converged, there was little or no difference in the point estimates of relative sensitivity and relative specificity between models (Figure 7.22 and Appendix D.4). However, precision of the estimates differed between the models with Model 2 having the least precision. The relative specificities of ID 44 and ID 27 had the largest difference in point estimates and precision respectively. For ID 27, the relative specificities for Models 1, 2, and 3 were 1.08 (1.03 to 1.14), 1.08 (0.96 to 1.21) and 1.08 (1.05 to 1.11), showing a significant difference in Models 1 and 3 but not in Model 2. This review was the example used in section 6.3.1.2 and the variances and correlations for the three models were presented in Table 6.2.

For IDs 31 and 32, though numerical differences were very small, statistical significance changed as a consequence of narrower confidence intervals for Model 1. Since relative sensitivity and relative specificity are functions of the pooled estimates, accounting for study level correlation between tests implicitly by assuming common variances as in Model 1 or explicitly by allowing for unequal variances and dependence between tests as in Model 3, led to more precise estimates. There was one (ID 56) test comparison with a qualitative change in relative specificity.

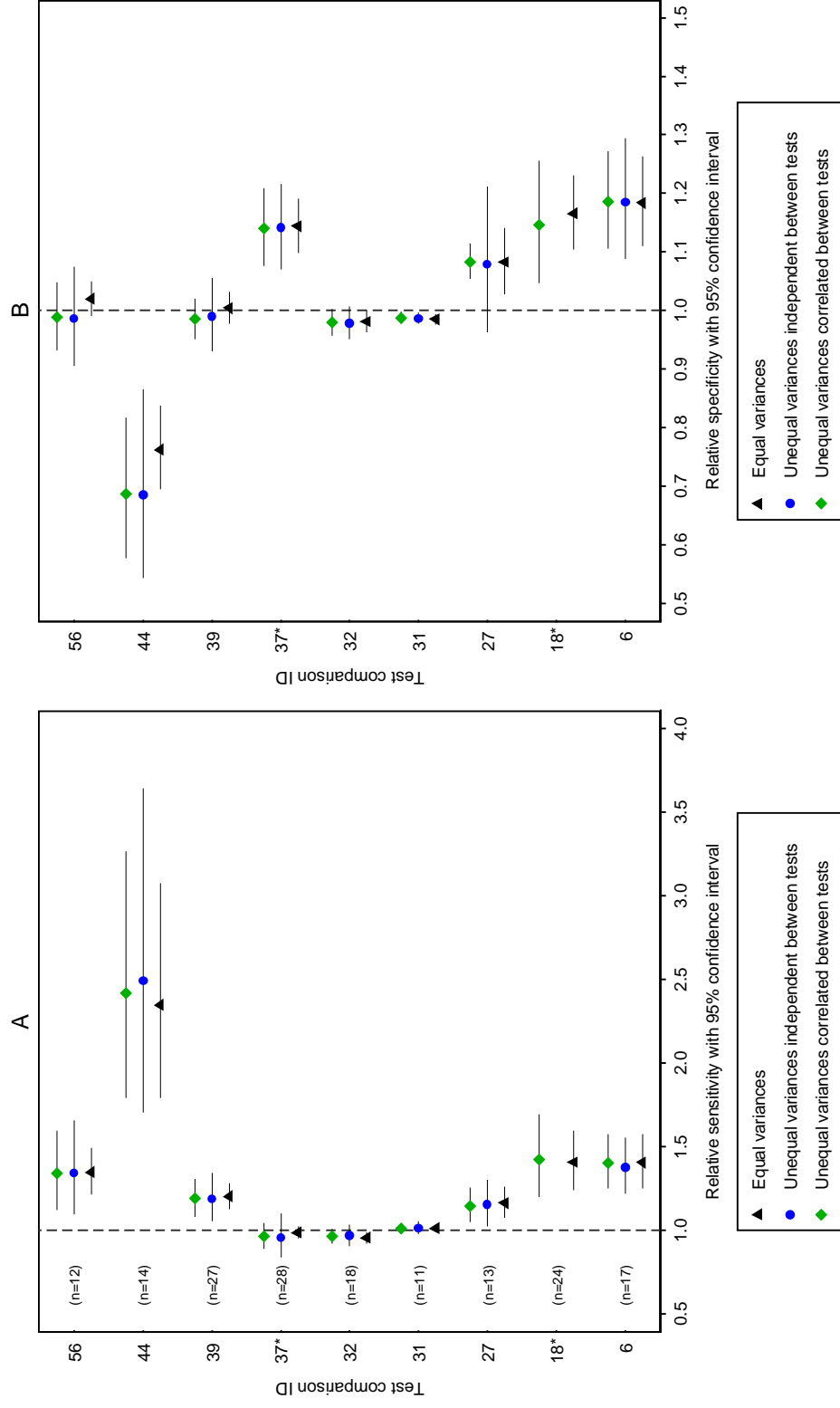


Figure 7.22 | Comparison of bivariate meta-regression models with different covariance structures fitted to direct test comparisons

*For these two test comparisons, direct comparisons were a subset of the data.

n is the number of comparative studies. For nine direct comparisons, the plots show estimates from bivariate models that assumed variances of the random effects for logit sensitivities and logit specificities were the same for both tests (black triangles), allowed for unequal variances and independence between tests (blue circles), and those that allowed for unequal variances and correlations between tests (green diamonds). The dashed line on plots A and B is the line of no difference in sensitivity and no difference in specificity between the index and comparator tests in a comparison.

7.6.3 Is it important to allow shape of SROC curves to differ between tests in HSROC meta-regression models?

HSROC models that assumed a common shape for SROC curves of both tests and HSROC models that allowed the SROC curve for each test to have its own shape were assessed. For four (IDs 4, 7, 31 and 52) of the 57 test comparisons, both models did not converge and only the common shape model converged for ID 29. However, five of the test comparisons (IDs 23, 43, 47, 51 and 54) had estimates with extremely wide and potentially unreliable confidence intervals, and so were excluded from further analyses to avoid misleading conclusions (see estimates in Appendix D.5). Thus, results from both models were compared for 47 test comparisons with the common shape model used as the reference category.

There were four test comparisons (IDs 8, 11, 42 and 44) with more than a 10% difference between the point estimates from both models (Figure 7.23 panel A). Across test comparisons, the median (interquartile range) ratio of relative sensitivities was 1.00 (0.98 to 1.01). One test comparison (ID 25) was excluded from the plot in panel B because the ratio of the standard errors was too large (4.42). There were differences in precision of the estimates though, on average, differences were small (Figure 7.23 panel B). The median (interquartile range) ratio of the standard errors of log relative sensitivities was 1.00 (0.93 to 1.10). Figure 7.24 shows the estimates of relative sensitivities computed for both models and the 16 (34%) test comparisons where likelihood ratio tests indicated statistical evidence ($P \leq 0.05$) of a difference in the shape of the SROC curves (see full results in Appendix D.5). There was a change in the statistical significance of the relative sensitivity of two test comparisons (IDs 5 and 25), and for one test comparison (ID 42), the ranking of tests altered between models.

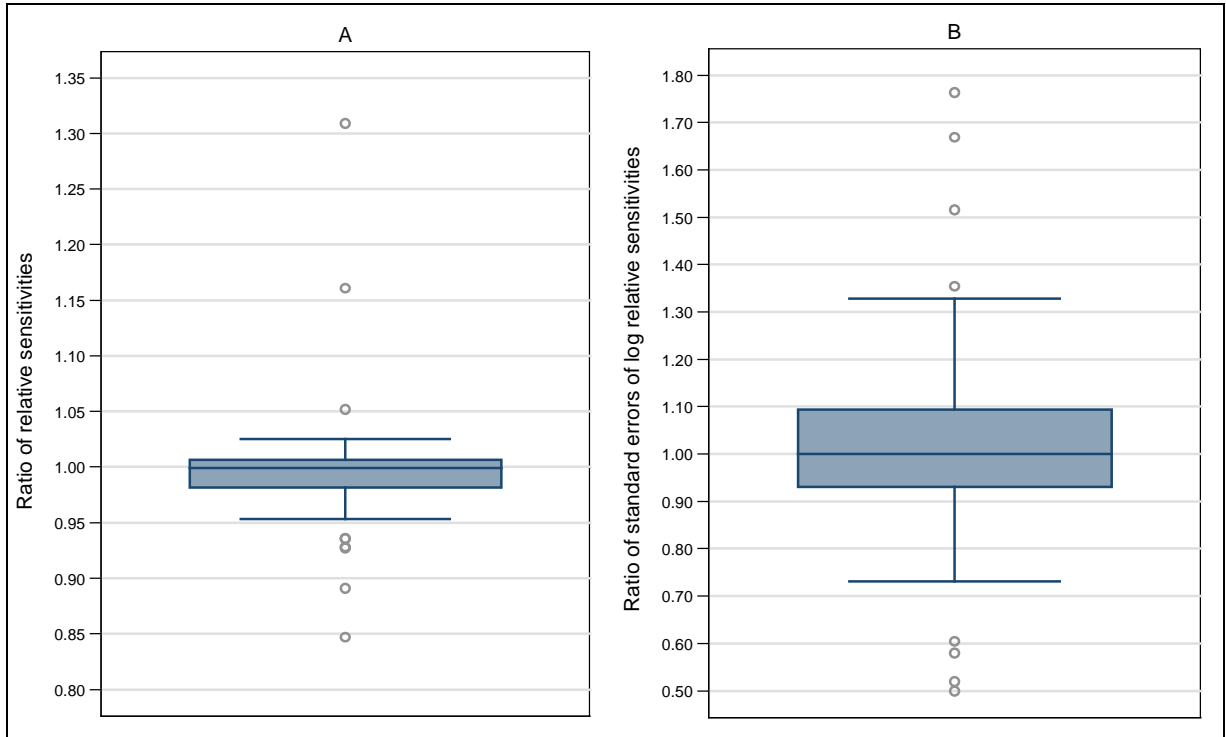


Figure 7.23| Differences in estimates from HSROC meta-regression models with common and different shape for SROC curves

Panel A shows the ratios of the relative sensitivities between both models with the common shape model as the reference category. Relative sensitivities were estimated on the log scale and so Panel B shows the ratios of the standard errors of the log relative sensitivities.

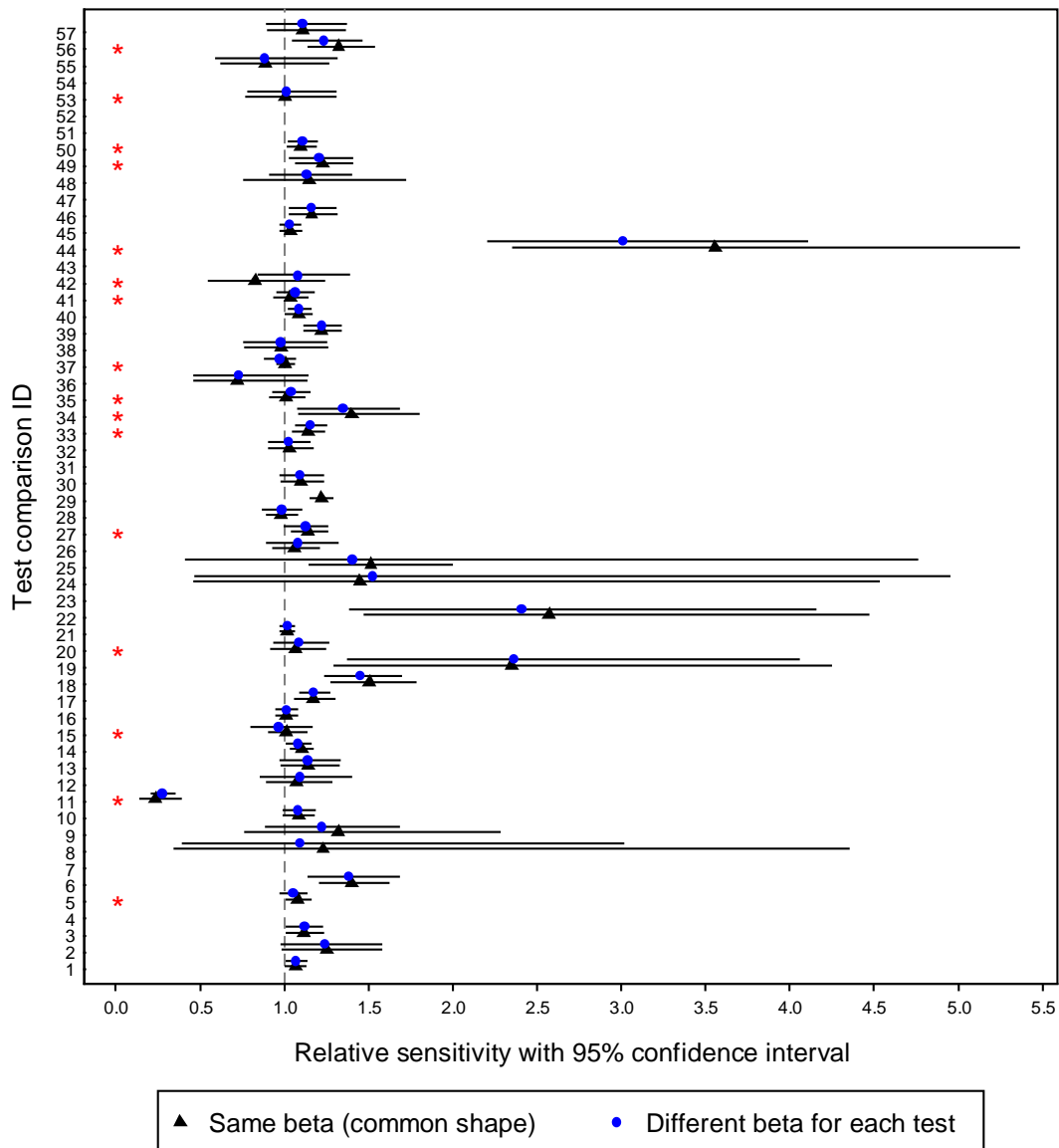


Figure 7.24| Comparison of relative sensitivity from HSROC meta-regression models with common and different shape between tests for SROC curves

The plot shows relative sensitivities derived from HSROC models that assumed the same shape for SROC curves of both tests (black triangles), and HSROC models that allowed the SROC curve for each test to have its own shape (blue circles). The estimates were computed at the median value of the specificities for each test in a test comparison. For five of the 57 test comparisons, one or both models did not converge. For five test comparisons, the confidence intervals were extremely wide and so the estimates were excluded from the plot and comparisons of both models. The red asterisks identify the 16 test comparisons where likelihood ratio tests indicated the differences between both models were statistically significant ($P \leq 0.05$). The dashed line is the line of no difference in sensitivity between the index and comparator tests in a test comparison.

7.7 Performance of different comparative meta-analysis methods

In Part II, removal of the correlation parameter from the bivariate model for meta-analysis of a single test (i.e. simplifying the bivariate model to two separate univariate models for sensitivity and specificity) was shown to have minimal effect on estimates of the variances or means and standard errors of both logit sensitivity and logit specificity. The findings from univariate and bivariate meta-regression models were compared to determine if the same was true for test comparisons. The results are presented in section 7.7.1. Section 7.7.2 compares findings from different Moses SROC meta-regression models while sections 7.7.3 and 7.7.4 compare findings from HSROC and Moses SROC meta-regression models.

7.7.1 Comparison of bivariate and univariate meta-regression

Estimates of relative sensitivity and relative specificity were obtained from univariate and bivariate models assuming equal variances across tests and also from models that allowed variances to differ by test (i.e. unequal variances). Due to similarity of the results of these two comparisons of bivariate and univariate models, only the comparison of models with unequal variances is presented in this section. For results from univariate and bivariate models that assumed variances were equal across tests in a test comparison, see the figure in Appendix D.6. Three test comparisons (IDs 7, 34 and 52) did not converge for the univariate model and seven (IDs 2, 7, 8, 11, 19, 23 and 42) did not converge for the bivariate model. Both models converged for 48 test comparisons and the univariate model was used as the reference category in comparisons of the two models.

Differences in relative sensitivity and relative specificity between both models were negligible (Figure 7.25 panel A and Appendix D.7). Across the 48 test comparisons, the

median (interquartile range) ratios of relative sensitivities and relative specificities were 1.00 (1.00 to 1.01) and 1.00 (1.00 to 1.00). Similarly, there were negligible differences in the precision of the estimates (Figure 7.25 panel B and Figure 7.26). Although standard errors for log relative sensitivities and log relative specificities tended to be higher for estimates from bivariate models relative to those from univariate models, the median (interquartile range) ratios of standard errors for log relative sensitivities and log relative specificities were 1.00 (1.00 to 1.05) and 1.00 (1.00 to 1.01).

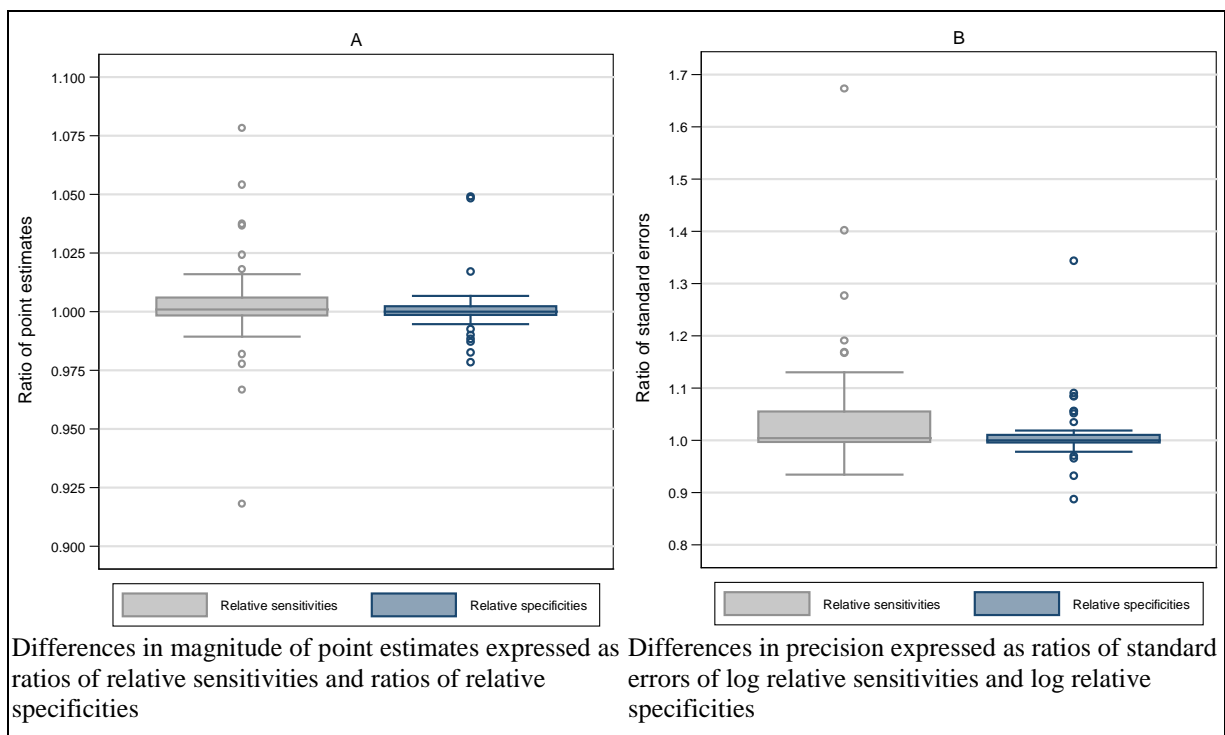


Figure 7.25| Differences in estimates from bivariate and univariate meta-regression models with unequal variances

Panel A shows the ratios of the relative sensitivities and relative specificities between both models with the univariate model as the reference category. Relative sensitivities and relative specificities were estimated on the log scale and so Panel B shows the ratios of the standard errors of log relative sensitivities and log relative specificities.

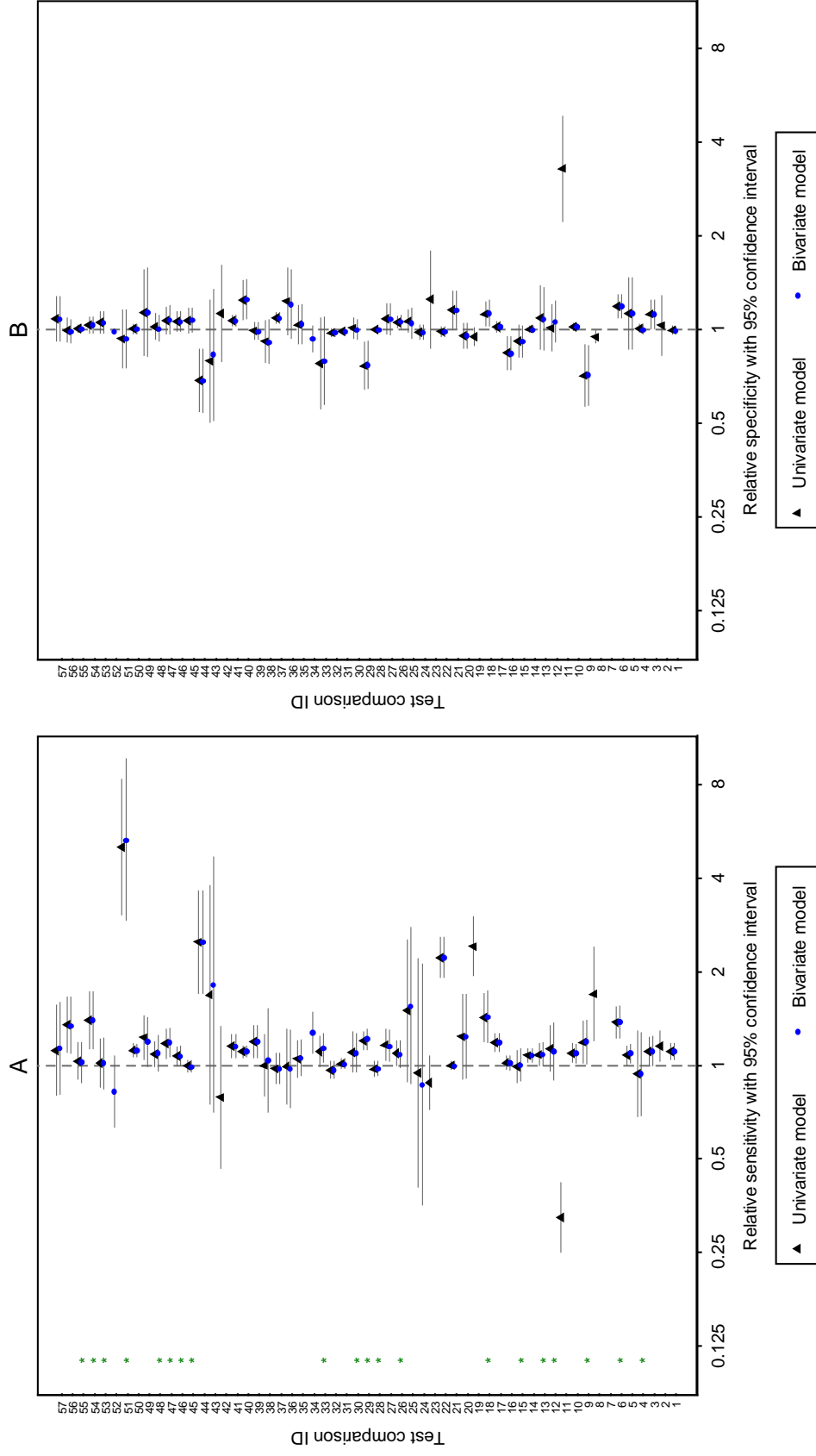


Figure 7.26| Comparison of relative sensitivity and relative specificity from bivariate and univariate models with unequal variances The plots show estimates from univariate models (black triangles) and bivariate models (blue circles) that allowed separate variances for the random effects of the logit sensitivities and specificities for each test in a comparative meta-analysis. For eight of the 57 test comparisons, only one of the two models converged and for test comparison 7, neither of the models converged. The green asterisks identify test comparisons where correlation was estimated as +1 or -1 for at least one of the tests. The dashed line on each plot is the line of no difference in test performance between index and comparator tests.

Estimates for the variance and correlation parameters from the models are given in Appendix D.8. In 20 of the 48 (42%) test comparisons, correlation of the logits was estimated as +1 or –1 for at least one of the tests (see test comparisons marked with an asterisk on Figure 7.26). In this subset, the median (interquartile range) ratios of relative sensitivities and relative specificities were 1.00 (1.00 to 1.01) and 1.00 (1.00 to 1.00). The corresponding median (interquartile range) ratios of standard errors for log relative sensitivities and log relative specificities were 0.99 (1.03 to 1.07) and 0.99 (1.00 to 1.01). Results from the subset of test comparisons were similar to those from the whole cohort.

Table 7.3 shows the eight test comparisons (IDs 1, 15, 18, 20, 29, 30, 33 and 44) where likelihood ratio tests indicated a statistically significant difference in model fit between univariate and bivariate models (see Appendix D.7 for full results). Figure 7.26 shows one change in the statistical significance of relative sensitivity (ID 33) and one qualitative change in relative sensitivity (ID 15).

Table 7.3| Bivariate and univariate (unequal variance) models with statistically significant differences in model fit

ID	Bivariate model		Univariate model		P value*
	Relative sensitivity (95% CI)	Relative specificity (95% CI)	Relative sensitivity (95% CI)	Relative specificity (95% CI)	
1	1.11 (1.04–1.18)	1.00 (0.97–1.02)	1.11 (1.04–1.18)	1.00 (0.97–1.02)	0.05
15	1.01 (0.89–1.13)	0.92 (0.81–1.03)	0.99 (0.88–1.12)	0.92 (0.82–1.03)	0.01
18	1.43 (1.19–1.74)	1.13 (1.03–1.24)	1.42 (1.19–1.71)	1.12 (1.03–1.23)	<0.0001
20	1.24 (0.91–1.69)	0.96 (0.87–1.05)	1.24 (0.91–1.69)	0.96 (0.87–1.05)	0.01
29	1.21 (1.12–1.31)	0.77 (0.65–0.92)	1.20 (1.12–1.28)	0.77 (0.64–0.91)	0.02
30	1.10 (0.95–1.26)	1.00 (0.93–1.07)	1.10 (0.95–1.28)	1.01 (0.94–1.09)	0.02
33	1.14 (1.02–1.26)	0.79 (0.57–1.10)	1.11 (0.99–1.24)	0.78 (0.56–1.09)	0.01
44	2.49 (1.71–3.64)	0.69 (0.54–0.87)	2.49 (1.70–3.65)	0.69 (0.54–0.87)	0.001

*P value from likelihood ratio tests comparing both models.

Univariate models were estimated by assuming an independent variance-covariance structure, i.e. correlation of the logits = 0.

7.7.2 Comparison of unweighted and weighted Moses SROC meta-regression

7.7.2.1 Same shape assumed for SROC curves

The rDORs from unweighted and weighted Moses SROC analyses in which the shape of the SROC curves were assumed to be the same for both tests, were compared using estimates from the weighted analyses as the reference category. There were large differences between estimates of rDORs from both analyses with more than a two-fold difference for eight (14%) test comparisons (IDs 7, 9, 19, 24, 28, 34, 42 and 54) (see Appendix D.9 for full results). On average, the unweighted analyses gave higher rDORs relative to those from weighted analyses (Figure 7.27 panel A). The median (interquartile range) ratio of rDORs was 1.14 (0.88 to 1.52). Similarly, unweighted analyses gave higher standard errors for log rDORs compared to weighted analyses (Figure 7.27 panel B); the median (interquartile range) ratio of standard errors for log rDORs was 1.16 (1.00 to 1.34).

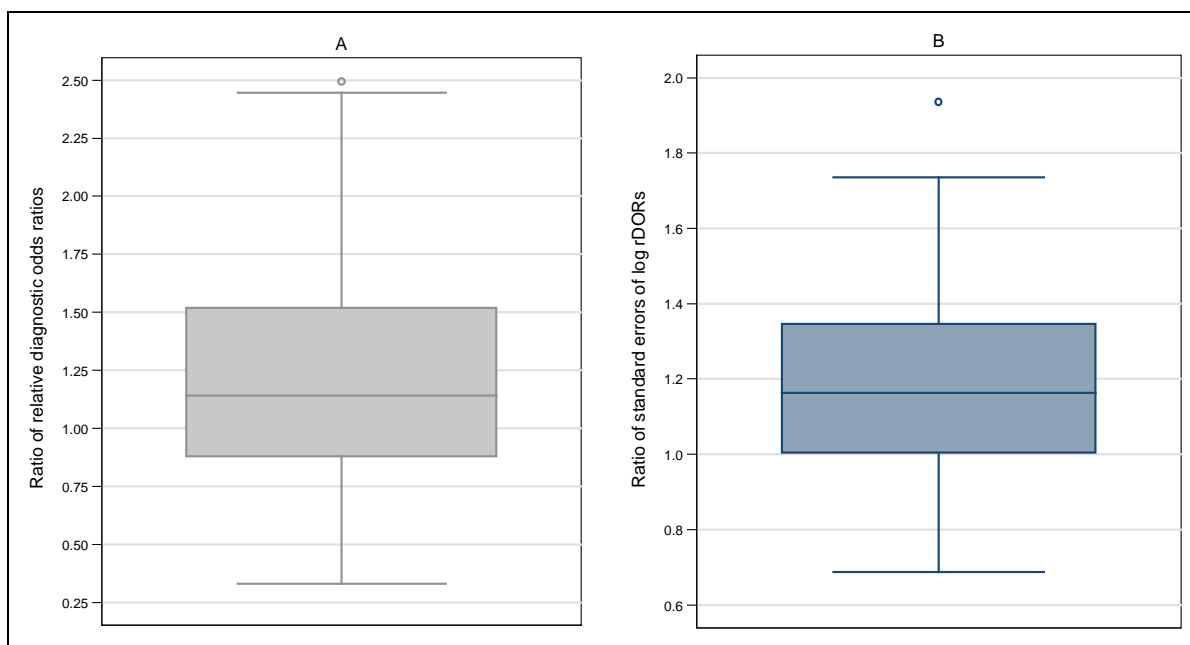


Figure 7.27| Differences in estimates from unweighted and weighted Moses SROC meta-regression (same shape) models

Panel A shows the ratios of the relative diagnostic odds ratios (rDORs) between both models with the weighted analysis as the reference category. Panel B shows the ratios of the standard errors of the log of the rDORs.

For eight (14%) test comparisons, the statistical significance of the rDORs differed between models as shown in Figure 7.28. Qualitative differences were observed for three test comparisons (IDs 8, 9 and 11).

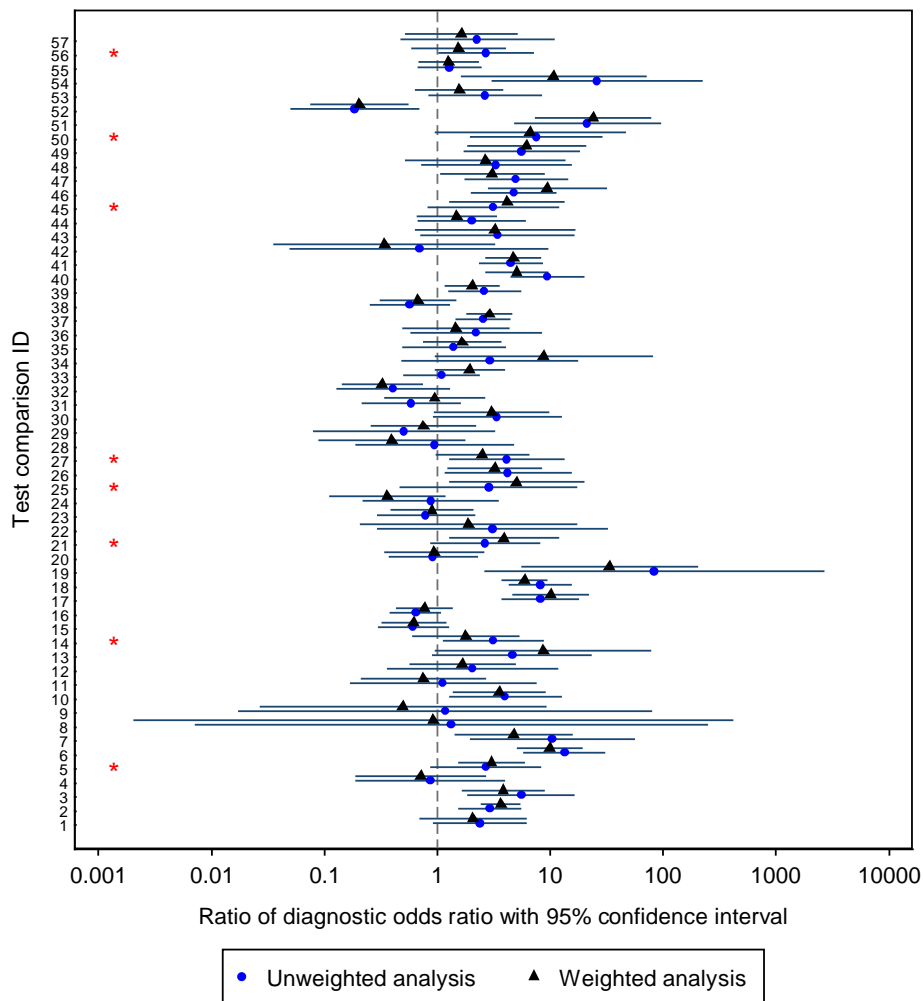


Figure 7.28| Comparison of unweighted and weighted Moses SROC meta-regression (same shape) models

The graph (plotted on the log scale) shows estimates of the ratio of diagnostic odds ratios from both models for each test comparison. The dashed line is the line of no difference in test accuracy between the index and comparator tests in a test comparison. The red asterisks identify the eight test comparisons where there was a change in statistical significance between the models.

7.7.2.2 Different shape for SROC curves

Unweighted and weighted meta-analyses in which the shape of SROC curves was allowed to differ by test were also performed. For nine test comparisons (IDs 4, 7, 22, 25, 31, 43, 47, 52 and 54), the relative sensitivities obtained from one or both of the analyses had exceptionally wide confidence intervals (see Appendix D.9) which may indicate poor estimation. In addition, relative sensitivities and their 95% CIs were not estimable for two test comparisons—ID 25 for the unweighted analysis and ID 28 for the weighted analysis. Therefore, the results from both models were compared for 47 test comparisons.

Unlike the comparison of rDORs from common shape models, the unweighted analyses for different shape models tended to give lower relative sensitivities compared to those from weighted analyses (Figure 7.29 panel A). The median (interquartile range) ratio of relative sensitivities was 0.95 (0.90 to 1.03). The unweighted analyses gave higher standard errors for the log relative sensitivities compared to those from weighted analyses (Figure 7.29 panel B); the median (interquartile range) ratio of standard errors was 1.09 (0.76 to 1.85).

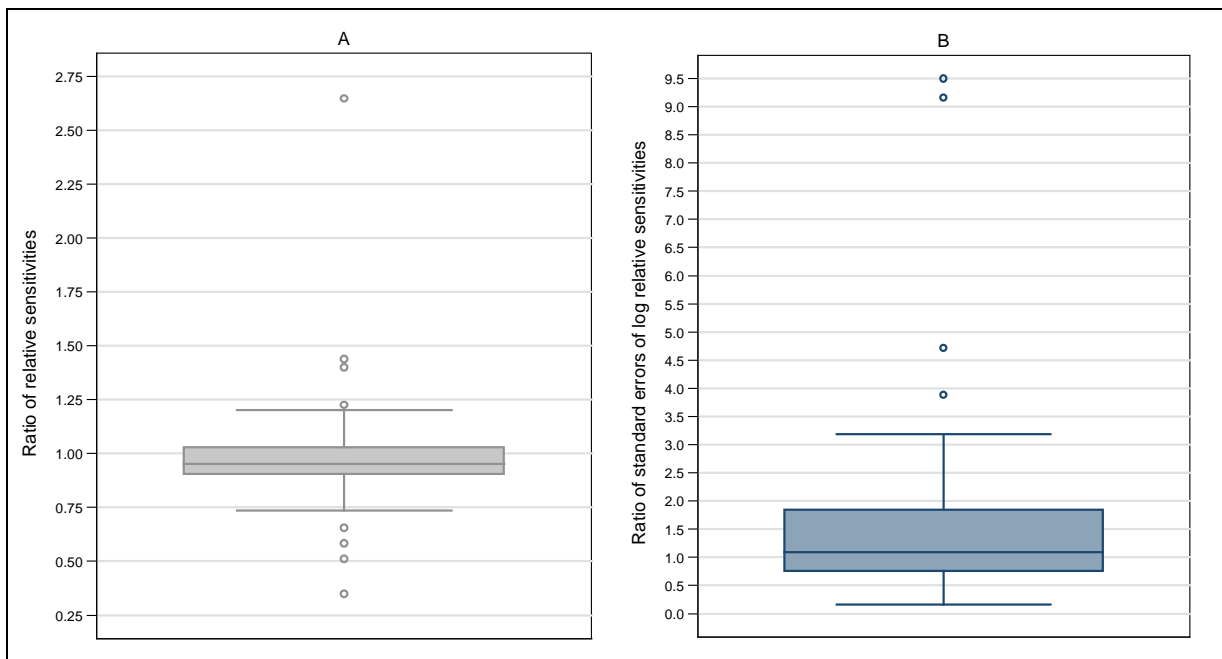


Figure 7.29| Differences in estimates from unweighted and weighted Moses SROC (different shape) meta-regression models

Panel A shows the ratios of the relative sensitivities between both models with the weighted analysis as the reference category. Panel B shows the ratios of the standard errors of the log of the relative sensitivities.

Figure 7.30 only shows results for 38 of the 47 test comparisons because the confidence limits for nine test comparisons (IDs 3, 11, 19, 24, 26, 35, 44, 51 and 57) were too large and made other estimates on the plot less visible. For nine (19%) of the 47 test comparisons, the statistical significance of the relative sensitivities differed between models as shown on Figure 7.30. There were qualitative differences for 10 test comparisons but the confidence intervals for relative sensitivities were often very wide (Table 7.4).

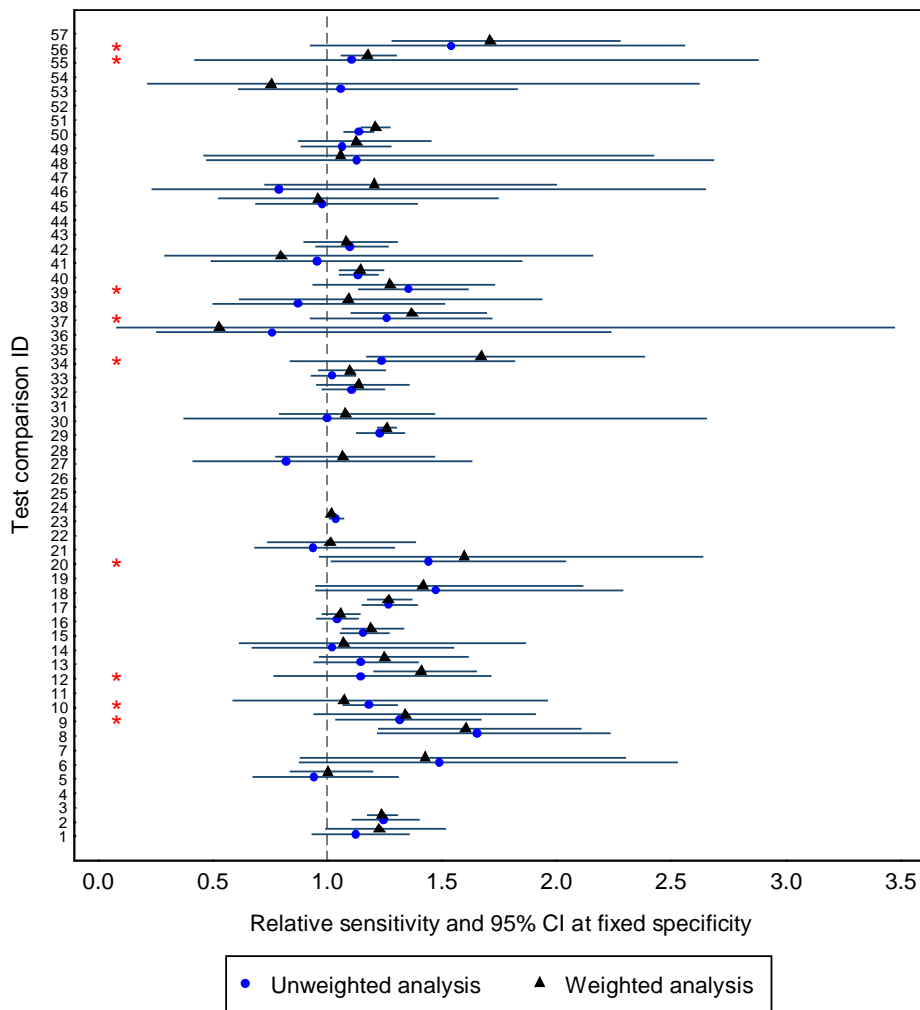


Figure 7.30| Comparison of unweighted and weighted Moses SROC meta-regression (different shape) models

The graph shows estimates of relative sensitivity from both models for each test comparison. The estimates were computed at the median value of the specificities for each test in a test comparison. The dashed line is the line of no difference in test accuracy between the index and comparator tests in a test comparison. The red asterisks identify the nine test comparisons where there was a change in statistical significance between the models. The plot shows results for 38 of the 47 test comparisons because the confidence limits for nine test comparisons were too large and made the plot uninformative.

Table 7.4| Qualitative differences between unweighted and weighted Moses SROC meta-regression models (shape allowed to differ by test)

ID	Relative sensitivity (95% CI)	
	Unweighted	Weighted
3	0.82 (0.12–5.57)	1.11 (0.68–1.82)
5	0.94 (0.68–1.31)	1.00 (0.84–1.20)
11	0.37 (0.03–5.24)	1.06 (0.79–1.42)
21	0.94 (0.69–1.29)	1.01 (0.74–1.38)
24	1.07 (0.02–55.7)	0.40 (0.01–11.1)
27	0.82 (0.41–1.63)	1.07 (0.78–1.47)
30	1.00 (0.37–2.65)	1.08 (0.79–1.47)
38	0.87 (0.50–1.51)	1.09 (0.62–1.93)
46	0.79 (0.24–2.65)	1.20 (0.73–2.00)
53	1.06 (0.61–1.83)	0.76 (0.22–2.62)

Relative sensitivity was computed using sensitivities estimated at the median value of specificity for each of the tests in a test comparison.

7.7.3 Comparison of HSROC and unweighted Moses SROC meta-regression

7.7.3.1 Same shape assumed for SROC curves

The HSROC model did not converge for ID 7. For IDs 19 and 29, estimates of rDORs from HSROC models were extremely large (2938 and 40438 respectively) and so they were excluded from comparisons of the two models. For the remaining 54 test comparisons, there were large differences between rDORs from the two models with more than a two-fold difference for 12 (22%) test comparisons (see Appendix D.10). HSROC models tended to give higher rDORs compared to Moses models (Figure 7.31 panel A), with median (interquartile range) ratio of rDORs of 1.11 (0.82 to 1.50). Estimates from HSROC models were more precise than those from Moses models (Figure 7.31 panel B); the median (interquartile range) ratio of standard errors for log rDORs was 0.89 (0.66 to 1.09). For 11 (20%) test comparisons, the statistical significance of the rDORs differed between models as shown in Figure 7.32. Qualitative changes were observed for six test comparisons (IDs 4, 8, 9, 11, 35 and 36).

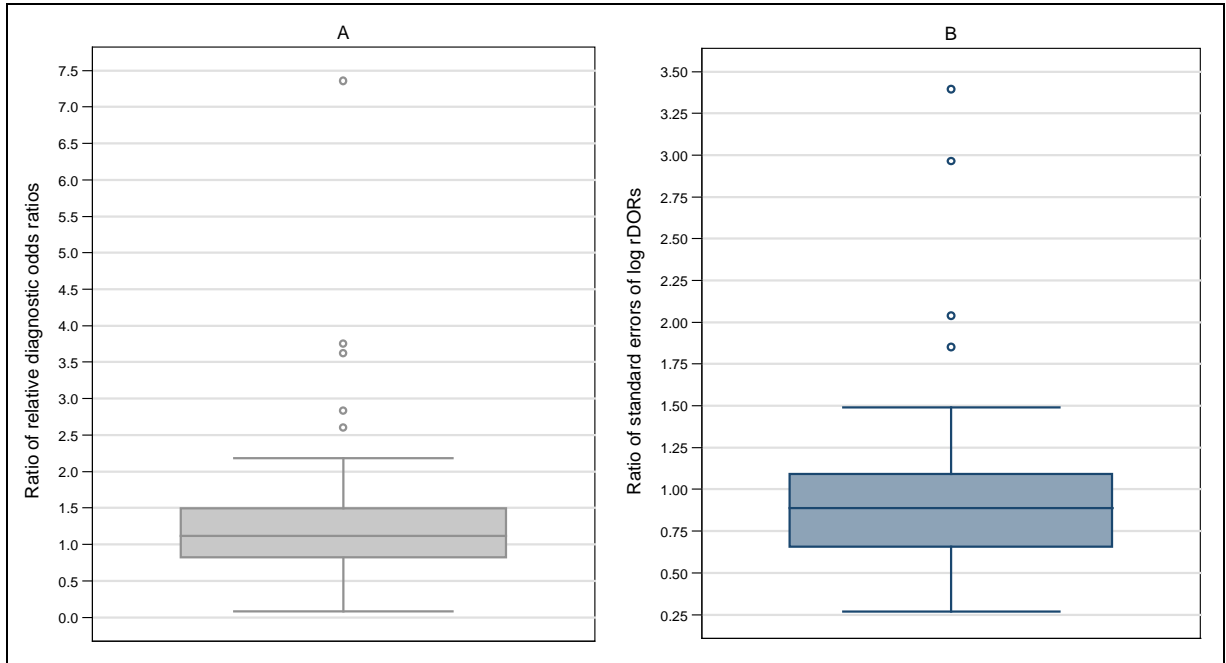


Figure 7.31| Differences in estimates from HSROC and unweighted Moses SROC meta-regression (same shape) models

Panel A shows the ratios of the relative diagnostic odds ratios (rDORs) between both models with the unweighted Moses SROC model as the reference category. Panel B shows the ratios of the standard errors of the log of the rDORs.

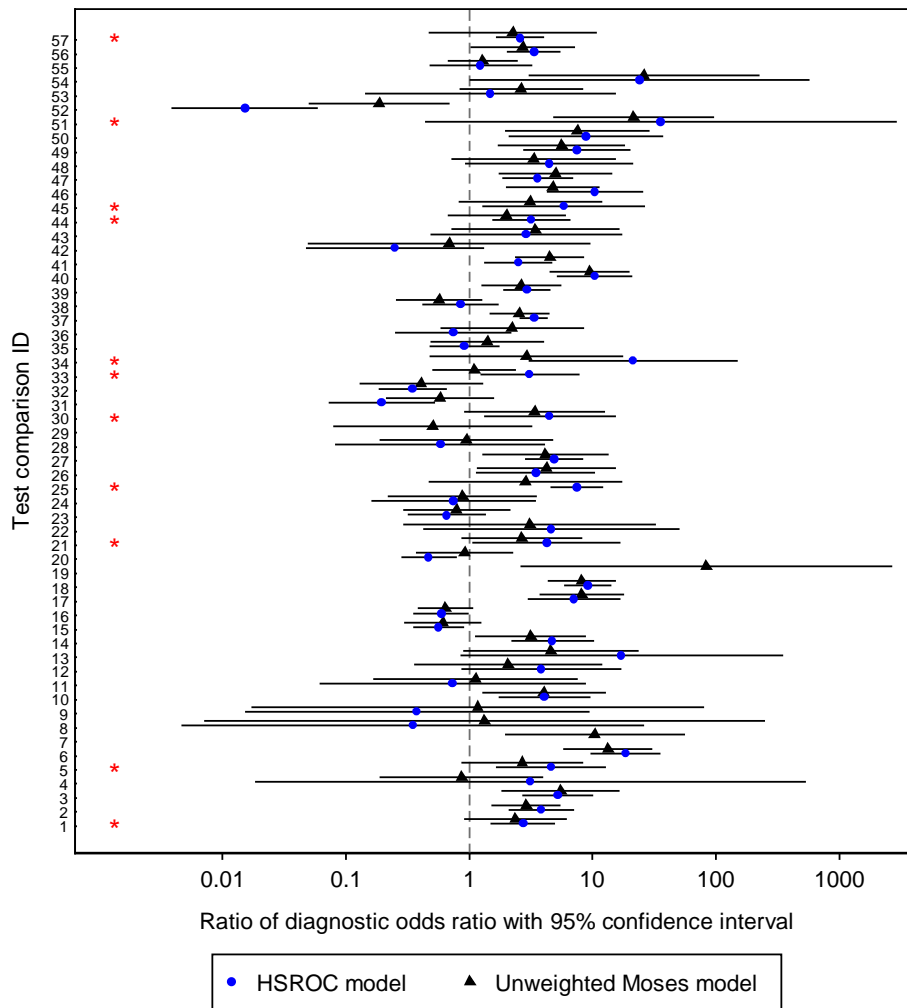


Figure 7.32| Comparison of unweighted Moses SROC and HSROC meta-regression (same shape) models

The graph (plotted on the log scale) shows estimates of the ratio of diagnostic odds ratios for each of the test comparisons. The HSROC model did not converge for ID 7. For IDs 19 and 29, the rDORs from HSROC models were extremely large and so they were excluded from the plot to enhance visibility of other estimates. The dashed line is the line of no difference in test accuracy between the index and comparator tests in a test comparison. The red asterisks identify the 11 test comparisons where there was a change in statistical significance between the models.

7.7.3.2 Different shape for SROC curves

The HSROC model did not converge for five test comparisons (IDs 4, 7, 29, 31 and 52) and relative sensitivity was not estimable from the Moses model for ID 25. For three (IDs 22, 43 and 47) test comparisons, relative sensitivity was poorly estimated in one or both models (see Appendix D.10). Therefore, estimates from 48 test comparisons were considered. The relative

sensitivities from HSROC models were on average slightly lower than those from Moses models (Figure 7.33 panel A). The median (interquartile range) ratio of relative sensitivities was 0.98 (0.93 to 1.11). Estimates from HSROC models were considerably more precise than those from Moses models (Figure 7.33 panel B); the median (interquartile range) ratio of standard errors for log relative sensitivities was 0.37 (0.15 to 0.77).

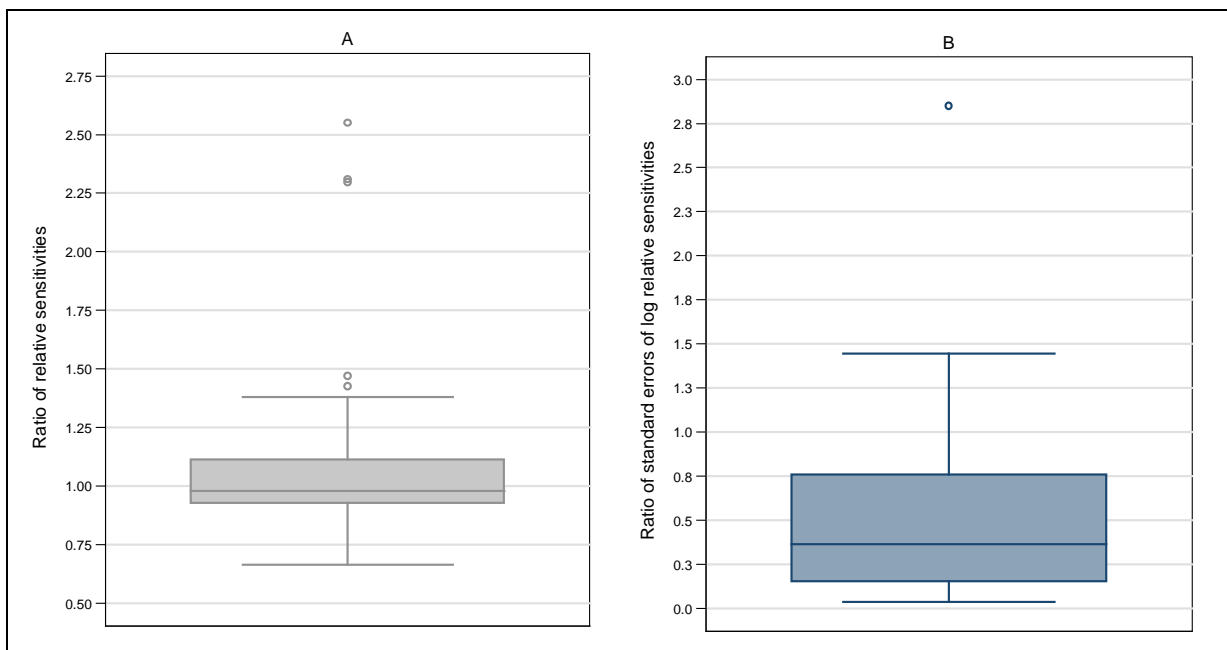


Figure 7.33| Differences in estimates from HSROC and unweighted Moses SROC (different shape) meta-regression models

Panel A shows the ratios of the relative sensitivities between both models with the unweighted analysis as the reference category. Panel B shows the ratios of the standard errors of the log of the relative sensitivities.

Estimates of relative sensitivities with their 95% CIs are shown in Figure 7.34 for both models. Five test comparisons (IDs 23, 24, 26, 51 and 54) were excluded from the plot because the estimates and/or confidence limits from one or both models were extremely large (see Appendix D.10). For 17 (35%) test comparisons, the statistical significance of the rDORs differed between models as shown in Figure 7.34. There were qualitative changes for 14 (29%) test comparisons (Table 7.5).

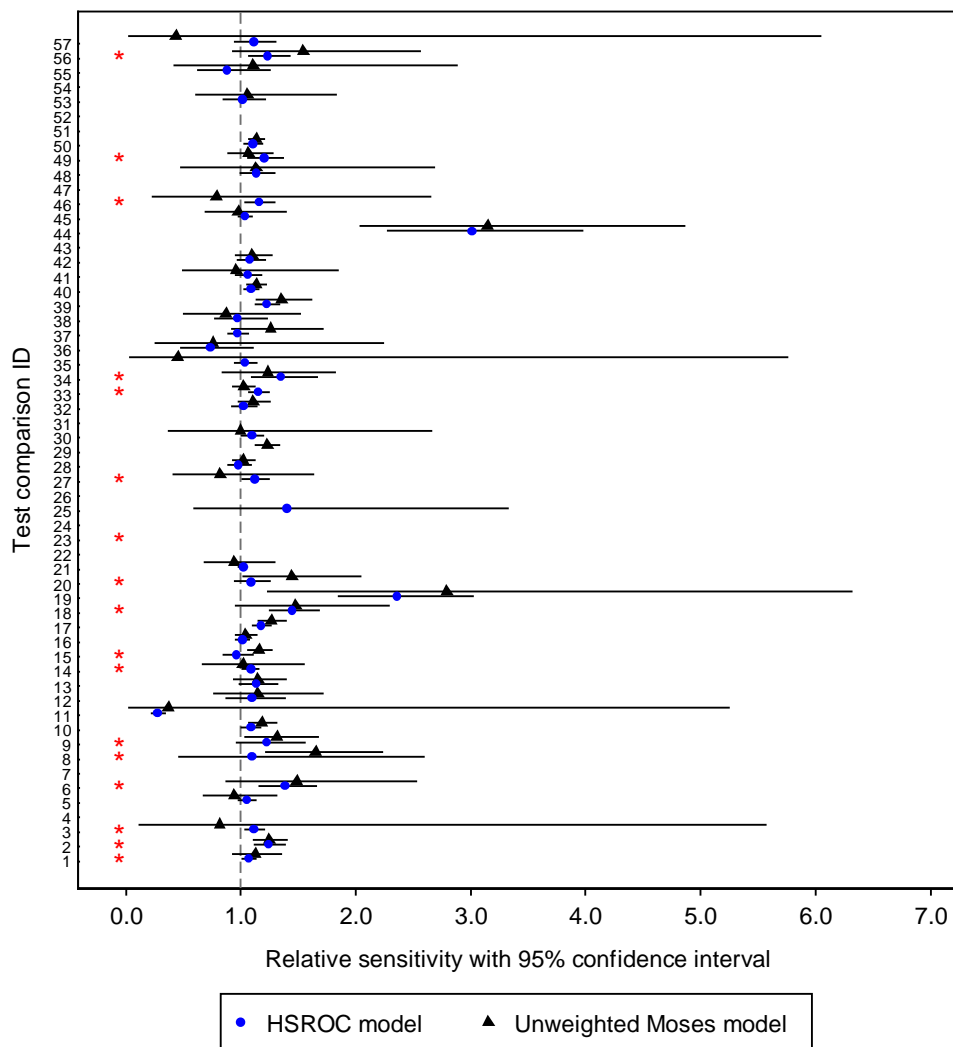


Figure 7.34| Comparison of unweighted Moses SROC and HSROC meta-regression models

The graph shows estimates of relative sensitivity for HSROC and Moses models that allowed the shape of the curves to differ. The estimates were computed at the median value of the specificities for each test in a test comparison. For five test comparisons, estimates were too large for inclusion on the plot. The dashed line on each plot is the line of no difference in test performance between the index and comparator tests. The red asterisks identify the 17 test comparisons where there was a change in statistical significance between the models.

Table 7.5| Qualitative differences between HSROC and unweighted Moses SROC meta-regression models (shape allowed to differ by test)

ID	Relative sensitivity (95% CI)	
	Unweighted Moses model	HSROC model
3	0.82 (0.12–5.57)	1.12 (1.02–1.22)
5	0.94 (0.68–1.31)	1.05 (0.98–1.13)
15	1.16 (1.06–1.27)	0.97 (0.80–1.16)
21	0.94 (0.69–1.29)	1.02 (0.98–1.06)
26	0.79 (0.06–11.2)	1.09 (0.90–1.32)
27	0.82 (0.41–1.63)	1.12 (1.00–1.26)
28	1.02 (0.93–1.12)	0.98 (0.88–1.10)
35	0.45 (0.04–5.75)	1.04 (0.94–1.15)
37	1.26 (0.93–1.72)	0.97 (0.89–1.06)
41	0.96 (0.50–1.85)	1.06 (0.96–1.18)
45	0.98 (0.69–1.39)	1.03 (0.98–1.10)
46	0.79 (0.24–2.65)	1.16 (1.03–1.31)
54	0.99 (0.00–204)	1.02 (0.04–26.5)
57	0.43 (0.03–6.05)	1.11 (0.90–1.37)

Relative sensitivity was computed using sensitivities estimated at the median value of specificity for each of the tests in a test comparison.

7.7.4 Comparison of HSROC and weighted Moses SROC meta-regression

Both Moses et al³⁷ and Irwig et al²¹⁷ recommended unweighted regression (see section 6.3.2.1). In section 7.7.2 (comparison of weighted and unweighted Moses models), the cost of no weighting in common shape Moses models was an average increase in rDORs with a decrease in precision. For different shape models, on average, there was a decrease in relative sensitivities and their precision. Since unweighted versus weighted Moses models, and HSROC versus unweighted Moses models have been dealt with in detail in the two preceding sections, for completeness, this section briefly considers HSROC versus weighted Moses models.

7.7.4.1 Same shape assumed for SROC curves

Using the same 54 test comparisons as in section 7.7.3.1, on average, common shape HSROC models gave higher rDORs (Figure 7.35 panel A) and slightly more precise estimates (Figure 7.35 panel B) in comparison to weighted Moses models. The median (interquartile range) ratio of rDORs and ratio of standard errors for log rDORs were 1.28 (0.92 to 1.67) and 0.96 (0.74 to 1.29).

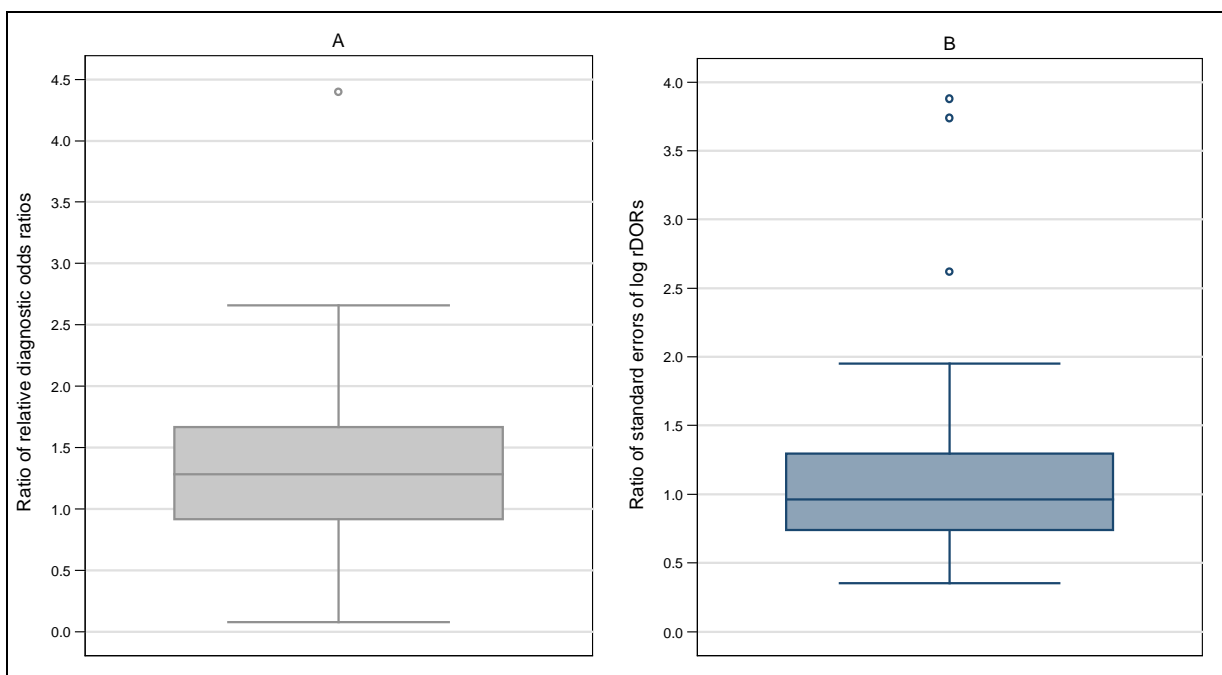


Figure 7.35| Differences in estimates from HSROC and weighted Moses SROC meta-regression (same shape) models

Panel A shows the ratios of the relative diagnostic odds ratios (rDORs) between both models with the weighted Moses SROC model as the reference category. Panel B shows the ratios of the standard errors of the log of the rDORs.

7.7.4.2 Different shape for SROC curves

In addition to the five HSROC models that did not converge (see section 7.7.3.2), relative sensitivity was not estimable or was poorly estimated in HSROC or weighted Moses models for four (IDs 23, 28, 47 and 54) test comparisons (see Appendices D.9 and D.10). Across the remaining 48 test comparisons, the median (interquartile range) ratio of relative sensitivities

and ratio of standard errors for log relative sensitivities were 1.00 (0.88 to 1.07) and 0.39 (0.16 to 0.70). These results were similar to those from the comparison of HSROC and unweighted Moses models (see section 7.7.3.2).

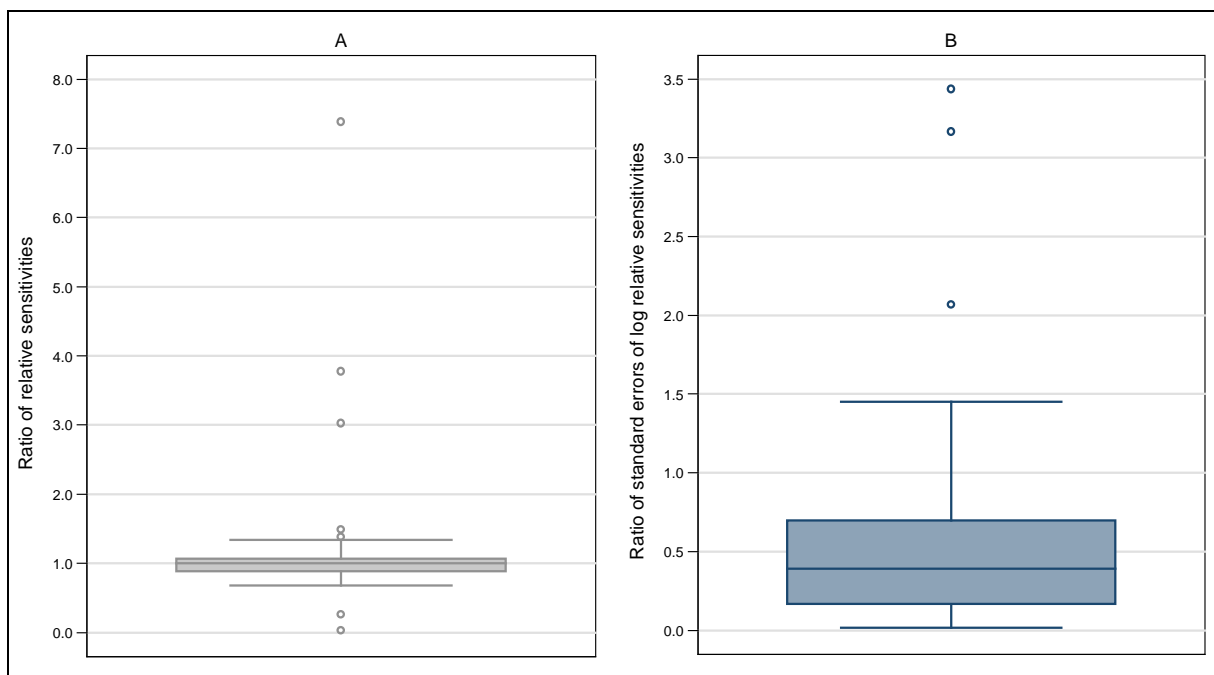


Figure 7.36| Differences in estimates from HSROC and weighted Moses SROC meta-regression (different shape) models

Panel A shows the ratios of the relative sensitivities between both models with the weighted Moses SROC model as the reference category. Panel B shows the ratios of the standard errors of the log of the relative sensitivities.

7.8 Discussion

7.8.1 Summary of findings

7.8.1.1 Modelling assumptions in hierarchical models

The two-stage approach used to analyse each test comparisons provided insight into the appropriateness of modelling assumptions. The first stage enabled assessments using meta-analyses of single tests while the second stage allowed for comparisons of relative test performance. In hierarchical meta-regression models, quantitative (magnitude, precision and statistical significance) and qualitative (ranking of test performance) differences in relative

test performance occurred with different modelling assumptions. The higher convergence failure rate observed with the between-study approach for dealing with comparative studies relative to the within-study approach can be attributed to more realizations of the random effects; each test result from a comparative study is treated as if obtained from a different study. Thus, the standard errors from the between-study approach are incorrect and the within-study approach should always be used. The within-study approach gave more precise estimates compared to those from the between-study approach.

Assumptions of equal variances have the advantage of simplifying estimation of hierarchical models. This may not always be appropriate, leading to incorrect standard errors and misleading inferences. For almost half (43%) of the 49 test comparisons where both bivariate models—equal variances and unequal variances with independence between tests—converged, there was statistical evidence of differences in model fit between the models. Although numerical differences between point estimates were on average very small, there were differences in precision and the ranking of tests.

Similarly, for the subset of *direct* comparisons, there were little or no differences in the point estimates of relative sensitivity and relative specificity between the three models fitted. However, precision of the estimates differed and models with equal variances tended to agree with more complex models that take into account correlation between tests at study level. Given the complexity of the latter and the paucity of comparative studies, assumption of equal variances may be a suitable simplification but requires further investigation because there were few direct comparisons with sufficient data for the analyses.

The degree to which variances differ between tests in a test comparison may depend on the number and sample size of included studies, type of comparison (direct or indirect) or the type of tests. For example, comparing CT and US for diagnosis of acute appendicitis (ID 46), the between-study variability in the performance of US appeared greater compared to that of CT. In separate meta-analysis of CT and US, the variance of the random effects for logit sensitivity and logit specificity, and the correlation were 0.016, 0.007 and +1 for CT, and 0.698, 0.921 and -0.034 for US. Differences in variation may be due to US being operator dependent while CT is not. If this explains the difference in variability, then the difference is likely to persist even if the meta-analysis is restricted to direct comparisons of US and CT. However if factors inducing variability between studies are not attributable to a test but to other aspects related to study design and execution, then it is possible that heterogeneity may be greater in indirect than in direct comparisons. This warrants further investigation in a future study.

Although there was statistical evidence of differences in the shape of SROC curves for some test comparisons, there were mainly very small differences in relative test performance. The number of studies seems to be a driver in the estimation of the shape parameter and so the estimation may be unreliable when there are few studies. Of the assumptions investigated, the assumption of equal shape appeared to have the least effect on findings.

7.8.1.2 Comparisons of univariate and bivariate models

Comparing bivariate and univariate models, irrespective of the covariance structure used to model the random effects of logit sensitivity and logit specificity, there were little or no differences in relative sensitivity and relative specificity across the test comparisons. Previous

empirical studies of meta-analysis of single tests also found agreement between univariate and bivariate models.^{36,225} Bivariate meta-analysis of diagnostic accuracy is an application of multivariate meta-analysis. Findings similar to those in this study have been shown for treatment effects where the borrowing of strength provided by multivariate meta-analysis has often been found to be small.^{269,270} Trikalinos et al compared univariate and multivariate meta-analyses of treatment effects using empirical and simulation studies and found small numerical differences in the summary effects and their confidence intervals.²⁷¹ However, when there are missing data for an outcome, estimates from multivariate meta-analyses have been shown to have better statistical properties than those from univariate meta-analyses.²¹⁴

In view of the findings of the present study and lack of missing data because complete data for both the diseased and non-diseased groups are used in test accuracy meta-analysis, some may infer that bivariate meta-regression is unjustified. On the contrary, bivariate analysis is useful because it enables estimation within a single modelling framework thus allowing for joint confidence and prediction regions around summary points. Notwithstanding, if estimation problems are encountered when fitting bivariate models, a univariate model is likely to be a valid alternative.

7.8.1.3 Comparisons of HSROC and Moses SROC models

Results and conclusions from weighted and unweighted Moses SROC models differed substantially, implying that the example illustrated in section 6.3.2.1 is not unique. For HSROC and unweighted Moses models which assumed a common shape for SROC curves, there were large differences between rDORs with more than a two-fold difference for several (22%) test comparisons. The ratios of rDORs were on average higher from the comparison of

HSROC and weighted Moses models than from the comparison of HSROC and unweighted Moses models. The tendency for the Moses model to underestimate accuracy can be partly explained by the use of zero cell corrections in a large number (53/57, 93%) of test comparisons. As described by Moses et al,³⁷ in a meta-analysis where a study had a zero in any of the cells of the 2x2 table, a zero cell correction of 0.5 was added to the cell counts of all studies in the meta-analysis, including the studies without a zero cell.

For HSROC and unweighted Moses models that allowed curves to differ by test, although differences were small on average, there were a large number (35%) of changes in statistical significance as well as several qualitative changes (29%). Estimates from HSROC models were on average considerably more precise than those from unweighted or weighted Moses models. In a small empirical evaluation of eight meta-analyses of single tests, Harbord et al also showed that SROC curves derived from Moses models can differ from those obtained from HSROC models.³⁵

7.8.2 Implications for research and practice

Due to the effect of different assumptions on meta-analytic findings and conclusions shown in this study, meta-analysts should carefully assess modelling assumptions when fitting hierarchical models. These modelling assumptions are briefly mentioned in the statistical chapter of the Cochrane Handbook for DTA reviews²³ but the advice needs to be strengthened in light of the new evidence. For example, there needs to be a shift in the analyses of test comparisons towards greater emphasis on the importance of exploring covariance structures.

The extent to which different assumptions can be explored will depend on the available data and software capability. As noted earlier in section 7.2.3.3, even different commands that purport to fit the same mathematical model within a software package can offer different options. Assessment of the models should not rely entirely on statistical tests of model fit such as likelihood ratio tests, but the estimates and model parameters should also be examined.

For *direct* comparisons, findings from bivariate models with the most complex covariance structure which allowed for separate variances and study level correlations between tests were similar to those from the simplest bivariate models which assumed common variances and correlation across tests and studies. This implies that the simplest bivariate meta-regression model may be valid for *direct* comparisons when comparative studies are few as is often the case in comparative meta-analysis. Nevertheless, it should be noted that there were few direct comparisons for this evaluation and the approach did not model within-study correlations. Further evaluation is needed in a simulation study before a definitive recommendation can be made.

Irrespective of the type of test comparison, there was evidence to suggest that univariate meta-regression can be an alternative when bivariate meta-regression is not feasible. Although mentioned earlier in the chapter, it should be stressed that these univariate models use a binomial likelihood to model within-study variability and are not the frequently used traditional univariate methods that use an approximate likelihood which require continuity corrections. The use of univariate random effects logistic regression models may therefore be a sensible alternative when bivariate models fail, but this will be formally examined via a simulation study in Chapter 8.

The methodological limitations of the Moses method are well documented and were discussed in section 1.4.3. Taking into account the limitations as well as the obvious differences between estimates from Moses SROC and HSROC meta-regression, the Moses model should not be used for test comparisons. This recommendation extends to investigations of heterogeneity which are typically performed using this same meta-regression approach.

7.8.3 Strengths and limitations

To the author's knowledge, this is the first empirical evaluation of comparative meta-analysis methods. The investigation not only compared the methods, but also assessed common assumptions made when fitting hierarchical meta-regression models. A key strength of this empirical evaluation is the relatively large cohort of test comparisons with a wide range of test types and target conditions. Therefore, the test comparisons are considered representative of the literature. Furthermore, there was variety in the type (comparative and non-comparative) and number of studies included, thus making both direct and indirect comparisons possible.

The preliminary meta-analyses of single tests are an important by product of this study. These analyses enabled empirical investigation of the performance of hierarchical models for the meta-analysis of a single test. While there have been previous empirical studies, they have mainly focused on the bivariate model^{36,272-274} or included a small number of meta-analyses.^{35,225} This study also appears to be the first to consider the HSROC model in great detail using a large cohort of meta-analyses.

There are limitations. First, the evaluation was limited to frequentist based methods. While all the methods identified in Chapter 6 may merit evaluation, only commonly used methods or theoretically rigorous classical methods were evaluated in order to keep the evaluation manageable within the scope of a thesis and relevant to most meta-analysts. Second, a limitation of an empirical study such as this is that true values of the estimates are unknown unlike in a simulation study. The simplistic philosophy adopted was to assume that where differences occurred, more complicated models were correct but this reasoning may not always be true. Third, data extraction was performed by a single person. Although data were subsequently double checked by the same person, the possibility of errors cannot be completely eliminated. Nevertheless, due to overlap with other review cohorts in the thesis that were randomly checked by another researcher, the risk of errors is small and unlikely to affect the conclusions of this study. Importantly, any errors will not affect the comparison of the models. Fourth, different covariance structures were not investigated for the HSROC model. Due to the extensive scope of the analyses in this chapter, focus was on the bivariate model since it is used more often than the HSROC model according to the findings in Chapter 4 and citations of the methods presented in Chapter 6. There is no reason to doubt applicability of the findings of bivariate models to HSROC models. However, because HSROC models are non-linear generalized mixed models, models with complex covariance structures are likely to be more challenging to fit than bivariate models.

7.8.4 Conclusions

Since findings from Moses SROC and HSROC meta-regression differ and the Moses model has methodological limitations, the Moses model should not be used for making inferences about relative test performance. In the recommended hierarchical models, different modelling

assumptions can lead to different conclusions about relative test performance and so assumptions should be thoroughly investigated where possible. In particular, assumptions about the covariance structure should not be taken for granted.

For estimation of SROC curves, assuming the same shape for the curves of different tests may be appropriate especially when there are few studies. For estimation of summary points, simplifying bivariate meta-regression models to univariate models may be a valid alternative for comparative meta-analyses but joint inferences cannot be made about sensitivity and specificity or their differences. Future research should investigate these simpler models for comparative meta-analyses in a simulation study as the study described in the next chapter is limited to simpler hierarchical models for meta-analysis of a single test.

8 PERFORMANCE OF METHODS FOR META-ANALYSIS WITH FEW STUDIES OR SPARSE DATA

A paper based on the content of this chapter has been published.

Citation: **Takwoingi Y**, Guo B, Riley R, Deeks J. Performance of methods for meta-analysis of diagnostic test accuracy with few studies or sparse data. *Statistical Methods in Medical Research*. Epub ahead of print June 26 2015.

8.1 Introduction

Hierarchical models possess theoretical advantages over simpler methods for meta-analysis of test accuracy studies but fitting them is not trivial as exemplified in earlier chapters. The models are often fitted using a frequentist approach that relies on likelihood based methods for the estimation of five parameters in a basic model without covariates. Solving the likelihood equations requires an iterative process and in certain circumstances, for instance when there are few studies and/or sparse data (e.g. zero cells due to perfect sensitivity and/or specificity) in the meta-analysis, the models either fail to converge or they converge but give unreliable parameter estimates (e.g. with one or more missing standard errors; see examples in Table 7.2). These issues are often encountered by meta-analysts³⁶ and there is uncertainty about how to proceed with meta-analysis in such situations.

Academic illustrations of the application of hierarchical methods have typically involved large meta-analyses.^{35,39,41,42,53,54,194,226,275} In contrast, experience of supporting Cochrane and non-Cochrane diagnostic test accuracy review authors, suggest that small meta-analyses or sparse data often occur and pose a challenge to these data hungry hierarchical models as shown in section 2.3.5 and Chapter 7. Others have also noted the problem of non-convergence.^{39,110,192,275,276} The preliminary meta-analyses of single tests and comparative

meta-analyses in the preceding chapter also revealed problems with model convergence and stability.

Despite the increasing uptake of these models, a recent survey has suggested a lack of clarity about recommended methods for meta-analysis and a need for guidance.²⁷⁶ Therefore, the aim of this chapter is to examine the performance of hierarchical models for meta-analysis of test accuracy studies with sparse data, and to provide recommendations for how to proceed in this situation. Sensitivity and specificity are the test accuracy measures most commonly used in meta-analyses,²⁷ and so only methods for synthesis of these measures was considered.

Although the issue of small number of studies or sparse data in comparative meta-analysis is of prime interest, as previously outlined in section 1.7.2, the scope of the simulation study in this thesis is limited to meta-analysis of a single test. Due to limited evidence on how well hierarchical models perform in meta-analysis of a single test, establishing the validity of simpler models for evaluation of a single test is the first step of a two-step process in terms of developing the evidence base. Furthermore, in situations with limited data for some tests in a comparative review, separate meta-analyses may be required for such tests individually, in addition to the comparative meta-analysis of tests with sufficient data. Also, meta-analyses of single tests are usually undertaken prior to comparative meta-analyses and such analyses can provide insight into whether or not the models will converge when they are extended to include a test comparison. This approach was adopted in Chapter 7 for the empirical evaluation of comparative meta-analysis methods. Thus, a comprehensive evaluation of issues in the meta-analysis of a single test is pertinent and is the challenge addressed in this chapter.

The understanding gained from the present study will aid the design of a future simulation study addressing test comparisons which is the second challenge.

The outline of this chapter is as follows. Section 8.2 outlines the relationship between the bivariate and HSROC model parameters, and describes various simplifications of both models. In section 8.3 two motivating examples where the bivariate model failed to converge are outlined and simpler forms of the hierarchical models applied to resolve this are presented. In section 8.4 the simulation study is described and the results for full and simplified hierarchical models are presented in section 8.5. Section 8.6 considers extensions to test comparisons using examples. In section 8.7 the findings of the simulation are discussed and conclude with recommendations for selecting an appropriate meta-analytic approach in practice.

8.2 Model parsimony

As shown by Harbord et al,⁴⁴ the five parameters of the bivariate model can be expressed in terms of those of the HSROC model as follows:

$$\mu_A = \exp\left(-\frac{\beta}{2}\right)\left(\Theta + \frac{\Lambda}{2}\right), \mu_B = -\exp\left(\frac{\beta}{2}\right)\left(\Theta - \frac{\Lambda}{2}\right) \quad (8.1)$$

$$\sigma_A^2 = \exp(-\beta)\left(\sigma_\theta^2 + \frac{1}{4}\sigma_\alpha^2\right) \quad (8.2)$$

$$\sigma_B^2 = \exp(\beta)\left(\sigma_\theta^2 + \frac{1}{4}\sigma_\alpha^2\right) \quad (8.3)$$

$$\sigma_{AB} = -\left(\sigma_\theta^2 - \frac{1}{4}\sigma_\alpha^2\right). \quad (8.4)$$

Parameters can be removed from HSROC and bivariate models to simplify the models. Based on the equations above, the equivalence of simplified HSROC and bivariate models can be

shown. For the HSROC model, the random effect for accuracy can be dropped thus assuming a fixed effect for the accuracy parameter and that only the threshold parameter varies between studies, or vice versa. An HSROC model with $\sigma_{\alpha}^2 = 0$ corresponds to a bivariate model with a correlation of -1 , while an HSROC model with $\sigma_{\theta}^2 = 0$ corresponds to a bivariate model with a correlation of $+1$. The HSROC framework also allows the assumption of a symmetric SROC curve with constant DOR by setting $\beta = 0$, which in the bivariate model corresponds to equal variances of logit sensitivity and logit specificity ($\sigma_A^2 = \sigma_B^2$).⁴⁴ This model is equivalent to assuming an exchangeable variance-covariance structure for the bivariate model as was done in the example in section 2.3.5.1. If both accuracy and threshold are modelled as fixed effect parameters, $\sigma_{\alpha}^2 = 0$ and $\sigma_{\theta}^2 = 0$, then based on equations 8.2, 8.3 and 8.4, $\sigma_A^2 = 0$, $\sigma_B^2 = 0$ and $\sigma_{AB} = 0$. Thus, fixed HSROC models with a symmetric or asymmetric curve are equivalent to simultaneously fitting two univariate fixed effect logistic regression models for sensitivity and specificity (see results for these models for the motivating examples in Table 8.1). Henceforth, they are referred to as fixed effect models; the models can be considered a special case of the random effects models where the variances of the random effects are zero.

For the bivariate model, dropping the correlation parameter or setting it equal to zero results in two univariate random effects logistic regression models for sensitivity and specificity. This model was applied to the example in section 2.3.2.2 and empirically evaluated in Chapter 7. Recall this is a simplification of the bivariate generalized mixed model achieved by setting the covariance or correlation parameter to zero (see section 1.4.4.1 and equation 1.12). The model is also equivalent to assuming an independent variance-covariance structure.

The following nine models obtained by simplifying bivariate and HSROC models will be fitted to motivating examples in the next section.

1. Univariate fixed effect logistic regression model – includes μ_A for the logit sensitivities, and μ_B for the logit specificities.
2. Univariate random effects logistic regression model – includes μ_A and σ_A^2 for the logit sensitivities, and μ_B and σ_B^2 for the logit specificities. For brevity, from here on this model will be referred to simply as the univariate random effects model (UREM).
3. Bivariate model – includes all five parameters μ_A , σ_A^2 , μ_B , σ_B^2 , and the covariance σ_{AB}^2 (see section 1.4.4.1 and equation 1.10).
4. Complete HSROC model – includes all five parameters Λ , Θ , β , σ_α^2 and σ_θ^2 (see section 1.4.4.2 and equation 1.13).
5. Symmetric HSROC model – includes Λ , Θ , σ_α^2 and σ_θ^2 .
6. HSROC model with fixed threshold – includes Λ , Θ , β and σ_α^2 .
7. HSROC model with fixed accuracy – includes Λ , Θ , β and σ_θ^2 .
8. HSROC model with fixed accuracy and threshold – includes Λ , Θ and β (allows for asymmetry in the SROC curve).
9. Symmetric HSROC model with fixed accuracy and threshold – includes only two parameters Λ and Θ .

Throughout the rest of this chapter, an HSROC model that contained all five parameters is referred to as a complete HSROC model.

8.3 Motivating examples of meta-analyses of a single test

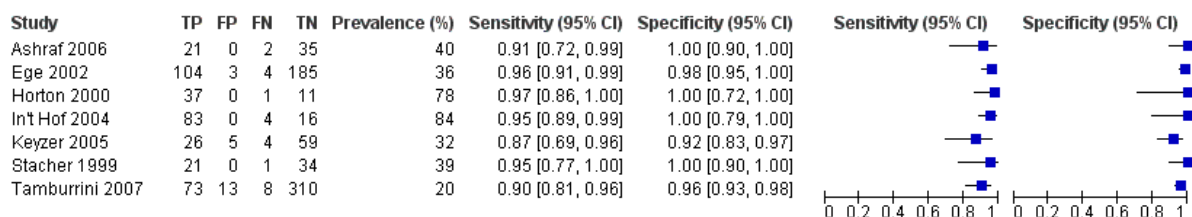
8.3.1 Non-contrast computed tomography for diagnosing appendicitis

Hlibczuk et al²⁷⁷ reviewed the diagnostic accuracy of non-contrast CT for emergency department evaluation of adults with suspected appendicitis. Seven studies, evaluating 1,060 patients of whom 389 had appendicitis, were included in the review. The prevalence of appendicitis in the studies ranged from 20% to 84%, with a median of 39%. The forest plot (Figure 8.1) shows between-study variation in the sensitivities and specificities, though specificity was perfect (100%) in four studies. The authors attempted to fit the bivariate model in SAS but the model failed to converge.

8.3.2 Computed tomography for diagnosing scaphoid fractures

Yin et al¹⁸⁷ assessed the diagnostic accuracy of CT for diagnosing suspected scaphoid fractures. Six studies, evaluating 211 patients of whom 44 had a scaphoid fracture, were included in the review. The prevalence of scaphoid fractures in the studies ranged from 12% to 38%, with a median of 20%. Figure 8.1 shows the estimates of sensitivity and specificity with almost no between-study variation; five of the six studies reported 100% sensitivity while all studies reported 100% specificities. The authors pooled sensitivity, specificity, and the DOR using a random effects model (method not specified).

Example 1: Non-contrast computed tomography for diagnosing appendicitis (Hlibczuk et al 2010)



Example 2: Computed tomography for diagnosing scaphoid fractures (Yin et al 2010)

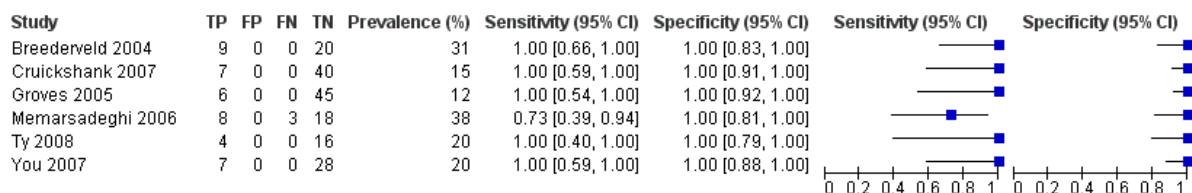


Figure 8.1| Forest plot of sensitivity and specificity estimates from studies included in the two motivating examples

FN=false negative; FP=false positive; TN=true negative; TP=true positive.

(Adapted from Takwoingi et al 2015²⁷⁸)

8.3.3 Results from reanalysis of the two example datasets

To examine the performance of different models, the two datasets were reanalysed by fitting univariate, bivariate and HSROC models using the SAS NLMIXED procedure. In total nine different versions of these models were considered (see models in section 8.2). Univariate random effects logistic regression models for sensitivity and specificity were simultaneously obtained by setting the covariance parameter in a bivariate generalized linear mixed model equal to zero (see equation 1.12). This is equivalent to a bivariate model with an assumed independent between-study variance-covariance structure. Additional summary measures such as likelihood ratios and DORs were produced using the ESTIMATE statement within NLMIXED. Despite numerous attempts with different starting values and optimization algorithms, the bivariate model failed to converge for both datasets. The models fitted and results obtained for both datasets are summarised in Table 8.1.

Table 8.1 | Summary accuracy measures obtained from different meta-analytic models applied to the two motivating examples

Meta-analytic model	Sensitivity (95% CI)	Specificity (95% CI)	LR+ (95% CI)	LR- (95% CI)	DOR (95% CI)
<i>Non-contrast CT for appendicitis (n = 7)</i>					
Univariate fixed effect logistic regression	93.8 (90.6–96.0)	96.9 (95.1–98.0)	30 (19–48)	0.06 (0.04–0.10)	471 (244–907)
Univariate random effects logistic regression	93.8 (90.8–95.9)	97.4 (93.7–99.0)	37 (14–93)	0.06 (0.04–0.10)	579 (205–1635)
Bivariate random effects model	NE	NE	NE	NE	NE
Complete HSROC	94.1 (88.5–97.0)	97.8 (92.6–99.4)	42 (12–148)	0.06 (0.03–0.12)	700 (130–3771)
Symmetric HSROC	94.1 (88.2–97.2)	97.5 (94.1–99.0)	38 (15–93)	0.06 (0.03–0.13)	628 (149–2657)
Fixed accuracy	93.8 (90.1–96.2)	96.9 (94.7–98.2)	30 (18–51)	0.06 (0.04–0.10)	471 (223–995)
Fixed threshold	94.1 (88.9–96.9)	97.8 (93.0–99.3)	42 (13–140)	0.06 (0.03–0.12)	701 (141–3485)
Fixed accuracy and threshold	93.8 (90.6–96.0)	96.9 (95.1–98.0)	30 (19–48)	0.06 (0.04–0.10)	471 (244–907)
Symmetric fixed accuracy and threshold	93.8 (90.6–96.0)	96.9 (95.1–98.0)	30 (19–48)	0.06 (0.04–0.10)	471 (244–907)
<i>CT for scaphoid fractures (n = 6)</i>					
Univariate fixed effect logistic regression	93.2 (78.8–98.1)	100	2.26E+07 (NE)	0.07 (0.02–0.23)	3.31E+08 (NE)
Univariate random effects logistic regression	99.0 (3.7–100)	100	3.25E+07 (NE)	0.01 (3.94E-06–24)	3.34E+09 (NE)
Bivariate random effects	NE	NE	NE	NE	NE
Complete HSROC	99.1 (2.2–100)	100	2.07E+09 (NE)	0.01 (2.14E-06–41)	2.21E+11 (NE)
Symmetric HSROC	98.6 (12.9–100)	100 (0–100)	54762 (1.29E-05–2.33E+14)	0.01 (3.36E-05–6.11)	3818334 (0.0001–1.33E+17)
Fixed accuracy	99.0 (6.4–100)	100	1.59E+11 (NE)	0.01 (7.03E-06–13)	1.64E+13 (NE)
Fixed threshold	99.0 (6.4–100)	100	2.37E+09 (NE)	0.01 (7.03E-06–13)	2.43E+11 (NE)
Fixed accuracy and threshold	93.2 (78.8–98.1)	100	7.80E+09 (NE)	0.07 (0.02–0.23)	1.14E+11 (NE)
Symmetric fixed accuracy and threshold	93.2 (78.8–98.1)	100	2.27E+07 (NE)	0.07 (0.02–0.23)	3.34E+08 (NE)

n = number of studies in the meta-analysis, NE = Not estimable.

(Adapted from Takwoingi et al 2015²⁷⁸)

For the appendicitis dataset, the complete HSROC model successfully converged and produced reliable estimates only when boundary constraints ($\sigma^2 \geq 0$) were specified for σ_α^2 and σ_θ^2 ; the boundary constraint for σ_θ^2 was activated (estimation truncated at zero) and the between-study correlation was estimated as +1. As previously mentioned in section 2.3.5, this is due to the maximum likelihood estimator truncating the between-study covariance matrix on the boundary of its parameter space.^{110,111}

Since the maximum likelihood estimation problems encountered with the bivariate model are most likely due to boundary estimation of the variance and/or covariance parameters, attempts were made to plot the profile log likelihood for the covariance parameter (maximized with respect to the other four parameters). It was not possible to produce a plot for the scaphoid fracture example because the bivariate model failed even with fixed values for the covariance. This is unsurprising since there was almost no between-study variation in sensitivity and specificity, and therefore a fixed, non-zero between-study covariance is not really meaningful.

Figure 8.2 shows the profile log likelihood for the covariance parameter for the appendicitis example. There is very little change in the profile log likelihood and no curved maximum point as one would usually expect from maximum likelihood estimation. The maximum of the profile log likelihood was achieved at a covariance of 0.02 (dashed line). For covariances above 0.02, the bivariate model failed to converge or was unstable, but values between -0.05 and 0.02 appear to be supported by the data. The dotted line shows the value of the log likelihood for a covariance of zero, i.e., independence between sensitivity and specificity. The change in the log likelihood is

negligible, suggesting that univariate random effects logistic regression models would be appropriate for pooling sensitivity and specificity in this example.

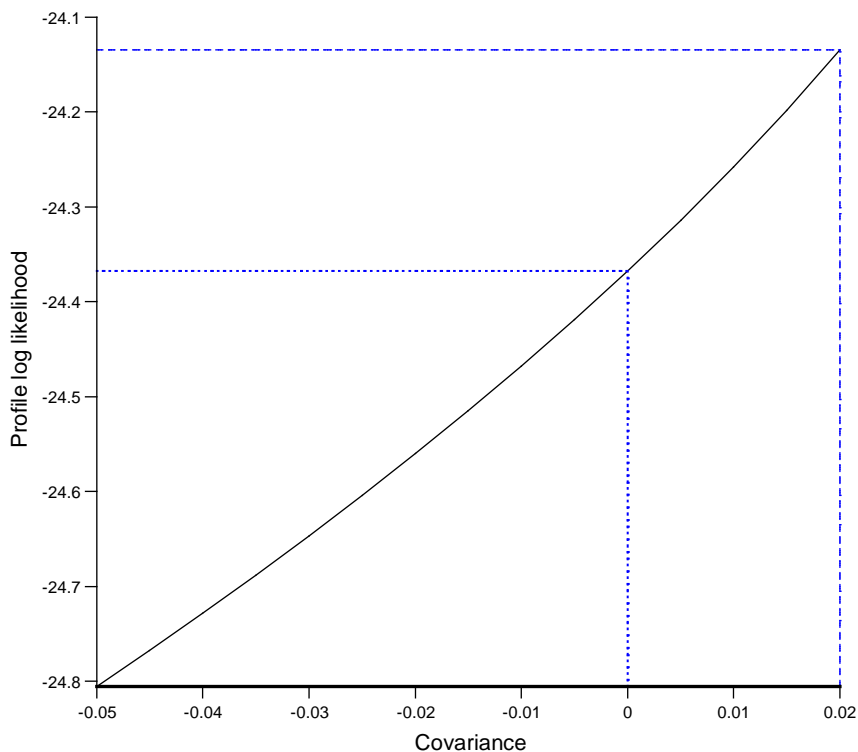


Figure 8.2| Profile log-likelihood function of the covariance parameter in the bivariate model applied to the appendicitis example
(Adapted from Takwoingi et al 2015²⁷⁸)

The two examples illustrate the problem of model convergence, poor parameter estimation and the need for simpler models. There were only subtle differences in summary estimates and 95% CI for sensitivity, specificity and the negative likelihood ratio between models fitted to the appendicitis dataset. In contrast, clear differences were observed for the positive likelihood ratio and the DOR. For example, the DORs (95% CI) for the univariate fixed effect model and the complete HSROC model were 471 (244 to 907) and 700 (130 to 3771). For the scaphoid fractures dataset, there were differences in summary estimates and 95% CI for sensitivity from the

univariate fixed effect model and the HSROC models with both fixed accuracy and threshold parameters compared to the other models. For example, the summary sensitivities (95% CI) for the univariate fixed effect model and the complete HSROC model were 93.2% (78.8% to 98.1%) and 99.1% (2.2% to 100%). These examples show that results can differ importantly between models. Therefore, in situations of non-convergence of the full hierarchical models, the identification of simpler meta-analytic methods that still give valid answers is critical.

8.4 Simulation study methods

A simulation study was conducted to compare the performance of a univariate random effects logistic regression model (UREM) and the HSROC model with various simplifications (by removing model parameters or setting them equal to zero). Given the mathematical equivalence of the HSROC and bivariate models when no covariate is included,⁴⁴ there was no need to examine the performance of both models. The HSROC model was chosen because it has greater flexibility for introducing model parsimony by dropping parameters than the bivariate model.⁴⁴ Since several authors^{53,110,226,275} have shown that approximate methods for modelling within-study variability are biased, only methods that use a binomial likelihood were investigated. The specifications for the scenarios were devised to replicate realistic situations encountered in meta-analysis of diagnostic accuracy studies. The effect of these factors was investigated: 1) number of studies; 2) magnitude of diagnostic accuracy (DOR); 3) prevalence of disease; 4) between-study variation in accuracy and threshold; and 5) asymmetry in the SROC curve. The simulation approach used in a previous study²²⁰ was modified to define the simulation scenarios and generate the simulated datasets described below.

8.4.1 Generation of simulated data

Meta-analyses with different number of studies ($N = 5, 10, 20$) were investigated. The size of a study in each meta-analysis, n_j , was randomly sampled from a uniform distribution, $U(20,200)$. Diagnostic accuracy studies are often small in size^{27,28} hence the reason for varying n_j between 20 and 200. To generate individual studies for each meta-analytic dataset with an underlying prevalence p , individuals within a study were randomly classified as diseased or non-diseased. Each individual was then assigned a continuous test result value, x , which was randomly sampled from logistic distributions with means μ_1 and μ_2 (where $\mu_2 > \mu_1$), and standard deviations σ_1 and σ_2 for non-diseased and diseased as shown in Figure 8.3. The diagnostic threshold, t , was chosen as the average of the means of the two distributions, i.e., $t = (\mu_1 + \mu_2)/2$. For each study i in the j th meta-analysis, t was used to determine the outcome of an individual's test result; positive if $x_{ij} > t$, or negative if $x_{ij} \leq t$. To create the 2x2 table for each study, individuals were then classified as true positives, false negatives, false positives or true negatives based on test result and disease status.

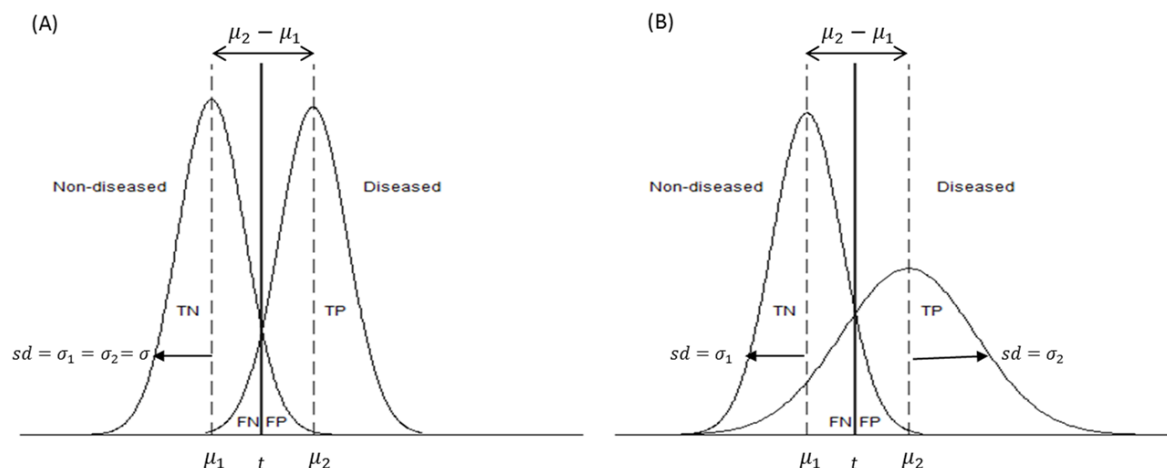


Figure 8.3| Underlying bilogistic distribution of diseased and non-diseased used in the simulation

FN = false negative; FP = false positive; sd = standard deviation; t = threshold; TN = true negative; TP = true positive

The distribution for the non-diseased group has mean μ_1 , and the distribution for the diseased group has mean μ_2 . The distributions have the same standard deviation ($\sigma_1 = \sigma_2 = \sigma$) in (A) but the standard deviations are different ($\sigma_1 \neq \sigma_2$) in (B). Diagnostic accuracy is the standardized difference in means ($(\mu_2 - \mu_1)/\sigma$).

(Adapted from Deeks et al 2005²²⁰)

The standardised distance between the means μ_1 and μ_2 was used to determine diagnostic accuracy (Figure 8.3). The DOR at t can be calculated as follows:

$$\text{DOR} = \exp \left[\sqrt{\frac{\pi^2}{3}} \left(\frac{\mu_2 - t}{\sigma_2} - \frac{\mu_1 - t}{\sigma_1} \right) \right] \quad (8.5)$$

The sensitivity and specificity at t can be obtained using the following:

$$\text{Sensitivity} = \frac{\exp \left[\sqrt{\frac{\pi^2}{3}} \left(\frac{\mu_2 - t}{\sigma_2} \right) \right]}{1 - \exp \left[\sqrt{\frac{\pi^2}{3}} \left(\frac{\mu_2 - t}{\sigma_2} \right) \right]} \quad (8.6)$$

$$\text{Specificity} = 1 - \frac{\exp\left[\sqrt{\frac{\pi^2}{3}}\left(\frac{\mu_1 - t}{\sigma_1}\right)\right]}{1 + \exp\left[\sqrt{\frac{\pi^2}{3}}\left(\frac{\mu_1 - t}{\sigma_1}\right)\right]} \quad (8.7)$$

When the distributions of test results for the diseased and non-diseased have the same standard deviation ($\sigma_1 = \sigma_2 = \sigma$) as in Figure 8.3 (A), sensitivity = specificity at t and the SROC curve has a symmetric shape. For scenarios with symmetric SROC curves, values of diagnostic accuracy that correspond to the following were investigated:

- I. $\mu_2 = 3, \mu_1 = 1$ and $\sigma = 1$. These values were chosen such that $(\mu_2 - \mu_1)/\sigma=2$ giving a log DOR of 3.63 (DOR = 38) and sensitivity = specificity = 0.86.
- II. $\mu_2 = 4, \mu_1 = 1$ and $\sigma = 1$. $(\mu_2 - \mu_1)/\sigma=3$ (log DOR = 5.44, DOR = 231; sensitivity = specificity = 0.94).

The values are arbitrary and were chosen to represent tests with moderate (DOR = 38) and high (DOR = 231) accuracy. The latter is of particular interest because sparse data are more likely to occur when test accuracy is high.

If two logistic distributions have different standard deviations ($\sigma_1 \neq \sigma_2$) as shown in Figure 8.3 (B), sensitivity \neq specificity at t and the SROC curve has an asymmetric shape (also see sections 1.3.2.2 for a discussion of the shape of ROC curves, and sections 1.4.3 and 1.4.4.2 for a discussion of the shape of SROC curves). For scenarios with asymmetric SROC curves where $\sigma_2 = 2\sigma_1$, using the same μ_2, μ_1 and σ_1 as in (I) and (II) above, the following values of diagnostic accuracy were investigated:

- I. DOR = 15, sensitivity = 0.71 and specificity = 0.86. The same values used for μ_2 , μ_1 and σ_1 in (I) above to obtain a DOR of 38 and sensitivity = specificity = 0.86, give a DOR of 15 and sensitivity of 0.71 based on substituting $\sigma_2 = 2$ into equations 8.5 and 8.6),
- II. DOR = 59 (reduces from a DOR of 231), sensitivity = 0.80 and specificity = 0.94.

To begin, zero between-study variation in both accuracy and threshold was assumed.

Subsequently, between-study variation in diagnostic accuracy was introduced by adding a value τ sampled from a normal distribution with zero mean and standard deviation $0.3\sigma_1$. This value was added to the difference in means ($\mu_2 - \mu_1$) for each study. Between-study variation in diagnostic threshold was introduced by also sampling from a normal distribution with the average threshold t as the mean and standard deviation $0.3\sigma_1$. For each scenario, 10 000 independent meta-analysis datasets were generated to enable precise estimation of model performance even if a large proportion of models fail to converge. If all 10 000 datasets for each scenario successfully converged, they will give a standard error of 0.0022 for the estimation of 95% confidence interval coverage probability.²⁷⁹ However if only 1000 datasets converged, the standard error will be 0.0069. The datasets were created using Stata version 10.1.

Table 8.2 summarises the different scenarios investigated. The meta-analysis dataset for the base scenario for each DOR contained five studies with an underlying prevalence of 5% and no heterogeneity in accuracy or threshold.

Table 8.2| Scenarios evaluated in the simulation

Scenario	Prevalence (%)	DOR	Heterogeneity in accuracy and threshold	Asymmetry in SROC curve
1–3	5	38	No	No
4–6	25	38	No	No
7–9	50	38	No	No
10–12	5	38	Yes	No
13–15	25	38	Yes	No
16–18	50	38	Yes	No
19–21	5	231	No	No
22–24	25	231	No	No
25–27	50	231	No	No
28–30	5	231	Yes	No
31–33	25	231	Yes	No
34–36	50	231	Yes	No
37–39	5	15	Yes	Yes
40–42	25	15	Yes	Yes
43–45	50	15	Yes	Yes
46–48	5	59	Yes	Yes
49–51	25	59	Yes	Yes
52–54	50	59	Yes	Yes

Each subset of 3 scenarios corresponds to 5, 10, and 20 studies.
(Adapted from Takwoingi et al 2015²⁷⁸)

8.4.2 Meta-analytic models fitted to each dataset

Of the nine models in section 8.2, the following seven models were fitted to each meta-analysis dataset:

1. Univariate random effects logistic regression model
2. Complete HSROC model

3. Symmetric HSROC model
4. HSROC model with fixed threshold
5. HSROC model with fixed accuracy
6. HSROC model with fixed accuracy and threshold
7. Symmetric HSROC model with fixed accuracy and threshold

The SAS NLMIXED procedure was used to fit the models because Stata does not have an inbuilt or user defined command for fitting non-linear generalized mixed models as mentioned earlier in section 2.2.5. Note that because of the mathematical relationship between the bivariate and HSROC model, it is possible in Stata to obtain estimates for the five parameters of the HSROC model using functions of parameters from the bivariate model fitted.⁴⁴ Additional estimates were computed by using the ESTIMATE statement in NLMIXED. The log DOR was computed at the average operating point (summary sensitivity and specificity). This log DOR is exactly the same value as Λ if the SROC curve is symmetric.

8.4.3 Facilitating convergence of hierarchical models

To aid convergence, a wide range of starting values for model parameters was provided by specifying a grid of points for a grid search of starting values. A quasi-Newton optimization technique (the NLMIXED default) was used because it provides an appropriate balance between computation speed and stability [SAS Institute Inc. SAS OnlineDoc® 9.1.3. Cary, North Carolina, 2004]. To prevent estimation of negative variances and to reduce computational problems, boundary constraints ($\sigma^2 \geq 0$) were specified for the variance parameters in the models. To reduce the number of models that failed to converge, models were refitted by trying a new set

of starting values and/or changing the optimization technique to a Newton-Raphson technique. To obtain a new set of starting values, a model without random effects was fitted and the new parameter estimates were used together with the original grid of points for the variance parameters. Thus for some datasets, up to four attempts were made to fit a hierarchical model.

8.4.4 Assessment of model convergence and stability

As explained in section 7.2.3.4, a model that meets a convergence criterion may be unstable or have missing standard errors due to issues with model identifiability. Therefore, convergence was assessed in two stages. First, by checking whether the convergence criterion was met and also whether the additional estimates defined in the ESTIMATE statements were produced. Second, because standard errors are computed from the final Hessian matrix, eigenvalues of the Hessian were calculated to detect if there were problems. At a true minimum, eigenvalues will all be positive, i.e., positive definite. Therefore, for convergence to be deemed successful, the model had to meet the convergence criterion, produce additional estimates, and the Hessian had to be positive definite.

8.4.5 Assessment of performance of meta-analytic models

The performance of the methods was assessed by examining estimates of the following measures of diagnostic accuracy: log DOR, logit sensitivity and logit specificity. Estimability was expressed in terms of the percentage of meta-analyses that successfully converged and the percentage where the between-study correlation was not estimated as -1 or $+1$. The latter was computed for only the complete HSROC model. For each scenario, only results from meta-analyses that successfully converged as defined above were used to calculate (a) the difference

between the average parameter estimate and the true parameter value to determine bias; (b) the average standard error and mean square error (MSE incorporates both bias and variability) to assess model accuracy; and (c) the coverage of the 95% CIs by computing the percentage of meta-analyses for which the true parameter value was within the 95% confidence interval.

8.5 Simulation results

Altogether 54 scenarios were explored. Only the results for the log DOR are shown in detail in this chapter; the results for logit sensitivity and logit specificity are only briefly mentioned and are also shown in appendices D.1 and D.2. Because homogeneous accuracy and threshold is the exception rather than the norm for meta-analysis of test accuracy studies, to illustrate key findings, results are presented mainly for scenarios with heterogeneity at a DOR of 231 (sparse data are of interest and zero false positives and/or false negatives are more likely to occur when diagnostic accuracy is high).

8.5.1 Estimability

Zero cells occurred frequently especially when diagnostic accuracy was high (Table 8.3).

Convergence rates were higher for the complete HSROC model in scenarios with heterogeneity compared to scenarios without heterogeneity. This is likely due to the inclusion of heterogeneity parameters in the HSROC model that become problematic to estimate when the true heterogeneity is zero. Convergence increased with increasing number of studies and prevalence, and with decreasing diagnostic accuracy. Convergence decreased in scenarios with asymmetry in the SROC curve (shown in Figure 8.4 for scenarios with heterogeneity). Across scenarios, non-convergence and problems with model identifiability were more common with the complete

HSROC and fixed threshold models compared to the other hierarchical models (Table 8.4); the symmetric fixed accuracy threshold model always converged. The complete HSROC model often poorly estimated the correlation between the logit transformed sensitivities and specificities as +1 or -1 (Table 8.3); estimation as -1 occurred much more frequently than +1. The correlation was more likely to be estimated more sensibly between -1 and +1 when there was heterogeneity in accuracy and threshold, greater prevalence of disease and more studies in a meta-analysis.

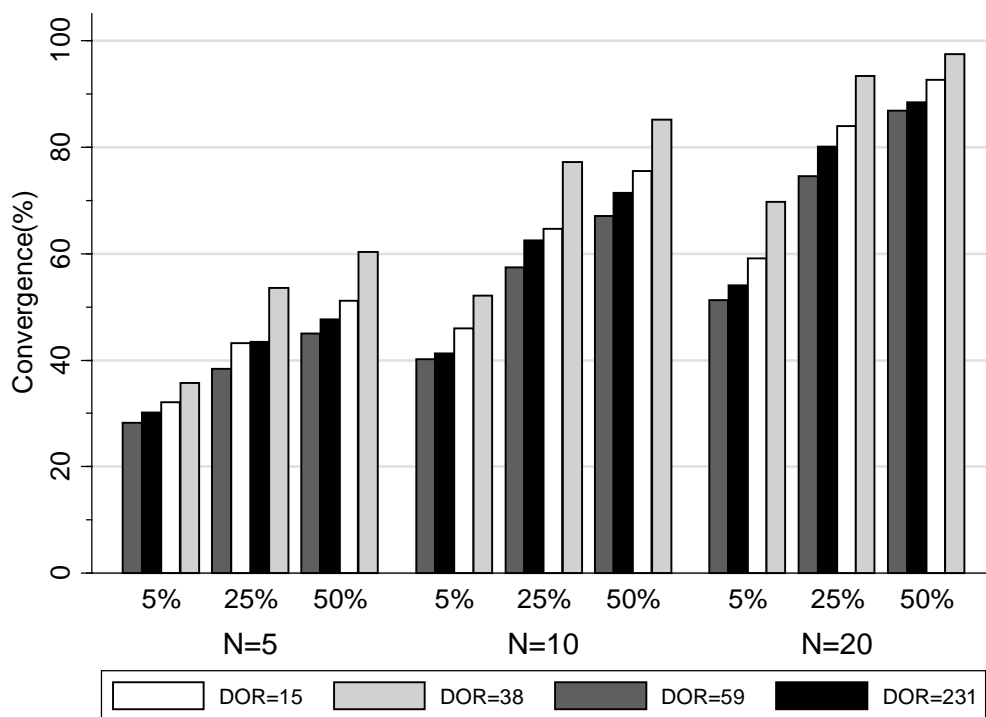


Figure 8.4| Proportion of meta-analyses that successfully converged for the complete HSROC model in 36 different scenarios with heterogeneity in accuracy and threshold
DOR=diagnostic odds ratio; N=number of studies.

Bars are grouped according to disease prevalence of 5%, 25% and 50% for each meta-analysis size (5, 10 and 20 studies). DORs of 15 and 59 correspond to scenarios with asymmetric SROC curves.

Table 8.3| Convergence and estimability of the complete HSROC model applied to 10 000 datasets in 36 different scenarios

DOR	N	Prevalence (%)	No heterogeneity in accuracy and threshold			Heterogeneity in accuracy and threshold						
			Meta-analyses with a zero cell* (%)	Successful model fit (positive definite) (%)	% $\hat{\rho}_{AB} = -1$ or $+1$	Meta-analyses with a zero cell* (%)	Successful model fit (positive definite) (%)	% $\hat{\rho}_{AB} = -1$ or $+1$				
38	5	5	48	18	14	2.6	1.8	50	36	21	0.6	14
38	5	25	50	18	15	1.7	1.6	51	54	31	0.2	22
38	5	50	52	18	15	1.4	2.0	53	60	34	0.2	26
38	10	5	60	25	17	3.7	4.4	60	52	21	0.2	31
38	10	25	65	24	18	2.4	3.8	67	77	24	0.0	54
38	10	50	72	25	18	2.6	4.1	73	85	24	0.0	61
38	20	5	75	32	20	5.8	6.4	77	70	20	0.0	50
38	20	25	77	30	20	3.7	6.3	78	93	12	0.0	82
38	20	50	82	28	20	3.1	5.8	84	97	10	0.0	88
231	5	5	96	18	11	5.7	1.4	97	30	18	2.4	10
231	5	25	97	21	14	5.0	1.6	98	43	29	2.1	13
231	5	50	99	21	15	4.7	1.9	99	48	31	2.3	14
231	10	5	99	23	13	7.7	3.0	99	41	21	1.3	19
231	10	25	99	29	17	7.6	4.1	100	62	29	0.9	33
231	10	50	100	29	18	6.7	4.1	100	71	32	0.9	39
231	20	5	100	29	15	9.7	4.8	100	54	20	0.3	34
231	20	25	100	35	20	9.2	6.2	100	80	23	0.2	57
231	20	50	100	35	20	9.2	6.5	100	88	22	0.1	66

$\hat{\rho}_{AB}$ = estimated between-study correlation; DOR = diagnostic odds ratio; N = number of studies.

*The percentage of meta-analyses out of 10 000 where at least one study included a zero cell.

All results are presented as percentages and are based on 10 000 meta-analysis datasets.

(Adapted from Takwoingi et al 2015²⁷⁸)

Table 8.4| Performance of all meta-analytic models in estimating the log DOR for scenarios with a DOR of 231

Studies	Heterogeneity*	Meta-analytic model†	5% prevalence				25% prevalence				50% prevalence			
			N	Bias (%)‡	MSE	Coverage (%)	N	Bias (%)	MSE	Coverage (%)	N	Bias (%)	MSE	Coverage (%)
5	No	Complete HSROC	1767	13.5	6.24	98.9	2086	4.20	0.41	98.8	2114	3.27	0.23	97.7
		Symmetric HSROC	1276	4.32	4.26	95.5	2753	2.28	0.26	97.1	4179	2.58	0.20	97.4
		FA	1932	36.2	31.9	98.7	2228	4.86	0.59	98.5	2232	3.64	0.26	97.4
		FT	1792	37.8	32.6	98.9	2174	4.77	0.63	99.3	2288	3.26	0.27	98.9
		FAT	9798	30.1	28.0	97.0	9556	1.82	0.37	95.9	4722	1.24	0.17	94.5
	Yes	SFAT	10000	37.0	40.7	88.5	10000	1.69	0.42	95.8	10000	1.19	0.15	95.5
		UREM	3173	11.2	5.67	98.5	2869	3.69	0.40	97.5	2883	2.71	0.22	97.4
		Complete HSROC	3020	40.2	41.6	97.6	4339	2.43	0.57	95.8	4772	1.26	0.30	94.4
		Symmetric HSROC	3442	20.6	20.6	94.2	6331	0.76	0.33	93.6	7594	0.58	0.25	93.3
		FA	5490	40.6	42.6	97.1	6222	2.97	0.82	95.4	6331	1.57	0.31	93.9
10	No	FT	2976	51.9	51.4	98.2	2266	3.66	1.12	97.7	2171	2.26	0.36	97.9
		FAT	9691	22.3	22.9	91.8	9288	-2.16	0.66	85.1	4833	-3.18	0.30	79.4
		SFAT	10000	28.3	33.0	84.2	10000	-2.38	0.77	84.6	10000	-3.23	0.30	80.7
		UREM	5833	11.8	6.53	97.8	6311	2.01	0.43	96.8	6573	1.10	0.30	96.4
		Complete HSROC	2325	7.70	0.94	99.0	2903	2.42	0.15	97.2	2862	1.90	0.10	97.0
	Yes	Symmetric HSROC	1924	1.70	0.37	97.3	3581	1.60	0.11	96.4	5383	1.50	0.09	96.9
		FA	2175	11.0	4.52	98.7	2569	2.58	0.15	97.3	2719	1.96	0.10	96.4
		FT	2177	10.6	4.55	98.9	2670	2.44	0.15	98.3	2666	1.72	0.10	98.2
		FAT	9881	5.55	3.45	96.9	9594	0.71	0.09	95.5	4596	0.55	0.07	95.5
		SFAT	10000	6.40	4.96	95.9	10000	0.62	0.09	95.4	10000	0.51	0.07	95.4
20	No	UREM	5612	6.59	1.02	98.4	5417	1.71	0.12	97.0	5311	1.30	0.08	96.7
		Complete HSROC	4129	9.55	4.58	98.1	6248	0.93	0.16	95.2	7136	0.61	0.13	94.5
		Symmetric HSROC	5895	2.72	1.89	95.8	8488	0.20	0.14	93.7	9387	0.35	0.12	93.4
		FA	6772	9.14	5.04	96.9	7776	0.45	0.15	93.6	8088	0.30	0.12	92.1
		FT	2759	10.8	6.67	98.0	1840	0.60	0.17	96.9	1579	0.27	0.13	97.5
	Yes	FAT	9775	-0.31	2.42	88.0	9301	-4.19	0.23	77.3	4765	-4.62	0.24	71.4
		SFAT	10000	0.15	3.34	87.0	10000	-4.38	0.24	76.7	10000	-4.45	0.23	72.2
		UREM	8302	5.30	1.42	97.4	8942	0.46	0.16	96.6	9237	0.36	0.12	97.1
		Complete HSROC	2915	4.87	0.40	99.4	3513	1.62	0.07	97.0	3513	1.28	0.05	95.9
		Symmetric HSROC	2654	1.16	0.16	96.1	4359	1.01	0.05	96.1	6149	1.04	0.04	96.3

Table 8.4 continued...

Studies	Heterogeneity*	Meta-analytic model [†]	5% prevalence				25% prevalence				50% prevalence			
			N	Bias (%) [‡]	MSE	Coverage (%)	N	Bias (%)	MSE	Coverage (%)	N	Bias (%)	MSE	Coverage (%)
		FA	2425	4.77	0.44	98.5	2969	1.60	0.06	96.5	3076	1.19	0.04	95.7
		FT	2439	4.47	0.39	99.1	2888	1.53	0.07	97.6	2996	1.23	0.05	97.1
		FAT	9917	1.25	0.17	95.7	9615	0.37	0.04	95.2	4528	0.34	0.03	94.7
		SFAT	10000	1.21	0.17	95.7	10000	0.34	0.04	95.1	10000	0.29	0.03	94.9
		UREM	8094	3.64	0.33	97.1	7930	1.03	0.05	96.5	7963	0.83	0.04	96.1
20	Yes	Complete HSROC	5406	3.36	0.39	97.2	8011	0.43	0.07	95.5	8843	0.14	0.06	94.3
		Symmetric HSROC	8040	0.51	0.17	95.0	9679	0.14	0.07	94.1	9905	0.06	0.06	93.7
		FA	7767	1.38	0.33	95.3	8930	-0.34	0.07	92.4	9179	-0.39	0.06	90.7
		FT	2054	0.95	0.24	95.9	992	-0.68	0.07	95.8	781	-0.95	0.06	95.6
		FAT	9758	-4.37	0.30	81.7	9293	-4.88	0.20	66.0	4559	-5.19	0.21	58.3
		SFAT	10000	-4.46	0.30	81.4	10000	-5.01	0.21	64.4	10000	-5.12	0.21	57.6
		UREM	9455	1.64	0.29	96.4	9869	0.11	0.07	97.4	9951	0.04	0.06	97.2

DOR = diagnostic odds ratio; FA = fixed accuracy HSROC model; FAT = fixed accuracy and threshold HSROC model; FT = fixed threshold HSROC model; MSE = mean square error; N = number of meta-analyses out of 10 000 where hierarchical models successfully converged; SFAT = symmetric fixed accuracy and threshold HSROC model; UREM = univariate random effects logistic regression model.

* Heterogeneity in accuracy and threshold.

† All simulated data models presented in the table were symmetric and so should agree with symmetric HSROC models when there is heterogeneity and with SFAT models when there is no heterogeneity. Fixed effect models (FA, FT and FAT) should agree partly with simulated data models when there is no heterogeneity. Complete HSROC models and univariate random effects models should not agree with any of the simulated data models.

‡ Bias is presented as a percentage of the true value of the log diagnostic odds ratio.

(Adapted from Takwoingi et al 2015²⁷⁸)

8.5.2 Bias

In the base scenario for a DOR of 231, the symmetric HSROC model gave the least percentage bias for the DOR (4.32%); bias was highest for the fixed threshold (37.8%) and fixed accuracy (36.2%) models (Table 8.4). These rankings were consistent as the number of studies increased. As prevalence increased, the two models (symmetric and asymmetric) with both accuracy and threshold as fixed effect became the least biased while the fixed accuracy model remained the most biased. When heterogeneity was introduced, each of the seven models produced the largest bias for the DOR at the lowest prevalence, though the univariate random effects model gave the least biased DOR (Figure 8.5). For all models, bias decreased as prevalence and the number of studies increased. However, the decrease in bias resulted in a change from overestimation to underestimation for the two fixed effect models. For bias in the estimates of sensitivity, the observed results were similar to those of the DOR, but the relationship with prevalence was reversed for bias in the estimates of specificity (appendices E.1 and E.2). Bias in specificity was very small compared to that of the DOR or sensitivity. For the three measures, in scenarios with heterogeneity and asymmetry in the SROC curve, bias was lower than in the corresponding symmetric model.

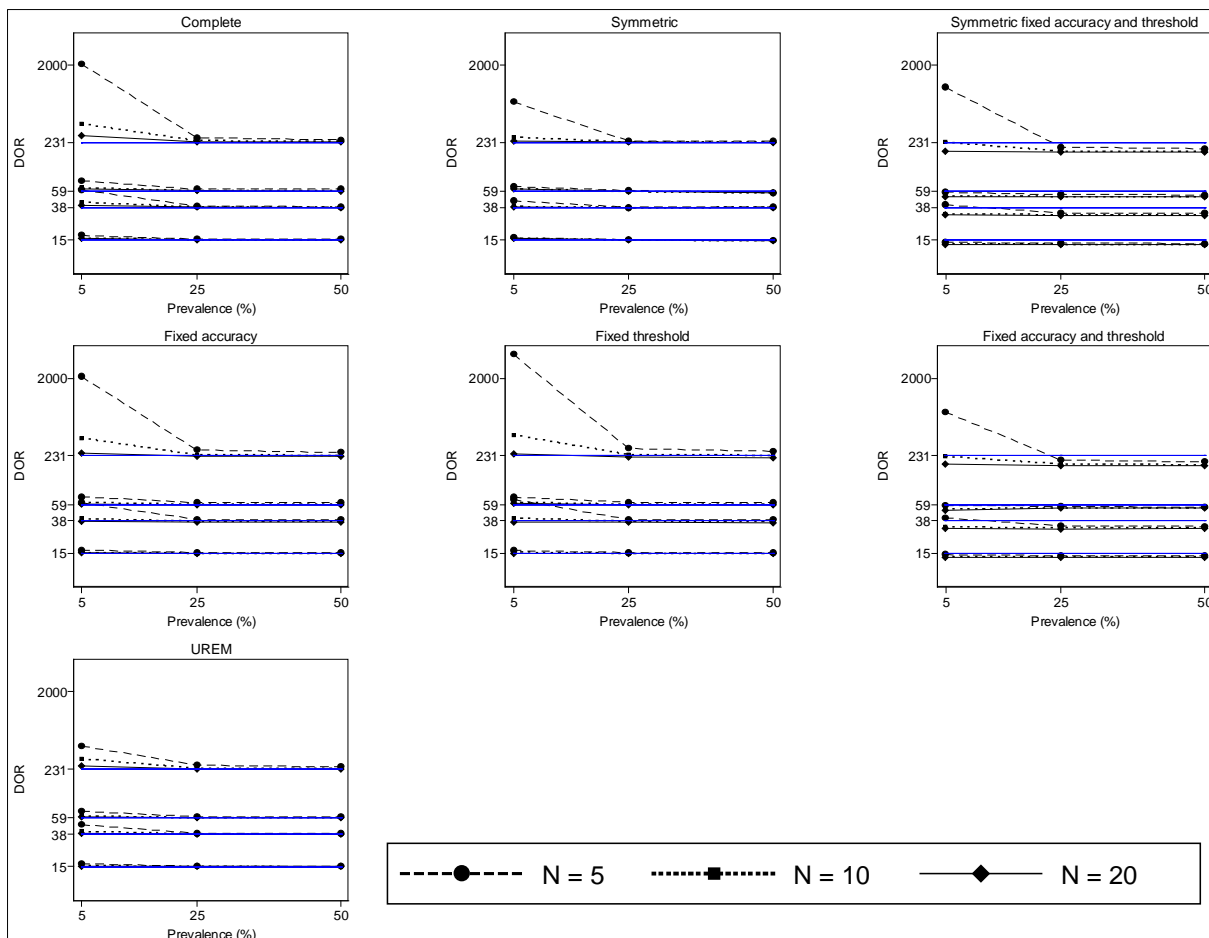


Figure 8.5| Bias in the estimated diagnostic odds ratio for 36 scenarios with heterogeneity in accuracy and threshold

DOR = diagnostic odds ratio; N = number of studies; UREM = univariate random effects logistic regression model.

The graphs are plotted on the log scale but show the corresponding values for the DOR. The blue horizontal lines correspond to the true DOR in each simulation scenario.

8.5.3 Model accuracy

A MSE of zero indicates that the model estimated the parameter of interest with perfect accuracy, i.e., no bias and no variability in the estimation. The MSE of the DOR was highest for the symmetric fixed accuracy threshold model (40.7) but lowest for the symmetric HSROC model (4.26) in the base scenario (Table 8.4). At higher prevalence, the two fixed effect models had the lowest MSE. For all models, the MSE of the DOR decreased as the number of studies and prevalence increased. When heterogeneity was introduced, the univariate random effects model had the lowest MSE at 5% prevalence but the symmetric

HSROC model had slightly lower MSE than the univariate random effects model at higher values of prevalence (see also Figure 8.6). As the number of studies and prevalence increased, the MSE for all models decreased and became almost identical except for those of the two fixed effect models. Results for sensitivity were similar to those for the DOR. The MSE for specificity was generally very low and increased slightly with increasing prevalence. For the asymmetric SROC curve scenarios, the findings for the three measures were similar to those of the corresponding symmetric scenarios.

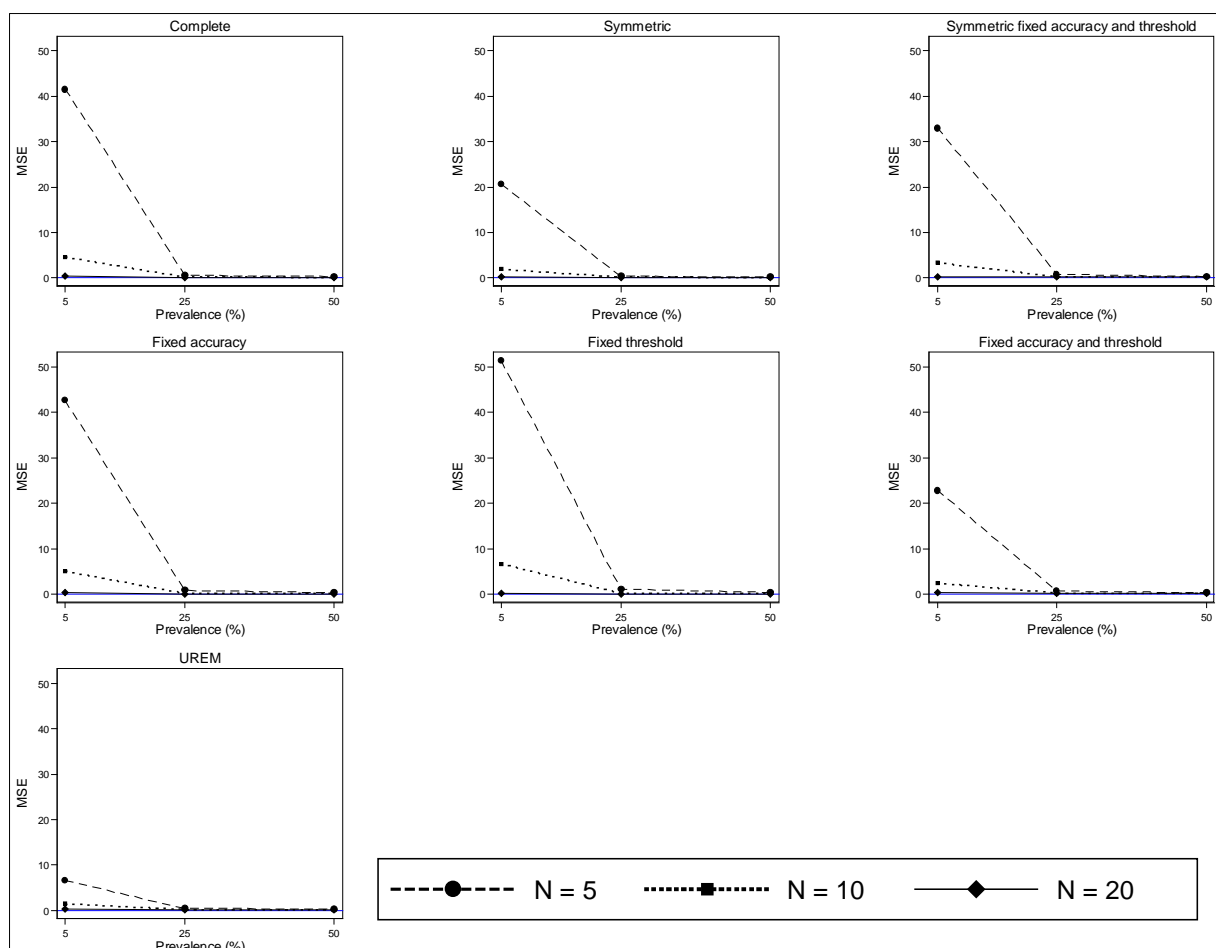


Figure 8.6| Mean square error for the estimated log diagnostic odds ratio for nine different scenarios with heterogeneity in accuracy and threshold and diagnostic odds ratio of 231

MSE = mean square error; N = number of studies; UREM = univariate random effects logistic regression model.

8.5.4 Coverage

For a DOR of 231, the symmetric HSROC models gave the best coverage of the 95% confidence intervals for estimation of the DOR (95.5%) in the base scenario (Table 8.4). With the exception of the symmetric fixed accuracy threshold model, all models were conservative as shown by coverage greater than 95%. The coverage of 88% for the symmetric fixed accuracy threshold model implied over-confidence in the estimates but coverage increased as prevalence or the number of studies increased. In contrast, introduction of heterogeneity led to very poor coverage for the two fixed effect models with coverage becoming lower as prevalence increased (Figure 8.7). The univariate random effects model and symmetric HSROC model often showed good coverage, although the latter tended to show under-coverage as prevalence increased. For sensitivity, the results were comparable to those of the DOR. Across all models, coverage was low for specificity when there was heterogeneity unlike scenarios without heterogeneity. The asymmetric SROC curve scenarios produced similar results to the symmetric SROC curve scenarios.

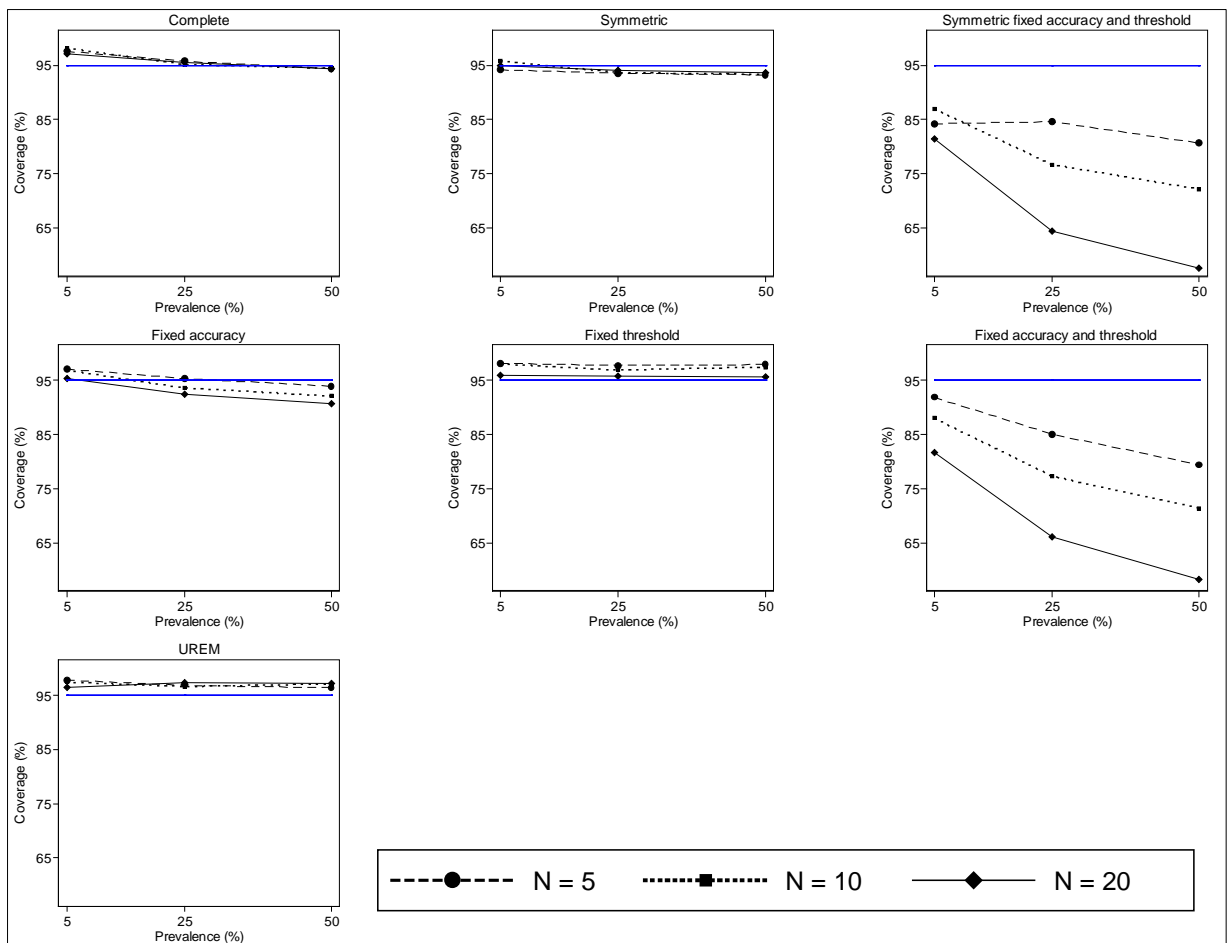


Figure 8.7| Coverage of 95% confidence intervals for the log diagnostic odds ratio for nine different scenarios with heterogeneity in accuracy and threshold and a diagnostic odds ratio of 231

N = number of studies; UREM = univariate random effects logistic regression model. The blue horizontal line corresponds to the 95% confidence interval coverage, i.e., the percentage of simulations for which the 95% confidence interval of the estimated summary log diagnostic odds ratio included the underlying value.

8.5.5 Summary of simulation results and application to motivating examples

The following key points were observed:

- Hierarchical models are more likely to converge if there is heterogeneity in accuracy and threshold.
- Convergence is also affected by number of studies, prevalence and magnitude of diagnostic accuracy.

- Correlation between sensitivity and specificity across studies is often poorly estimated as +1 or -1, especially when heterogeneity is zero or small, the number of studies is small, and the prevalence was low.
- In the absence of heterogeneity, the two fixed effect models were the least biased with low MSE and good coverage properties for studies with moderate to high prevalence. The symmetric fixed accuracy threshold model may be of greater utility because it always converged. The symmetric HSROC model performed better than both fixed effect models when prevalence was low and there were few studies, *but* this finding was based on a convergence rate as low as 13%.
- When heterogeneity was present, the univariate random effects logistic regression model and the symmetric HSROC model were often the least biased with low MSE and good coverage (however, there is a *risk of selection bias* in these results for scenarios with lower prevalence with smaller numbers of studies where as few as 34% of simulations converged).

In the simulation, the fixed threshold model often gave biased and imprecise results.

However, for the appendicitis example, the fixed threshold model gave results similar to the complete HSROC model. The results can be explained by the fact that the estimation of σ_{θ}^2 was truncated at zero in the complete model and so removing σ_{θ}^2 from the HSROC model was appropriate in this example unlike in the simulation scenarios. The results in Table 8.1 indicate that while the univariate random effects model and symmetric HSROC model appear to be generally applicable when there is heterogeneity, other models like the fixed threshold or fixed accuracy can be considered if it is apparent the variance parameter for threshold or accuracy cannot be estimated.

For the scaphoid fractures example, the results of the simulation indicate that using a univariate fixed effect model (including the equivalent fixed accuracy threshold and symmetric fixed accuracy threshold models) was valid because there was no heterogeneity in the specificities (all six studies reported 100% specificity) and very limited heterogeneity in the sensitivities (five of the studies reported 100% sensitivity). Even for the fixed effect models, computation of the positive likelihood ratio and DOR were problematic because of the perfect specificity.

8.6 Application to test comparisons

Although not discussed in Chapter 2, analyses of single tests with small number of studies were performed in some of the reviews. For example, the meta-analysis of Type 5 tests in the *P. falciparum* malaria review⁹⁵ and the meta-analysis of the MDQ in primary care settings in the bipolar disorder review.⁵⁶ Meta-analyses with small numbers of studies were also encountered in the evaluations reported in Chapter 7. Of the 114 meta-analyses of a single test conducted as preliminary analyses of the 57 test comparisons, 60 (53%) had less than 10 studies. The results of this simulation study are directly applicable to all of these analyses.

For comparative meta-analyses, the simulation findings can also be applied especially for indirect comparisons with few or no paired studies. For instance, in section 2.3.5.1, data were sparse and there were no comparative studies for the comparison of ERCP and IOC for detection of common bile duct stones. Alternative models similar to models considered in the simulation were applied to each test separately before a simplified model was selected for the test comparison. The comparative meta-analysis was performed using a bivariate model with an exchangeable covariance structure. This model is equivalent to a symmetric HSROC

model since the shape of the SROC curve depends on the ratio of the variances of the random effects for logit sensitivity and logit specificity ($\beta = \log(\sigma_B/\sigma_A)$) as shown by Harbord et al.⁴⁴ Another situation where a simpler model is appropriate is exemplified in section 2.3.2.2. In this example, summary points were estimated at a 5% FPR in a univariate logistic regression analysis comparing nine tests in the Down's syndrome screening review. Clearly, a bivariate approach and estimation of the between-study correlation is futile in such situations. The findings of the empirical comparison of bivariate and univariate models for test comparisons in Chapter 7 also support the use of simpler models.

For tests based on similar biologic mechanisms, it is logical to expect test results to be conditionally dependent. Thus, the simulated datasets in this study may be unrepresentative and the impact of this dependence on simulation results may not be negligible. Conditional dependence may occur for test results in the non-diseased group, the diseased group or both. A simulation study based only on paired data with different concordance probabilities ($P(Y_A = 1, Y_B = 1|D = 1)$ and $P(Y_A = 0, Y_B = 0|D = 0)$; $Y=1$ if test result is positive and $Y = 0$ if test result is negative) can be used to investigate the impact of correlated data on the performance of the models. Also, this will enable evaluation of the joint meta-analytic model proposed by Trikalinos et al²⁰⁹ (see section 6.3.4.5). For simplicity the possibility of asymmetry in test dependence between the diseased and non-diseased groups may be ignored.

8.7 Discussion

8.7.1 Principal findings and recommendations

In this study meta-analyses were simulated under a number of scenarios and hierarchical models for meta-analysis of diagnostic accuracy studies were evaluated. The findings indicate

that simplifying hierarchical models is valid when there are few studies or sparse data.

Recommendations for selecting alternative models when bivariate or HSROC models fail to converge or converge but give unreliable estimates, are outlined in the Box 8.1. If estimation of an average operating point is of interest instead of a SROC curve, a univariate logistic regression approach is recommended with or without random effects depending on the extent to which sensitivity and/or specificity vary between studies. These methods are an appropriate alternative for obtaining independent summaries of sensitivity and specificity with confidence intervals. However, joint inferences cannot be made about sensitivity and specificity through confidence and prediction regions around the average operating point. These regions account for correlation between sensitivity and specificity, and are useful for illustrating uncertainty around the average operating point and the extent of heterogeneity. If interest lies in the estimation of a SROC curve, the symmetric HSROC model or its fixed effect equivalent should be considered instead. In extreme situations with no heterogeneity and sparse data, such as the scaphoid fractures example, even the simplest models may fail to produce usable summary estimates.

Box 8.1| Recommendations for selecting alternative models when bivariate or HSROC models fail*

Plot the data

Visual inspection of forest plots and SROC plots may help to identify whether heterogeneity exists. For example, one may observe complete or near complete lack of variability between estimates of sensitivity and/or specificity, indicating no heterogeneity in one or both parameters (sensitivity and/or specificity equal to 100%), or conversely wide variability in observed estimates (e.g. non-overlapping confidence intervals) indicating large heterogeneity.

Analyses

Select a simpler hierarchical fixed effect or random effects model based on inference of interest (summary points or SROC curve), observation from the data plot, and previous output from the failed bivariate or HSROC model

Note: when prevalence is very low and the number of studies is very small, there is potential for bias in estimates of test performance and the results of the meta-analysis should be interpreted with caution.

<i>Heterogeneity</i>	<i>Focus of inference</i>	
	<i>Summary point (summary sensitivity and specificity)</i>	<i>SROC curve</i>
Variability in sensitivity and/or specificity between studies observed on the plot	Univariate random effects logistic regression models	Symmetric HSROC model
Minimal or no variability in sensitivity and/or specificity between studies observed on the plot	Univariate fixed effect logistic regression models†	Symmetric fixed accuracy and threshold model

A symmetric SROC curve can be described using the diagnostic odds ratio (exponent of the value of the accuracy parameter).

Section 8.4.3 contains suggestions for facilitating convergence of hierarchical models.

*Bivariate or HSROC models either failed to converge or converged (i.e. met the convergence criterion) but gave unreliable estimates (e.g. with no standard errors, or dependent on starting values).

† The symmetric fixed accuracy threshold model is equivalent to simultaneously fitting two univariate fixed effect logistic regression models for sensitivity and specificity.

(Adapted from Takwoingi et al 2015²⁷⁸)

Given the poor performance of simpler models like the fixed accuracy and fixed threshold models in the simulation, meta-analysts are advised to carefully explore their data and visually inspect forest plots and SROC plots before undertaking meta-analyses. Such preliminary analyses will provide an indication of the degree of heterogeneity and the pattern of scatter of the study points in ROC space. These analyses and the output from unstable or failed models should inform the approach for simplifying hierarchical models as shown by the appendicitis example. Although more complex and seldom used in practice, a Bayesian approach is an alternative to the maximum likelihood approach. In an empirical evaluation, both approaches were found to be similar although Bayesian methods suggested greater uncertainty (wide credible intervals) around the point estimates.²⁷

8.7.2 Comparison with previous research

A normal distribution is typically assumed for the random effects in hierarchical meta-analytic models; violation of this assumption may contribute to non-convergence. Heavy tailed distributions such as t or Cauchy distributions may be used instead of a normal distribution,^{42,194} but random effects are restricted to be normally distributed in SAS NLMIXED and Stata. A Bayesian approach allows alternative distributions though a normal distribution is often assumed in practice.⁵⁷ As the models are often fitted using a maximum likelihood approach, the objective of this study was to offer solutions within the hierarchical framework recommended for meta-analysis, using one of the software packages that have made meta-analysis of test accuracy studies more accessible to meta-analysts. A composite likelihood approach (implemented in R using the glmmML package) that offers some robustness to model misspecifications was recently proposed.²⁸⁰ Results from the simulation study where the composite likelihood method and the bivariate generalized mixed model were

applied to data generated from a bivariate t distribution suggested the methods were insensitive to the heavy tailed distribution under the logit link function. Only the logit link was used in all the models presented in this thesis.

The simulations and application to motivating examples support and extend empirical evidence suggesting that univariate methods generate summary results similar to those derived using full hierarchical methods.^{35,36,225} The findings also agree with a recent simulation study evaluating the performance of the bivariate model.²⁸¹ However, the present study is more comprehensive including application to real motivating examples, investigation of a broad array of possible models, suggestions for improving model convergence and guidance on how to select an appropriate model. Furthermore, no limit is prescribed on the number of studies required to fit a hierarchical model, rather the merit of applying a particular model should be carefully assessed as the motivating examples have shown.

8.7.3 Strengths and limitations

The study has some limitations. First, because of the number of scenarios investigated, it was not possible to fully explore the effect of heterogeneity or varying the threshold. Factors considered vital were addressed, and the sample size of studies in a meta-analysis was varied to reflect reality. According to Begg,²⁸² the statistical properties of hierarchical models are likely to be most vulnerable when the number of studies is small, and also when sample sizes are highly variable.

Second, analyses of the simulated datasets were conducted only in SAS and convergence rates may differ between software packages because of differences in obtaining starting values and

model fitting options. Nonetheless, SAS is the software most often used to fit HSROC models in frequentist analyses and several options for improving convergence were explored.

Third, when comparing models, analyses were not limited to datasets that converged across all models and there is potential for selection bias. Estimates of model performance (bias/MSE/coverage) can only be interpreted with confidence when convergence has been achieved for a high proportion of simulations. In the tables the number that converged are reported along with measures of performance so that interpretation can be made with knowledge of convergence rates. Non-convergence occurred more frequently in challenging datasets where poor model performance (bias, MSE and coverage) can be expected. Therefore, more complex methods with poor convergence rates may be biased or give imprecise estimates. The performance of simpler models with better convergence rates should also be affected but if the models give unbiased and precise estimates, then simpler models are robust and applicable in such situations.

Fourth, the simulation was limited to meta-analysis of a single test though the results may be applicable to test comparisons in certain contexts as discussed in section 8.6. Future work will seek to extend the simulation as suggested in that section.

8.7.4 Conclusions

In summary, random effects logistic models should be the default approach for test accuracy meta-analyses. Univariate random effects logistic regression models for sensitivity and specificity are recommended if a bivariate model fails, or a symmetric HSROC model if estimation of a SROC curve is required and the HSROC model fails. If homogeneity can be

assumed, the two models can be further simplified to their fixed effect equivalent. However, when prevalence is very low and the number of studies is very small, the results of any meta-analysis should be interpreted with caution.

9 THESIS DISCUSSION AND CONCLUSIONS

Systematic reviews and meta-analyses of diagnostic accuracy studies can provide answers to important clinical questions but producing the reviews is widely acknowledged to be more complex than reviews of interventions. With the increasing publication of DTA reviews and their potential impact on clinical and policy decision making, it is vital to ensure appropriate use of methods to avoid misleading results and conclusions. Study designs and methods recommended for meta-analysis of diagnostic accuracy are not as well understood as their equivalents for evaluations of interventions. Therefore, methods for DTA reviews addressing two domains— study design (type of included studies) and analysis (strategies for test comparisons in addition to methods for comparative meta-analysis) —were the focus of this thesis.

The thesis addressed four main questions centred on the overarching aim of assessing the reliability and transparency of evidence derived from systematic reviews and meta-analyses of comparative accuracy, including availability and validity of meta-analytic methods. The evidence base for test accuracy meta-analysis has been significantly expanded, especially with respect to test comparisons. Research presented in this thesis has provided evidence for the first time about the methodological and reporting characteristics of comparative reviews; availability of comparative studies; differences between direct and indirect test comparisons; and meta-analytic methods and how they should be used to obtain reliable answers. Here, in this final chapter, an overview of the thesis is provided in relation to the key findings and implications for practice, and further research recommendations.

9.1 Overview of thesis

The four main aims addressed in this thesis were to (1) investigate the methods and reporting of comparative reviews (Chapter 4); (2) assess availability of primary comparative accuracy studies and their importance in test comparisons by comparing meta-analyses of comparative (direct comparisons) and non-comparative (indirect comparisons) studies (Chapter 5); (3) identify and evaluate comparative meta-analysis methods (Chapters 6 and 7); and (4) determine how meta-analyses should be performed when there are few studies or sparse data (Chapter 8).

The first two chapters of the thesis set the scene by providing background and examples of different clinical scenarios and important methodological issues. Specifically, Chapter 1 explained key concepts; described various study designs for comparing test accuracy; introduced meta-analytic methods both for the analysis of a single test and for test comparisons; and defined the scope of the thesis.

In Chapter 2, real life case studies were used to highlight key methodological issues across multiple clinical scenarios. The seven systematic reviews highlighted the complexity of test comparisons. A common issue was the scarcity of well-designed comparative studies and so meta-analyses were based mainly on indirect comparisons. Hierarchical models with different modelling assumptions of varying complexity were illustrated, and it was clear that availability of data is likely to be a driver for choosing an analytic approach. Simplifying hierarchical models to avoid overfitting and estimation problems seemed a realistic and reasonable prospect, but there was uncertainty about validity due to lack of evidence-based guidance. Therefore, the thesis sought to understand whether the issues were common and to

formulate evidence-based solutions to support reviewers and meta-analysts faced with similar issues in the future.

In Chapter 3, the data sources and search strategies used to obtain materials for subsequent chapters of the thesis were described. Identification of systematic reviews and comparative meta-analytic methods were an essential component of the thesis that enabled two methodological reviews (Chapter 4 and Chapter 6) and two empirical studies (Chapter 5 and Chapter 7). Altogether, 269 of the 286 systematic reviews that met the main inclusion criteria were included across the three cohorts for Chapters 4, 5 and 7. The large number of reviews encompassing a diverse array of test types and clinical topics facilitated detailed examination of the issues addressed in the methodological review of reviews and the empirical studies. The empirical evaluation of meta-analytic methods in Chapter 7 then provided a foundation and context for the simulation study in Chapter 8. The research findings of Chapters 4 to 8 and their implications are summarised below.

9.1.1 Are methods and reporting of comparative reviews adequate?

To understand the current landscape for comparative accuracy reviews, a descriptive survey of systematic reviews of multiple tests published over a five-year period between 2008 and 2012 was undertaken. The methodological review in Chapter 4 highlighted the common use of less robust methods for test comparisons and poor reporting, but examples of good practice were also noted. To improve quality and transparency, and to increase confidence in decision making informed by these reviews, the development of a reporting guideline for test accuracy reviews was recommended as an urgent priority. Following completion of this work, the author became aware of an initiative to develop a PRISMA extension for DTA reviews

(PRISMA-DTA) and is now a member of the advisory group.¹⁴⁴ In the interim, the checklist developed and presented in Chapter 4, though focused on items of relevance to comparative accuracy reviews, can aid better reporting. The checklist includes a rationale for each item and the reasoning behind the recommendations. The findings of the survey and the checklist will also facilitate the development of the component of PRISMA-DTA that deals with comparative reviews.

9.1.2 Are comparative accuracy studies essential for test comparisons?

In chapter 5 it was shown that direct comparisons are seldom feasible in a meta-analysis setting due to the paucity of comparative studies. Yet evidence derived from indirect comparisons often differed from that derived from direct comparisons. Existence of bias in indirect comparisons cannot be proven in a meta-epidemiological study, but there are theoretical reasons why the results from the comparative studies should give reliable estimates of relative test performance than from indirect comparisons. Therefore, robustly designed studies in which all patients receive all tests or are randomly assigned to receive one or other of the tests should be more routinely undertaken and are preferred for evidence to guide test selection.

There was no statistical evidence of a direction in the differences observed in the main analysis that included multiple estimates from some reviews. In contrast, the sensitivity analysis that included one estimate per review gave a larger and significant difference, favouring the newer test relative to the older test or current practice. Given the difference in conclusions between the two analyses, further investigation with a larger cohort of reviews is recommended in a future update of this empirical study.

9.1.3 What methods are available for comparative meta-analyses?

Thirteen comparative meta-analysis methods, varying in complexity and methodological rigour, were identified following a thorough literature search and contact with experts. The methods were reviewed in Chapter 6 to highlight their advantages and limitations. Of all the methods, hierarchical models (both frequentist and Bayesian methods) are the most statistically rigorous. Methods are still evolving and so bivariate and HSROC meta-regression models remain the most sophisticated and theoretically sound approach in use.

9.1.4 Do comparative meta-analysis methods give the same results?

In Chapter 7, assumptions commonly made in hierarchical meta-regression models were investigated using an empirical cohort of 57 pairwise meta-analyses. In addition, meta-analytic methods identified in Chapter 6 that were deemed to be methodologically robust and/or commonly used were also empirically evaluated. The analyses were limited to classical methods. In hierarchical models, numerical and qualitative differences occurred with different modelling assumptions. Although assumptions of equal variances have the advantage of simplifying estimation of hierarchical models, it was shown that the assumption is often unjustified. The effect of assumptions about the shape of SROC curves in HSROC meta-regression models was less dramatic compared to assumptions about the covariance structure in bivariate meta-regression models. Meta-analysts are encouraged to carefully assess modelling assumptions when fitting these models. Univariate random effects logistic regression models were shown to give similar results to bivariate models for meta-analysis of a single test or test comparisons, implying that univariate models may be suitable alternatives when bivariate models are unstable or unnecessary (e.g. when all or most studies report the

same sensitivity or specificity). There were obvious differences between estimates from Moses SROC and HSROC meta-regression models. Given the known methodological limitations of the Moses model,^{39,40} it should not be used for test comparisons. The recommendation also applies to investigations of heterogeneity performed using this same meta-regression approach.

9.1.5 How should meta-analyses be undertaken with few studies or sparse data?

Meta-analyses of a single test were simulated under a number of realistic scenarios that reflect practice. The findings indicated that simplifying hierarchical models is valid when there are few studies or sparse data. Recommendations for selecting alternative models when bivariate or HSROC models fail were developed. Generally, univariate random effects logistic regression models were recommended if a bivariate model fails, or a symmetric HSROC model if a HSROC model fails. The findings are likely to be applicable to comparative meta-analyses, especially for indirect comparisons with few or no paired studies. However, further research is needed to confirm the validity of such simplifications in comparative meta-analyses.

9.2 Validity of test comparisons

Given the scarcity of comparative studies, differences between direct and indirect comparisons, use of inadequate methods and poor reporting of published reviews, as well as the challenges of fitting complex hierarchical models, some may argue there is little value in comparative accuracy reviews. On the contrary, examples of good practice were demonstrated and with proper guidance and tools to aid review authors, comparative reviews can be

improved and made fit for purpose. The reporting checklist and recommendations about meta-analytic methods that were developed in the thesis are a useful starting point.

The magnitude of the differences observed between direct and indirect comparisons suggests the need for review authors to carefully consider the design of studies included in comparative test accuracy reviews and to reflect study design weaknesses in the interpretation of their findings. This will enable users of reviews, especially those who rely on such evidence to guide clinical practice and formulate policy, to judge the reliability of the evidence.

Inappropriate modelling assumptions used in meta-analysis also impact on estimates of test performance, perhaps to the same or greater degree than the issue of study design. Since direct comparisons often have limited data, there is a gain by borrowing information from indirect comparisons to allow more precise summary estimates of test performance, in addition to reliable estimation of correlations in the bivariate model or shape of SROC curves in the HSROC model, and variances in both models. However, as indirect comparisons may be prone to bias, any gain in information that arises by incorporating indirect information also increases the risk of bias. Thus, there is a potential trade-off between increased precision but increased bias.

With few studies and large numbers of model parameters to estimate in complex models, model fitting is a challenge. Having shown that univariate models and symmetric HSROC models can give valid answers in challenging scenarios and beyond, the question then arises—should simpler models fitted to limited data in direct comparisons be preferred to more complex data hungry models fitted to more data from indirect comparisons? Moreover, simpler models are readily fitted using binomial within-study likelihoods while very complex

models may require approximations of the within-study likelihood or assumptions about the covariance structure.²²¹ There is no definite universally applicable meta-analytic solution; each meta-analysis merits individual consideration and simpler models should be considered a special case.

For pairwise direct test comparisons, a departure from bivariate meta-analysis and confidence regions offers the potential to produce forest plots of relative sensitivities and relative specificities that include pooled estimates such as the diamond shown on traditional forest plots of interventions. This is a familiar concept for meta-analysts that is not currently used in Cochrane DTA reviews. This is appealing but some may call for caution because the evidence is based entirely on an empirical evaluation. However, the cohort of reviews was large and the results are backed up by those from multivariate meta-analysis of interventions; a simulation study may only serve to confirm the findings and identify special circumstances where only a bivariate model is applicable.

There appears to be a trade-off between study design and analysis leading to a conundrum with unanswered questions. The most obvious and simplest way forward is to recommend sensitivity analyses to check the robustness of indirect comparisons and modelling assumptions. When results from different analyses disagree, another dilemma ensues. The bottom line is to stress to investigators and funders of future test accuracy studies the value of asking comparative questions of clinical importance, and for journals to demand studies are reported in a manner that ensures they are fit for purpose and avoids research waste. After all, as quoted earlier in the thesis,

“...the hundreds of hours spent conducting a scientific study ultimately contributes only a piece of an enormous puzzle. The value of any single study is derived from how it fits with and

expands previous work, as well as from the study's intrinsic properties. Through systematic review the puzzle's intricacies may be disentangled.”³¹

9.3 Strengths and limitations

The questions in this thesis were addressed using methodological reviews of systematic reviews and statistical methods, empirical studies and simulation. Thus, a broad range of research methods appropriate to each question was used to ensure rigorous and comprehensive evaluations, and to assure confidence in the research findings. Where necessary, the strengths and limitations of each piece of work were considered within the respective chapter but more general issues are discussed below.

The review of reviews and empirical studies were based on an extensive database of systematic reviews of diagnostic tests. DARE is based on extensive searches of a wide array of databases and also includes grey literature. Further searching of other databases is unlikely to yield more good quality reviews or provide better coverage of clinical topics and test types. Thus the review of reviews (Chapter 4) and empirical research (Chapters 5 and 7) are likely to represent close to complete available data to answer these questions from the worldwide literature up to October 2012. Although only two databases were searched for the review of comparative meta-analysis methods, both databases combined constitute a comprehensive collection of methodology papers and conference abstracts relevant to healthcare. Furthermore, to augment the database searches, methodological experts and research groups known to have an interest in test accuracy meta-analysis were contacted. This proved to be a valuable strategy as relevant ongoing work or papers in press were identified. Thus, the work in this thesis represents an up-to-date synopsis of comparative meta-analysis methods.

On the contrary, the most recent systematic reviews in the empirical cohort were published in 2012. The time lag reflects the duration of this PhD. It is possible for improvements in methods and reporting to have occurred in this period. Yet this effort appears to be the only methodological review of comparative reviews. The message that better reporting is needed is clearly of immediate relevance as evidenced by the initiation of PRISMA-DTA. This should enhance ongoing efforts by members of the Cochrane Screening and Diagnostic Test Methods Group to improve the methodological quality and reporting of DTA reviews. As more Cochrane reviews become available, it is expected that these reviews will serve as exemplars and the average quality of DTA reviews will improve. As of Issue 1 of 2016, there were 57 full reviews (including four updates) and 95 protocols of DTA reviews published in the Cochrane Library.²⁸³

Data extraction of the methodological and reporting characteristics of systematic reviews was sometimes challenging due to poor reporting and inconsistent use of terminology in test research. Therefore, subjective judgement was applied in disentangling the required information. The effect on conclusions is likely to be negligible due to a second assessor conducting checks of random subsets of reviews. However, a residual effect cannot be totally ruled out.

9.4 Implications of thesis findings for scientific research and practice

A clear framework for test comparisons is lacking in both primary and secondary test accuracy research and there is inadequate coverage of comparative accuracy in textbooks. To the author's knowledge, the Cochrane Handbook for Systematic Reviews of Diagnostic Accuracy appears to be only guidance available for conducting comparative meta-analysis.

The handbook chapter on statistical analysis was published in 2010 and is in need of updating to reflect methodological developments that have occurred since its publication. If review authors are not well informed and are unaware of the methodological challenges and potential solutions, it is unsurprising to find wide variation in the conduct and reporting of DTA reviews.

The paucity of comparative studies indicates an urgent need to educate trialists, clinical investigators, funders and ethics committees about the merit of such studies for obtaining reliable evidence about the relative performance of competing diagnostic tests. In particular there is a need to train statisticians in clinical trials units, many of whom have no experience in test research. Since RCTs of the clinical effectiveness of tests are rarely available and most tests are not evaluated beyond their clinical performance, standards similar to those for RCTs and systematic reviews of RCTs should also be applied in the evaluation of diagnostic accuracy. Yet there is a preponderance of single test evaluations in the literature.

Approximately half (49%) of the test accuracy reviews identified in Chapter 3 addressed the accuracy of a single test. Furthermore, about two thirds of the primary studies identified in Chapter 5 did not address the comparative question of the reviews. Are single test evaluations in primary studies or systematic reviews of any value or does the research agenda need to be overhauled to address comparative questions? Primary and secondary research focussed on the evaluation of a single test or which do not address clinically important comparative questions, constitute research waste and should be actively discouraged. On the other hand, exceptional situations can arise where a test is the first to be introduced into practice for a particular condition or where the appropriate comparator is unclear. In such situations, evaluations of a single test are justified.

The thesis has shown that important differences can occur between direct and indirect comparisons. Consequently, clinicians and policy makers need to interpret and apply the findings of comparative meta-analyses with caution. The highest level of evidence is a meta-analysis of robust comparative studies provided the studies reflect the setting and patient spectrum of interest. If such meta-analyses are unavailable, meta-analyses that contain only (or mainly) non-comparative studies may still be valuable but should be used with caution due to the potential for confounding. As evidence accrues and comparative studies become available, the validity of such meta-analyses should be assessed.

To facilitate the assessment of the validity of reviews, they need to be well reported. For this to be a reality, appropriate reporting guidelines endorsed by journal editors are needed. Journal editors have a role to play as gatekeepers by demanding that review authors adhere to such reporting guidelines. As different meta-analytic models and modelling assumptions can affect results and conclusions of a review, peer reviewers have a responsibility to ensure rigorous methods are used and that the methods are also clearly reported. Tutorial guides should be developed by methodologists to assist meta-analysts and authors in navigating the complexity of the methods. Enhancements to existing software programs and macros are also needed to make the methods more accessible and to encourage appropriate use. It is evident that good quality reviews require collective effort by methodologist, authors, editors and peer reviewers. This can be achieved as demonstrated by the rigorous editorial process implemented for Cochrane DTA reviews. The process involves clinicians and methodologists in order to ensure the clinical relevance and methodological rigour of the reviews.

Interest in synthesis of systematic reviews (known as overviews or umbrella reviews) has grown with increasing publication of systematic reviews. While overviews have so far mainly focused on therapeutic interventions, DTA overviews are being funded by the NIHR with the intention of producing them as Cochrane DTA overviews. Yet methodological guidance for their production is lacking. A DTA overview may include DTA reviews that compare the accuracy of more than one test and/or reviews that focus on the accuracy of one test at a time. How these overviews differ from comparative reviews that include a large number of tests or from multiple test reviews is unclear. Given the complexity inherent in comparative reviews that has been shown in this thesis, overviews of DTA reviews should not be undertaken naïvely and research is needed to ensure they are a valid reflection of the body of evidence.

9.5 Future research

There is scope for future research to extend the work in this thesis and in other related areas as follows.

9.5.1 Sources of bias and variation in comparative studies

In stark contrast to test accuracy studies of a single test, there is no evidence about potential sources of bias and variation in comparative accuracy studies. This is one of the reasons that QUADAS-2, the quality assessment tool for DTA studies recommended by Cochrane, does not include criteria for assessing studies that compare multiple index tests.¹³⁵ This lack of evidence also precluded making definitive statements about risk of bias of comparative studies relative to non-comparative studies in this thesis. As highlighted in Chapter 5, comparative accuracy studies should provide the most reliable evidence on relative test performance but they may not be devoid of bias. Consequently, the term bias was used

cautiously when investigating differences between direct and indirect comparisons. To provide evidence for informing risk of bias assessments as well as the design of future primary studies, research into potential sources of bias and variation in comparative studies is a priority.

9.5.2 Heterogeneity in relative test performance

Heterogeneity is expected in DTA reviews and typically higher than in intervention reviews. This may be partly due to estimation of point estimates of test accuracy which are proportions akin to absolute risks in treatment or control groups in RCTs. This could potentially make the estimates more heterogeneous than estimates of relative measures, in addition to differences in patient characteristics and other factors. In network meta-analysis, heterogeneity of the contrasts of treatments is typically assumed to be the same across a network. While that may be acceptable, it is clearly untenable for test comparisons where the random effects are on point estimates rather than relative estimates. Future research should seek to understand whether the magnitude of heterogeneity observed in conventional comparative meta-analysis of point estimates persist in meta-analysis of relative differences.

9.5.3 Performance of comparative meta-analysis methods with few studies or sparse data

The focus of the thesis was primarily on comparative meta-analysis methods and their performance in situations with few studies or sparse data was also of interest. However, because of limited evidence on the performance of hierarchical models for the meta-analysis of a single test, developing the evidence base in this context was a first priority for the simulation study described in Chapter 8. To build on the work in Chapter 8, the next challenge is a simulation study of test comparisons which may well yield similar conclusions

given the concordance of results of the empirical and simulation studies of a single test. However, different conclusions are plausible. In addition to issues addressed in Chapter 8, other issues raised in the comparative meta-analyses in Chapter 7, particularly regarding the intricacies of covariance structures in direct (paired data analyses) and indirect comparisons should be considered. Section 8.6 briefly outlined the issue of conditional dependence between tests that should also be considered in extending the simulation to test comparisons. The simulation will also have implications for investigations of heterogeneity because the same meta-regression approach is used for such evaluations. A question often posed by meta-analysts and review authors is “*What is the minimum number of studies needed for meta-regression?*” A simulation study can provide answers.

9.5.4 Comparing test accuracy across multiple thresholds per study

The thesis only considered methods applicable in the common situation where a single 2x2 table is available, or can be derived for each study included in the meta-analysis. However, as mentioned earlier in section 1.3.1, thresholds are needed to dichotomise the data for certain test measurements. Thus, information from multiple thresholds is sometimes available for some included studies. Methods have been proposed which allow inclusion of data from multiple thresholds per study for a single test⁴⁶⁻⁴⁸ but these have not been applied to test comparisons. Since ranking of test performance is not consistent across thresholds if accuracy depends on threshold and selection of a single threshold to include in a meta-analysis is usually arbitrary or data driven, methods that allow simultaneous comparisons of tests across multiple thresholds should make optimal use of the available data and are worth exploring.

9.5.5 Evaluation of Bayesian comparative meta-analysis methods

Bayesian comparative meta-analysis methods were not evaluated in the thesis. Since completing the methodological review in Chapter 6, a new Bayesian method has been published.²⁸⁴ The model allows for the comparison of two tests using direct and indirect comparisons via a third test, an approach similar to mixed treatment comparison. The model can be extended to more than two tests and also accounts for imperfect reference standards. This approach and others in Chapter 6 that allow for complex modelling of different data structures require evaluation to assess their validity as well as their performance against frequentist methods.

9.6 Conclusions

The questions posed in the thesis indicated gaps in the evidence base about the characteristics of systematic reviews and meta-analyses that compare test accuracy, and the impact of study design and meta-analytic approaches on the findings of these reviews. There was wide variation in methods and reporting of comparative reviews, casting doubt on the utility of many reviews. Nonetheless, a few good examples were available. The widely held view that comparative accuracy studies are seldom available has been substantiated in the thesis. While the lack of comparative studies is a potential threat to the validity of comparative reviews as evidenced by differences between results from direct and indirect comparisons, methods used for meta-analysis also have implications for review findings. Different methods can lead to differences in the magnitude, precision, direction and/or importance of meta-analytic findings, and so choosing suitable methods and applying them appropriately is essential to avoid misleading conclusions. In addition, when appropriate methods are used, valid answers can also be obtained from the synthesis of few studies or sparse data which is not uncommon in

DTA meta-analysis of a single test or test comparisons. From the thesis, one can deduce that study design and analytic approaches are likely to have a synergistic effect on the reliability of comparative reviews.

The findings of the thesis have important implications for the design of future primary studies of test accuracy, for systematic reviews and meta-analyses of test comparisons, and also for clinical practice. In the absence of evidence of clinical effectiveness, comparative accuracy provides the best available evidence for guiding test selection and decision making. Ideally direct comparisons should be prioritised, though the use of indirect information is the only option in situations where direct evidence is very limited or unavailable. The issue of test selection is common and critical to health technology assessment, and so it is vital that well-designed comparative accuracy studies are available for systematic reviews and meta-analyses that provide evidence of comparative test accuracy. Furthermore, to prevent inappropriate recommendations, it is imperative that the methods underpinning the reviews are robust and clearly reported to avoid ambiguity and to increase confidence in the utility of comparative reviews. Challenges remain and methods are still evolving, but the thesis is undoubtedly a significant contribution to the expansion of the evidence base and a major step towards evidence based guidance for the conduct of comparative accuracy reviews as well as raising awareness of the need for better primary research.

APPENDICES

Appendix A Software programs

A.1| SAS program for fitting HSROC model with common or separate variance parameters across tests for Type 1 and Type 4 RDTs

A.2| Stata program for fitting the bivariate model with different covariance structures to ERCP versus IOC data

Appendix B Forms, statistical methods and examples for Chapter 4

B.1| Screening form for reviews

B.2| Data extraction form for reviews

B.3| Statistical methods used for test comparisons

B.4| Summary of methodological and reporting characteristics of five exemplar comparative reviews

Appendix C Sensitivity analysis for one-sided contour-enhanced funnel plot of the ratio of relative diagnostic odds ratio

Appendix D Datasets and additional figures for Chapter 7

D.1| Characteristics of meta-analyses for empirical evaluation of methods

D.2| Datasets with convergence and estimation issues in HSROC models applied to individual tests

D.3| Estimates for bivariate models with and without assumption of equal variances across tests

D.4| Comparison of bivariate models with different covariance structures fitted to direct test comparisons

D.5| Estimates of relative sensitivity from HSROC models with common and different shape between tests for SROC curves

D.6| Comparison of relative sensitivity and relative specificity from bivariate and univariate models with equal variances

D.7| Estimates of relative accuracy from bivariate and univariate models with unequal variances

D.8| Estimates of variance and correlation parameters from bivariate and univariate models with unequal variances

D.9 | Estimates from unweighted and weighted Moses SROC meta-regression models

D.10| Estimates from unweighted Moses SROC and HSROC meta-regression models

Appendix E Additional simulation results

E.1| Performance of all meta-analytic models in estimating sensitivity for scenarios with a DOR of 231

E.2| Performance of all meta-analytic models in estimating specificity for scenarios with a DOR of 231

Appendix A: Software programs

A.1| SAS program for fitting HSROC model with common or separate variance parameters across tests for Type 1 and Type 4 RDTs

```
/* Import data from the excel data file */
proc import out=types1v4
  datafile='\\Mds\user\S-Z\takwoiny\types1v4.csv'
  dbms=csv
  replace;
  getnames=yes;
run;

/* Create two separate records for the true results in each study, the first for
the diseased group (dis=0.5), and the second for the non-diseased group (dis = -0.5)*/
data types1v4;
  set types1v4;
  t4=0;
  if test = "Type 4 tests" then t4=1;
  dis=0.5; pos=tp; n=tp+fn; output;
  dis=-0.5; pos=fp; n=tn+fp; output;
run;

/* Ensure that both records for a study are clustered within-study */
proc sort data=types1v4;
  by study_id test;
run;

/* HSROC model with common shape and common variance across tests */
proc nlmixed data=types1v4 cov cov start alpha=0.05 hess;
  parms alpha = 5 theta = 1 beta = 1 s2ua = 2 s2ut =1 alpha_t4=1 theta_t4=0;
  bounds s2ua>=0;
  bounds s2ut>=0;
```

Appendix A: Software programs

```

logitp= (theta+ut+theta_t4*t4+(alpha+ua+alpha_t4*t4)*dis)*exp(-(beta)*dis);
p = exp(logitp)/(1+exp(logitp));
model pos ~ binomial(n,p);
random ut ua ~normal([0,0],[s2ut,0,s2ua])
subject=study_id out =randeffects;
estimate 'E(logitSe)' exp(-beta*0.5)*(theta+0.5*alpha);
estimate 'E(logitSp)' exp(beta*0.5)*(theta-0.5*alpha);
estimate 'Var(logitSe)' exp(beta)*(s2ut+0.25*s2ua);
estimate 'Var(logitSp)' exp(beta)*(s2ut+0.25*s2ua);
estimate 'Cov(logits)' -(s2ut-0.25 * s2ua);
estimate 'Corr(logits)' (-s2ut-0.25*s2ua)/(sqrt(exp(-beta)*(s2ut+0.25*s2ua))
*sqrt(exp(beta)*(s2ut+0.25*s2ua)));
estimate "E(logitSe)_t4" exp(-beta*0.5)*(theta+theta_t4+0.5*(alpha+alpha_t4));
estimate "E(logitSp)_t4" exp(beta*0.5)*(theta+theta_t4-0.5*(alpha+alpha_t4));
estimate "logRelative sensitivity cv level 4 vs 1" log(exp(exp(-beta*0.5)*
(theta+theta_t4+0.5*(alpha+alpha_t4)))/(1+exp(exp(-beta*0.5)*(theta+theta_t4+0.5
*(alpha+alpha_t4)))))-log(exp(exp(-beta*0.5)*(theta+0.5*alpha)))/(1+exp(exp(-beta*0.5)
*(theta+0.5*alpha))));
estimate "logRelative specificity cv level 4 vs 1" log(exp(-exp(beta*0.5)*
(theta+theta_t4-0.5*(alpha+alpha_t4)))/(1+exp(-exp(beta*0.5)*(theta+theta_t4 -
0.5*(alpha+alpha_t4)))))-log(exp(-exp(beta*0.5)*(theta-0.5*alpha)))/(1+exp(-exp(beta*0.5)
*(theta-0.5*alpha))));
estimate "alpha_t4" alpha+alpha_t4 ;
estimate "theta_t4" theta+theta_t4 ;

run;

/* FINAL MODEL: model with separate variances and common shape */
proc nlmixed data=types1v4 cov ecov start alpha=0.05 hess;
parms alpha = 5 theta = 1 beta = 1 s2ua = 1 s2ut = 2 s2ut = 1
alpha_t4=1 theta_t4=0 s2ua4=1 s2ut4=1 covt=0 covs=0;
bounds s2ua>=0;
bounds s2ut>=0;
bounds s2ua4>=0;
bounds s2ut4>=0;
logitp= ((theta + ut) +(theta_t4+ut4)*t4 +((alpha + ua) + (alpha_t4+ua4)*t4) * dis)
* exp(-(beta)*dis);
p = exp(logitp)/(1+exp(logitp));
model pos ~ binomial(n,p);
random ut ua ut4 ua4~normal([0,0,0,0],[s2ut,0,s2ua,covt,0,s2ut4,0,cova,0,s2ua4])

```

```

subject=study_id out =randeffects;
estimate 'E(logitSe)' exp(-beta*0.5)*(theta+0.5*alpha);
estimate 'E(logitSp)' -exp(beta*0.5)*(theta-0.5*alpha);
estimate 'Var(logitSe)' exp(-beta)*(s2ut+0.25*s2ua);
estimate 'Var(logitSp)' exp(beta)*(s2ut+0.25*s2ua);
estimate 'Cov(logits)' -(s2ut-0.25 * s2ua);
estimate 'Corr(logits)' (- (s2ut-0.25*s2ua)) / (sqrt( exp(-beta)*(s2ut+0.25*s2ua))
*sqrt( exp(beta)*(s2ut+0.25*s2ua))));
estimate "E(logitSe)_t4" exp(-beta*0.5)*(theta+theta_t4+0.5*(alpha+alpha_t4));
estimate "E(logitSp)_t4" -exp(beta*0.5)*(theta+theta_t4-0.5*(alpha+alpha_t4));
estimate "logRelative sensitivity cv level 4 vs 1" log( exp( exp(-beta*0.5)*
(theta+theta_t4+0.5*(alpha+alpha_t4)))/(1+exp( exp(-beta*0.5)*(theta+theta_t4+0.5
*(alpha+alpha_t4)) ) ) -log( exp( exp(-beta*0.5)*(theta+theta_t4+0.5
*(theta+0.5*alpha)) ) ) );
estimate "logRelative specificity cv level 4 vs 1" log( exp( -exp(beta*0.5)*
(theta+theta_t4-0.5*(alpha+alpha_t4)))/(1+exp( -exp(beta*0.5)*(theta+theta_t4 -
0.5*(alpha+alpha_t4)) ) ) -log( exp( -exp(beta*0.5)*(theta-0.5*alpha)))/(1+exp( -exp(beta*0.5)
*(theta-0.5*alpha)) ) ) );
estimate "alpha_t4" alpha+alpha_t4 ;
estimate "theta_t4" theta+theta_t4 ;

```

```
run;
```

A.2| Stata program for fitting the bivariate model with different covariance structures to ERCP versus IOC data

```
/******  
*  
* TITLE:      Endoscopic retrograde cholangiopancreatography versus intraoperative *  
*            cholangiography for diagnosis of common bile duct stones           *  
* AUTHOR:      Yemisi Takwoingi                                                *  
* DATE CREATED: 15/07/2014                                                       *  
* DATE MODIFIED: 15/07/2014                                                    *  
* PURPOSE:    Bivariate model for meta-analysis of ERCP and IOC for diagnosis of CBD*  
*            stones                                                             *  
*  
*****/  
  
*** Read in the data from the .csv file ***  
insheet using "ERCP vs IOC for CBD stones.csv", comma clear  
  
*** Produce a summary of the dataset to check data import was ok ***  
describe  
  
*** Convert the string variable 'test' to a numeric variable named 'testtype' ***  
encode test, gen(testtype)  
  
*** List the numeric value assigned to each test ***  
label list testtype  
  
/*Set up the data  
Generate 5 new variables of type long. This is needed before reshaping the data.  
• n1 is number diseased  
• n0 is number without disease  
• true1 is number of true positives  
• true0 is the number of true negatives  
• study is the unique identifier for each study. _n will generate a sequence of numbers.*/  
gen long n1=tp+fn  
gen long n0=fp+tn  
gen long true1=tp  
gen long true0=tn  
gen long recordid=_n  
  
*** Convert data from wide to long form ***  
reshape long n true, i(recordid) j(sens)  
  
/**** Generate a new binary variable spec of type byte that takes the value 0  
when sens=1 and vice versa ***/  
gen byte spec=1-sens  
  
/**** Sort data to ensure studies are clustered together first by study and then
```

by values of the covariate testtype ***/
sort studyid testtype

*** Create dummy variables for the covariate testtype ***

```
gen seERCP=0
gen spERCP=0
gen seIOC=0
gen spIOC=0
replace seERCP=1 if testtype==1 & sens==1
replace spERCP=1 if testtype==1 & spec==1
replace seIOC=1 if testtype==2 & sens==1
replace spIOC=1 if testtype==2 & spec==1
```

***** Meta-analysis of ERCP *****

*** Model A: unstructured variance-covariance matrix ***

```
xtmelogit true sens spec if testtype==1, nocons || studyid: sens spec, nocons cov(un) ///
binomial(n) refineopts(iterate(3)) inpoints(5) variance
```

*** Run again without varaince option to get the correlation ***

```
xtmelogit true sens spec if testtype==1, nocons || studyid: sens spec, nocons cov(un) ///
binomial(n) refineopts(iterate(3)) inpoints(5)
```

*** Model B: exchangeable variance-covariance matrix ***

```
xtmelogit true sens spec if testtype==1, nocons || studyid: sens spec, nocons cov(exc) ///
binomial(n) refineopts(iterate(3)) inpoints(5) variance
```

```
xtmelogit true sens spec if testtype==1, nocons || studyid: sens spec, nocons cov(exc) ///
binomial(n) refineopts(iterate(3)) inpoints(5)
```

*** Model C: independent variance-covariance matrix ***

```
xtmelogit true sens spec if testtype==1, nocons || studyid: sens spec, nocons cov(ind) ///
binomial(n) refineopts(iterate(3)) inpoints(5) variance
```

```
xtmelogit true sens spec if testtype==1, nocons || studyid: sens spec, nocons cov(ind) ///
binomial(n) refineopts(iterate(3)) inpoints(5)
```

*** Model D: fixed specificity but random effects included for sensitivity ***

```
xtmelogit true sens spec if testtype==1, nocons || studyid: sens, nocons cov(ind) ///
binomial(n) refineopts(iterate(3)) inpoints(5) variance
```

```
xtmelogit true sens spec if testtype==1, nocons || studyid: sens, nocons cov(ind) ///
binomial(n) refineopts(iterate(3)) inpoints(5)
```

***** Meta-analysis of IOC *****

*** Model A: unstructured variance-covariance matrix ***

```
xtmelogit true sens spec if testtype==2, nocons || studyid: sens spec, nocons cov(un) ///
```

Appendix A: Software programs

```
binomial(n) refineopts(iterate(3)) intpoints(5) variance
```

```
xtmelogit true sens spec if testtype==2, nocons || studyid: sens spec, nocons cov(un) ///  
binomial(n) refineopts(iterate(3)) intpoints(5)
```

```
*** Model B: exchangeable variance-covariance matrix ***
```

```
xtmelogit true sens spec if testtype==2, nocons || studyid: sens spec, nocons cov(exc) ///  
binomial(n) refineopts(iterate(3)) intpoints(5) variance
```

```
xtmelogit true sens spec if testtype==2, nocons || studyid: sens spec, nocons cov(exc) ///  
binomial(n) refineopts(iterate(3)) intpoints(5)
```

```
*** Model C: independent variance-covariance matrix ***
```

```
xtmelogit true sens spec if testtype==2, nocons || studyid: sens spec, nocons cov(ind) ///  
binomial(n) refineopts(iterate(3)) intpoints(5) variance
```

```
xtmelogit true sens spec if testtype==2, nocons || studyid: sens spec, nocons cov(ind) ///  
binomial(n) refineopts(iterate(3)) intpoints(5)
```

```
*** Model D: fixed sensitivity but random effects included for specificity ***
```

```
xtmelogit true sens spec if testtype==2, nocons || studyid: spec, nocons cov(ind) ///  
binomial(n) refineopts(iterate(3)) intpoints(5) variance
```

```
xtmelogit true sens spec if testtype==2, nocons || studyid: spec, nocons cov(ind) ///  
binomial(n) refineopts(iterate(3)) intpoints(5)
```

```
*****TEST COMPARISON*****
```

```
*** Fit the model without the covariate ***
```

```
xtmelogit true sens spec, nocons || studyid: sens spec, nocons cov(exc) ///  
binomial(n) refineopts(iterate(3)) intpoints(5) variance nolr
```

```
/** Store the estimates of the log likelihood from the model above for doing the likelihood  
ratio test **/
```

```
estimates store A
```

```
/** Add covariate terms to the model for both logit sensitivity and logit specificity.  
This model assumes equal variances for both tests. **/
```

```
xtmelogit true seERCP seIOC spERCP spIOC, nocons || studyid: sens spec, nocons ///  
cov(exc) binomial(n) refineopts(iterate(3)) intpoints(5) variance nolr
```

```
estimates store B
```

```
/* Perform a likelihood ratio test comparing the model (A) without covariate with the model  
(B) that includes the covariate testtype and assumes equal variances for each test. Use the  
stored values in A and B. */
```

```
lrtest A B
```


Appendix A: Software programs

```
*** Fit model with covariate testtype and separate variances for the tests ***
xtmelogit true seERCP seIOC spERCP spIOC, nocons || studyid: seERCP spERCP, ///
nocons cov(exc) || study: seIOC spIOC, nocons cov(exc) binomial(n) refineopts(iterate(3))
intpoints(5) variance nolr

estimates store C

lrtest A C

/* To find the covariance between the expected (mean) logit sensitivity and expected logit
specificity, display contents of the variance-covariance matrix: */
matrix list e(V)

*** Estimate relative sensitivity and relative specificity ***
nlcom log_relative_sensitivity: log(invlogit(_b[seERCP]))-log(invlogit(_b[seIOC]))
nlcom log_relative_specificity: log(invlogit(_b[spERCP]))-log(invlogit(_b[spIOC]))

*** Delete the program from Stata's memory if it exists already ***
capture program drop renamematrix

*** Rename the elements of the coefficient and variance matrices ***
program define renamematrix, eclass
    matrix mb = e(b)
    matrix mv = e(V)
    matrix colnames mb = logitseERCP:_cons logitseIOC:_cons logitspERCP:_cons
    logitspIOC:_cons
    matrix colnames mv = logitseERCP:_cons logitseIOC:_cons logitspERCP:_cons
    logitspIOC:_cons
    matrix rownames mv = logitseERCP:_cons logitseIOC:_cons logitspERCP:_cons
    logitspIOC:_cons
    ereturn post mb mv
end

*** Run the program ***
renamematrix

/** Display summary estimates by taking the inverse logits of the mean logit sensitivity and
mean logit specificity for each test ***/
_diparm logitseERCP, label(Sensitivity ERCP) invlogit
_diparm logitseIOC, label(Sensitivity IOC) invlogit
_diparm logitspERCP, label(Specificity ERCP) invlogit
_diparm logitspIOC, label(Specificity IOC) invlogit

/** Display other summary estimates derived using functions of the mean logit sensitivities
and mean logit specificities ***/
_diparm logitseERCP logitspERCP, label(LR+ ERCP) ci(log) function(invlogit(@1)/(1-
invlogit(@2))) derivative(exp(@2-1)*invlogit(@1)^2/invlogit(@2) exp(@2)*invlogit(@1))
```

Appendix A: Software programs

```
_diparm logitseIOC logitspIOC, label(LR+ IOC) ci(log) function(invlogit(@1)/(1-  
invlogit(@2))) derivative(exp(@2-1)*invlogit(@1)^2/invlogit(@2) exp(@2)*invlogit(@1))  
_diparm logitseERCP logitspERCP, label(LR- ERCP) ci(log) function((1-  
invlogit(@1))/invlogit(@2)) derivative(exp(-@1)*invlogit(@1)^2/invlogit(@2) exp(-@1-  
@2)*invlogit(@1))  
_diparm logitseIOC logitspIOC, label(LR- IOC) ci(log) function((1-  
invlogit(@1))/invlogit(@2)) derivative(exp(-@1)*invlogit(@1)^2/invlogit(@2) exp(-@1-  
@2)*invlogit(@1))
```

*****END*****

Appendix B: Forms, statistical methods and examples for Chapter 4

B.1| Screening form for reviews

REVIEW SCREENING FORM			
Date:	Review ID:		
Review identifier: <i>(Surname of first author + year of publication)</i>			
1. Is it a test accuracy review?	Yes ↓	Unclear ↓	No ↓ Exclude
2. Did the review evaluate more than one staging, diagnostic or screening tests (comparative review or multiple index tests for the same condition)?	Yes ↓	Unclear ↓	No ↓ Exclude
3. Was a meta-analysis performed?	Yes ↓	Unclear ↓	No ↓ Exclude
4. Full text of article is available?	Yes ↓	Unclear ↓	No ↓ Exclude
5. Data available to assess study design (comparative and non-comparative)?	Yes ↓	Unclear ↓	No ↓ Exclude

Final Decision: **Include** **Exclude**

Reasons for exclusion to be recorded in spreadsheet:

1. Not a test accuracy review
2. Review of a single test
3. Narrative synthesis
4. Full text unavailable
5. Unable to assess study design

B.2| Data extraction form for reviews

REVIEW DATA EXTRACTION FORM

SECTION A: GENERAL CHARACTERISTICS

Review identifier: _____

(Surname of first author & year of publication)

Review title:

Journal:

Publication type: Cochrane review General medical Specialist TAR

English language publication: Yes No Unclear

Target condition(s):

ICD-10 code:

Reference standard(s):

Number of tests evaluated in the review: _____

Type of included studies: Comparative only Any study type

Number of comparative studies: _____ Number of non-comparative studies: _____

Clinical purpose of the tests: Staging Diagnosis Screening Unclear Other

If unclear or other, please give details:

Type of tests: Clinical and physical examination Imaging Laboratory Unclear Other

If unclear or other, please give details:

SECTION B: STATISTICAL METHODS

Summary measures

Number of measures: _____

Summary measures used:

- | | | | |
|------------------------------|------------------------------|-----------------------------|----------------------------------|
| Area under the curve: | Yes <input type="checkbox"/> | No <input type="checkbox"/> | Unclear <input type="checkbox"/> |
| Diagnostic odds ratio: | Yes <input type="checkbox"/> | No <input type="checkbox"/> | Unclear <input type="checkbox"/> |
| Likelihood ratios: | Yes <input type="checkbox"/> | No <input type="checkbox"/> | Unclear <input type="checkbox"/> |
| Predictive values: | Yes <input type="checkbox"/> | No <input type="checkbox"/> | Unclear <input type="checkbox"/> |
| Q* statistic: | Yes <input type="checkbox"/> | No <input type="checkbox"/> | Unclear <input type="checkbox"/> |
| Sensitivity and specificity: | Yes <input type="checkbox"/> | No <input type="checkbox"/> | Unclear <input type="checkbox"/> |
| Other: | Yes <input type="checkbox"/> | No <input type="checkbox"/> | Unclear <input type="checkbox"/> |

If unclear or other, please give details: _____

Relative measures used to summarise differences in accuracy? Yes No Unclear

Meta-analysis

Test comparison feasible? Yes No

Direct comparison done: Yes No Not applicable

(Not applicable if no comparative studies included)

Hierarchical model used for meta-analysis? Yes No Method not specified:

Meta-analytic method used: _____

Test comparison method: Meta-regression Univariate pooling of sensitivity and specificity or DORs

Z-test Paired t-test Unpaired t-test Chi-squared test

Comparison of Q* Overlapping CIs Narrative None

Other Unclear

If unclear or other, please give details:

Multiple cut-offs used? Yes No

If yes, were they accounted for in the comparative analyses? Yes No Unclear

(Meta-analysis at each cut-off or fitted appropriate model)

Investigation of heterogeneity

Heterogeneity investigated? Yes No Unclear

If investigated, method used: Meta-regression Subgroup analyses Unclear

If unclear, please give details:

SECTION C: REPORTING

Reporting guideline used? Yes No Unclear

If yes, which guideline(s)?

Review objectives and role of test in diagnostic pathway

Clear comparative objective stated? Yes No

Role of index test(s): Replacement Triage Add on Unclear

If unclear or other, please give details:

Study identification and description

Flow diagram presented? Yes No

If yes, did it include number of studies per test? Yes No

Were comparative studies identified? Yes No Not applicable

(Not applicable if no comparative studies included)

Study characteristics presented? Yes No

Strategy for comparing test accuracy (direct and/or indirect comparisons)

Was the strategy reported? Yes No Not applicable

(Not applicable if only comparative studies included)

Results of data analyses

2x2 data available? Yes No

Individual study estimates of test accuracy? Yes No

Forest plot(s) presented? Yes No

SROC plot(s): Compared points or curves for at least 2 tests Separate plot per test None

Discussion

Limitations of indirect comparison acknowledged? Yes No Not applicable

(Not applicable if only comparative studies included)

SECTION D: ADDITIONAL COMMENTS

B.3| Statistical methods used for test comparisons

Statistical method*	Year	Citation
ANCOVA	2008	Meserve B B, Cleland J A, Boucher T R. A meta-analysis examining clinical test utilities for assessing meniscal injury. <i>Clinical Rehabilitation</i> 2008; 22(2): 143-161
ANCOVA	2009	Meserve BB, Cleland JA, Boucher TR. A meta-analysis examining clinical test utility for assessing superior labral anterior posterior lesions. <i>American Journal of Sports Medicine</i> 2009; 37(11): 2252-2258
Bivariate meta-regression	2009	Dijkers R, van der Zaag-Loonen HJ, Willems TP, Post WJ, Oudkerk M. Is there an indication for computed tomography and magnetic resonance imaging in the evaluation of coronary artery bypass grafts? <i>Journal of Computer Assisted Tomography</i> 2009; 33(3): 317-327
Bivariate meta-regression	2009	Menke J. Diagnostic accuracy of contrast-enhanced MR angiography in severe carotid stenosis: meta-analysis with meta-regression of different techniques. <i>European Radiology</i> 2009; 19(9): 2204-2216
Bivariate meta-regression	2010	Di Nisio M, van Sluis GL, Bossuyt PM, Buller HR, Porreca E, Rutjes AW. Accuracy of diagnostic tests for clinically suspected upper extremity deep vein thrombosis: a systematic review. <i>Journal of Thrombosis and Haemostasis</i> 2010; 8(4): 684-692
Bivariate meta-regression	2010	Pennant M, Takwoingi Y, Pennant L, Davenport C, Fry-Smith A, Eisinga A, Andronis L, Arvanitis T, Deeks J, Hyde C. A systematic review of positron emission tomography (PET) and positron emission tomography/computed tomography (PET/CT) for the diagnosis of breast cancer recurrence. <i>Health Technology Assessment</i> 2010; 14(50): 1-74
Bivariate meta-regression	2010	Schuetz GM, Zacharopoulou NM, Schlattmann P, Dewey M. Meta-analysis: noninvasive coronary angiography using computed tomography versus magnetic resonance imaging. <i>Annals of Internal Medicine</i> 2010; 152(3): 167-177
Bivariate meta-regression	2011	Lucassen W, Geersing GJ, Erkens PM, Reitsma JB, Moons KG, Buller H, van Weert HC. Clinical decision rules for excluding pulmonary embolism: a meta-analysis. <i>Annals of Internal Medicine</i> 2011; 155(7): 448-460
Bivariate meta-regression	2011	Xing Y, Bronstein Y, Ross MI, Askew RL, Lee JE, Gershenwald JE, Royal R, Cormier JN. Contemporary diagnostic imaging modalities for the staging and surveillance of melanoma patients: a meta-analysis. <i>Journal of the National Cancer Institute</i> 2011; 103(2): 129-142
Chi-squared test	2010	Mitchell AJ, Bird V, Rizzo M, Meader N. Diagnostic validity and added value of the Geriatric Depression Scale for depression in primary care: a meta-analysis of GDS30 and GDS15. <i>Journal of Affective Disorders</i> 2010; 125(1-3): 10-17
HSROC meta-regression	2008	Kyzas PA, Evangelou E, Denaxa-Kyza D, Ioannidis JP. 18F-fluorodeoxyglucose positron emission tomography to evaluate cervical node metastases in patients with head and neck squamous cell carcinoma: a meta-analysis. <i>Journal of the National Cancer Institute</i> 2008; 100(10): 712-720
HSROC meta-regression	2008	Mowatt G, Burr JM, Cook JA, Siddiqui MA, Ramsay C, Fraser C, Azuara-Blanco A, Deeks JJ, OAG Screening Project. Screening tests for detecting open-angle glaucoma: systematic review and meta-analysis. <i>Investigative Ophthalmology and Visual Science</i> 2008; 49(12): 5373-5385
HSROC meta-	2008	Vestergaard M E, Macaskill P, Holt P E, Menzies S W. Dermoscopy compared with naked eye examination for the diagnosis

B.3 continued...

Statistical method*	Year	Citation
regression		of primary melanoma: a meta-analysis of studies performed in a clinical setting. <i>British Journal of Dermatology</i> 2008; 159(3): 669-676
HSROC meta-regression	2009	Mant J, Doust J, Roaife A, Barton P, Cowie MR, Glasziou P, Mant D, McManus RJ, Holder R, Deeks J, Fletcher K, Qume M, Sohanpal S, Sanders S, Hobbs FD. Systematic review and individual patient data meta-analysis of diagnosis of heart failure, with modelling of implications of different diagnostic strategies in primary care. <i>Health Technology Assessment</i> 2009; 13(32): 1-207
HSROC meta-regression	2010	Mowatt G, N'Dow J, Vale L, Nabi G, Boachie C, Cook JA, Fraser C, Griffiths TR, Aberdeen Technology Assessment Review (TAR) Group. Photodynamic diagnosis of bladder cancer compared with white light cystoscopy: systematic review and meta-analysis. <i>International Journal of Technology Assessment in Health Care</i> 2011; 27(1): 3-10.
HSROC meta-regression	2010	Williams GJ, Macaskill P, Chan SF, Turner RM, Hodson E, Craig JC. Absolute and relative accuracy of rapid urine tests for urinary tract infection in children: a meta-analysis. <i>Lancet Infectious Diseases</i> 2010; 10(4): 240-250
HSROC meta-regression	2011	Abba Katharine, Deeks Jonathan J, Olliaro Piero L, Naing Cho-Min, Jackson Sally M, Takwoingi Yemisi, Donegan Sarah, Garner Paul. Rapid diagnostic tests for diagnosing uncomplicated <i>P. falciparum</i> malaria in endemic countries. <i>Cochrane Database of Systematic Reviews</i> : Reviews 2011; Issue 7
HSROC meta-regression	2011	Hodgkinson J, Mant J, Martin U, Guo B, Hobbs FD, Deeks JJ, Heneghan C, Roberts N, McManus RJ. Relative effectiveness of clinic and home blood pressure monitoring compared with ambulatory blood pressure monitoring in diagnosis of hypertension: systematic review. <i>BMJ</i> 2011; 342:d3621
HSROC meta-regression	2011	Wang Louis W, Fahim Magid A, Hayen Andrew, Mitchell Ruth L, Baines Laura, Lord Stephen, Craig Jonathan C, Webster Angela C. Cardiac testing for coronary artery disease in potential kidney transplant recipients. <i>Cochrane Database of Systematic Reviews</i> : Reviews 2011; Issue 12
HSROC meta-regression	2012	Allred S Kate, Deeks Jonathan J, Guo Boliang, Neilson James P, Alfirevic Zarko. Second trimester serum tests for Down's Syndrome screening. <i>Cochrane Database of Systematic Reviews</i> : Reviews 2012; Issue 6
Logistic regression	2008	Planche T, Aghaizu A, Holliman R, Riley P, Poloniecki J, Breathnach A, Krishna S. Diagnosis of <i>Clostridium difficile</i> infection by toxin detection kits: a systematic review. <i>Lancet Infectious Diseases</i> 2008; 8(12): 777-784
Pooled difference in OR	2010	Floriani I, Torri V, Rulli E, Garavaglia D, Compagnoni A, Salvolini L, Giovagnoni A. Performance of imaging modalities in diagnosis of liver metastases from colorectal cancer: a systematic review and meta-analysis. <i>Journal of Magnetic Resonance Imaging</i> 2010; 31(1): 19-31
Pooled difference in OR	2010	Zhu MM, Xu XT, Nie F, Tong JL, Xiao SD, Ran ZH. Comparison of immunochemical and guaiac-based fecal occult blood test in screening and surveillance for advanced colorectal neoplasms: a meta-analysis. <i>Journal of Digestive Diseases</i> 2010; 11(3): 148-160
Pooled difference in Se	2010	Cattamanichi A, Davis JL, Pai M, Huang L, Hopewell PC, Steingart KR. Does bleach processing increase the accuracy of sputum smear microscopy for diagnosing pulmonary tuberculosis? <i>Journal of Clinical Microbiology</i> 2010; 48(7): 2433-2439

B.3 continued...

Statistical method*	Year	Citation
Pooled difference in Se and Sp	2012	Alrajhi K, Woo MY, Vaillancourt C. Test characteristics of ultrasonography for the detection of pneumothorax: a systematic review and meta-analysis. <i>Chest</i> 2012; 141(3): 703-708
Pooled RDOR	2008	Ewald B, Ewald D, Thakkinstian A, Attia J. Meta-analysis of B type natriuretic peptide and N-terminal pro B natriuretic peptide in the diagnosis of clinical heart failure and population screening for left ventricular systolic dysfunction. <i>Internal Medicine Journal</i> 2008; 38(2): 101-113
Pooling ratios of Se and Sp	2011	Sun J, Garfield DH, Lam B, Yan J, Gu A, Shen J, Han B. The value of autofluorescence bronchoscopy combined with white light bronchoscopy compared with white light alone in the diagnosis of intraepithelial neoplasia and invasive lung cancer: a meta-analysis. <i>Journal of Thoracic Oncology</i> 2011; 6(8): 1336-1344
SROC meta-regression	2008	van Vliet E P, Heijnenbrok-Kal M H, Hunink M G, Kuipers E J, Siersema P D. Staging investigations for oesophageal cancer: a meta-analysis. <i>British Journal of Cancer</i> 2008; 98(3): 547-557
SROC meta-regression	2010	Yin ZG, Zhang JB, Kan SL, Wang XG. Diagnosing suspected scaphoid fractures: a systematic review and meta-analysis. <i>Clinical Orthopaedics and Related Research</i> 2010; 468(3): 723-734
SROC regression (Q*)	2008	Hovels AM, Heesakkers RA, Adang EM, Jager GJ, Strum S, Hoogveen YL, Severens JL, Barentsz JO. The diagnostic accuracy of CT and MRI in the staging of pelvic lymph nodes in patients with prostate cancer: a meta-analysis. <i>Clinical Radiology</i> 2008; 63(4): 387-395
t-test	2009	Rajpara SM, Botello AP, Townend J, Ormerod AD. Systematic review of dermoscopy and digital dermoscopy/artificial intelligence for the diagnosis of melanoma. <i>British Journal of Dermatology</i> 2009; 161(3): 591-604
t-test (paired)	2010	Lewis NR, Scott BB. Meta-analysis: deamidated gliadin peptide antibody and tissue transglutaminase antibody compared as screening tests for coeliac disease. <i>Alimentary Pharmacology and Therapeutics</i> 2010; 31(1): 73-81
Z-test	2008	Horsthuis K, Bipat S, Bennink R J, Stoker J. Inflammatory bowel disease diagnosed with US, MR, scintigraphy, and CT: meta-analysis of prospective studies. <i>Radiology</i> 2008; 247(1): 64-79
Z-test	2008	Raijmakers P G, Paul M A, Lips P. Sentinel node detection in patients with thyroid carcinoma: a meta-analysis. <i>World Journal of Surgery</i> 2008; 32(9): 1961-1967
Z-test	2008	Sosna J, Sella T, Sy O, Lavin PT, Eliahou R, Fraifeld S, Libson E. Critical analysis of the performance of double-contrast barium enema for detecting colorectal polyps > or = 6 mm in the era of CT colonography. <i>AJR American Journal of Roentgenology</i> 2008; 190(2): 374-385
Z-test	2008	van Randen A, Bipat S, Zwinderman AH, Ubbink DT, Stoker J, Boermeester MA. Acute appendicitis: meta-analysis of diagnostic performance of CT and graded compression US related to prevalence of disease. <i>Radiology</i> 2008; 249(1): 97-106
Z-test	2009	Gu P, Pan LL, Wu SQ, Sun L, Huang G. CA 125, PET alone, PET-CT, CT and MRI in diagnosing recurrent ovarian carcinoma: a systematic review and meta-analysis. <i>European Journal of Radiology</i> 2009; 71(1): 164-174

B.3 continued...

Statistical method*	Year	Citation
Z-test	2010	Choi HJ, Ju W, Myung SK, Kim Y. Diagnostic performance of computer tomography, magnetic resonance imaging, and positron emission tomography or positron emission tomography/computer tomography for detection of metastatic lymph nodes in patients with cervical cancer: meta-analysis. <i>Cancer Science</i> 2010; 101(6): 1471-1479
Z-test	2010	Ngamruengphong S, Sharma VK, Nguyen B, Das A. Assessment of response to neoadjuvant therapy in esophageal cancer: an updated systematic review of diagnostic accuracy of endoscopic ultrasonography and fluorodeoxyglucose positron emission tomography. <i>Diseases of the Esophagus</i> 2010; 23(3): 216-231
Z-test	2010	Niekel MC, Bipat S, Stoker J. Diagnostic imaging of colorectal liver metastases with CT, MR imaging, FDG PET, and/or FDG PET/CT: a meta-analysis of prospective studies including patients who have not previously undergone treatment. <i>Radiology</i> 2010; 257(3): 674-684
Z-test	2010	Pan L, Han Y, Sun X, Liu J, Gang H. FDG-PET and other imaging modalities for the evaluation of breast cancer recurrence and metastases: a meta-analysis. <i>Journal of Cancer Research and Clinical Oncology</i> 2010; 136(7): 1007-1022
Z-test	2011	Liu T, Cheng T, Xu W, Yan WL, Liu J, Yang HL. A meta-analysis of 18FDG-PET, MRI and bone scintigraphy for diagnosis of bone metastases in patients with breast cancer. <i>Skeletal Radiology</i> 2011; 40(5): 523-531
Z-test	2011	Liu T, Xu JY, Xu W, Bai YR, Yang WL, Yang HL. Fluorine-18 deoxyglucose Positron Emission Tomography, Magnetic Resonance Imaging and Bone Scintigraphy for the Diagnosis of Bone Metastases in Patients with Lung Cancer: Which One is the Best? - a Meta-analysis. <i>Clinical Oncology</i> 2011; 23(5): 350-358
Z-test	2011	Tang S, Huang G, Liu J, Liu T, Treven L, Song S, Zhang C, Pan L, Zhang T. Usefulness of 18F-FDG PET, combined FDG-PET/CT and EUS in diagnosing primary pancreatic carcinoma: a meta-analysis. <i>European Journal of Radiology</i> 2011; 78(1): 142-150
Z-test	2011	Xu GZ, Zhu XD, Li MY. Accuracy of whole-body PET and PET-CT in initial M staging of head and neck cancer: a meta-analysis. <i>Head and Neck</i> 2011; 33(1): 87-94
Z-test	2011	Yang HL, Liu T, Wang XM, Xu Y, Deng SM. Diagnosis of bone metastases: a meta-analysis comparing 18FDG PET, CT, MRI and bone scintigraphy. <i>European Radiology</i> 2011; 21(12): 2604-2617
Z-test	2012	Wu LM, Chen FY, Jiang XX, Gu HY, Yin Y, Xu JR. 18F-FDG PET, combined FDG-PET/CT and MRI for evaluation of bone marrow infiltration in staging of lymphoma: a systematic review and meta-analysis. <i>European Journal of Radiology</i> 2012; 81(2): 303-311
Unclear (HSROC model)	2008	Kriston L, Holzel L, Weiser A-K, Berner M M, Harter M. Meta-analysis: are 3 questions enough to detect unhealthy alcohol use? <i>Annals of Internal Medicine</i> 2008; 149(12): 879-888
Unclear (Bivariate model)	2008	Selman TJ, Mann CH, Zamora J, Khan KS. A systematic review of tests for lymph node status in primary endometrial cancer. <i>BMC Women's Health</i> 2008; 8:8

B.3 continued...

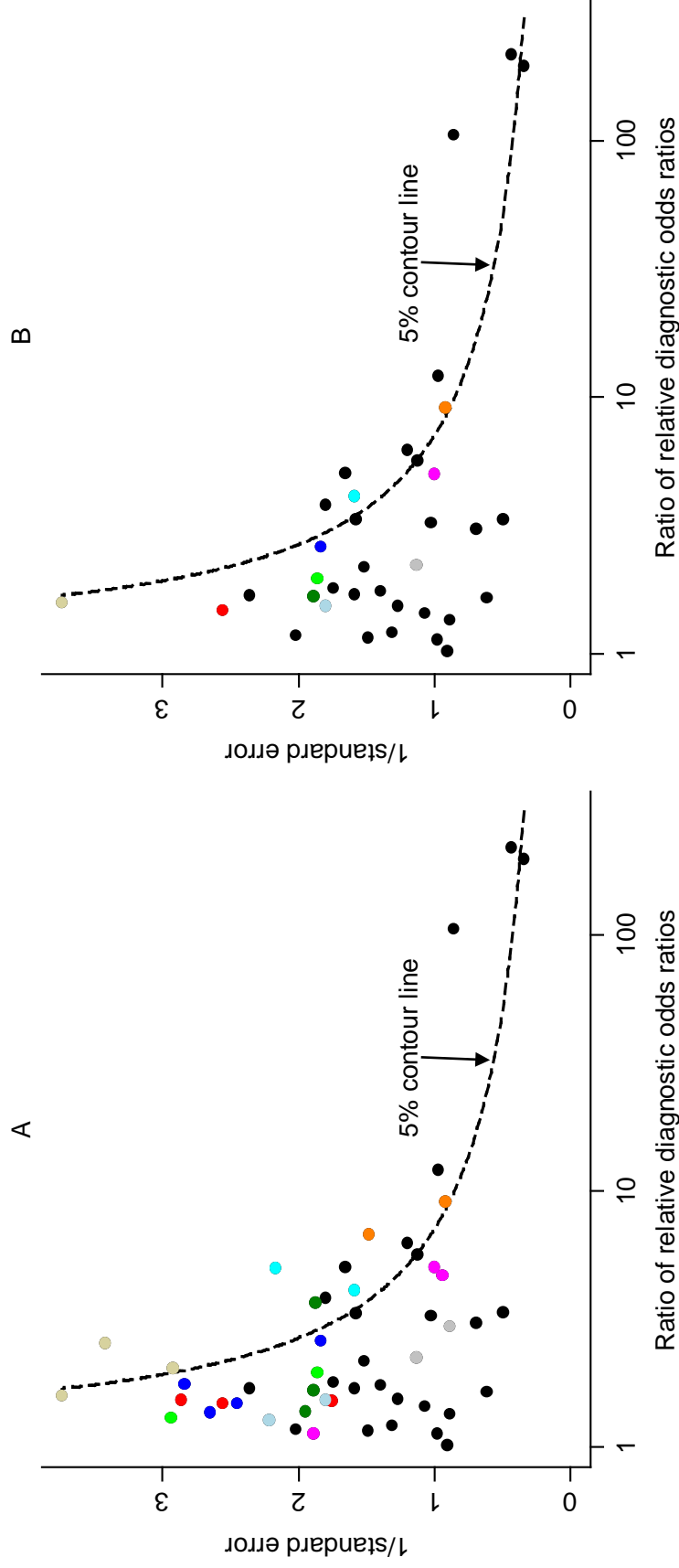
Statistical method*	Year	Citation
Unclear (SROC and univariate models)	2010	Mahajan N, Polavaram L, Vankayala H, Ference B, Wang Y, Ager J, Kovach J, Afonso L. Diagnostic accuracy of myocardial perfusion imaging and stress echocardiography for the diagnosis of left main and triple vessel coronary artery disease: a comparative meta-analysis. <i>Heart</i> 2010; 96(12): 956-966
Unclear (Simple pooling)	2011	Kraft M, Betz CS, Leunig A, Arens C. Value of fluorescence endoscopy for the early diagnosis of laryngeal cancer and its precursor lesions. <i>Head and Neck</i> 2011; 33(7): 941-948
Unclear (SROC and univariate models)	2011	Qu X, Huang X, Wu L, Huang G, Ping X, Yan W. Comparison of virtual cystoscopy and ultrasonography for bladder cancer detection: a meta-analysis. <i>European Journal of Radiology</i> 2011; 80(2): 188-197
Unclear (Bivariate model)	2012	Kobayashi Y, Hayashino Y, Jackson JL, Takagaki N, Hinoitsu S, Kawakami K. Diagnostic performance of chromoendoscopy and narrow band imaging for colonic neoplasms: a meta-analysis. <i>Colorectal Disease</i> 2012; 14(1): 18-28

*Although the meta-analytic method was stated, indicated in (), the approach used for comparing tests was unclear in 8 reviews. Reviews are sorted according to statistical method, year of publication and first author.

B.4| Summary of methodological and reporting characteristics of five exemplar comparative reviews

Review	Publication type	Review objectives	Test comparison and meta-analysis	Comments
Allred 2012 ¹⁰³	Cochrane	To estimate and compare the accuracy of second trimester serum markers for the detection of Down's syndrome, both as individual markers and as combinations of markers.	Comprehensive description of statistical methods for both direct and indirect comparisons, including the strategy for handling multiple thresholds. HSRoc meta-regression models were used to formally compare test accuracy. Relative accuracy was expressed in terms of the relative diagnostic odds ratio as appropriate.	A well-structured, large and complex review due to the number of tests, test combinations and thresholds included.
Pennant 2010 ¹⁴⁵	Technology assessment report	To assess the incremental diagnostic accuracy of PET and PET/CT compared with existing diagnostic strategies and to compare the diagnostic accuracy of PET and PET/CT for the diagnosis of breast cancer recurrence	Rationale given for the test comparisons performed. For each pairwise comparison of imaging modalities, a bivariate meta-regression model was used to compare test accuracy. Relative accuracy reported using relative sensitivities and relative specificities.	Forest plots showing the pair of test accuracy estimates from each study were presented.
Schuetz 2010 ¹⁴⁷	General medical journal	To compare CT and MRI for ruling out clinically significant coronary artery disease (CAD) in adults with suspected or known CAD.	A bivariate meta-regression model was used to compare test accuracy. Both direct and indirect comparisons were done but the test comparison strategy was not specified in the methods.	Review findings interpreted with caution due to limited evidence from comparative studies.
Wang 2011 ¹²⁹	Cochrane	To assess the diagnostic accuracy of non-invasive cardiac screening tests versus coronary angiography in potential kidney transplant recipients. Diagnostic accuracy was compared between screening tests.	Clear and detailed description of test comparison strategy and methods. Both direct and indirect comparisons were planned. Test accuracy was statistically compared in a HSRoc meta-regression model.	Included an exemplary flow diagram.
Williams 2010 ¹⁴⁶	Specialist	To assess whether rapid urine tests were sufficiently sensitive to avoid urine culture in children with negative results and to compare the accuracy of dipsticks with microscopy.	Detailed test comparison strategy and methods. Differences in thresholds for test positivity were accounted for. Direct comparisons were performed using a HSRoc model and meta-regression. Where appropriate, the relative diagnostic odds ratio was used to summarise relative accuracy.	A very detailed description of the methods.

Appendix C: Sensitivity analysis for one-sided contour-enhanced funnel plot of the ratio of relative diagnostic odds ratio



Reviews that contributed a single pairwise meta-analysis are represented by black points. Reviews that contributed multiple pairwise meta-analyses are represented by points in other colours; points that belong to the same review are represented by the same colour. Altogether there are 52 points from 36 reviews in plot A. After random selection of one pairwise meta-analysis from each review, eight out of 36 points were above the contour line in plot B instead of 13 out of 52 in plot A.

Appendix D: Datasets and additional figures for Chapter 7

D.1| Characteristics of meta-analyses for empirical evaluation of methods

ID	Reference	Target condition	Index test	Comparator	Test comparison	N_{IC}^1	N_I^2	N_C^3	N_T^4	Unit of analysis	Thresholds
1	Alkhalayal 2007 ¹⁵²	Equivocal acute appendicitis	CT	US	Indirect	3	25	25	47	Patient	No
2	Arbyn 2004	Cervical lesions	HPV triage test	Cytology triage test	Direct only	4	4	4	4	Patient	No
3	Bafounta 2001	Melanoma	Dermoscopy	Clinical examination	Direct only	8	8	8	8	Lesion	Yes
4	Basaran 2009	Acute appendicitis	MRI	CT	Indirect	0	4	3	7	Patient	No
5	Battaglia 2006	Congestive heart failure	ELISA or ELISA NT-proBNP ⁵	RIA	Indirect	0	6	13	19	Patient	Yes
6	Birim 2005	Mediastinal lymph node metastases in non-small cell lung cancer	PET	CT	Direct only	17	17	17	17	Patient	No
7	Brazzelli 2009	Ischaemic stroke	MRI	Non-contrast CT	Direct only	7	7	7	7	Patient	No
8	Carlson 1994	Ovarian cancer	US	CA 125	Indirect	0	7	3	10	Patient	No (only 35 U/ml for CA 125)
9	Cavallazzi 2008	Right ventricular dysfunction	NT-proBNP	BNP	Indirect	1	3	5	7	Patient	Yes
10	de Vries 1996	Peripheral arterial disease	Color-guided duplex US	Duplex US	Indirect	0	6	8	14	Segment	Yes (different ultrasound criteria)
11	Deville 2000 ²³⁶	Herniated discs	Cross straight leg raising	Straight leg raising	Indirect	6	6	11	11	Patient	Yes
12	Dong 2008	Carcinoma of unknown primary	FDG-PET/CT	FDG-PET	Indirect	1	8	21	28	Patient	No
13	Dong 2009	Thyroid carcinoma	FDG-PET/CT	FDG-PET	Indirect	0	4	14	18	Patient	No
14	Doria 2006	Appendicitis	CT	US	Indirect	5	8	23	26	Patient	No
15	Ewald 2008	Heart Failure	NT-proBNP	BNP	Direct only	7	7	7	7	Patient	Yes

D.1 continued...

ID	Reference	Target condition	Index test	Comparator	Test comparison	N_{IC}^1	N_C^3	N_T^4	Unit of analysis	Thresholds
16	Fleischmann 1998 ¹⁵⁷	Coronary artery disease	Exercise SPECT	Exercise ECHO	Indirect	6	23	44	Patient	Yes (different non-numeric positivity criteria)
17	Gisbert 2006 ²⁴⁰	Helicobacter pylori infection	Monoclonal stool antigen test	Polyclonal stool antigen test	Indirect	9	9	16	Patient	No
18	Gould 2003 ²⁴¹	Mediastinal staging in patients with non-small-cell lung cancer	PET	CT	Indirect	24	24	33	Patient	No
19	Granader 2008	Breast cancer	MRI	Mammography	Direct only	4	4	4	Lesion	Yes (different BIRAD cut-offs)
20	Gu 2007 ¹⁶⁰	Pleural effusion	Cytokeratin fragment 19 (CYFRA 21-1)	Carcinoembryonic antigen	Indirect	8	15	19	Patient	Yes
21	Hamon 2007	Coronary artery disease	64-section CT	16-section CT	Indirect	0	12	17	Patient	No
22	Hayashino 2005	Pulmonary embolism	Helical CT	Ventilation perfusion scanning	Indirect	2	9	5	Patient	Yes
23	Hodgkinson 2011	Hypertension	Clinic blood pressure monitoring	Home blood pressure monitoring	Indirect	1	3	9	Patient	No
24	Hovels 2008	Staging of pelvic lymph nodes in patients with prostate cancer	MRI	CT	Indirect	3	10	17	Patient	No
25	Karger 2007	Disorders of primary haemostasis	PFA-EPI	PFA-ADP	Indirect	5	6	6	Patient	No
26	Kearon 1998	Deep vein thrombosis	Venous ultrasonography	Plethysmography	Indirect	1	18	6	Patient	No
27	Kittler 2002	Melanoma	Dermoscopy	Unaided eye	Direct only	13	13	13	Image /Patient	Yes
28	Koumans 1998	Gonococcal infections of the endocervix	Nuclei acid hybridization (Gen-Probe Pace 2)	Nuclei acid amplification (Abbott LCR)	Indirect	0	10	3	Patient	No
29	Kriston 2008	Risky drinking	AUDIT-C	AUDIT	Direct only	5	5	5	Patient	Yes
30	Ledro-Cano 2007	Cholelithiasis	EUS	MRCP	Direct only	7	7	7	Patient	No

D.1 continued...

ID	Reference	Target condition	Index test	Comparator	Test comparison	N_{IC}^1	N_T^2	N_C^3	N_T^4	Unit of analysis	Thresholds
31	Lewis 2006	Coeliac disease	tTG antibody	EMA	Direct only	34	34	34	34	Patient	No
32	Lewis 2010	Coeliac disease	IgA-DGP	IgA-tTG	Direct only	11	11	11	11	Patient	No
33	Mahajan 2010 ¹⁶⁵	Left main and triple vessel coronary artery disease	SE	MPI	Indirect	6	14	15	23	Patient	No
34	Mirza 2010 ¹⁶⁵	Endoleak after endovascular aneurysm repair	Contrast enhanced US	Duplex US	Indirect	5	7	21	23	Patient	No
35	Mitchell 2010 ¹⁶⁶	Depression	GDS15	GDS30	Indirect	4	10	7	13	Patient	Yes
36	Ngamruengphong 2010 ¹⁶⁸	Esophageal cancer	EUS	FDG-PET	Indirect	3	7	15	19	Patient	No
37	Nishimura 2007 ¹⁷⁰	Rheumatoid arthritis	Anti-CCP antibody	Rheumatoid factor	Indirect	28	37	50	59	Patient	Yes
38	Olatidoye 1998	Predicting myocardial infarction and cardiac death in patients with unstable angina pectoris	Cardiac Troponin T	Cardiac Troponin I	Indirect	2	12	9	19	Patient	Yes
39	Roos 2007	Renal dysfunction	Cystatin C	Serum creatinine	Direct only	27	27	27	27	Patient	Yes
40	Schuetz 2010 ¹⁴⁷	Coronary artery disease	CT	MRI	Indirect	5	89	19	103	Patient	No
41	Schuijf 2006	Coronary artery stenoses	MSCT	MRI	Indirect	1	24	28	51	Segment	No
42	Shie 2008	Bone metastases in patients with breast cancer	FDG-PET	Bone scintigraphy	Direct only	4	4	4	4	Lesion	No
43	Smith 2011 ¹⁷⁷	Acetabular labral tears	MR arthrography	MRI	Indirect	4	15	8	19	Hips	No
44	Sun 2011	Intraepithelial neoplasia and invasive lung cancer	Auto-fluorescence bronchoscopy combined with white light bronchoscopy	White light bronchoscopy alone	Direct only	14	14	14	14	Biopsy	No
45	Tan 2002	Renal artery stenosis	Gadolinium-enhanced MRA	Non-enhanced MRA	Indirect	2	12	15	25	Artery	No

D.1 continued...

ID	Reference	Target condition	Index test	Comparator	Test comparison	N _{IC} ¹	N _T ²	N _C ³	N _T ⁴	Unit of analysis	Thresholds
46	Terasawa 2004 ¹⁸⁰	Acute appendicitis	CT	US	Indirect	4	12	14	22	Patient	No
47	van Randen 2008	Appendicitis	CT	Graded compression US	Direct only	6	6	6	6	Patient	No
48	Verma 2006	Cholecholelithiasis	EUS	MRCP	Direct only	5	5	5	5	Patient	No
49	Vestergaard 2008	Primary melanoma	Dermoscopy	Eye examination	Direct only	9	9	9	9	Lesion	Yes (criteria varied)
50	Visser 2000	Peripheral arterial disease	Gadolinium-enhanced MRA	Color-guided duplex US	Indirect	0	10	21	31	Segment	No
51	Wang 2005	Gastroesophageal reflux disease in patients with non-cardiac chest pain	Proton pump inhibitor	Placebo	Direct only	6	6	6	6	Patient	No
52	Wiese 2000 ²⁶⁵	Vaginal trichomoniasis	Papanicolaou smear	Wet mount	Indirect	7	7	30	30	Patient	No
53	Wijnberger 2001	Neonatal respiratory distress	Lamellar body count	Lecithin/sphingomyelin ratio	Direct only	6	6	6	6	Patient	Yes
54	Worster 2002 ²⁰⁴	Acute urolithiasis	Non-contrast helical CT	Intravenous pyelography	Direct only	4	4	4	4	Patient	No
55	Xu 2011 ¹⁸⁶	Head and neck cancer	PET-CT	PET	Indirect	3	7	8	12	Patient	No
56	Yang 2009	Bladder cancer	FISH	Cytology	Direct only	12	12	12	12	Patient	Yes
57	Zhu 2010	Advanced colorectal neoplasms	Immunochemical faecal occult blood tests	Guaiaac-based faecal occult blood tests	Direct only	7	7	7	7	Patient	No

Anti-CCP = anti-cyclic citrullinated peptide; AUDIT = alcohol use disorders identification test; AUDIT-C = alcohol use disorders identification test-consumption; BNP = B-type natriuretic peptide; CA 125 = cancer antigen 125; CT = computed tomography; ECHO = echocardiography; ELISA = enzyme-linked immunosorbent assay; EmA = IgA antiendomyosial antibodies; EUS = endoscopic ultrasonography; FDG-PET = fluorine 18 fluorodeoxyglucose positron emission tomography; FISH = fluorescence in situ hybridization; GDS15 = geriatric depression scale (15-item questionnaire); GDS30 = geriatric depression scale (30-item questionnaire); HPV = Human papillomavirus; IgA-DGP = IgA deamidated gliadin peptides; IgA-tTG = IgA tissue transglutaminase; MPI = myocardial perfusion imaging; MR = magnetic resonance; MRA = magnetic resonance angiography; MRCP = magnetic resonance cholangiopancreatography; MRI = magnetic resonance imaging; MSCT = multislice computed tomography; NT-proBNP = N-terminal pro-B-type natriuretic peptide; PET = positron emission tomography; PFA-ADP = Platelet function analyser using collagen/adenosinediphosphate; PFA-EPI = Platelet

Appendix D: Datasets and additional figures for Chapter 7

function analyser using collagen/epinephrine; RIA = radioimmunosorbent assay; SE = stress echocardiography; SPECT = single photon emission computed tomography; tTG = tissue transglutaminase; US = ultrasound.

¹N_{IC} = Number of studies comparing index test and comparator (i.e. comparative studies).

²N_I = Index test – experimental or newer test.

³N_C = Comparator – current practice or older test.

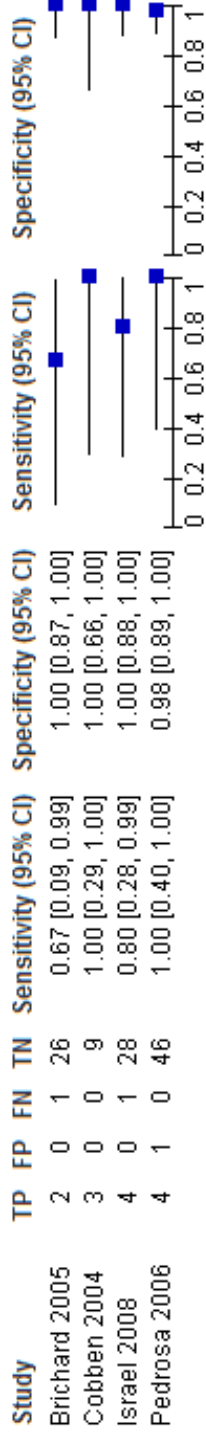
⁴N_T = Total number of studies

⁵One ELISA study reported data for 4 subgroups. Hence the total number of study cohorts that evaluated the test is 9.

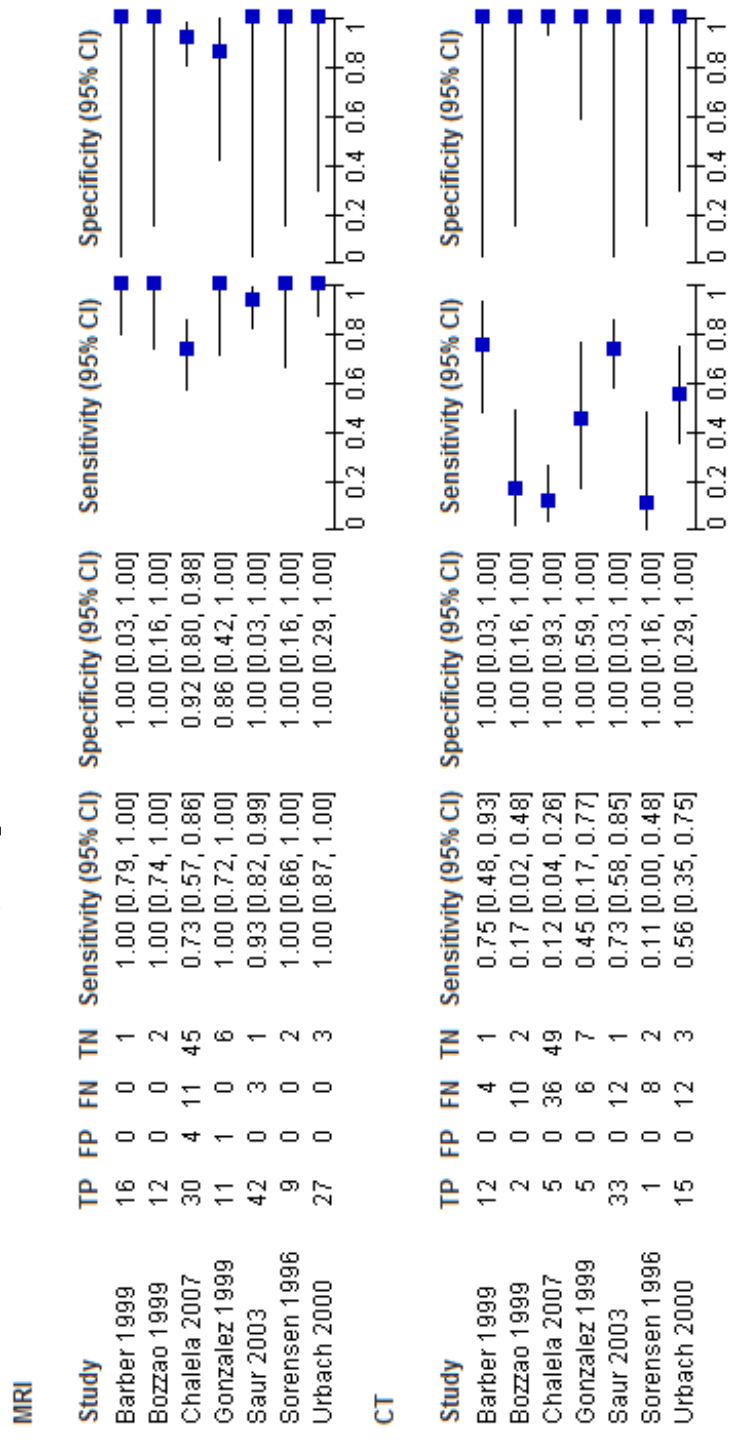
ID uniquely identifies each test comparison. The table was sorted according to ID.

D.2| Datasets with convergence and estimation issues in HSROC models applied to individual tests

Basaran 2009 (ID 4): index test = MRI



Brazzelli 2009 (ID 7): index test= MRI, comparator = CT



D.2 continued...

Hayashino 2005 (ID 22): comparator = ventilation perfusion scanning

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Garg 1998	0	0	7	18	0.00 [0.00, 0.41]	1.00 [0.81, 1.00]
Goodman 1995	1	0	10	9	0.09 [0.00, 0.41]	1.00 [0.66, 1.00]
PIOPED 1990	102	14	149	466	0.41 [0.35, 0.47]	0.97 [0.95, 0.98]
Trujillo 1997	72	6	100	277	0.42 [0.34, 0.50]	0.98 [0.95, 0.99]
Woods 1989	6	1	7	24	0.46 [0.19, 0.75]	0.96 [0.80, 1.00]

Koumans 1998 (ID 28): comparator = nuclei acid amplification (LCR) test

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Buimer 1996	10	11	1	361	0.91 [0.59, 1.00]	0.97 [0.95, 0.99]
Ching 1995	91	23	3	1422	0.97 [0.91, 0.99]	0.98 [0.98, 0.99]
Stary 1997	16	2	0	108	1.00 [0.79, 1.00]	0.98 [0.94, 1.00]

Worster 2002 (ID 54): index test = non-contrast helical CT

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Miller 1998	72	0	3	31	0.96 [0.89, 0.99]	1.00 [0.89, 1.00]
Niall 1999	28	1	0	11	1.00 [0.88, 1.00]	0.92 [0.62, 1.00]
Sourtzis 1999	36	0	0	17	1.00 [0.90, 1.00]	1.00 [0.80, 1.00]
Yilmaz 1998	60	1	4	32	0.94 [0.85, 0.98]	0.97 [0.84, 1.00]

D.3| Estimates from bivariate models with and without assumption of equal variances across tests

ID	N1	N2	NC	Variances	Log relative sensitivity	SE of log relative sensitivity	Log relative specificity	SE of log relative specificity	σ_A^2	σ_B^2	ρ_{AB}	σ_{A1}^2	σ_{B1}^2	ρ_{A1B1}	σ_{A2}^2	σ_{B2}^2	ρ_{A2B2}	
1	25	3	3	Equal	0.079	0.019	0.005	0.009	0.994	0.751	0.338	-	-	-	-	-	-	-
				Unequal	0.104	0.031	-0.005	0.014	-	-	-	1.193	0.528	-0.020	0.962	0.909	0.533	-
2	4	4	4	Equal	0.186	0.051	0.051	0.020	0.141	0.135	-1.000	-	-	-	-	-	-	-
				Unequal	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3	8	8	8	Equal	0.097	0.036	0.139	0.033	0.106	0.454	-0.563	-	-	-	-	-	-	-
				Unequal	0.106	0.055	0.114	0.054	-	-	-	0.121	0.601	0.266	0.261	0.476	-0.651	-
4	4	3	0	Equal	-0.004	0.132	0.006	0.015	0.668	0.360	-1.000	-	-	-	-	-	-	-
				Unequal	-0.065	0.156	0.004	0.017	-	-	-	0.000	0.000	-0.777	1.610	0.466	-1.000	-
5	6	13	0	Equal	0.082	0.033	0.105	0.132	0.268	1.399	-0.473	-	-	-	-	-	-	-
				Unequal	0.089	0.035	0.119	0.136	-	-	-	0.433	2.946	-0.727	0.198	0.475	-0.071	-
6	17	17	17	Equal	0.341	0.058	0.169	0.033	0.152	0.473	-0.558	-	-	-	-	-	-	-
				Unequal	0.320	0.061	0.171	0.044	-	-	-	0.268	0.760	0.045	0.099	0.503	-1.000	-
7	7	7	7	Equal	-	-	-	-	-	-	-	-	-	-	-	-	-	-
				Unequal	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8	7	3	0	Equal	-	-	-	-	-	-	-	-	-	-	-	-	-	-
				Unequal	-	-	-	-	-	-	-	-	-	-	-	-	-	-
9	3	5	1	Equal	0.171	0.074	-0.303	0.098	0.192	0.085	1.000	-	-	-	-	-	-	-
				Unequal	0.173	0.082	-0.338	0.113	-	-	-	0.529	0.092	0.406	0.103	0.361	1.000	-
10	6	8	0	Equal	0.091	0.033	0.019	0.014	0.141	0.337	-0.413	-	-	-	-	-	-	-
				Unequal	0.089	0.035	0.019	0.014	-	-	-	0.131	0.291	-0.277	0.150	0.361	-0.455	-
11	6	11	6	Equal	-1.782	0.183	1.231	0.168	0.513	0.503	-0.532	-	-	-	-	-	-	-
				Unequal	-	-	-	-	-	-	-	-	-	-	-	-	-	-
12	8	21	1	Equal	0.119	0.068	0.080	0.073	0.985	1.502	0.422	-	-	-	-	-	-	-
				Unequal	0.102	0.109	0.059	0.078	-	-	-	3.369	0.631	1.000	0.433	1.626	0.158	-
13	4	14	0	Equal	0.089	0.040	0.105	0.108	0.643	2.148	0.022	-	-	-	-	-	-	-

D.3 continued...

ID	N1	N2	NC	Variances	Log relative sensitivity	SE of log relative sensitivity	Log relative specificity	SE of log relative specificity	σ_A^2	σ_B^2	ρ_{AB}	σ_{A1}^2	σ_{B1}^2	ρ_{A1B1}	σ_{A2}^2	σ_{B2}^2	ρ_{A2B2}
				Unequal	0.083	0.044	0.081	0.118	-	-	-	0.324	1.198	-1.000	0.740	2.374	0.128
14	8	23	5	Equal	0.106	0.019	0.010	0.009	0.452	0.215	0.384	-	-	-	-	-	-
				Unequal	0.074	0.027	0.003	0.013	-	-	-	0.264	0.093	0.031	0.471	0.238	0.204
15	7	7	7	Equal	0.018	0.027	-0.117	0.024	0.305	0.201	-0.906	-	-	-	-	-	-
				Unequal	0.007	0.061	-0.088	0.060	-	-	-	0.382	0.123	-0.810	0.406	0.568	-1.000
16	27	23	6	Equal	0.013	0.020	-0.174	0.055	0.335	0.574	-0.310	-	-	-	-	-	-
				Unequal	0.019	0.027	-0.174	0.063	-	-	-	0.374	0.512	-0.375	0.284	0.570	-0.282
17	16	9	9	Equal	0.177	0.039	0.007	0.010	0.373	0.952	-0.452	-	-	-	-	-	-
				Unequal	0.172	0.034	0.022	0.021	-	-	-	0.530	0.915	-0.834	0.087	1.029	-0.484
18	33	24	24	Equal	0.327	0.058	0.157	0.028	0.342	0.346	-0.735	-	-	-	-	-	-
				Unequal	0.361	0.097	0.122	0.049	-	-	-	0.222	0.315	0.213	0.881	0.968	-1.000
19	4	4	4	Equal	0.874	0.115	-0.051	0.019	0.002	0.539	1.000	-	-	-	-	-	-
				Unequal	-	-	-	-	-	-	-	-	-	-	-	-	-
20	12	15	8	Equal	-0.034	0.049	-0.039	0.022	0.711	3.959	-0.773	-	-	-	-	-	-
				Unequal	0.216	0.159	-0.044	0.048	-	-	-	1.243	5.299	-0.780	0.452	2.010	-0.379
21	12	17	0	Equal	0.009	0.015	0.162	0.070	0.882	0.968	-0.091	-	-	-	-	-	-
				Unequal	0.000	0.016	0.147	0.072	-	-	-	0.044	0.101	0.924	1.427	1.224	-0.184
22	9	5	2	Equal	1.064	0.325	-0.023	0.016	0.842	0.370	-1.000	-	-	-	-	-	-
				Unequal	0.799	0.076	-0.014	0.019	-	-	-	1.037	0.587	-0.108	0.000	0.000	-0.998
23	7	3	1	Equal	-0.084	0.057	0.045	0.131	0.595	1.903	-0.969	-	-	-	-	-	-
				Unequal	-	-	-	-	-	-	-	-	-	-	-	-	-
24	10	17	3	Equal	0.254	0.336	-0.026	0.029	2.001	1.676	-0.374	-	-	-	-	-	-
				Unequal	-0.142	0.456	-0.018	0.026	-	-	-	1.658	1.565	0.342	2.419	2.135	-0.798
25	6	5	5	Equal	0.475	0.166	0.081	0.025	1.404	0.292	-1.000	-	-	-	-	-	-
				Unequal	0.443	0.296	0.045	0.057	-	-	-	2.187	0.374	0.534	1.578	0.342	-0.787
26	18	6	1	Equal	0.089	0.058	0.018	0.022	0.504	1.162	-0.259	-	-	-	-	-	-

D.3 continued...

ID	N1	N2	NC	Variances	Log relative sensitivity	SE of log relative sensitivity	Log relative specificity	SE of log relative specificity	σ_A^2	σ_B^2	ρ_{AB}	σ_{A1}^2	σ_{B1}^2	ρ_{A1B1}	σ_{A2}^2	σ_{B2}^2	ρ_{A2B2}
				Unequal	0.081	0.050	0.058	0.024	-	-	-	0.606	2.025	-0.068	0.301	0.235	-1.000
27	13	13	13	Equal	0.153	0.039	0.080	0.026	0.312	1.394	0.053	-	-	-	-	-	-
				Unequal	0.145	0.060	0.077	0.058	-	-	-	0.339	3.366	0.530	0.308	1.075	-0.244
28	10	3	0	Equal	-0.024	0.030	0.000	0.009	0.353	0.306	0.569	-	-	-	-	-	-
				Unequal	-0.022	0.027	-0.001	0.006	-	-	-	0.471	0.415	0.500	0.000	0.000	1.000
29	5	5	5	Equal	0.204	0.019	-0.433	0.071	0.000	0.220	1.000	-	-	-	-	-	-
				Unequal	0.193	0.039	-0.261	0.089	-	-	-	5.512	0.290	-0.653	0.137	0.461	-1.000
30	7	7	7	Equal	0.084	0.044	0.011	0.019	1.095	1.560	-1.000	-	-	-	-	-	-
				Unequal	0.091	0.072	0.001	0.036	-	-	-	1.687	1.471	-1.000	1.532	1.280	-1.000
31	18	18	18	Equal	0.012	0.006	-0.015	0.004	1.042	0.528	0.165	-	-	-	-	-	-
				Unequal	0.014	0.018	-0.013	0.005	-	-	-	0.831	0.586	0.134	1.154	2.888	0.079
32	11	11	11	Equal	-0.046	0.017	-0.019	0.009	0.709	0.204	0.316	-	-	-	-	-	-
				Unequal	-0.032	0.032	-0.022	0.014	-	-	-	0.745	0.193	0.492	0.508	0.348	-0.258
33	14	15	6	Equal	0.103	0.041	0.039	0.075	0.524	0.733	-0.912	-	-	-	-	-	-
				Unequal	0.128	0.053	-0.232	0.166	-	-	-	0.116	0.291	-1.000	0.672	1.022	-0.835
34	7	21	5	Equal	0.255	0.070	-0.006	0.027	1.355	1.612	0.083	-	-	-	-	-	-
				Unequal	0.242	0.077	-0.066	0.049	-	-	-	0.141	0.578	-1.000	1.468	2.134	0.097
35	10	7	4	Equal	0.024	0.047	-0.059	0.023	0.264	0.516	0.007	-	-	-	-	-	-
				Unequal	0.058	0.068	0.040	0.073	-	-	-	0.018	0.522	0.446	0.376	0.324	-0.470
36	7	15	3	Equal	-0.154	0.094	0.078	0.090	2.001	0.794	-0.382	-	-	-	-	-	-
				Unequal	-0.025	0.147	0.188	0.129	-	-	-	2.676	1.863	-0.714	1.375	0.432	-0.203
37	37	50	28	Equal	-0.012	0.016	0.107	0.014	0.571	0.892	-0.248	-	-	-	-	-	-
				Unequal	-0.026	0.057	0.087	0.021	-	-	-	0.550	0.538	-0.377	0.572	1.095	-0.226
38	12	9	2	Equal	0.034	0.125	-0.067	0.029	0.293	0.404	-0.336	-	-	-	-	-	-
				Unequal	0.037	0.197	-0.089	0.083	-	-	-	0.189	0.346	0.216	0.301	0.815	-0.860
39	27	27	27	Equal	0.185	0.031	0.004	0.014	0.389	1.000	-0.305	-	-	-	-	-	-

D.3 continued...

ID	N1	N2	NC	Variances	Log relative sensitivity	SE of log relative sensitivity	Log relative specificity	SE of log relative specificity	σ_A^2	σ_B^2	ρ_{AB}	σ_{A1}^2	σ_{B1}^2	ρ_{A1B1}	σ_{A2}^2	σ_{B2}^2	ρ_{A2B2}
				Unequal	0.174	0.060	-0.010	0.032	-	-	-	0.464	0.695	-0.252	0.607	1.519	-0.472
40	89	19	5	Equal	0.077	0.025	0.211	0.060	0.887	0.863	0.218	-	-	-	-	-	-
				Unequal	0.103	0.020	0.223	0.075	-	-	-	1.125	0.901	0.318	0.112	0.727	-0.513
41	24	28	1	Equal	0.160	0.045	-0.016	0.014	0.332	1.060	-0.140	-	-	-	-	-	-
				Unequal	0.143	0.043	0.066	0.020	-	-	-	0.320	1.144	-0.093	0.328	0.381	-0.101
42	4	4	4	Equal	-0.411	0.189	0.121	0.133	1.334	5.302	-0.827	-	-	-	-	-	-
				Unequal	-	-	-	-	-	-	-	-	-	-	-	-	-
43	14	8	4	Equal	0.603	0.225	-0.298	0.179	2.382	2.190	-0.535	-	-	-	-	-	-
				Unequal	0.598	0.483	-0.186	0.249	-	-	-	0.731	1.514	0.089	5.695	2.097	-0.800
44	14	14	14	Equal	0.853	0.138	-0.271	0.047	0.855	0.766	-0.664	-	-	-	-	-	-
				Unequal	0.914	0.193	-0.378	0.119	-	-	-	1.509	1.010	-0.633	0.941	0.894	-0.750
45	12	15	2	Equal	0.016	0.018	0.084	0.034	1.824	1.163	0.362	-	-	-	-	-	-
				Unequal	-0.006	0.021	0.071	0.047	-	-	-	0.095	0.674	1.000	3.250	1.519	0.360
46	12	14	4	Equal	0.106	0.032	0.086	0.028	0.419	0.573	0.107	-	-	-	-	-	-
				Unequal	0.069	0.036	0.061	0.036	-	-	-	0.016	0.007	1.000	0.698	0.920	-0.034
47	6	6	6	Equal	0.092	0.036	0.090	0.033	0.109	0.087	0.502	-	-	-	-	-	-
				Unequal	0.168	0.054	0.075	0.053	-	-	-	0.569	0.053	1.000	0.051	0.305	-0.548
48	5	5	5	Equal	0.076	0.052	0.037	0.035	0.379	1.176	-0.886	-	-	-	-	-	-
				Unequal	0.089	0.068	0.009	0.049	-	-	-	0.464	1.466	-1.000	0.465	1.146	-1.000
49	9	9	9	Equal	0.216	0.052	0.032	0.016	0.260	2.111	-0.028	-	-	-	-	-	-
				Unequal	0.173	0.093	0.129	0.166	-	-	-	0.140	2.499	0.886	0.595	4.868	-0.608
50	10	21	0	Equal	0.099	0.025	0.013	0.020	0.585	1.588	0.259	-	-	-	-	-	-
				Unequal	0.111	0.025	0.007	0.019	-	-	-	1.880	0.511	0.204	0.459	2.074	0.307
51	6	6	6	Equal	-	-	-	-	-	-	-	-	-	-	-	-	-
				Unequal	1.670	0.305	-0.067	0.110	-	-	-	0.052	0.127	-1.000	0.223	0.531	-1.000
52	7	30	7	Equal	-0.414	0.076	-0.008	0.007	0.624	3.902	0.265	-	-	-	-	-	-

D.3 continued...

ID	N1	N2	NC	Variances	Log relative sensitivity	SE of log relative sensitivity	Log relative specificity	SE of log relative specificity	σ_A^2	σ_B^2	ρ_{AB}	σ_{A1}^2	σ_{B1}^2	ρ_{A1B1}	σ_{A2}^2	σ_{B2}^2	ρ_{A2B2}
				Unequal	-0.194	0.134	-0.013	0.009	-	-	-	0.500	2.875	-0.584	0.678	7.921	0.995
53	6	6	6	Equal	0.014	0.091	0.047	0.022	0.008	0.528	-1.000	-	-	-	-	-	-
				Unequal	0.017	0.096	0.053	0.041	-	-	-	0.033	1.137	-1.000	0.080	0.232	-1.000
54	4	4	4	Equal	0.333	0.089	0.033	0.029	0.283	0.025	1.000	-	-	-	-	-	-
				Unequal	0.334	0.108	0.033	0.029	-	-	-	0.000	0.000	-0.451	0.374	0.002	-1.000
55	7	8	3	Equal	-0.016	0.071	0.010	0.015	0.303	0.090	-1.000	-	-	-	-	-	-
				Unequal	0.022	0.076	0.009	0.015	-	-	-	0.549	0.152	-1.000	0.019	0.015	-1.000
56	12	12	12	Equal	0.299	0.052	0.020	0.014	0.262	0.671	-0.057	-	-	-	-	-	-
				Unequal	0.298	0.105	-0.014	0.043	-	-	-	0.581	0.473	0.421	0.314	0.824	-0.679
57	7	7	7	Equal	0.177	0.074	0.073	0.024	0.458	0.630	-0.247	-	-	-	-	-	-
				Unequal	0.127	0.175	0.081	0.084	-	-	-	0.400	1.144	0.123	0.739	0.758	-0.360

N1 = number of studies for index test; N2 = number of studies for comparator; NC = number of comparative studies; SE = standard error; σ_A^2 = variance of logit sensitivity across both index and comparator tests; σ_{A1}^2 = variance of logit sensitivity for index test; σ_{A2}^2 = variance of logit sensitivity for comparator; σ_B^2 = variance of logit specificity across both index and comparator tests; σ_{B1}^2 = variance of logit specificity for index test; σ_{B2}^2 = variance of logit specificity for comparator; ρ_{AB} = correlation of logit sensitivity and logit specificity across both index and comparator tests; ρ_{A1B1} = correlation of logit sensitivity and logit specificity for index test; ρ_{A2B2} = correlation of logit sensitivity and logit specificity for comparator.

D.4| Comparison of bivariate models with different covariance structures fitted to direct test comparisons

ID	Number of studies	Model	Relative sensitivity (95% CI)	Relative specificity (95% CI)
6	17	Equal	1.41 (1.26–1.57)	1.18 (1.11–1.26)
		Unequal & independent	1.38 (1.22–1.55)	1.19 (1.09–1.29)
		Unequal & correlated	1.40 (1.25–1.57)	1.19 (1.11–1.27)
18	24	Equal	1.41 (1.24–1.60)	1.17 (1.10–1.23)
		Unequal & independent	–	–
		Unequal & correlated	1.42 (1.20–1.69)	1.15 (1.05–1.26)
27	13	Equal	1.16 (1.08–1.26)	1.08 (1.03–1.14)
		Unequal & independent	1.16 (1.03–1.30)	1.08 (0.96–1.21)
		Unequal & correlated	1.15 (1.05–1.25)	1.08 (1.05–1.11)
31	11	Equal	1.01 (1.00–1.02)	0.99 (0.98–0.99)
		Unequal & independent	1.01 (0.98–1.05)	0.99 (0.98–1.00)
		Unequal & correlated	1.01 (0.99–1.04)	0.99 (0.98–1.00)
32	18	Equal	0.96 (0.92–0.99)	0.98 (0.96–1.00)
		Unequal & independent	0.97 (0.91–1.03)	0.98 (0.95–1.01)
		Unequal & correlated	0.96 (0.92–1.01)	0.98 (0.96–1.00)
37	28	Equal	0.99 (0.95–1.02)	1.14 (1.10–1.19)
		Unequal & independent	0.96 (0.84–1.10)	1.14 (1.07–1.22)
		Unequal & correlated	0.96 (0.89–1.04)	1.14 (1.08–1.21)
39	27	Equal	1.20 (1.13–1.28)	1.00 (0.98–1.03)
		Unequal & independent	1.19 (1.06–1.34)	0.99 (0.93–1.05)
		Unequal & correlated	1.19 (1.08–1.31)	0.99 (0.95–1.02)
44	14	Equal	2.35 (1.79–3.07)	0.76 (0.69–0.84)
		Unequal & independent	2.49 (1.71–3.64)	0.69 (0.54–0.87)
		Unequal & correlated	2.42 (1.79–3.27)	0.69 (0.58–0.82)
56	12	Equal	1.35 (1.22–1.49)	1.02 (0.99–1.05)
		Unequal & independent	1.35 (1.10–1.66)	0.99 (0.91–1.07)
		Unequal & correlated	1.34 (1.13–1.60)	0.99 (0.93–1.05)

The table shows estimates from three bivariate models that assumed variances of the random effects for logit sensitivities and logit specificities were the same for both tests, allowed for unequal variances and independence between tests, and those that allowed for unequal variances and correlations between tests.

D.5| Estimates of relative sensitivity from HSROC models with common and different shape between tests for SROC curves

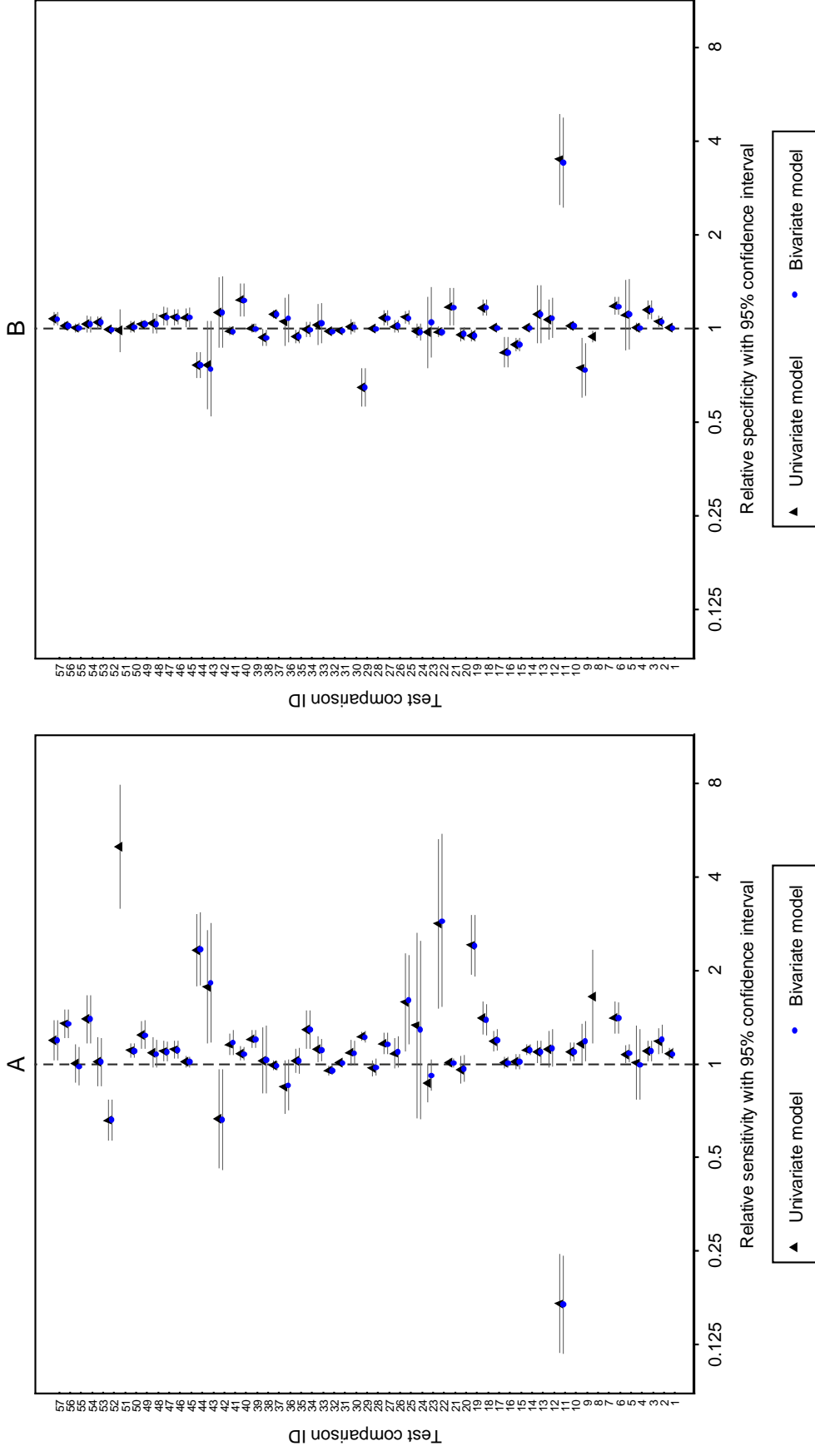
ID	Number of studies			Relative sensitivity (95% CI) at median specificity		P value*
	Index	Comparator	Total	Common shape	Different shape	
1	25	25	47	1.06 (1.01–1.13)	1.07 (1.01–1.13)	0.12
2	4	4	4	1.25 (0.99–1.58)	1.24 (0.98–1.58)	0.76
3	8	8	8	1.12 (1.02–1.23)	1.12 (1.02–1.22)	0.50
5	6	13	19	1.08 (1.01–1.15)	1.05 (0.98–1.13)	0.04
6	17	17	17	1.40 (1.21–1.62)	1.39 (1.14–1.68)	0.87
8	7	3	10	1.23 (0.35–4.35)	1.10 (0.40–3.01)	0.20
9	3	5	7	1.32 (0.77–2.28)	1.22 (0.89–1.68)	0.15
10	6	8	14	1.08 (1.00–1.18)	1.08 (1.00–1.18)	0.93
11	6	11	11	0.24 (0.14–0.39)	0.27 (0.21–0.35)	0.01
12	8	21	28	1.07 (0.90–1.28)	1.10 (0.86–1.40)	0.13
13	4	14	18	1.14 (0.98–1.32)	1.14 (0.98–1.33)	0.91
14	8	23	26	1.10 (1.04–1.17)	1.08 (1.01–1.16)	0.27
15	7	7	7	1.01 (0.91–1.13)	0.97 (0.80–1.16)	<0.001
16	27	23	44	1.01 (0.95–1.08)	1.01 (0.95–1.08)	0.59
17	16	9	16	1.17 (1.06–1.30)	1.18 (1.09–1.27)	0.16
18	33	24	33	1.51 (1.28–1.78)	1.45 (1.24–1.69)	0.22
19	4	4	4	2.35 (1.30–4.24)	2.36 (1.38–4.05)	0.44
20	12	15	19	1.07 (0.92–1.24)	1.09 (0.94–1.26)	0.01
21	12	17	29	1.02 (0.98–1.06)	1.02 (0.98–1.06)	0.76
22	9	5	12	2.57 (1.48–4.47)	2.41 (1.39–4.15)	0.71
23 [†]	7	3	9	0.97 (0.86–1.09)	2.40 (0.09–67.0)	0.02
24	10	17	24	1.45 (0.46–4.53)	1.52 (0.47–4.94)	0.72
25	6	5	6	1.51 (1.15–1.99)	1.40 (0.41–4.76)	0.09
26	18	6	23	1.06 (0.93–1.20)	1.09 (0.90–1.32)	0.25
27	13	13	13	1.14 (1.04–1.25)	1.12 (1.00–1.26)	<0.001
28	3	10	13	0.98 (0.90–1.07)	0.98 (0.88–1.10)	0.44
29	5	5	5	1.22 (1.16–1.29)	–	
30	7	7	7	1.10 (0.98–1.23)	1.10 (0.98–1.23)	0.97
32	11	11	11	1.03 (0.91–1.17)	1.03 (0.91–1.15)	0.80
33	14	15	23	1.14 (1.05–1.23)	1.15 (1.07–1.25)	0.002
34	7	21	23	1.40 (1.09–1.80)	1.35 (1.08–1.68)	0.02
35	10	7	13	1.01 (0.91–1.12)	1.04 (0.94–1.15)	0.001
36	7	15	19	0.72 (0.46–1.13)	0.73 (0.47–1.14)	0.91
37	37	50	59	1.01 (0.96–1.06)	0.97 (0.89–1.06)	<0.001
38	12	9	19	0.98 (0.77–1.25)	0.97 (0.76–1.25)	0.76
39	27	27	27	1.22 (1.12–1.34)	1.22 (1.12–1.34)	1.00
40	89	19	103	1.09 (1.01–1.16)	1.09 (1.03–1.16)	0.20

D.5 continued...

ID	Number of studies			Relative sensitivity (95% CI) at median specificity		P value*
	Index	Comparator	Total	Common shape	Different shape	
41	24	28	51	1.04 (0.94–1.14)	1.06 (0.96–1.18)	<0.001
42	4	4	4	0.83 (0.55–1.24)	1.08 (0.85–1.38)	<0.001
43 [†]	14	8	18	8.67 (0.71–106)	35.11 (0.20–6058)	0.19
44	14	14	14	3.56 (2.36–5.36)	3.01 (2.21–4.10)	<0.001
45	12	15	25	1.03 (0.97–1.10)	1.03 (0.98–1.10)	0.65
46	12	14	22	1.16 (1.03–1.31)	1.16 (1.03–1.31)	0.85
47 [†]	6	6	6	1.09 (0.93–1.28)	0.52 (0–29800000)	0.001
48	5	5	5	1.14 (0.76–1.72)	1.13 (0.92–1.40)	0.28
49	9	9	9	1.23 (1.07–1.43)	1.20 (1.04–1.40)	<0.001
50	10	21	31	1.13 (1.02–1.18)	1.11 (1.03–1.19)	0.02
51 [†]	6	6	6	7.96 (1.19–53.3)	5.33 (2.38–11.9)	0.47
53	6	6	6	1.01 (0.77–1.31)	1.01 (0.79–1.30)	0.05
54 [†]	4	4	4	1.13 (0.08–16.1)	1.02 (0.04–26.5)	0.16
55	7	8	12	0.89 (0.63–1.26)	0.88 (0.60–1.31)	0.91
56	12	12	12	1.32 (1.14–1.53)	1.24 (1.05–1.46)	0.02
57	7	7	7	1.11 (0.90–1.36)	1.11 (0.90–1.37)	0.74

[†]Five test comparisons (IDs 23, 43, 47, 51 and 54) had estimates with extremely wide confidence intervals and were considered to be potentially unreliable. Therefore, they were excluded from comparisons of the common and different shape HSROC models in section 7.6.

D.6| Comparison of relative sensitivity and relative specificity from bivariate and univariate models with equal variances



The plots show estimates from univariate models (black triangles) and bivariate models (blue circles) that assumed variances of the random effects for the logit sensitivities and specificities were the same for both tests in a comparative meta-analysis. For IDs 8 and 51, only the univariate model converged. For ID 7, neither of the models converged. The dashed line on each plot is the line of no difference in test performance between the index and comparator tests in a test comparison.

D.7| Estimates of relative accuracy from bivariate and univariate models with unequal variances

ID	Bivariate model		Univariate model		P value*
	Relative sensitivity (95% CI)	Relative specificity (95% CI)	Relative sensitivity (95% CI)	Relative specificity (95% CI)	
1	1.11 (1.04–1.18)	1.00 (0.97–1.02)	1.11 (1.04–1.18)	1.00 (0.97–1.02)	0.05
2	–	–	1.15 (1.03–1.29)	1.03 (0.82–1.29)	–
3	1.11 (1.00–1.24)	1.12 (1.01–1.25)	1.11 (1.00–1.23)	1.12 (1.01–1.25)	0.34
4	0.94 (0.69–1.27)	1.00 (0.97–1.04)	0.94 (0.68–1.29)	1.01 (0.98–1.04)	0.68
5	1.09 (1.02–1.17)	1.13 (0.86–1.47)	1.08 (1.01–1.16)	1.13 (0.86–1.47)	0.26
6	1.38 (1.22–1.55)	1.19 (1.09–1.29)	1.38 (1.23–1.54)	1.19 (1.09–1.29)	0.06
7	–	–	–	–	–
8	–	–	1.70 (1.20–2.4)	0.95 (0.90–0.99)	–
9	1.19 (1.01–1.40)	0.71 (0.57–0.89)	1.19 (1.02–1.39)	0.71 (0.57–0.89)	0.54
10	1.09 (1.02–1.17)	1.02 (0.99–1.05)	1.10 (1.02–1.17)	1.02 (0.99–1.05)	0.62
11	–	–	0.32 (0.25–0.42)	3.29 (2.23–4.87)	–
12	1.11 (0.89–1.37)	1.06 (0.91–1.24)	1.13 (0.96–1.34)	1.01 (0.85–1.20)	0.07
13	1.09 (1.00–1.18)	1.08 (0.86–1.37)	1.08 (0.99–1.18)	1.09 (0.86–1.38)	0.71
14	1.08 (1.02–1.13)	1.00 (0.98–1.03)	1.08 (1.02–1.13)	1.00 (0.98–1.03)	0.85
15	1.01 (0.89–1.13)	0.92 (0.81–1.03)	0.99 (0.88–1.12)	0.92 (0.82–1.03)	0.01
16	1.02 (0.97–1.07)	0.84 (0.74–0.95)	1.02 (0.97–1.07)	0.84 (0.75–0.95)	0.22
17	1.19 (1.11–1.27)	1.02 (0.98–1.06)	1.18 (1.11–1.27)	1.02 (0.98–1.06)	0.08
18	1.43 (1.19–1.74)	1.13 (1.03–1.24)	1.42 (1.19–1.71)	1.12 (1.03–1.23)	<0.0001
19	–	–	2.41 (1.94–3.00)	0.95 (0.89–1.02)	–
20	1.24 (0.91–1.69)	0.96 (0.87–1.05)	1.24 (0.91–1.69)	0.96 (0.87–1.05)	0.01
21	1.00 (0.97–1.03)	1.16 (1.01–1.33)	1.00 (0.97–1.03)	1.16 (1.01–1.33)	0.86
22	2.22 (1.92–2.58)	0.99 (0.95–1.02)	2.22 (1.91–2.58)	0.99 (0.95–1.02)	0.99
23	–	–	0.88 (0.72–1.07)	1.25 (0.87–1.79)	–
24	0.87 (0.35–2.12)	0.98 (0.93–1.03)	0.95 (0.40–2.21)	0.98 (0.93–1.03)	0.06
25	1.56 (0.87–2.78)	1.05 (0.94–1.17)	1.50 (0.89–2.54)	1.06 (0.98–1.16)	0.33
26	1.08 (0.98–1.20)	1.06 (1.01–1.11)	1.09 (0.99–1.20)	1.06 (1.01–1.10)	0.06
27	1.16 (1.03–1.30)	1.08 (0.96–1.21)	1.16 (1.03–1.31)	1.08 (0.97–1.21)	0.26
28	0.98 (0.93–1.03)	1.00 (0.99–1.01)	0.97 (0.92–1.03)	1.00 (0.99–1.01)	0.67
29	1.21 (1.12–1.31)	0.77 (0.65–0.92)	1.20 (1.12–1.28)	0.77 (0.64–0.91)	0.02
30	1.10 (0.95–1.26)	1.00 (0.93–1.07)	1.10 (0.95–1.28)	1.01 (0.94–1.09)	0.02
31	1.01 (0.98–1.05)	0.99 (0.98–1.00)	1.01 (0.98–1.05)	0.99 (0.98–1.00)	0.93
32	0.97 (0.91–1.03)	0.98 (0.95–1.01)	0.97 (0.91–1.03)	0.98 (0.95–1.00)	0.55
33	1.14 (1.02–1.26)	0.79 (0.57–1.10)	1.11 (0.99–1.24)	0.78 (0.56–1.09)	0.01
34	1.27 (1.09–1.48)	0.94 (0.85–1.03)	–	–	–
35	1.06 (0.93–1.21)	1.04 (0.90–1.20)	1.05 (0.91–1.20)	1.04 (0.90–1.2)	0.57
36	0.98 (0.73–1.30)	1.21 (0.94–1.55)	0.99 (0.75–1.31)	1.23 (0.96–1.58)	0.23
37	0.97 (0.87–1.09)	1.09 (1.05–1.14)	0.98 (0.87–1.09)	1.09 (1.05–1.14)	0.06
38	1.04 (0.71–1.53)	0.92 (0.78–1.08)	1.00 (0.79–1.26)	0.92 (0.78–1.07)	0.35

D.7 continued...

ID	Bivariate model		Univariate model		P value*
	Relative sensitivity (95% CI)	Relative specificity (95% CI)	Relative sensitivity (95% CI)	Relative specificity (95% CI)	
39	1.19 (1.06–1.34)	0.99 (0.93–1.05)	1.19 (1.06–1.34)	0.99 (0.93–1.06)	0.19
40	1.11 (1.06–1.15)	1.25 (1.08–1.45)	1.11 (1.06–1.15)	1.25 (1.08–1.44)	0.09
41	1.15 (1.06–1.25)	1.07 (1.03–1.11)	1.15 (1.06–1.25)	1.07 (1.03–1.11)	0.88
42	–	–	0.79 (0.47–1.33)	1.13 (0.79–1.61)	–
43	1.82 (0.71–4.68)	0.83 (0.51–1.35)	1.69 (0.75–3.79)	0.79 (0.50–1.24)	0.21
44	2.49 (1.71–3.64)	0.69 (0.54–0.87)	2.49 (1.70–3.65)	0.69 (0.54–0.87)	0.001
45	0.99 (0.95–1.04)	1.07 (0.98–1.18)	1.00 (0.96–1.04)	1.07 (0.98–1.17)	0.52
46	1.07 (1.00–1.15)	1.06 (0.99–1.14)	1.07 (1.00–1.15)	1.06 (0.99–1.14)	0.98
47	1.18 (1.06–1.31)	1.08 (0.97–1.19)	1.18 (1.05–1.31)	1.07 (0.97–1.18)	0.43
48	1.09 (0.96–1.25)	1.01 (0.92–1.11)	1.09 (0.99–1.20)	1.02 (0.93–1.13)	0.17
49	1.19 (0.99–1.43)	1.14 (0.82–1.57)	1.23 (1.05–1.45)	1.13 (0.83–1.56)	0.11
50	1.12 (1.06–1.17)	1.01 (0.97–1.04)	1.12 (1.07–1.17)	1.01 (0.97–1.04)	0.47
51	5.31 (2.92–9.66)	0.93 (0.75–1.16)	5.04 (3.05–8.33)	0.94 (0.75–1.16)	0.35
52	0.82 (0.63–1.07)	0.99 (0.97–1.01)	–	–	–
53	1.02 (0.84–1.23)	1.05 (0.97–1.14)	1.02 (0.85–1.21)	1.06 (0.98–1.14)	0.61
54	1.40 (1.13–1.73)	1.03 (0.98–1.09)	1.40 (1.13–1.73)	1.03 (0.98–1.09)	0.99
55	1.02 (0.88–1.19)	1.01 (0.98–1.04)	1.03 (0.90–1.18)	1.01 (0.98–1.04)	0.18
56	1.35 (1.10–1.66)	0.99 (0.91–1.07)	1.35 (1.10–1.66)	1.00 (0.91–1.09)	0.09
57	1.14 (0.81–1.60)	1.08 (0.92–1.28)	1.12 (0.80–1.56)	1.08 (0.92–1.28)	0.69

*P value from likelihood ratio tests comparing both models.

– indicates missing estimates due to lack of convergence of the model.

Univariate models were estimated by assuming an independent variance-covariance structure, i.e. correlation of the logits = 0.

D.8| Estimates of variance and correlation parameters from bivariate and univariate models with unequal variances

ID	Bivariate model						Univariate model			
	σ_{A1}^2	σ_{B1}^2	ρ_{A1B1}	σ_{A2}^2	σ_{B2}^2	ρ_{A2B2}	σ_{A1}^2	σ_{B1}^2	σ_{A2}^2	σ_{B2}^2
1	1.193	0.528	-0.020	0.962	0.909	0.533	1.190	0.527	0.947	0.877
2	-	-	-	-	-	-	0.176	0.158	0.050	0.159
3	0.121	0.601	0.266	0.261	0.476	-0.651	0.109	0.604	0.258	0.474
4	0.000	0.000	-0.777	1.610	0.466	-1.000	0.000	0.000	1.643	0.000
5	0.433	2.946	-0.727	0.198	0.475	-0.071	0.333	2.909	0.197	0.477
6	0.268	0.760	0.045	0.099	0.503	-1.000	0.263	0.757	0.048	0.503
7	-	-	-	-	-	-	-	-	-	-
8	-	-	-	-	-	-	3.212	0.940	0.000	0.000
9	0.529	0.092	0.406	0.103	0.361	1.000	0.479	0.107	0.077	0.332
10	0.131	0.291	-0.277	0.150	0.361	-0.455	0.150	0.305	0.137	0.357
11	-	-	-	-	-	-	0.147	0.000	0.600	0.611
12	3.369	0.631	1.000	0.433	1.626	0.158	2.113	0.612	0.448	1.625
13	0.324	1.198	-1.000	0.740	2.374	0.128	0.000	1.329	0.770	2.370
14	0.264	0.093	0.031	0.471	0.238	0.204	0.264	0.093	0.475	0.241
15	0.382	0.123	-0.810	0.406	0.568	-1.000	0.297	0.121	0.381	0.533
16	0.374	0.512	-0.375	0.284	0.570	-0.282	0.381	0.509	0.282	0.554
17	0.530	0.915	-0.834	0.087	1.029	-0.484	0.508	0.689	0.104	1.029
18	0.222	0.315	0.213	0.881	0.968	-1.000	0.207	0.303	0.745	0.855
19	-	-	-	-	-	-	0.000	0.722	0.000	2.305
20	1.243	5.299	-0.780	0.452	2.010	-0.379	1.226	4.869	0.456	1.911
21	0.044	0.101	0.924	1.427	1.224	-0.184	0.047	0.106	1.353	1.235
22	1.037	0.587	-0.108	0.000	0.000	-0.998	1.033	0.586	0.000	0.000
23	-	-	-	-	-	-	0.870	2.324	0.164	0.060
24	1.658	1.565	0.342	2.419	2.135	-0.798	1.376	1.425	2.403	1.826
25	2.187	0.374	0.534	1.578	0.342	-0.787	2.026	0.308	1.343	0.086
26	0.606	2.025	-0.068	0.301	0.235	-1.000	0.604	2.060	0.208	0.162
27	0.339	3.366	0.530	0.308	1.075	-0.244	0.319	3.395	0.316	1.076
28	0.471	0.415	0.500	0.000	0.000	1.000	0.404	0.412	0.000	0.000
29	5.512	0.290	-0.653	0.137	0.461	-1.000	7.114	0.296	0.088	0.494
30	1.687	1.471	-1.000	1.532	1.280	-1.000	1.042	0.912	1.149	0.708
31	0.831	0.586	0.134	1.154	2.888	0.079	0.841	0.579	1.155	2.869
32	0.745	0.193	0.492	0.508	0.348	-0.258	0.726	0.175	0.525	0.357
33	0.116	0.291	-1.000	0.672	1.022	-0.835	0.000	0.294	1.112	1.168
34	0.141	0.578	-1.000	1.468	2.134	0.097	-	-	-	-
35	0.018	0.522	0.446	0.376	0.324	-0.470	0.019	0.526	0.437	0.335
36	2.676	1.863	-0.714	1.375	0.432	-0.203	2.852	2.173	1.376	0.424
37	0.550	0.538	-0.377	0.572	1.095	-0.226	0.562	0.548	0.566	1.089
38	0.189	0.346	0.216	0.301	0.815	-0.860	0.195	0.348	0.000	0.672
39	0.464	0.695	-0.252	0.607	1.519	-0.472	0.490	0.724	0.585	1.578
40	1.125	0.901	0.318	0.112	0.727	-0.513	1.114	0.886	0.139	0.752

D.8 continued...

ID	Bivariate model						Univariate model			
	σ_{A1}^2	σ_{B1}^2	ρ_{A1B1}	σ_{A2}^2	σ_{B2}^2	ρ_{A2B2}	σ_{A1}^2	σ_{B1}^2	σ_{A2}^2	σ_{B2}^2
41	0.320	1.144	-0.093	0.328	0.381	-0.101	0.324	1.145	0.329	0.382
42	–	–	–	–	–	–	2.950	4.987	0.062	7.887
43	0.731	1.514	0.089	5.695	2.097	-0.800	0.750	1.564	4.867	1.981
44	1.509	1.010	-0.633	0.941	0.894	-0.750	1.510	1.014	0.948	0.911
45	0.095	0.674	1.000	3.250	1.519	0.360	0.000	0.667	2.766	1.516
46	0.016	0.007	1.000	0.698	0.920	-0.034	0.000	0.000	0.695	0.917
47	0.569	0.053	1.000	0.051	0.305	-0.548	0.598	0.000	0.059	0.270
48	0.464	1.466	-1.000	0.465	1.146	-1.000	0.000	1.178	0.000	1.021
49	0.140	2.499	0.886	0.595	4.868	-0.608	0.109	2.498	0.380	4.746
50	1.880	0.511	0.204	0.459	2.074	0.307	1.905	0.501	0.474	2.046
51	0.052	0.127	-1.000	0.223	0.531	-1.000	0.000	0.069	0.029	0.469
52	0.500	2.875	-0.584	0.678	7.921	0.995	–	–	–	–
53	0.033	1.137	-1.000	0.080	0.232	-1.000	0.000	1.206	0.000	0.227
54	0.000	0.000	-0.451	0.374	0.002	-1.000	0.000	0.000	0.374	0.000
55	0.549	0.152	-1.00	0.019	0.015	-1.00	0.511	0.145	0.000	0.000
56	0.581	0.473	0.421	0.314	0.824	-0.679	0.566	0.506	0.292	0.761
57	0.400	1.144	0.123	0.739	0.758	-0.360	0.380	1.140	0.736	0.759

σ_{A1}^2 = variance of logit sensitivity for index test; σ_{A2}^2 = variance of logit sensitivity for comparator; σ_{B1}^2 = variance of logit specificity for index test; σ_{B2}^2 = variance of logit specificity for comparator; ρ_{A1B1} = correlation of logit sensitivity and logit specificity for index test; ρ_{A2B2} = correlation of logit sensitivity and logit specificity for comparator.

D.9 | Estimates from unweighted and weighted Moses SROC meta-regression models

ID	Same shape across tests		Different shape per test	
	Relative diagnostic odds ratio (95% CI)		Relative sensitivity (95% CI)	
	Unweighted	Weighted	Unweighted	Weighted
1	2.37 (0.92–6.07)	2.07 (0.70–6.13)	1.12 (0.93–1.35)	1.23 (0.99–1.52)
2	2.90 (1.54–5.43)	3.65 (2.49–5.35)	1.25 (1.11–1.40)	1.24 (1.18–1.30)
3	5.50 (1.86–16.3)	3.84 (1.68–8.79)	0.82 (0.12–5.57)	1.11 (0.68–1.82)
4	0.87 (0.19–3.92)	0.71 (0.19–2.68)	2.08 (0–5453332)*	1.42 (0–6788074)*
5	2.69 (0.88–8.23)	3.03 (1.57–5.85)	0.94 (0.68–1.31)	1.00 (0.84–1.20)
6	13.3 (5.89–30.1)	9.93 (5.19–19.0)	1.49 (0.88–2.53)	1.43 (0.89–2.30)
7	10.4 (1.97–54.8)	4.75 (1.44–15.7)	367 (0.07–1945765)*	210 (1.19–36975)*
8	1.33 (0.01–243)	0.92 (0.00–408)	1.65 (1.22–2.23)	1.60 (1.22–2.10)
9	1.17 (0.02–79.1)	0.50 (0.03–9.07)	1.32 (1.04–1.67)	1.34 (0.94–1.91)
10	4.02 (1.29–12.5)	3.54 (1.40–8.94)	1.18 (1.07–1.31)	1.07 (0.59–1.96)
11	1.13 (0.17–7.45)	0.75 (0.21–2.67)	0.37 (0.03–5.24)	1.06 (0.79–1.42)
12	2.05 (0.36–11.7)	1.67 (0.58–4.85)	1.15 (0.77–1.71)	1.41 (1.20–1.65)
13	4.59 (0.91–23.1)	8.69 (0.97–77.5)	1.15 (0.94–1.39)	1.25 (0.97–1.61)
14	3.15 (1.14–8.71)	1.78 (0.61–5.22)	1.02 (0.67–1.55)	1.07 (0.62–1.86)
15	0.61 (0.30–1.24)	0.62 (0.33–1.18)	1.16 (1.06–1.27)	1.19 (1.06–1.33)
16	0.64 (0.39–1.06)	0.77 (0.44–1.35)	1.04 (0.96–1.13)	1.06 (0.98–1.14)
17	8.14 (3.75–17.7)	10.2 (4.73–21.8)	1.27 (1.15–1.39)	1.27 (1.18–1.37)
18	8.16 (4.37–15.3)	5.96 (3.80–9.34)	1.48 (0.95–2.29)	1.42 (0.95–2.11)
19	83.4 (2.64–2633)	33.4 (5.61–199)	2.79 (1.23–6.31)	2.28 (1.43–3.63)
20	0.92 (0.38–2.24)	0.94 (0.35–2.54)	1.44 (1.02–2.04)	1.60 (0.97–2.64)
21	2.66 (0.88–8.05)	3.89 (1.29–11.7)	0.94 (0.69–1.29)	1.01 (0.74–1.38)
22	3.09 (0.30–31.9)	1.89 (0.21–17.1)	0.00 (0–1.5E+28)*	0.33 (0–3.17E+11)*
23	0.79 (0.30–2.12)	0.90 (0.39–2.05)	1.04 (1.01–1.07)	1.02 (1.00–1.03)
24	0.88 (0.22–3.46)	0.36 (0.11–1.16)	1.07 (0.02–55.7)	0.40 (0.01–11.1)
25	2.86 (0.48–17.2)	5.05 (1.29–19.8)	–	0.46 (0–54500000)*
26	4.22 (1.18–15.1)	3.24 (1.26–8.35)	0.79 (0.06–11.2)	0.95 (0.20–4.57)
27	4.14 (1.29–13.3)	2.51 (0.98–6.42)	0.82 (0.41–1.63)	1.07 (0.78–1.47)
28	0.95 (0.19–4.71)	0.40 (0.09–1.74)	1.02 (0.93–1.12)	–
29	0.51 (0.08–3.20)	0.75 (0.26–2.18)	1.23 (1.13–1.34)	1.26 (1.22–1.30)
30	3.40 (0.93–12.5)	3.04 (0.95–9.72)	1.00 (0.37–2.65)	1.08 (0.79–1.47)
31	0.58 (0.22–1.58)	0.95 (0.34–2.62)	110 (0.46–26502)*	1.64 (0.46–5.81)
32	0.41 (0.13–1.27)	0.33 (0.14–0.73)	1.11 (0.98–1.25)	1.14 (0.95–1.35)
33	1.10 (0.51–2.35)	1.93 (0.96–3.89)	1.02 (0.93–1.12)	1.10 (0.96–1.25)
34	2.91 (0.49–17.4)	8.77 (0.97–79.6)	1.24 (0.84–1.82)	1.67 (1.17–2.38)
35	1.40 (0.49–3.98)	1.66 (0.77–3.60)	0.45 (0.04–5.75)	0.88 (0.52–1.51)
36	2.22 (0.59–8.32)	1.46 (0.50–4.26)	0.76 (0.26–2.24)	0.53 (0.08–3.48)
37	2.55 (1.48–4.39)	2.90 (1.85–4.56)	1.26 (0.93–1.72)	1.37 (1.11–1.69)
38	0.57 (0.26–1.26)	0.67 (0.32–1.44)	0.87 (0.50–1.51)	1.09 (0.62–1.93)
39	2.63 (1.27–5.47)	2.04 (1.19–3.50)	1.35 (1.14–1.61)	1.27 (0.94–1.73)
40	9.47 (4.55–19.7)	5.06 (2.69–9.51)	1.13 (1.05–1.22)	1.14 (1.05–1.24)

D.9 continued...

ID	Same shape across tests		Different shape per test	
	Relative diagnostic odds ratio (95% CI)		Relative sensitivity (95% CI)	
	Unweighted	Weighted	Unweighted	Weighted
41	4.49 (2.39–8.45)	4.71 (2.70–8.21)	0.96 (0.50–1.85)	0.80 (0.29–2.16)
42	<i>0.69 (0.05–9.51)</i>	<i>0.34 (0.04–3.20)</i>	1.10 (0.95–1.26)	1.08 (0.90–1.31)
43	3.42 (0.73–16.1)	3.25 (0.64–16.4)	<i>137 (1.59–11805)*</i>	<i>1270 (0.24–6855656)*</i>
44	2.01 (0.68–6.00)	1.48 (0.67–3.30)	3.15 (2.04–4.86)	2.67 (1.84–3.87)
45	3.13 (0.83–11.8)	4.15 (1.30–13.3)	0.98 (0.69–1.39)	0.96 (0.53–1.74)
46	4.77 (2.03–11.2)	9.42 (2.85–31.1)	0.79 (0.24–2.65)	1.20 (0.73–2.00)
47	5.00 (1.77–14.2)	3.06 (1.07–8.75)	<i>1.34 (1.17–1.53)</i>	<i>0 (0–2.96E+15)*</i>
48	3.33 (0.73–15.2)	2.67 (0.53–13.5)	1.13 (0.48–2.68)	1.06 (0.46–2.42)
49	5.60 (1.73–18.1)	6.16 (1.86–20.5)	1.06 (0.89–1.28)	1.13 (0.87–1.45)
50	7.56 (2.00–28.7)	6.66 (0.96–46.2)	1.14 (1.07–1.20)	1.21 (1.15–1.27)
51	21.5 (4.90–94.1)	24.1 (7.45–77.6)	6.28 (3.76–10.5)	6.81 (3.66–12.7)
52	0.19 (0.05–0.68)	0.20 (0.08–0.54)	<i>58.0 (1.74–1929)*</i>	<i>18.8 (0.44–809)*</i>
53	2.64 (0.84–8.25)	1.57 (0.64–3.80)	1.06 (0.61–1.83)	0.76 (0.22–2.62)
54	<i>26.1 (3.10–220)</i>	<i>10.7 (1.64–70.0)</i>	<i>0.99 (0.00–204)</i>	<i>0 (0–4.36E+16)*</i>
55	1.28 (0.68–2.41)	1.25 (0.69–2.28)	1.11 (0.42–2.88)	1.18 (1.06–1.30)
56	2.72 (1.05–7.08)	1.54 (0.60–3.96)	1.54 (0.93–2.56)	1.71 (1.28–2.28)
57	2.26 (0.48–10.7)	1.64 (0.53–5.11)	0.43 (0.03–6.05)	0.74 (0.12–4.44)

*Potentially unreliable estimates.

– indicates the relative sensitivity and its 95% CI were not estimable using the *nlcom* command in Stata post estimation of the Moses model. Estimates in italics were those with more than a two-fold difference between both models.

D.10| Estimates from unweighted Moses SROC and HSROC meta-regression models

ID	Same shape across tests		Different shape per test	
	Relative diagnostic odds ratio (95% CI)		Relative sensitivity (95% CI)	
	Unweighted Moses model	HSROC model	Unweighted Moses model	HSROC model
1	2.37 (0.92–6.07)	2.72 (1.52–4.86)	1.12 (0.93–1.35)	1.07 (1.01–1.13)
2	2.90 (1.54–5.43)	3.84 (2.11–6.99)	1.25 (1.11–1.40)	1.24 (0.98–1.58)
3	5.50 (1.86–16.3)	5.22 (2.74–9.94)	0.82 (0.12–5.57)	1.12 (1.02–1.22)
4	0.87 (0.19–3.92)	3.14 (0.02–528)	2.08 (0–5453332)*	–
5	2.69 (0.88–8.23)	4.60 (1.67–12.7)	0.94 (0.68–1.31)	1.05 (0.98–1.13)
6	13.3 (5.89–30.1)	18.5 (9.81–34.8)	1.49 (0.88–2.53)	1.39 (1.14–1.68)
7	10.4 (1.97–54.8)	–	367 (0.07–1945765)*	–
8	1.33 (0.01–243)	0.35 (0.00–25.9)	1.65 (1.22–2.23)	1.10 (0.40–3.01)
9	1.17 (0.02–79.1)	0.38 (0.02–9.29)	1.32 (1.04–1.67)	1.22 (0.89–1.68)
10	4.02 (1.29–12.5)	4.08 (1.77–9.36)	1.18 (1.07–1.31)	1.08 (1.00–1.18)
11	1.13 (0.17–7.45)	0.73 (0.06–8.65)	0.37 (0.03–5.24)	0.27 (0.21–0.35)
12	2.05 (0.36–11.7)	3.86 (0.88–16.9)	1.15 (0.77–1.71)	1.10 (0.86–1.40)
13	4.59 (0.91–23.1)	17.2 (0.87–342)	1.15 (0.94–1.39)	1.14 (0.98–1.33)
14	3.15 (1.14–8.71)	4.73 (2.22–10.1)	1.02 (0.67–1.55)	1.08 (1.01–1.16)
15	0.61 (0.30–1.24)	0.57 (0.36–0.90)	1.16 (1.06–1.27)	0.97 (0.80–1.16)
16	0.64 (0.39–1.06)	0.59 (0.36–0.98)	1.04 (0.96–1.13)	1.01 (0.95–1.08)
17	8.14 (3.75–17.7)	7.07 (3.03–16.5)	1.27 (1.15–1.39)	1.18 (1.09–1.27)
18	8.16 (4.37–15.3)	9.14 (5.97–14.0)	1.48 (0.95–2.29)	1.45 (1.24–1.69)
19	83.4 (2.64–2633)	2938 (0–3.03E+14)*	2.79 (1.23–6.31)	2.36 (1.38–4.05)
20	0.92 (0.38–2.24)	0.47 (0.29–0.78)	1.44 (1.02–2.04)	1.09 (0.94–1.26)
21	2.66 (0.88–8.05)	4.23 (1.08–16.5)	0.94 (0.69–1.29)	1.02 (0.98–1.06)
22	3.09 (0.30–31.9)	4.62 (0.43–49.4)	0.00 (0–1.5E+28)*	2.41 (1.39–4.15)
23	0.79 (0.30–2.12)	0.66 (0.32–1.34)	1.04 (1.01–1.07)	2.40 (0.09–67.0)
24	0.88 (0.22–3.46)	0.75 (0.16–3.40)	1.07 (0.02–55.7)	1.52 (0.47–4.94)
25	2.86 (0.48–17.2)	7.42 (4.59–12.0)	–	1.40 (0.41–4.76)
26	4.22 (1.18–15.1)	3.45 (1.16–10.2)	0.79 (0.06–11.2)	1.09 (0.90–1.32)
27	4.14 (1.29–13.3)	4.86 (2.86–8.26)	0.82 (0.41–1.63)	1.12 (1.00–1.26)
28	0.95 (0.19–4.71)	0.58 (0.08–4.04)	1.02 (0.93–1.12)	0.98 (0.88–1.10)
29	0.51 (0.08–3.20)	40438 (1.94–8.42E+8)*	1.23 (1.13–1.34)	–
30	3.40 (0.93–12.5)	4.51 (1.33–15.3)	1.00 (0.37–2.65)	1.10 (0.98–1.23)
31	0.58 (0.22–1.58)	0.20 (0.07–0.52)	110 (0.46–26502)*	–
32	0.41 (0.13–1.27)	0.35 (0.19–0.64)	1.11 (0.98–1.25)	1.03 (0.91–1.15)
33	1.10 (0.51–2.35)	3.11 (1.26–7.67)	1.02 (0.93–1.12)	1.15 (1.07–1.25)
34	2.91 (0.49–17.4)	21.4 (3.10–148)	1.24 (0.84–1.82)	1.35 (1.08–1.68)
35	1.40 (0.49–3.98)	0.91 (0.48–1.72)	0.45 (0.04–5.75)	1.04 (0.94–1.15)
36	2.22 (0.59–8.32)	0.74 (0.25–2.14)	0.76 (0.26–2.24)	0.73 (0.47–1.14)
37	2.55 (1.48–4.39)	3.34 (2.60–4.28)	1.26 (0.93–1.72)	0.97 (0.89–1.06)
38	0.57 (0.26–1.26)	0.85 (0.43–1.69)	0.87 (0.50–1.51)	0.97 (0.76–1.25)

D.10 continued...

ID	Same shape across tests		Different shape per test	
	Relative diagnostic odds ratio (95% CI)		Relative sensitivity (95% CI)	
	Unweighted Moses model	HSROC model	Unweighted Moses model	HSROC model
39	2.63 (1.27–5.47)	2.92 (1.90–4.49)	1.35 (1.14–1.61)	1.22 (1.12–1.34)
40	9.47 (4.55–19.7)	10.4 (5.22–20.8)	1.13 (1.05–1.22)	1.09 (1.03–1.16)
41	4.49 (2.39–8.45)	2.50 (1.35–4.64)	0.96 (0.50–1.85)	1.06 (0.96–1.18)
42	<i>0.69 (0.05–9.51)</i>	<i>0.25 (0.05–1.30)</i>	1.10 (0.95–1.26)	1.08 (0.85–1.38)
43	3.42 (0.73–16.1)	2.89 (0.49–16.9)	<i>137 (1.59–11805)*</i>	<i>35.1 (0.20–6058)*</i>
44	2.01 (0.68–6.00)	3.19 (1.55–6.55)	3.15 (2.04–4.86)	3.01 (2.21–4.10)
45	3.13 (0.83–11.8)	5.84 (1.30–26.3)	0.98 (0.69–1.39)	1.03 (0.98–1.10)
46	<i>4.77 (2.03–11.2)</i>	<i>10.4 (4.31–25.1)</i>	0.79 (0.24–2.65)	1.16 (1.03–1.31)
47	5.00 (1.77–14.2)	3.57 (1.88–6.79)	<i>1.34 (1.17–1.53)</i>	<i>0.52 (0–29800000)*</i>
48	3.33 (0.73–15.2)	4.46 (0.94–21.1)	1.13 (0.48–2.68)	1.13 (0.92–1.40)
49	5.60 (1.73–18.1)	7.45 (2.78–20.0)	1.06 (0.89–1.28)	1.20 (1.04–1.40)
50	7.56 (2.00–28.7)	8.78 (2.11–36.4)	1.14 (1.07–1.20)	1.11 (1.03–1.19)
51	21.5 (4.90–94.1)	35.6 (0.45–2841)	6.28 (3.76–10.5)	5.33 (2.38–11.9)
52	<i>0.19 (0.05–0.68)</i>	<i>0.02 (0.00–0.06)</i>	58.0 (1.74–1929)*	–
53	2.64 (0.84–8.25)	1.48 (0.14–15.1)	1.06 (0.61–1.83)	1.01 (0.79–1.30)
54	26.1 (3.10–220)	23.9 (1.02–558)	0.99 (0.00–204)	1.02 (0.04–26.5)
55	1.28 (0.68–2.41)	1.24 (0.48–3.18)	1.11 (0.42–2.88)	0.88 (0.60–1.31)
56	2.72 (1.05–7.08)	3.34 (2.07–5.41)	1.54 (0.93–2.56)	1.24 (1.05–1.46)
57	2.26 (0.48–10.7)	2.58 (1.66–4.00)	<i>0.43 (0.03–6.05)</i>	<i>1.11 (0.90–1.37)</i>

*Potentially unreliable estimates.

– indicates the HSROC model did not converge or relative sensitivity and its 95% CI were not estimable using the *nlcom* command in Stata post estimation of the Moses model. Estimates in italics were those with more than a two-fold difference between both models.

Appendix E: Additional simulation results

E.1| Performance of all meta-analytic models in estimating sensitivity for scenarios with a DOR of 231

Studies	Heterogeneity*	Meta-analytic model	5% prevalence				25% prevalence				50% prevalence			
			N	Bias (%)†	MSE	Coverage (%)	N	Bias (%)	MSE	Coverage (%)	N	Bias (%)	MSE	Coverage (%)
5	No	Complete HSROC	1767	25.4	6.06	98.5	2086	6.18	0.31	98.7	2114	3.25	0.10	98.4
		Symmetric HSROC	1276	6.34	4.10	94.7	2753	1.30	0.15	97.4	4179	2.60	0.08	98.3
		FA	1932	70.1	31.4	98.1	2228	7.17	0.47	98.9	2232	3.49	0.11	98.6
		FT	1792	73.6	32.1	98.5	2174	7.29	0.52	98.4	2288	3.25	0.12	97.8
		FAT	9798	59.7	27.9	96.7	9556	2.83	0.32	96.0	4722	1.45	0.08	94.7
		SFAT	10000	73.5	40.6	88.1	10000	2.64	0.36	95.8	10000	1.17	0.08	95.5
		UREM	3173	20.9	5.53	98.0	2869	5.70	0.31	97.1	2883	2.64	0.10	97.4
		Complete HSROC	3020	80.3	41.8	97.5	4339	3.99	0.52	94.9	4772	1.21	0.20	92.9
5	Yes	Symmetric HSROC	3442	39.5	20.5	93.6	6331	-0.45	0.26	93.5	7594	0.53	0.15	93.0
		FA	5490	81.0	42.9	96.8	6222	4.98	0.77	94.4	6331	1.46	0.20	92.4
		FT	2976	103	52.0	96.5	2266	6.42	1.08	93.7	2171	1.82	0.22	89.3
		FAT	9691	48.4	23.7	92.8	9288	-1.07	0.60	84.7	4833	-2.24	0.20	73.4
		SFAT	10000	60.4	34.0	85.3	10000	-1.35	0.71	83.9	10000	-3.32	0.18	75.1
		UREM	5833	23.7	6.56	96.2	6311	3.41	0.39	93.0	6573	0.99	0.20	91.0
		Complete HSROC	2325	14.3	0.88	98.7	2903	3.75	0.11	98.4	2862	1.68	0.04	97.3
		Symmetric HSROC	1924	1.95	0.34	96.4	3581	1.48	0.07	97.3	5383	1.39	0.04	97.3
10	No	FA	2175	20.7	4.42	98.3	2569	3.92	0.11	98.6	2719	1.82	0.04	97.2
		FT	2177	19.9	4.45	98.7	2670	3.51	0.11	98.0	2666	1.62	0.04	97.3
		FAT	9881	10.8	3.43	96.8	9594	1.07	0.07	95.8	4596	0.12	0.03	96.5
		SFAT	10000	12.5	4.93	95.7	10000	0.94	0.07	95.7	10000	0.45	0.03	95.2
		UREM	5612	12.4	0.98	98.3	5417	2.58	0.09	97.1	5311	1.23	0.04	96.4
		Complete HSROC	4129	18.8	4.63	97.8	6248	1.74	0.14	95.2	7136	0.65	0.09	93.4
		Symmetric HSROC	5895	4.74	1.88	95.9	8488	0.12	0.11	93.4	9387	0.35	0.08	92.9
		Complete HSROC	4129	18.8	4.63	97.8	6248	1.74	0.14	95.2	7136	0.65	0.09	93.4

E.1 continued...

Studies	Heterogeneity*	Meta-analytic model	5% prevalence			25% prevalence			50% prevalence					
			N	Bias (%)†	MSE	Coverage (%)	N	Bias (%)	MSE	Coverage (%)	N	Bias (%)	MSE	Coverage (%)
		FA	6772	18.0	5.08	96.5	7776	1.03	0.14	93.3	8088	0.35	0.09	92.3
		FT	2759	21.4	6.75	95.7	1840	1.13	0.15	90.3	1579	0.22	0.09	89.1
		FAT	9775	3.97	2.46	91.2	9301	-3.70	0.13	80.4	4765	-4.46	0.11	70.6
		SFAT	10000	4.95	3.40	90.3	10000	-3.96	0.13	79.8	10000	-4.46	0.11	70.6
		UREM	8302	10.4	1.43	95.8	8942	0.95	0.14	93.4	9237	0.42	0.09	92.4
20	No	Complete HSROC	2915	9.13	0.38	99.1	3513	2.44	0.05	97.0	3513	1.26	0.02	96.8
		Symmetric HSROC	2654	1.52	0.14	96.0	4359	0.99	0.03	96.0	6149	1.03	0.02	97.0
		FA	2425	8.71	0.41	98.2	2969	2.37	0.05	97.1	3076	1.19	0.02	96.8
		FT	2439	8.15	0.36	98.5	2888	2.33	0.05	97.1	2996	1.19	0.02	96.9
		FAT	9917	2.33	0.16	95.6	9615	0.57	0.03	95.3	4528	-0.10	0.02	96.0
		SFAT	10000	2.26	0.16	95.6	10000	0.52	0.03	95.2	10000	0.27	0.02	95.4
		UREM	8094	6.86	0.31	97.0	7930	1.56	0.04	96.4	7963	0.79	0.02	96.2
20	Yes	Complete HSROC	5406	6.67	0.38	97.0	8011	0.75	0.06	95.4	8843	0.25	0.04	93.6
		Symmetric HSROC	8040	0.83	0.15	95.1	9679	0.22	0.05	94.7	9905	0.19	0.04	93.6
		FA	7767	2.79	0.32	95.0	8930	-0.58	0.06	92.4	9179	-0.26	0.04	91.9
		FT	2054	2.15	0.22	94.4	992	-0.97	0.06	87.7	781	-0.68	0.04	89.2
		FAT	9758	-3.52	0.19	88.6	9293	-4.70	0.08	76.1	4559	-4.72	0.07	68.4
		SFAT	10000	-3.68	0.20	88.5	10000	-4.87	0.09	75.7	10000	-4.97	0.08	64.9
		UREM	9455	3.34	0.28	95.1	9869	0.20	0.06	94.2	9951	0.19	0.04	93.2

DOR = diagnostic odds ratio; FA = fixed accuracy HSROC model; FAT = fixed accuracy and threshold HSROC model; FT = fixed threshold HSROC model; MSE = mean square error; N = number of meta-analyses out of 10,000 where hierarchical models successfully converged; SFAT = symmetric fixed accuracy and threshold HSROC model; UREM = univariate random effects logistic regression model.

* Heterogeneity in accuracy and threshold.

† Bias is presented as a percentage of the true value of the log diagnostic odds ratio.

E.2| Performance of all meta-analytic models in estimating specificity for scenarios with a DOR of 231

Studies	Heterogeneity*	Meta-analytic model	5% prevalence			25% prevalence			50% prevalence					
			N	Bias (%)†	MSE	Coverage (%)	N	Bias (%)	MSE	Coverage (%)	N	Bias (%)	MSE	Coverage (%)
5	No	Complete HSROC	1767	1.68	0.05	97.4	2086	2.23	0.06	97.6	2114	3.28	0.10	98.6
		Symmetric HSROC	1276	2.31	0.05	97.6	2753	3.26	0.08	98.0	4179	2.56	0.08	98.3
		FA	1932	2.22	0.05	98.1	2228	2.55	0.07	98.2	2232	3.80	0.12	98.3
		FT	1792	2.10	0.05	97.9	2174	2.24	0.07	97.8	2288	3.26	0.11	98.6
		FAT	9798	0.57	0.04	95.2	9556	0.80	0.05	95.7	4722	1.04	0.07	96.0
		SFAT	10000	0.56	0.04	95.2	10000	0.73	0.05	95.6	10000	1.21	0.07	95.9
		UREM	3173	1.47	0.04	96.8	2869	1.69	0.06	97.1	2883	2.77	0.09	97.4
5	Yes	Complete HSROC	3020	0.14	0.13	88.4	4339	0.86	0.15	90.6	4772	1.32	0.19	92.8
		Symmetric HSROC	3442	1.71	0.14	91.6	6331	1.97	0.15	91.6	7594	0.64	0.16	93.0
		FA	5490	0.11	0.13	91.2	6222	0.97	0.16	91.0	6331	1.68	0.20	92.8
		FT	2976	-0.03	0.13	89.7	2266	0.90	0.15	90.0	2171	2.69	0.23	90.8
		FAT	9691	-3.79	0.15	65.9	9288	-3.24	0.16	69.5	4833	-4.13	0.18	76.1
		SFAT	10000	-3.89	0.15	65.7	10000	-3.42	0.16	69.4	10000	-3.13	0.18	75.7
		UREM	5833	-0.12	0.13	88.8	6311	0.62	0.15	90.0	6573	1.22	0.20	91.3
10	No	Complete HSROC	2325	1.14	0.02	97.8	2903	1.09	0.03	97.4	2862	2.11	0.05	97.2
		Symmetric HSROC	1924	1.44	0.02	98.0	3581	1.73	0.03	97.2	5383	1.62	0.04	97.2
		FA	2175	1.32	0.02	97.9	2569	1.24	0.03	98.1	2719	2.11	0.05	97.3
		FT	2177	1.26	0.02	97.9	2670	1.37	0.03	97.0	2666	1.81	0.04	97.0
		FAT	9881	0.32	0.02	95.1	9594	0.35	0.02	95.2	4596	0.98	0.04	94.2
		SFAT	10000	0.31	0.02	95.1	10000	0.29	0.02	95.2	10000	0.57	0.03	95.3
		UREM	5612	0.74	0.02	96.1	5417	0.85	0.03	96.3	5311	1.37	0.04	96.4
10	Yes	Complete HSROC	4129	0.26	0.06	91.4	6248	0.13	0.07	92.0	7136	0.57	0.09	93.6
		Symmetric HSROC	5895	0.70	0.06	93.1	8488	0.29	0.07	92.6	9387	0.35	0.08	93.1
		FA	6772	0.28	0.06	92.0	7776	-0.12	0.07	91.7	8088	0.24	0.08	92.2
		FT	2759	0.10	0.06	91.4	1840	0.07	0.07	89.0	1579	0.32	0.09	88.0

E.2 continued...

Studies	Heterogeneity*	Meta-analytic model	5% prevalence			25% prevalence			50% prevalence						
			N	Bias (%)†	MSE	Coverage (%)	N	Bias (%)	MSE	Coverage (%)	N	Bias (%)	MSE	Coverage (%)	
20	No	Complete HSROC Symmetric HSROC	FAT	9775	-4.59	0.10	60.8	9301	-4.68	0.10	64.5	4765	-4.80	0.11	68.9
			SFAT	10000	-4.65	0.10	60.5	10000	-4.81	0.10	64.4	10000	-4.44	0.11	70.6
			UREM	8302	0.19	0.06	91.2	8942	-0.04	0.07	91.2	9237	0.30	0.09	92.1
			Complete HSROC	2915	0.62	0.01	97.4	3513	0.80	0.01	97.2	3513	1.30	0.02	96.7
			Symmetric HSROC	2654	0.80	0.01	97.6	4359	1.04	0.10	97.5	6149	1.06	0.02	96.4
			FA	2425	0.82	0.01	97.4	2969	0.84	0.10	97.2	3076	1.18	0.02	96.5
20	Yes	Complete HSROC Symmetric HSROC	FA	2439	0.79	0.01	97.0	2888	0.74	0.10	96.9	2996	1.26	0.02	96.8
			FAT	9917	0.17	0.01	95.3	9615	0.17	0.10	95.5	4528	0.77	0.02	93.9
			SFAT	10000	0.16	0.01	95.3	10000	0.16	0.10	95.5	10000	0.32	0.02	94.8
			UREM	8094	0.42	0.01	96.2	7930	0.50	0.10	96.5	7963	0.86	0.02	96.0
			Complete HSROC	5406	0.04	0.03	93.2	8011	0.11	0.03	93.3	8843	0.03	0.04	93.5
			Symmetric HSROC	8040	0.18	0.03	93.8	9679	0.05	0.03	93.2	9905	-0.08	0.04	93.6
20	Yes	Complete HSROC Symmetric HSROC	FA	7767	-0.03	0.03	93.0	8930	-0.10	0.03	92.8	9179	-0.52	0.04	92.1
			FT	2054	-0.24	0.03	91.9	992	-0.39	0.03	90.0	781	-1.22	0.04	88.2
			FAT	9758	-5.23	0.07	52.2	9293	-5.06	0.07	56.7	4559	-5.66	0.09	59.8
			SFAT	10000	-5.25	0.07	52.6	10000	-5.16	0.07	56.3	10000	-5.28	0.08	63.5
			UREM	9455	-0.07	0.03	92.9	9869	0.03	0.03	92.9	9951	-0.11	0.04	92.8
			Complete HSROC	5406	0.04	0.03	93.2	8011	0.11	0.03	93.3	8843	0.03	0.04	93.5

DOR = diagnostic odds ratio; FA = fixed accuracy HSROC model; FAT = fixed accuracy and threshold HSROC model; FT = fixed threshold HSROC model; MSE = mean square error; N = number of meta-analyses out of 10,000 where hierarchical models successfully converged; SFAT = symmetric fixed accuracy and threshold HSROC model; UREM = univariate random effects logistic regression model.

* Heterogeneity in accuracy and threshold.

† Bias is presented as a percentage of the true value of the log diagnostic odds ratio.

REFERENCES

1. Zumla A, Abubakar I, Raviglione M, et al. Drug-resistant tuberculosis--current dilemmas, unanswered questions, challenges, and priority needs. *J Infect Dis.* 2012;205 Suppl 2:S228-40.
2. WHO Tuberculosis Fact Sheet Number 104. Updated October 2015. Accessed at <http://www.who.int/mediacentre/factsheets/fs104/en/> on 4 December 2015.
3. Steingart KR, Schiller I, Horne DJ, Pai M, Boehme CC, Dendukuri N. Xpert® MTB/RIF assay for pulmonary tuberculosis and rifampicin resistance in adults. *Cochrane Database Syst Rev.* 2014;1:CD009593.
4. Boehme CC, Nicol MP, Nabeta P, et al. Feasibility, diagnostic accuracy, and effectiveness of decentralised use of the Xpert MTB/RIF test for diagnosis of tuberculosis and multidrug resistance: a multicentre implementation study. *Lancet.* 2011;377(9776):1495-1505.
5. Burns PB, Rohrich RJ, Chung KC. The levels of evidence and their role in evidence-based medicine. *Plast Reconstr Surg.* 2011;128(1):305-10.
6. Alvarado A. A practical score for the early diagnosis of acute appendicitis. *Ann Emerg Med.* 1986;15(5):557-64.
7. Silverstein MD, Boland BJ. Conceptual framework for evaluating laboratory tests: case-finding in ambulatory patients. *Clin Chem.* 1994;40(8):1621-7.
8. Haddow JE, Palomaki GE. ACCE: A model process for evaluating data on emerging genetic tests. Oxford: Oxford University Press; 2003.
9. Horvath AR, Lord SJ, StJohn A, et al. From biomarkers to medical tests: the changing landscape of test evaluation. *Clin Chim Acta.* 2014;427:49-57.

References

10. Freedman LS. Evaluating and comparing imaging techniques: a review and classification of study designs. *Br J Radiol.* 1987;60(719):1071-81.
11. Lijmer JG, Leeflang M, Bossuyt PM. Proposals for a phased evaluation of medical tests. *Med Decis Making.* 2009;29(5):E13-21.
12. Sackett DL, Haynes RB. The architecture of diagnostic research. *BMJ.* 2002;324(7336):539-41.
13. Ferrante di Ruffano L, Hyde CJ, McCaffery KJ, Bossuyt PM, Deeks JJ. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. *BMJ.* 2012;344:e686.
14. Pepe MS. Evaluating technologies for classification and prediction in medicine. *Stat Med.* 2005;24(24):3687-96.
15. Zweig MH, Robertson EA. Why we need better test evaluations. *Clin Chem.* 1982;28(6):1272-6.
16. Taube SE, Jacobson JW, Lively TG. Cancer diagnostics: decision criteria for marker utilization in the clinic. *Am J Pharmacogenomics.* 2005;5(6):357-64.
17. National Institute for Health and Clinical Excellence. Guide to the methods of technology appraisal; June 2008. Accessed at <http://www.nice.org.uk/media/B52/A7/TAMethodsGuideUpdatedJune2008.pdf> on 9 August 2010.
18. Medical Services Advisory Committee. Guidelines for the assessment of diagnostic technologies. Australian Government Department of Health and Ageing; August 2005. Accessed at www.msac.gov.au/internet/msac/publishing.nsf/Content/guidelines-1 on 3 June 2012.

References

19. European Medicines Agency. Guideline on clinical evaluation of diagnostic agents; July 2009. Accessed at http://www.ema.europa.eu/ema/pages/includes/document/open_document.jsp?webContentId=WC500003580 on 21 July 2015.
20. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? *Ann Intern Med.* 2006;144(11):850-5.
21. Lord SJ, Irwig L, Bossuyt PM. Using the principles of randomized controlled trial design to guide test evaluation. *Med Decis Making.* 2009;29(5):E1-E12.
22. Ferrante di Ruffano L, Davenport C, Eisinga A, Hyde C, Deeks JJ. A capture-recapture analysis demonstrated that randomized controlled trials evaluating the impact of diagnostic tests on patient outcomes are rare. *J Clin Epidemiol.* 2012;65(3):282-7.
23. Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Chapter 10: Analysing and presenting results. In: Deeks JJ, Bossuyt PM, Gatsonis C, eds. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0*. The Cochrane Collaboration; 2010. Accessed at <http://srdta.cochrane.org/> on 1 December 2011.
24. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology.* 2010;21(1):128-38.
25. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making.* 2006;26(6):565-74.

References

26. Glas AS, Lijmer JG, Prins MH, Bossel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol.* 2003;56(11):1129-35.
27. Dahabreh IJ, Chung M, Kitsios GD, Terasawa T, Raman G, Tatsioni A, et al. Comprehensive overview of methods and reporting of meta-analyses of test accuracy. AHRQ publication no. 12-EHC044-EF. Rockville, MD: Agency for Healthcare Research and Quality; 2012.
28. Bachmann LM, Puhan MA, ter Riet G, Bossuyt PM. Sample sizes of studies on diagnostic accuracy: literature survey. *BMJ.* 2006;332(7550):1127-9.
29. Steingart KR, Sohn H, Schiller I, et al. Xpert® MTB/RIF assay for pulmonary tuberculosis and rifampicin resistance in adults. *Cochrane Database of Syst Rev.* 2013;1:CD009593.
30. Automated real-time nucleic acid amplification technology for rapid and simultaneous detection of tuberculosis and rifampicin resistance: Xpert MTB/RIF assay for the diagnosis of pulmonary and extrapulmonary TB in adults and children. Policy update. Geneva, World Health Organization, 2013. Accessed at http://apps.who.int/iris/bitstream/10665/112472/1/9789241506335_eng.pdf?ua=1 on 25 May 2015.
31. Mulrow CD. Rationale for systematic reviews. *BMJ.* 1994;309(6954):597-9.
32. Leeflang MM, Deeks JJ, Takwoingi Y, Macaskill P. Cochrane diagnostic test accuracy reviews. *Syst Rev.* 2013;2:82.
33. Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. Designing studies to ensure that estimates of test accuracy are transferable. *BMJ.* 2002;324(7338):669-71.

References

34. Leeflang MM, Deeks JJ, Rutjes AW, Reitsma JB, Bossuyt PM. Bivariate meta-analysis of predictive values of diagnostic tests can be an alternative to bivariate meta-analysis of sensitivity and specificity. *J Clin Epidemiol.* 2012;65(10):1088-97.
35. Harbord RM, Whiting P, Sterne JA, et al. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. *J Clin Epidemiol.* 2008;61(11):1095-1103.
36. Dahabreh IJ, Trikalinos TA, Lau J, Schmid C. An empirical assessment of bivariate methods for meta-analysis of test accuracy. AHRQ publication no. 12(13)-EHC136-EF. Rockville, MD: Agency for Healthcare Research and Quality; 2012.
37. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Stat Med.* 1993;12(14):1293-1316.
38. Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Stat Med.* 2002;21(9):1237-56.
39. Arends LR, Hamza TH, van Houwelingen JC, Heijnenbroek-Kal MH, Hunink MG, Stijnen T. Bivariate random effects meta-analysis of ROC curves. *Med Decis Making.* 2008;28(5):621-38.
40. Ma X, Nie L, Cole SR, Chu H. Statistical methods for multivariate meta-analysis of diagnostic tests: An overview and tutorial. *Stat Methods Med Res.* Epub ahead of print June 26 2013.
41. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol.* 2005;58(10):982-90.

References

42. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med.* 2001;20(19):2865-84.
43. Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM, Cochrane Diagnostic Test Accuracy Working Group. Systematic reviews of diagnostic test accuracy. *Ann Intern Med.* 2008;149(12):889-97.
44. Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics.* 2007;8(2):239-51.
45. Takwoingi Y, Riley RD, Deeks JJ. Meta-analysis of diagnostic accuracy studies in mental health. *Evid Based Ment Health.* 2015;18(4):103-9.
46. Dukic V, Gatsonis C. Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds. *Biometrics.* 2003;59(4):936-46.
47. Kester AD, Buntinx F. Meta-analysis of ROC curves. *Med Decis Making.* 2000;20(4):430-9.
48. Hamza T, Arends L, van Houwelingen H, Stijnen T. Multivariate random effects meta-analysis of diagnostic tests with multiple thresholds. *BMC Med Res Methodol.* 2009;9(1):73.
49. Riley RD, Ahmed I, Ensor J, et al. Meta-analysis of test accuracy studies: an exploratory method for investigating the impact of missing thresholds. *Syst Rev.* 2015;4(1):12.
50. Turner RM, Davey J, Clarke MJ, Thompson SG, Higgins JP. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews. *Int J Epidemiol.* 2012;41(3):818-27.
51. Harbord RM. Commentary on 'Multivariate meta-analysis: potential and promise'. *Stat Med.* 2011;30(20):2507-8; discussion 2509-10.

References

52. van Houwelingen HC, Zwinderman KH, Stijnen T. A bivariate approach to meta-analysis. *Stat Med.* 1993;12(24):2273-84.
53. Chu H, Cole SR. Bivariate meta-analysis for sensitivity and specificity with sparse data: a generalized linear mixed model approach (letter to the Editor). *J Clin Epidemiol.* 2006;59:1331-2.
54. Chu H, Guo H, Zhou Y. Bivariate random effects meta-analysis of diagnostic studies using generalized linear mixed models. *Med Decis Making.* 2010;30(4):499-508.
55. Zwinderman AH, Bossuyt PM. We should not pool diagnostic likelihood ratios in systematic reviews. *Stat Med.* 2008;27(5):687-97.
56. Carvalho AF, Takwoingi Y, Sales PM, et al. Screening for bipolar spectrum disorders: A comprehensive meta-analysis of accuracy studies. *J Affect Disord.* 2014;172C:337-46.
57. Macaskill P. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. *J Clin Epidemiol.* 2004;57(9):925-32.
58. Chappell FM, Raab GM, Wardlaw JM. When are summary ROC curves appropriate for diagnostic meta-analyses? *Stat Med.* 2009;28(21):2653-68.
59. Macaskill P, Walter SD, Irwig L, Franco EL. Assessing the gain in diagnostic performance when combining two diagnostic tests. *Stat Med.* 2002;21(17):2527-46.
60. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ.* 2006;332(7549):1089-92.
61. Grandjean H, Sarramon MF. Sonographic measurement of nuchal skinfold thickness for detection of Down syndrome in the second-trimester fetus: a multicenter prospective study. The AFDPHE Study Group. Association Francaise pour le

References

- Depistage et la Prevention des Handicaps de l'Enfant. *Obstet Gynecol.* 1995;85(1):103-6.
62. Grandjean H, Sarramon MF. Femur/foot length ratio for detection of Down syndrome: results of a multicenter prospective study. The Association Francaise pour le Depistage et la Prevention des Handicaps de l'Enfant Study Group. *Am J Obstet Gynecol.* 1995;173(1):16-9.
63. Merlin T, Lehman S, Hiller JE, Ryan P. The "linked evidence approach" to assess medical tests: a critical analysis. *Int J Technol Assess Health Care.* 2013;29(3):343-50.
64. Trikalinos TA, Siebert U, Lau J. Decision-analytic modeling to evaluate benefits and harms of medical tests: uses and limitations. *Med Decis Making.* 2009;29(5):E22-29.
65. Garcia Pena BM, Mandl KD, Kraus SJ, et al. Ultrasonography and limited computed tomography in the diagnosis and management of appendicitis in children. *JAMA.* 1999;282(11):1041-6.
66. Kaiser S, Frenckner B, Jorulf HK. Suspected appendicitis in children: US and CT--a prospective randomized study. *Radiology.* 2002;223(3):633-8.
67. Smith MP, Katz DS, Lalani T, et al. ACR appropriateness criteria® right lower quadrant pain--suspected appendicitis. *Ultrasound Q.* 2015;31(2):85-91.
68. Gurusamy KS, Giljaca V, Takwoingi Y, et al. Ultrasound versus liver function tests for diagnosis of common bile duct stones. *Cochrane Database of Syst Rev.* 2015;2:CD011548.
69. Giljaca V, Gurusamy KS, Takwoingi Y, et al. Endoscopic ultrasound versus magnetic resonance cholangiopancreatography for common bile duct stones. *Cochrane Database of Syst Rev.* 2015;2:CD011549.

References

70. Gurusamy KS, Giljaca V, Takwoingi Y, et al. Endoscopic retrograde cholangiopancreatography versus intraoperative cholangiography for diagnosis of common bile duct stones. *Cochrane Database of Syst Rev.* 2015;2:CD010339.
71. Takwoingi Y, Leeflang MM, Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. *Ann Intern Med.* 2013;158(7):544-54.
72. Zhou XH, Obuchowski NA, McClish DK. *Statistical Methods in Diagnostic Medicine.* John Wiley & Sons; 2011.
73. Poortman P, Lohle PN, Schoemaker CM, et al. Comparison of CT and sonography in the diagnosis of acute appendicitis: a blinded prospective study. *AJR. Am J Roentgenol.* 2003;181(5):1355-9.
74. Pickuth D, Heywang-Kobrunner SH, Spielmann RP. Suspected acute appendicitis: is ultrasonography or computed tomography the preferred imaging technique? *Eur J Surg.* 2000;166(4):315-9.
75. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction.* First ed. Oxford, UK: Oxford University Press; 2003.
76. Kernan WN, Viscoli CM, Makuch RW, Brass LM, Horwitz RI. Stratified randomization for clinical trials. *J Clin Epidemiol.* 1999;52(1):19-26.
77. Tsai S, Coughlin B, Hampf F, Munshi I, Wolfe J. Diagnosing appendicitis with CT and ultrasound using prospective patient stratification by body mass index. *Emerg. Radiol.* 2001;8(5):267-271.
78. Sivit CJ, Applegate KE, Stallion A, et al. Imaging evaluation of suspected appendicitis in a pediatric population: effectiveness of sonography versus CT. *AJR Am J Roentgenol.* 2000;175(4):977-80.

References

79. Lowe LH, Penney MW, Stein SM, et al. Unenhanced limited CT of the abdomen in the diagnosis of appendicitis in children: comparison with sonography. *AJR Am J Roentgenol*. 2001;176(1):31-5.
80. Hayen A, Macaskill P, Irwig L, Bossuyt P. Appropriate statistical methods are required to assess diagnostic tests for replacement, add-on, and triage. *J Clin Epidemiol*. 2010;63(8):883-91.
81. Altman DG, Deeks JJ, Sackett DL. Odds ratios should be avoided when events are common. *BMJ*. 1998;317(7168):1318.
82. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med*. 2000;342(25):1887-92.
83. Ioannidis JP, Haidich AB, Pappa M, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA*. 2001;286(7):821-30.
84. Deeks JJ, Dinnes J, D'Amico R, et al. Evaluating non-randomised intervention studies. *Health Technol Assess*. 2003;7(27):iii-x, 1-173.
85. Reeves BC, Deeks JJ, Higgins JPT, Wells GA. Including nonrandomized studies. In: Higgins JPT, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0 (Updated March 2011). The Cochrane Collaboration, 2011. Accessed at www.cochrane-handbook.org on 1 December 2011.
86. Glenny AM, Altman DG, Song F, et al. Indirect comparisons of competing interventions. *Health Technol Assess*. 2005;9(26):1-148.
87. Macaskill P. Contributions to the analysis and meta-analysis of diagnostic test comparisons [PHD thesis]. Sydney, University of Sydney; 2003.

References

88. Menke J, Larsen J. Meta-analysis: Accuracy of contrast-enhanced magnetic resonance angiography for assessing steno-occlusions in peripheral arterial disease. *Ann Intern Med.* 2010;153(5):325-34.
89. European Stroke Organisation, Tendera M, Aboyans V, et al. ESC Guidelines on the diagnosis and treatment of peripheral artery diseases: Document covering atherosclerotic disease of extracranial carotid and vertebral, mesenteric, renal, upper and lower extremity arteries: the Task Force on the Diagnosis and Treatment of Peripheral Artery Diseases of the European Society of Cardiology (ESC). *Eur Heart J.* 2011;32(22):2851-2906.
90. Collins R, Cranny G, Burch J, et al. A systematic review of duplex ultrasound, magnetic resonance angiography and computed tomography angiography for the diagnosis and assessment of symptomatic, lower limb peripheral arterial disease. *Health Technol Assess.* 2007;11(20):iii-iv, xi-xiii, 1-184.
91. Lumley T. Network meta-analysis for indirect treatment comparisons. *Stat Med.* 2002;21(16):2313-24.
92. Sutton A, Ades AE, Cooper N, Abrams K. Use of indirect and mixed treatment comparisons for technology assessment. *Pharmacoeconomics.* 2008;26(9):753-67.
93. Hutton B, Salanti G, Caldwell DM, et al. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. *Ann Intern Med.* 2015;162(11):777-84.
94. Dinnes J, Deeks J, Kirby J, Roderick P. A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy. *Health Technol Assess.* 2005;9(12):1-113, iii.

References

95. Abba K, Deeks JJ, Olliaro P, et al. Rapid diagnostic tests for diagnosing uncomplicated *P. falciparum* malaria in endemic countries. *Cochrane Database of Syst Rev.* 2011(7):CD008122.
96. Abba K, Kirkham AJ, Olliaro PL, et al. Rapid diagnostic tests for diagnosing uncomplicated non-falciparum or *Plasmodium vivax* malaria in endemic countries. *Cochrane Database of Syst Rev.* 2014;12:CD011431.
97. Alldred SK, Takwoingi Y, Guo B, et al. First trimester serum tests for Down's syndrome screening. *Cochrane Database of Syst Rev.* 2015(11):CD011975.
98. Price RN, Tjitra E, Guerra CA, Yeung S, White NJ, Anstey NM. Vivax malaria: neglected and not benign. *Am J Trop Med Hyg.* 2007;77(6 Suppl):79-87.
99. Takwoingi Y, Abba K, Garner P. Rapid diagnostic testing for *Plasmodium vivax* and non-falciparum malaria in endemic areas. *JAMA.* 2015; 314(10):1065-6.
100. Culpepper L. Misdiagnosis of bipolar depression in primary care practices. *J Clin Psychiatry.* 2014;75(3):e05.
101. Drancourt N, Etain B, Lajnef M, et al. Duration of untreated bipolar disorder: missed opportunities on the long road to optimal treatment. *Acta Psychiatr Scand.* 2013;127(2):136-44.
102. Penrose LS. The relative effects of paternal and maternal age in mongolism. 1933. *J Genet.* 2009;88(1):9-14.
103. Alldred SK, Deeks JJ, Guo B, Neilson JP, Alfirevic Z. Second trimester serum tests for Down's Syndrome screening. *Cochrane Database of Syst Rev.* 2012;6:CD009925.
104. Takwoingi Y, Deeks J. MetaDAS: A SAS macro for meta-analysis of diagnostic accuracy studies. User Guide Version 1.3; 2010. Accessed at

References

- <http://dta.cochrane.org/sites/dta.cochrane.org/files/uploads/MetaDAS%20Readme%20v1.3%20May%202012.pdf> on 11 August 2015.
105. BIO Ventures for Global Health. What are rapid diagnostic tests? 2015. Accessed at <http://www.bvgh.org/Current-Programs/Neglected-Disease-Product-Pipelines/Global-Health-Primer/Targets/cid/ViewDetails/ItemID/16.aspx> on 14 November 2015.
 106. UNICEF. Malaria diagnosis: A guide for selecting rapid diagnostic test (RDT) kits - 1st editio; 2007. Accessed at http://www.unicef.org/french/supply/files/Guidance_for_malaria_rapid_tests.pdf on 29 January 2016.
 107. Hirschfeld RM, Williams JB, Spitzer RL, et al. Development and validation of a screening instrument for bipolar spectrum disorder: the Mood Disorder Questionnaire. *Am J Psychiatry*. 2000;157(11):1873-5.
 108. Ghaemi SN, Miller CJ, Berv DA, Klugman J, Rosenquist KJ, Pies RW. Sensitivity and specificity of a new bipolar spectrum diagnostic scale. *J Affect Disord*. 2005;84(2-3):273-7.
 109. Angst J, Adolfsson R, Benazzi F, et al. The HCL-32: towards a self-assessment tool for hypomanic symptoms in outpatients. *J Affect Disord*. 2005;88(2):217-33.
 110. Riley R, Abrams K, Sutton A, Lambert P, Thompson J. Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC Med Res Methodol*. 2007;7(1):3.
 111. Chung Y, Rabe-Hesketh S, Choi IH. Avoiding zero between-study variance estimates in random-effects meta-analysis. *Stat Med*. 2013;32(23):4071-89.

References

112. Jellema P, van Tulder MW, van der Horst HE, Florie J, Mulder CJ, van der Windt DA. Inflammatory bowel disease: a systematic review on the value of diagnostic testing in primary care. *Colorectal Dis.* 2011;13(3):239-54.
113. Centre for Reviews and Dissemination. University of York. 2010. Accessed at <http://www.crd.york.ac.uk/CRDWeb/AboutPage.asp> on 28 February 2015.
114. Whiting P, Rutjes AW, Dinnes J, Reitsma J, Bossuyt PM, Kleijnen J. Development and validation of methods for assessing the quality of diagnostic accuracy studies. *Health Technol Assess.* 2004;8(25):iii, 1-234.
115. Cochrane Methodology Register (CMR) <http://community.cochrane.org/editorial-and-publishing-policy-resource/cochrane-methodology-register-cmr>. Accessed 20/06/2015.
116. About the Cochrane Methodology Register. Accessed at <http://www.cochranelibrary.com/help/the-cochrane-methodology-register-july-issue-2012.html> on 21 May 2016.
117. About the Cochrane Library. Accessed at <http://cochrane.demo.wiley.com/about/about-the-cochrane-library.html> on 21 July 2015.
118. Deeks JJ, Bossuyt PM, Gatsonis C, eds. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*. The Cochrane Collaboration. Accessed at <http://dta.cochrane.org/handbook-dta-reviews> on 25 November 2015.
119. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ.* 2009;339:b2535.

References

120. Stewart LA, Clarke M, Rovers M, et al. Preferred Reporting Items for Systematic Review and Meta-Analyses of individual participant data: the PRISMA-IPD Statement. *JAMA*. 2015;313(16):1657-65.
121. Zorzela L, Golder S, Liu Y, et al. Quality of reporting in systematic reviews of adverse events: systematic review. *BMJ*. 2014;348:f7668.
122. Zorzela L, Loke YK, Ioannidis JP, et al. PRISMA harms checklist: improving harms reporting in systematic reviews. *BMJ*. 2016;352:i157.
123. Hartling L, Chisholm A, Thomson D, Dryden DM. A descriptive analysis of overviews of reviews published between 2000 and 2011. *PLoS One*. 2012;7(11):e49667.
124. Hutton B, Salanti G, Chaimani A, et al. The quality of reporting methods and results in network meta-analyses: an overview of reviews and suggestions for improvement. *PLoS One*. 2014;9(3):e92508.
125. Ewald B, Ewald D, Thakkinstian A, Attia J. Meta-analysis of B type natriuretic peptide and N-terminal pro B natriuretic peptide in the diagnosis of clinical heart failure and population screening for left ventricular systolic dysfunction. *Intern Med J*. 2008;38(2):101-13..
126. Geersing GJ, Janssen KJ, Oudega R, et al. Excluding venous thromboembolism using point of care D-dimer tests in outpatients: a diagnostic meta-analysis. *BMJ*. 2009;339:b2990.
127. Minion J, Leung E, Menzies D, Pai M. Microscopic-observation drug susceptibility and thin layer agar assays for the detection of drug resistant tuberculosis: a systematic review and meta-analysis. *Lancet Infect Dis*. 2010;10(10):688-98.

References

128. Lucassen W, Geersing GJ, Erkens PM, et al. Clinical decision rules for excluding pulmonary embolism: a meta-analysis. *Ann Intern Med.* 2011;155(7):448-60.
129. Wang LW, Fahim MA, Hayen A, et al. Cardiac testing for coronary artery disease in potential kidney transplant recipients. *Cochrane Database of Syst Rev.* 2011(12):CD008691.
130. Bossuyt P, Davenport C, Deeks J, Hyde C, Leeflang M, Scholten R. Chapter 11: Interpreting results and drawing conclusions. In: Deeks JJ, Bossuyt PM, Gatsonis C, eds. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 0.9.* The Cochrane Collaboration; 2013. Accessed at <http://dta.cochrane.org/handbook-dta-reviews> on 29 January 2016.
131. Reid MC, Lane DA, Feinstein AR. Academic calculations versus clinical judgments: practicing physicians' use of quantitative measures of test accuracy. *Am J Med.* 1998;104(4):374-80.
132. Leeflang MM, Rutjes AW, Reitsma JB, Hooft L, Bossuyt PM. Variation of a test's sensitivity and specificity with disease prevalence. *CMAJ.* 2013;185(11):E537-44.
133. Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. *Med Decis Making.* 1993;13(4):313-21.
134. Wang J, Bossuyt P, Geskus R, et al. Using individual patient data to adjust for indirectness did not successfully remove the bias in this case of comparative test accuracy. *J Clin Epidemiol.* 2015;68(3):290-8.
135. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med.* 2011;155(8):529-36.

References

136. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA*. 1999;282(11):1061-6.
137. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ*. 2006;174(4):469-76.
138. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med*. 2004;140(3):189-202.
139. Willis BH, Quigley M. The assessment of the quality of reporting of meta-analyses in diagnostic research: a systematic review. *BMC Med Res Methodol*. 2011;11:163.
140. Mallett S, Deeks JJ, Halligan S, Hopewell S, Cornelius V, Altman DG. Systematic reviews of diagnostic tests in cancer: review of methods and reporting. *BMJ*. 2006;333(7565):413.
141. Cruciani M, Mengoli C. An overview of meta-analyses of diagnostic tests in infectious diseases. *Infect Dis Clin North Am*. 2009;23(2):225-67.
142. Naaktgeboren CA, van Enst WA, Ochodo EA, et al. Systematic overview finds variation in approaches to investigating and reporting on sources of heterogeneity in systematic reviews of diagnostic studies. *J Clin Epidemiol*. 2014;67(11):1200-9.
143. Whiting P, Rutjes AW, Dinnes J, Reitsma JB, Bossuyt PM, Kleijnen J. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. *J Clin Epidemiol*. 2005;58(1):1-12.
144. McInnes M, Moher D, Bossuyt P. Development and implementation of a reporting guideline for systematic reviews and meta-analyses of diagnostic accuracy studies: The PRISMA-DTA initiative; 2015. Accessed at <http://www.equator-network.org/wp->

References

- content/uploads/2009/02/PRISMA-DTA-Executive-Summary.pdf on 23 November 2015.
145. Pennant M, Takwoingi Y, Pennant L, et al. A systematic review of positron emission tomography (PET) and positron emission tomography/computed tomography (PET/CT) for the diagnosis of breast cancer recurrence. *Health Technol Assess.* 2010;14(50):1-103.
 146. Williams GJ, Macaskill P, Chan SF, Turner RM, Hodson E, Craig JC. Absolute and relative accuracy of rapid urine tests for urinary tract infection in children: a meta-analysis. *Lancet Infect Dis.* 2010;10(4):240-50.
 147. Schuetz GM, Zacharopoulou NM, Schlattmann P, Dewey M. Meta-analysis: noninvasive coronary angiography using computed tomography versus magnetic resonance imaging. *Ann Intern Med.* 2010;152(3):167-77.
 148. Giraudeau B, Higgins JP, Tavernier E, Trinquart L. Sample size calculation for meta-epidemiological studies. *Stat Med.* 2016;35(2):239-50.
 149. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *J Clin Epidemiol.* 1997;50(6):683-91.
 150. Salanti G, Del Giovane C, Chaimani A, Caldwell DM, Higgins JP. Evaluating the quality of evidence from a network meta-analysis. *PloS one.* 2014;9(7):e99682.
 151. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *J Clin Epidemiol.* 2008;61(10):991-6.

References

152. Al-Khayal KA, Al-Omran MA. Computed tomography and ultrasonography in the diagnosis of equivocal acute appendicitis. a meta-analysis. *Saudi Med J*. 2007;28(2):173-80.
153. Balk EM, Ioannidis JPA, Salem D, Chew PW, Lau J. Accuracy of biomarkers to diagnose acute cardiac ischemia in the emergency department: A meta-analysis. *Ann Emerg Med*. 2001;37(5):478-94.
154. Choi HJ, Ju W, Myung SK, Kim Y. Diagnostic performance of computer tomography, magnetic resonance imaging, and positron emission tomography or positron emission tomography/computer tomography for detection of metastatic lymph nodes in patients with cervical cancer: meta-analysis. *Cancer Sci*. 2010;101(6):1471-9.
155. de Bondt RBJ, Nelemans PJ, Hofman PAM, et al. Detection of lymph node metastases in head and neck cancer: A meta-analysis comparing US, USgFNAC, CT and MR imaging. *Eur J Radiol*. 2007;64(2):266-72.
156. Elamin MB, Murad MH, Mullan R, et al. Accuracy of Diagnostic Tests for Cushing's Syndrome: A Systematic Review and Metaanalyses. *J Clin Endocrinol Metab*. 2008;93(5):1553-62.
157. Fleischmann KE, Hunink MG, Kuntz KM, Douglas PS. Exercise echocardiography or exercise SPECT imaging: a meta analysis of diagnostic test performance. *JAMA*. 1998;280(10):913-20.
158. Gisbert JP, Abaira V. Accuracy of Helicobacter pylori diagnostic tests in patients with bleeding peptic ulcer: a systematic review and meta-analysis. *Am J Gastroenterol*. 2006;101(4):848-63.

References

159. Glas AS, Roos D, Deutekom M, Zwinderman AH, Bossuyt PM, Kurth KH. Tumor markers in the diagnosis of primary bladder cancer. A systematic review. *J Urol.* 2003;169(6):1975-82.
160. Gu P, Huang G, Chen Y, Zhu C, Yuan J, Sheng S. Diagnostic utility of pleural fluid carcinoembryonic antigen and CYFRA 21-1 in patients with pleural effusion: a systematic review and meta-analysis. *J Clin Lab Anal.* 2007;21(6):398-405.
161. Gu P, Pan LL, Wu SQ, Sun L, Huang G. CA 125, PET alone, PET-CT, CT and MRI in diagnosing recurrent ovarian carcinoma: A systematic review and meta-analysis. *Eur J Radiol.* 2009;71(1):164-74.
162. Heim SW, Schectman JM, Siadaty MS, Philbrick JT. D-dimer testing for deep venous thrombosis: a metaanalysis. *Clin Chem.* 2004;50(7):1136-47.
163. Mahajan N, Polavaram L, Vankayala H, et al. Diagnostic accuracy of myocardial perfusion imaging and stress echocardiography for the diagnosis of left main and triple vessel coronary artery disease: a comparative meta-analysis. *Heart.* 2010;96(12):956-66.
164. Mant J, Doust J, Roalfe A, et al. Systematic review and individual patient data meta-analysis of diagnosis of heart failure, with modelling of implications of different diagnostic strategies in primary care. *Health Technol Assess.* 2009;13(32):1-232.
165. Mirza TA, Karthikesalingam A, Jackson D, et al. Duplex ultrasound and contrast-enhanced ultrasound versus computed tomography for the detection of endoleak after EVAR: systematic review and bivariate meta-analysis. *Eur J Vasc Endovas Surg.* 2010;39(4):418-28.

References

166. Mitchell AJ, Bird V, Rizzo M, Meader N. Diagnostic validity and added value of the Geriatric Depression Scale for depression in primary care: a meta-analysis of GDS30 and GDS15. *J Affect Disord.* 2010;125(1-3):10-7.
167. Mowatt G, Zhu S, Kilonzo M, et al. Systematic review of the clinical effectiveness and cost-effectiveness of photodynamic diagnosis and urine biomarkers (FISH, ImmunoCyt, NMP22) and cytology for the detection and follow-up of bladder cancer. *Health Technol Assess.* 2010;14(4):1-331, iii-iv.
168. Ngamruengphong S, Sharma VK, Nguyen B, Das A. Assessment of response to neoadjuvant therapy in esophageal cancer: an updated systematic review of diagnostic accuracy of endoscopic ultrasonography and fluorodeoxyglucose positron emission tomography. *Dis Esophagus.* 2010;23(3):216-31.
169. Niekel MC, Bipat S, Stoker J. Diagnostic imaging of colorectal liver metastases with CT, MR imaging, FDG PET, and/or FDG PET/CT: a meta-analysis of prospective studies including patients who have not previously undergone treatment. *Radiology.* 2010;257(3):674-84
170. Nishimura K, Sugiyama D, Kogata Y, et al. Meta-analysis: diagnostic accuracy of anti-cyclic citrullinated peptide antibody and rheumatoid factor for rheumatoid arthritis. *Ann Intern Med.* 2007;146(11):797-808.
171. Noguchi Y, Nagata-Kobayashi S, Stahl JE, Wong JB. A meta-analytic comparison of echocardiographic stressors. *Int J Cardiovasc Imaging.* 2005;21(2-3):209-11.
172. Safdar N, Fine JP, Maki DG. Meta-analysis: methods for diagnosing intravascular device-related bloodstream infection. *Ann Intern Med.* 2005;142(6):451-66.

References

173. Scheidler J, Hricak H, Yu KK, Subak L, Segal MR. Radiological evaluation of lymph node metastases in patients with cervical cancer: a meta-analysis. *JAMA*. 1997;278(13):1016-1101.
174. Scherer K, Bedlack RS, Simel DL. Does this patient have myasthenia gravis? *JAMA*. 2005;293(15):1906-14.
175. Selman TJ, Luesley DM, Acheson N, Khan KS, Mann CH. A systematic review of the accuracy of diagnostic tests for inguinal lymph node status in vulvar cancer. *Gynecol Oncol*. 2005;99(1):206-14.
176. Shiga T, Wajima Zi, Apfel CC, Inoue T, Ohe Y. Diagnostic accuracy of transesophageal echocardiography, helical computed tomography, and magnetic resonance imaging for suspected thoracic aortic dissection: systematic review and meta-analysis. *Arch Intern Med*. 2006;166(13):1350-6.
177. Smith TO, Hilton G, Toms AP, Donell ST, Hing CB. The diagnostic accuracy of acetabular labral tears using magnetic resonance imaging and magnetic resonance arthrography: a meta-analysis. *Eur Radiol*. 2011;21(4):863-74.
178. Smith-Bindman R, Hosmer W, Feldstein VA, Deeks JJ, Goldberg JD. Second-trimester ultrasound to detect fetuses with Down syndrome: a meta-analysis. *JAMA*. 2001;285(8):1044-55.
179. St John A, Boyd JC, Lowes AJ, Price CP. The use of urinary dipstick tests to exclude urinary tract infection: a systematic review of the literature. *Am J Clin Pathol*. 2006;126(3):428-36.
180. Terasawa T, Blackmore C, Bent S, Kohlwes RJ. Systematic review: computed tomography and ultrasonography to detect acute appendicitis in adults and adolescents. *Ann Intern Med*. 2004;141(7):537-46.

References

181. Tian XY, Zhu H, Zhao J, She Q, Zhang GX. Diagnostic performance of urea breath test, rapid urea test, and histology for *Helicobacter pylori* infection in patients with partial gastrectomy: a meta-analysis. *J Clin Gastroenterol*. 2012;46(4):285-92.
182. Toloza EM, Harpole L, McCrory DC. Noninvasive staging of non-small cell lung cancer: a review of the current evidence. *Chest*. 2003;123(1 Suppl):137S-146S.
183. van der Windt DA, Jellema P, Mulder CJ, Frank Kneepkens CM, van der Horst HE. Diagnostic testing for celiac disease among patients with abdominal symptoms: a systematic review. *JAMA*. 2010;303(17):1738-46.
184. van Vliet EP, Heijnenbroek-Kal MH, Hunink MG, Kuipers EJ, Siersema PD. Staging investigations for oesophageal cancer: a meta-analysis. *Br J Cancer*. 2008;98(3):547-57.
185. Wardlaw JM, Chappell FM, Best JJ, Wartolowska K, Berry E. Non-invasive imaging compared with intra-arterial angiography in the diagnosis of symptomatic carotid stenosis: a meta-analysis. *Lancet*. 2006;367(9521):1503-12.
186. Xu GZ, Zhu XD, Li MY. Accuracy of whole-body PET and PET-CT in initial M staging of head and neck cancer: a meta-analysis. *Head Neck*. 2011;33(1):87-94.
187. Yin ZG, Zhang JB, Kan SL, Wang XG. Diagnosing suspected scaphoid fractures: a systematic review and meta-analysis. *Clin Orthop Relat Res*. 2010;468(3):723-34.
188. Sheps SB, Schechter MT. The Assessment of Diagnostic Tests. *JAMA*. 1984;252(17):2418-22.
189. Song F, Altman DG, Glenny AM, Deeks JJ. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. *BMJ*. 2003;326(7387):472.

References

190. Leeflang M, Reitsma J, Scholten R, et al. Impact of adjustment for quality on results of metaanalyses of diagnostic accuracy. *Clin Chem.* 2007;53(2):164-72.
191. Chu H, Nie L, Cole SR, Poole C. Meta-analysis of diagnostic accuracy studies accounting for disease prevalence: Alternative parameterizations and model selection. *Stat Med.* 2009;28:2384-99.
192. Paul M, Riebler A, Bachmann LM, Rue H, Held L. Bayesian bivariate meta-analysis of diagnostic test studies using integrated nested Laplace approximations. *Stat Med.* 2010;29(12):1325-39.
193. Putter H, Fiocco M, Stijnen T. Meta-analysis of diagnostic test accuracy studies with multiple thresholds using survival methods. *Biom J.* 2010;52(1):95-110.
194. Verde PE. Meta-analysis of diagnostic test data: a bivariate Bayesian modeling approach. *Stat.Med.* 2010;29(30):3088-3102.
195. Nikoloulopoulos AK. A vine copula mixed effect model for trivariate meta-analysis of diagnostic test accuracy studies accounting for disease prevalence. *Stat Methods Med Res.* Epub ahead of print 11 August 2015.
196. Nikoloulopoulos AK. A mixed effect model for bivariate meta-analysis of diagnostic test accuracy studies using a copula representation of the random effects distribution. *Stat Med.* 2015;34(29):3842-65.
197. Hoyer A, Kuss O. Meta-analysis of diagnostic tests accounting for disease prevalence: a new model using trivariate copulas. *Stat Med.* 2015;34(11):1912-24.
198. Kuss O, Hoyer A, Solms A. Meta-analysis for diagnostic accuracy studies: a new statistical model using beta-binomial distributions and bivariate copulas. *Stat Med.* 2014;33(1):17-30.

References

199. Zapf A, Hoyer A, Kramer K, Kuss O. Nonparametric meta-analysis for diagnostic accuracy studies. *Stat Med.* 2015;34(29):3831-41.
200. About Scopus. Accessed at <https://www.elsevier.com/solutions/scopus> on 8 January 2016.
201. Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. *Psychol Bull.* 1995;117(1):167-78.
202. Kowalski J, Tu XM, Jia G, Pagano M. A comparative meta-analysis on the variability in test performance among FDA-licensed enzyme immunosorbent assays for HIV antibody testing. *J Clin Epidemiol.* 2001;54(5):448-61.
203. Lijmer JG, Bossuyt PM, Heisterkamp SH. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. *Stat Med.* 2002;21(11):1525-37.
204. Worster A, Preyra I, Weaver B, Haines T. The accuracy of noncontrast helical computed tomography versus intravenous pyelography in the diagnosis of suspected acute urolithiasis: a meta-analysis. *Ann Emerg Med.* 2002;40(3):280-6.
205. Suzuki S, Moro-oka T, Choudhry NK. The conditional relative odds ratio provided less biased results for comparing diagnostic test accuracy in meta-analyses. *J Clin Epidemiol.* 2004;57(5):461-69.
206. Siadaty MS, Shu J. Proportional odds ratio model for comparison of diagnostic tests in meta-analysis. *BMC Med Res Methodol.* 2004;4(1):27.
207. Siadaty MS, Philbrick JT, Heim SW, Schectman JM. Repeated-measures modeling improved comparison of diagnostic tests in meta-analysis of dependent studies. *J Clin Epidemiol.* 2004;57(7):698-711.

References

208. Hamza TH, van Houwelingen HC, Heijnenbrok-Kal MH, Stijnen T. Associating explanatory variables with summary receiver operating characteristic curves in diagnostic meta-analysis. *J Clin Epidemiol.* 2009;62(12):1284-91.
209. Trikalinos TA, Hoaglin DC, Small KM, Terrin N, Schmid CH. Methods for the joint meta-analysis of multiple tests. *Res Synth Methods.* 2014;5(4):294-312.
210. Cheng W, Schmid CH, Trikalinos TA, Gatsonis C. Network meta-analysis modeling for diagnostic accuracy studies. Society for Research Synthesis Methodology Annual Meeting, Providence, RI, USA. 2013.
211. Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J Clin Epidemiol.* 1991;44(8):763-70.
212. Deeks JJ. Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. *BMJ.* 2001;323(7305):157-62.
213. Riley RD, Price MJ, Jackson D, et al. Multivariate meta-analysis using individual participant data. *Res Synth Methods.* 2015;6(2):157-74.
214. Riley RD, Thompson JR, Abrams KR. An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. *Biostatistics.* 2008;9(1):172-86.
215. Efthimiou O, Mavridis D, Riley RD, Cipriani A, Salanti G. Joint synthesis of multiple correlated outcomes in networks of interventions. *Biostatistics.* 2015;16(1):84-97.
216. Kittler H, Pehamberger H, Wolff K, Binder M. Diagnostic accuracy of dermoscopy. *Lancet Oncol.* 2002;3(3):159-65.
217. Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. *J Clin Epidemiol.* 1995;48(1):119-130; discussion 131-2.

References

218. de Vries SO, Hunink MG, Polak JF. Summary receiver operating characteristic curves as a technique for meta-analysis of the diagnostic performance of duplex ultrasonography in peripheral arterial disease. *Acad Radiol.* 1996;3(4):361-9.
219. Battaglia M, Pewsner D, Juni P, Egger M, Bucher HC, Bachmann LM. Accuracy of B-type natriuretic peptide tests to exclude congestive heart failure: systematic review of test accuracy studies. *Arch Intern Med.* 2006;166(10):1073-80.
220. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol.* 2005;58(9):882-93.
221. Riley RD, Takwoingi Y, Trikalinos TA, et al. Meta-analysis of test accuracy studies with multiple and missing thresholds: A multivariate-normal model. *J Biomet Biostat.* 2014;5(3):196.
222. Verde PE. Meta-analysis of paired-comparison studies of diagnostic test data: A Bayesian modeling approach. Bayes 2013, Rotterdam. 2013.
223. Caldwell DM, Ades AE, Higgins JP. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *BMJ.* 2005;331(7521):897-900.
224. Salanti G, Higgins JP, Ades AE, Ioannidis JP. Evaluation of networks of randomized trials. *Stat Methods Med Res.* 2008;17(3):279-301.
225. Simel DL, Bossuyt PM. Differences between univariate and bivariate models for summarizing diagnostic accuracy may not be large. *J Clin Epidemiol.* 2009;62(12):1292-1300.
226. Hamza TH, Van Houwelingen HC, Stijnen T. The binomial distribution of meta-analysis was preferred to model within-study variability. *J.Clin.Epidemiol.* 2008;61(1):41-51.

References

227. Rabe-Hesketh S, Skrondal A, Pickles A. "GLLAMM Manual". U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 160. 2004. Accessed at <http://biostats.bepress.com/ucbbiostat/paper160> on 31 October 2013.
228. Brown H, Prescott R. *Applied Mixed Models in Medicine*. Second edition. England: John Wiley and Sons Ltd; 2006.
229. Arbyn M, Buntinx F, Van Ranst M, Paraskevaidis E, Martin-Hirsch P, Dillner J. Virologic versus cytologic triage of women with equivocal Pap smears: a meta-analysis of the accuracy to detect high-grade intraepithelial neoplasia. *J Natl Cancer Inst*. 2004;96(4):280-93.
230. Bafounta ML, Beauchet A, Aegerter P, Saiag P. Is dermoscopy (epiluminescence microscopy) useful for the diagnosis of melanoma? Results of a meta-analysis using techniques adapted to the evaluation of diagnostic tests. *Arch Dermatol*. 2001;137(10):1343-50.
231. Basaran A, Basaran M. Diagnosis of acute appendicitis during pregnancy: a systematic review. *Obstet Gynecol Surv*. 2009;64(7):481-8; quiz 499.
232. Birim O, Kappetein AP, Stijnen T, Bogers AJ. Meta-analysis of positron emission tomographic and computed tomographic imaging in detecting mediastinal lymph node metastases in nonsmall cell lung cancer. *Ann Thorac Surg*. 2005;79(1):375-82.
233. Brazzelli M, Sandercock PA, Chappell FM, et al. Magnetic resonance imaging versus computed tomography for detection of acute vascular lesions in patients presenting with stroke symptoms. *Cochrane Database Syst Rev*. 2009(4):CD007424.
234. Carlson KJ, Skates SJ, Singer DE. Screening for ovarian cancer. *Ann Intern Med*. 1994;121(2):124-32.

References

235. Cavallazzi R, Nair A, Vasu T, Marik PE. Natriuretic peptides in acute pulmonary embolism: a systematic review. *Intensive Care Med.* 2008;34(12):2147-56.
236. Deville WL, van der Windt DA, Dzaferagic A, Bezemer PD, Bouter LM. The test of Lasegue: systematic review of the accuracy in diagnosing herniated discs. *Spine.* 2000;25(9):1140-7.
237. Dong MJ, Zhao K, Lin XT, Zhao J, Ruan LX, Liu ZF. Role of fluorodeoxyglucose-PET versus fluorodeoxyglucose-PET/computed tomography in detection of unknown primary tumor: a meta-analysis of the literature. *Nucl Med Commun.* 2008;29(9):791-802.
238. Dong MJ, Liu ZF, Zhao K, et al. Value of 18F-FDG-PET/PET-CT in differentiated thyroid carcinoma with radioiodine-negative whole-body scan: a meta-analysis. *Nucl Med Commun.* 2009;30(8):639-50.
239. Doria AS, Moineddin R, Kellenberger CJ, et al. US or CT for diagnosis of appendicitis in children and adults? A meta-analysis. *Radiology.* 2006;241(1):83-94.
240. Gisbert JP, de la Morena F, Abaira V. Accuracy of monoclonal stool antigen test for the diagnosis of H. pylori infection: a systematic review and meta-analysis. *Am J Gastroenterol.* 2006;101(8):1921-30.
241. Gould MK, Kuschner WG, Rydzak CE, et al. Test performance of positron emission tomography and computed tomography for mediastinal staging in patients with non-small-cell lung cancer: a meta-analysis. *Ann Intern Med.* 2003;139(11):879-892.
242. Granader EJ, Dwamena B, Carlos RC. MRI and mammography surveillance of women at increased risk for breast cancer: recommendations using an evidence-based approach. *Acad Radiol.* 2008;15(12):1590-5.

References

243. Hamon M, Morello R, Riddell JW, Hamon M. Coronary arteries: diagnostic performance of 16- versus 64-section spiral CT compared with invasive coronary angiography—meta-analysis. *Radiology*. 2007;245(3):720-31.
244. Hayashino Y, Goto M, Noguchi Y, Fukui T. Ventilation-perfusion scanning and helical CT in suspected pulmonary embolism: meta-analysis of diagnostic performance. *Radiology*. 2005;234(3):740-8.
245. Hodgkinson J, Mant J, Martin U, et al. Relative effectiveness of clinic and home blood pressure monitoring compared with ambulatory blood pressure monitoring in diagnosis of hypertension: systematic review. *BMJ*. 2011;342:d3621.
246. Hovels AM, Heesakkers RA, Adang EM, et al. The diagnostic accuracy of CT and MRI in the staging of pelvic lymph nodes in patients with prostate cancer: a meta-analysis. *Clin Radiol*. 2008;63(4):387-95.
247. Karger R, Donner-Banzhoff N, Muller HH, Kretschmer V, Hunink M. Diagnostic performance of the platelet function analyzer (PFA-100) for the detection of disorders of primary haemostasis in patients with a bleeding history—a systematic review and meta-analysis. *Platelets*. 2007;18(4):249-60.
248. Kearon C, Julian JA, Newman TE, Ginsberg JS. Noninvasive diagnosis of deep venous thrombosis. McMaster Diagnostic Imaging Practice Guidelines Initiative. *Ann Intern Med*. 1998;128(8):663-77.
249. Koumans EH, Johnson RE, Knapp JS, St Louis ME. Laboratory testing for *Neisseria gonorrhoeae* by recently introduced nonculture tests: a performance review with clinical and public health considerations. *Clin Infect Dis*. 1998;27(5):1171-80.

References

250. Kriston L, Holzel L, Weiser AK, Berner MM, Harter M. Meta-analysis: are 3 questions enough to detect unhealthy alcohol use? *Ann Intern Med.* 2008;149(12):879-88.
251. Ledro-Cano D. Suspected choledocholithiasis: endoscopic ultrasound or magnetic resonance cholangio-pancreatography? A systematic review. *Eur J Gastroenterol Hepatol.* 2007;19(11):1007-11.
252. Lewis NR, Scott BB. Systematic review: the use of serology to exclude or diagnose coeliac disease (a comparison of the endomysial and tissue transglutaminase antibody tests). *Aliment Pharmacol Ther.* 2006;24(1):47-54.
253. Lewis NR, Scott BB. Meta-analysis: deamidated gliadin peptide antibody and tissue transglutaminase antibody compared as screening tests for coeliac disease. *Aliment Pharmacol Ther.* 2010;31(1):73-81.
254. Olatidoye AG, Wu AH, Feng YJ, Waters D. Prognostic role of troponin T versus troponin I in unstable angina pectoris for cardiac events with meta-analysis comparing published studies. *Am J Cardiol.* 1998;81(12):1405-10.
255. Roos JF, Doust J, Tett SE, Kirkpatrick CM. Diagnostic accuracy of cystatin C compared to serum creatinine for the estimation of renal dysfunction in adults and children—a meta-analysis. *Clin Biochem.* 2007;40(5-6):383-91.
256. Schuijf JD, Bax JJ, Shaw LJ, et al. Meta-analysis of comparative diagnostic performance of magnetic resonance imaging and multislice computed tomography for noninvasive coronary angiography. *Am Heart J.* 2006;151(2):404-11.
257. Shie P, Cardarelli R, Brandon D, Erdman W, Abdulrahim N. Meta-analysis: comparison of F-18 Fluorodeoxyglucose-positron emission tomography and bone

References

- scintigraphy in the detection of bone metastases in patients with breast cancer. *Clin Nucl Med.* 2008;33(2):97-101.
258. Sun J, Garfield DH, Lam B, et al. The value of autofluorescence bronchoscopy combined with white light bronchoscopy compared with white light alone in the diagnosis of intraepithelial neoplasia and invasive lung cancer: a meta-analysis. *J Thorac Oncol.* 2011;6(8):1336-44.
259. Tan KT, van Beek EJ, Brown PW, van Delden OM, Tijssen J, Ramsay LE. Magnetic resonance angiography for the diagnosis of renal artery stenosis: a meta-analysis. *Clin Radiol.* 2002;57(7):617-24.
260. van Randen A, Bipat S, Zwinderman AH, Ubbink DT, Stoker J, Boermeester MA. Acute appendicitis: meta-analysis of diagnostic performance of CT and graded compression US related to prevalence of disease. *Radiology.* 2008;249(1):97-106.
261. Verma D, Kapadia A, Eisen GM, Adler DG. EUS vs MRCP for detection of choledocholithiasis. *Gastrointest Endosc.* 2006;64(2):248-54.
262. Vestergaard ME, Macaskill P, Holt PE, Menzies SW. Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting. *Br J Dermatol.* 2008;159(3):669-76.
263. Visser K, Hunink MG. Peripheral arterial disease: gadolinium-enhanced MR angiography versus color-guided duplex US—a meta-analysis. *Radiology.* 2000;216(1):67-77.
264. Wang WH, Huang JQ, Zheng GF, et al. Is proton pump inhibitor testing an effective approach to diagnose gastroesophageal reflux disease in patients with noncardiac chest pain?: a meta-analysis. *Arch Intern Med.* 2005;165(11):1222-8.

References

265. Wiese W, Patel SR, Patel SC, Ohl CA, Estrada CA. A meta-analysis of the Papanicolaou smear and wet mount for the diagnosis of vaginal trichomoniasis. *Am J Med.* 2000;108(4):301-8.
266. Wijnberger LD, Huisjes AJ, Voorbij HA, Franx A, Bruinse HW, Mol BW. The accuracy of lamellar body count and lecithin/sphingomyelin ratio in the prediction of neonatal respiratory distress syndrome: a meta-analysis. *BJOG.* 2001;108(6):583-8.
267. Yang MG, Zhao XK, Hou Y, Xiao N. [Meta-analysis of fluorescence in situ hybridization and cytology for diagnosis of bladder cancer]. *Ai Zheng.* 2009;28(6):655-62.
268. Zhu MM, Xu XT, Nie F, Tong JL, Xiao SD, Ran ZH. Comparison of immunochemical and guaiac-based fecal occult blood test in screening and surveillance for advanced colorectal neoplasms: a meta-analysis. *J Dig Dis.* 2010;11(3):148-60.
269. Trikalinos TA, Hoaglin DC, Schmid CH. An empirical comparison of univariate and multivariate meta-analyses for categorical outcomes. *Stat Med.* 2014;33(9):1441-59.
270. Chen Y, Cai Y, Hong C, Jackson D. Inference for correlated effect sizes using multiple univariate meta-analyses. *Stat Med.* Epub ahead of print 4 November 2015.
271. Trikalinos TA, Hoaglin DC, Schmid CH. Empirical and simulation-based comparison of univariate and multivariate meta-analysis for binary outcomes. AHRQ publication no. 13-EHC066-EF. Rockville, MD: Agency for Healthcare Research and Quality; 2013.
272. Menke J. Bayesian bivariate meta-analysis of sensitivity and specificity: summary of quantitative findings in 50 meta-analyses. *J Eval Clin Pract.* 2014;20(6):844-52.

References

273. Menke J. Bivariate random-effects meta-analysis of sensitivity and specificity with the Bayesian SAS PROC MCMC: methodology and empirical evaluation in 50 meta-analyses. *Med Decis Making*. 2013;33(5):692-701.
274. Menke J. Bivariate random-effects meta-analysis of sensitivity and specificity with SAS PROC GLIMMIX. *Methods Inf Med*. 2010;49(1):54-62, 62-4.
275. Hamza TH, Reitsma JB, Stijnen T. Meta-analysis of diagnostic studies: a comparison of random intercept, normal-normal and binomial-normal bivariate Summary ROC approaches. *Med Decis Making*. 2008;28(5):639-49.
276. Ochodo EA, Reitsma JB, Bossuyt PM, Leeflang MM. Survey revealed a lack of clarity about recommended methods for meta-analysis of diagnostic accuracy data. *J Clin Epidemiol*. 2013;66(11):1281-8.
277. Hlibczuk V, Dattaro JA, Jin Z, Falzon L, Brown MD. Diagnostic accuracy of noncontrast computed tomography for appendicitis in adults: a systematic review. *Ann Emerg Med*. 2010;55(1):51-59.e1
278. Takwoingi Y, Guo B, Riley RD, Deeks JJ. Performance of methods for meta-analysis of diagnostic test accuracy with few studies or sparse data. *Stat Methods Med Res*. Epub ahead of print 26 June 2015.
279. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Stat Med*. 2006;25(24):4279-92.
280. Chen Y, Liu Y, Ning J, Nie L, Zhu H, Chu H. A composite likelihood method for bivariate meta-analysis in diagnostic systematic reviews. *Stat Methods Med Res*. Epub ahead of print 14 December 2014.
281. Diaz M. Performance measures of the bivariate random effects model for meta-analyses of diagnostic accuracy. *Comput Stat Data Anal*. 2015;83:82-90.

References

282. Begg CB. Meta-analysis methods for diagnostic accuracy. *J Clin Epidemiol.* 2008;61(11):1081-2.
283. Cochrane Screening and Diagnostic Tests Methods Group. Accessed at <http://dta.cochrane.org/welcome> on 25 February 2016.
284. Menten J, Lesaffre E. A general framework for comparative Bayesian meta-analysis of diagnostic studies. *BMC Med Res Methodol.* 2015;15:70.