A COMPARISON OF METHODS TO ASSESS DIAGNOSTIC PERFORMANCE WHEN USING IMPERFECT REFERENCE STANDARDS

CHINYEREUGO MILLICENT UMEMNEKU CHIKERE

DOCTOR OF PHILOSOPHY

FACULTY OF MEDICAL SCIENCES

POPULATION HEALTH SCIENCES INSTITUTE

DECEMBER 2020

Abstract

Background:

Estimating the diagnostic accuracy (sensitivity and specificity) of a new medical test in the absence of a gold standard or perfect reference standard is a common problem in diagnostic accuracy studies. Failing to correct for this imperfection risks under- or over-estimating the accuracy measures of the index test.

Aim:

To identify and compare methods employed to evaluate the diagnostic accuracy of medical tests in the absence of a gold standard.

Methodology:

A systematic review was conducted to identify methods employed to evaluate the diagnostic accuracy in the absence of a gold standard. Promising correction methods and latent class models were explored and compared using simulation studies and clinical datasets.

Results:

The methods identified from the systematic review were classified into four main groups: methods employed when there is a missing gold standard; when there are multiple imperfect reference standards; correction methods; and other methods such as the test positivity rate. Following the simulation studies undertaken to compare the correction methods, the Staquet *et al* method was found to outperform the Brenner method. Investigation of the latent class models alongside the analysis of a clinical dataset indicates that the assumptions made on the tests being evaluated affect the estimates obtained and clinical decisions. Given three conditionally dependent tests, the fixed effect model and random effect model via logit link tended to be preferred to the finite mixture model and random effect model via probit link because they are less impacted by the choice of priors.

Conclusion:

Many methods have been developed to estimate the diagnostic accuracy of a medical test in the absence of a gold standard. The choice of method employed depends on

the varying assumptions or characteristics of the tests under investigation as this can affect the estimates obtained and the decisions made in practice.

Acknowledgement

Firstly, I will like to express my profound gratitude to my Supervisors, Professor Luke Vale, Dr Kevin Wilson, Dr Joy Allen and Dr Sara Graziadio (maternity cover supervisor) for their immense support and guidance throughout this PhD research programme. They are totally AMAZING.

Secondly, I am grateful to Newcastle University Research Excellence Award, the School of Mathematics, Statistics and Physics, the Health Economics Group in the Population Health Sciences Institute (previously known as the Institute of Health and Society), and the National Institute for Health Research, Newcastle In-Vitro Diagnostics Co-operative Newcastle University for providing the fund required to undertake this research. In addition, I would like to thank Dr Dennis Lendrem and Professor John Isaacs for granting me the permission to use the key clinical dataset employed in this research study which is the RA-MAP dataset on rheumatoid arthritis patients in North East England. Employing this clinical dataset in my research provided a clinical application of some of the statistical methods I explored within my research.

To my assessors, Professor John Mathews and Dr Thomas Chadwick, I say a big thank you. Their constructive criticism, comments and suggestions have helped to put this study into perspective and ultimately achieve the objectives. To my Examiners, Dr Clare Lendrem (Newcastle University) and Professor Yemisi Takwoingi (Birmingham University), thank you for the feedbacks, corrections and suggestions. They were very helpful.

I also thank the postgraduate research community, the Postgraduate Head, Professor Elaine McColl and her team, and the Baddiley Clark Building Hot Desk PhD students' cohort (2017 – 2020) for providing a viable and friendly environment for successful study.

A special thank you to Professor Patrick Bossuyt, for finding time to review the protocol and manuscript for the systematic review published which is also an integral part of this Thesis.

I am immensely grateful to my family, my husband (Dr Chikere Nkwonta), my children (ChukwudiEbube and NgoziChukwu) and parents (Mr. and Mrs. Simeon and Rose-Tina Umemneku). Their daily encouragements, emotional and psychological support,

have worked together to keep me motivated and bring this research to a successful completion. Thank you for believing in me.

Finally, thank you my Heavenly Father for making this dream come true.

TO GOD BE THE GLORY

Table of Contents

Abstract.....i

Acknowledgement..... iii

List of Tables x

List of Figures..... xiv

List of Abbreviations.....xviii

1.1. Introduction.....1

1.1.1. Medical test 1

1.1.2. Diagnostic Accuracy.....4

1.1.3. Reference Standard 5

1.1.4. Gold Standard 6

1.1.5. Diagnostic Accuracy Statistics.....7

1.2. Statement of Problem 11

1.2.1. Review of types of methods.....14

1.3. Scope of the research study.....17

1.4. Aim of the PhD research study 17

1.4.1. Objectives.....17

1.5. Research methodology.....17

1.6. Significance of the research study 18

2.1. Introduction.....19

2.2. Methodology 20

2.2.1. Eligibility Criteria 20

2.2.2. Search strategies and selection of articles 21

2.2.3. Data synthesis.....22

2.3. Results 22

2.3.1. Methods employed when gold standard is missing 27

2.3.2. Correction methods.....31

2.3.3. Methods with multiple imperfect reference standards 31
2.3.4. Other designs methods 32
2.4. Guidance to researchers 33
2.5. Discussion 35
2.6. Conclusion 36
3.1. Introduction 38
3.1.1. Gart & Buck Correction method 41
3.1.2. Staquet et al correction method 42
3.1.3. Brenner correction method 44
3.1.4. Emerson et al correction method 45
3.2. Aims of the simulation study 46
3.3. Methodology 46
3.3.1. Comparison of correction methods – conditional independence 48
3.3.2. Comparison of correction methods – conditional dependence 78
3.4. Application of methods to a clinical dataset 87
3.5. Summary 97
4.1. Introduction 100
4.2. Basic notation 100
4.3. Latent class model 101
4.3.1. Traditional latent class model 102
4.3.2. Fixed effect latent class model 104
4.3.3. Random effect latent class model 109
4.3.4. Finite mixture latent class model 111
4.4. Bayesian approach 112
4.4.1. Specification of prior information 113
4.4.2. Inference using the posterior distribution 115

4.4.3. Advantages and disadvantage of Bayesian approach over frequentist approach 117

4.5. Simulation 118

4.5.1. Simulated values 118

4.5.2. Conditional independence assumption 120

4.5.3. Conditional dependence assumption 126

4.6. Limitations 139

4.7. Summary 140

4.7.1. Conditional Independence 140

4.7.2. Conditional dependence 141

4.8. Revisiting the clinical dataset from chapter three 147

5.1. Description of the clinical data 149

5.1.1. Missing data 151

5.2. Exploration of clinical dataset 152

5.3. Aims of the clinical dataset analysis 155

5.4. Methodology for analysing the clinical datasets 155

5.4.1. Prior information on the disease activities scores (SDAI, CDAI and DAS28-ESR$_4$) 156

5.5. Analysis of the RA-MAP baseline clinical data 160

5.5.1. Analysis of the baseline data assuming DAS28-ESR$_4$ is a gold standard 162

5.5.2. Analysis of the baseline data assuming SDAI, CDAI and DAS28-ESR$_4$ are conditionally independent and none of the scores is a gold standard ... 162

5.5.3. Analysis of the baseline data assuming SDAI, CDAI and DAS28-ESR$_4$ are conditionally dependent and none of the scores are a gold standard 164

5.6. Discussion 171

6.1. Summary of the research study 174

6.2. Contributions of the research study 175

6.3. Strength of the research study.....178

6.4. Limitations of the research study.....180

6.5. Further research.....181

6.6. Conclusion.....182

References.....183

Appendices.....222

A.1. PRISMA Checklist.....223

A.2. Example of search on SCOPUS database.....226

A.3. Data extraction Sheet.....227

A.4. Data Extraction Sheet (Example).....228

A.5. Supplementary information.....229

Update of the systematic review till December 2020.....230

A.5.1: Tables of methods employed in evaluating medical test(s) with missing gold standard in a binary-class diagnostic outcome.....231

A.5.2: Tables of methods employed to evaluate medical test when there is missing gold standard and the diagnostic outcomes is classified into three. Hence, focusing on ROC surface and volume of surface (VUS).....250

A.5.3: Tables of methods employed in evaluating medical test(s) with an imperfect reference standard or no gold standard. 254

B.1. R Code Chapter three – comparison of correction methods.....263

C.1. R Code Chapter Four – Simulation of datasets for investigation of LCMs 279

C.2. Openbugs code employed to analyse the simulated dataset.....284

C.3. Diagnostic plots of 23CD dataset under CI assumption.....297

C.4. Diagnostic plots of 23CD dataset under the assumption of CD (FEM)...300

C.5. Diagnostic plots of 23CD dataset under the assumption of CD (REML) 303

C.6. Diagnostic plots of 23CD under the assumption of CD using informative priors not centred on the truth (FEM).....306

C.7. Diagnostic plots of 123CD under the assumption of CI (FEM). 309

C.8. Diagnostic plots of 123CD under the assumption of CD (REML).....311

C.9. Diagnostic plots of 123CD under the assumption of CD (FEM_W) using informative priors centred on the simulated truth 313

C.10. Diagnostic plots of 123CD under the assumption of CD (REML).....314

C.11. Diagnostic plots of 123CD under CD assumption and the priors are not centred on the simulated truth.**.....316

D.1. Diagnostic plots of the RABR dataset under CI assumption.....318

D.2. Density plots of the prior and posterior distribution of DAS28-ESR_4, SDAI and CDAI using the REML on the RABR dataset.....321

D.3. Diagnostic plots of the RABR dataset under CD assumption 322

D.4. Diagnostic plots of the sensitivity analysis of RABR.....327

D.5. R-Code for clinical dataset analysis 328

List of Tables

Table 1: Description of different roles performed by different medical tests. 2
Table 2: Summary of the methods employed to evaluate medical tests in the absence of a gold standard.....15
Table 3: Summary of classification of methods employed when there is missing or no gold standard.....24
Table 4: 2 by 2 contingency table of the index test and imperfect reference standard 40
Table 5: Cell probabilities of the 4 x 2 and 2 x 2 classification of participants 50
Table 6: Estimates from unadjusted and corrected sensitivities and specificities of the index test under the conditional independence assumption when the reference standard is perfect.....55
Table 7: Unadjusted and corrected sensitivities and specificities of the index test when the reference standard is imperfect and better than the index test.....59
Table 8: Unadjusted and corrected sensitivities and specificities of the index test when the reference standard is imperfect and the index test is better than the reference standard 63
Table 9: Unadjusted and corrected sensitivities and specificities of the index test when the reference standard is imperfect and has same sensitivity and specificity as the index test. 67
Table 10: Number of illogical and undefined results obtained at various sample sizes and prevalences.....75
Table 11: 4 x 2 and 2 x 2 tables of cell probabilities classified by the true disease, reference standard and index tests results.....81
Table 12: Results of HRA cytology and punch biopsy in classifying patients into high grade and non-high grade squamous intraepithelial lesion 87
Table 13: Unadjusted and corrected sensitivities and specificities of HRA cytology .88
Table 14: Results of the visual inspection (reference standard) and fluorescence - based devices (LFpen and FC) by two separate examiners 90
Table 15: Sensitivity and Specificity of LFpen and FC stratified by examiner 1 and 2 with the NC detection. 91
Table 16: Results of the operative intervention (reference standard) and fluorescence - based devices (LF and FC) classified by examiners.....93

Table 17: Sensitivity and Specificity of LFpen and FC stratified by examiner 1 and 2 with the dentine caries lesions detection. 94
Table 18: Simulated dataset of 500 participants assuming conditional independence 120
Table 19: Estimated prevalence, sensitivities and specificities of the three tests from the different LCMs under the conditional independence assumption. 122
Table 20: Simulated dataset of 500 participants assuming conditional dependence between two tests (test 2 and test 3). 127
Table 21: Estimated prevalence, and sensitivities and specificities of the three tests from the different LCMs under the conditional independence assumption. 128
Table 22: Estimated prevalence, sensitivities and specificities of the three tests from the different LCMs assuming conditional dependence of two tests (test 2 and test 3) 130
Table 23: Simulated dataset of 500 participants assuming conditional dependence between two tests (test 2 and test 3) with specificities close to one. 131
Table 24: Estimated prevalence, sensitivities and specificities of the three tests from the different LCMs assuming conditional dependence of two tests (test 2 and test 3). 132
Table 25: Estimated prevalence, sensitivities and specificities of the three tests from the different LCMs assuming conditional dependence of two tests (test 2 and test 3). 133
Table 26: Simulated dataset of 500 participants assuming conditional dependence among all tests. 134
Table 27: Estimated sensitivities and specificities of the three tests from the different LCMs under the conditional independence assumption. 135
Table 28: Estimated prevalence, and the sensitivities and specificities of the three tests from the different LCMs assuming conditional dependence among all three tests..136
Table 29: Simulated dataset of 500 participants assuming conditional dependence among all tests and the specificities of all the tests are close to one.....137
Table 30: Estimated prevalence, and the sensitivities and specificities of the three tests from the different LCMs assuming conditional dependence among all three tests..138
Table 31: Estimated prevalence, and the sensitivities and specificities of the three tests from the different LCMs assuming conditional dependence among all three tests..139
Table 32: Comparison of the different LCMs explored in this chapter 144

Table 33: Estimated sensitivities and specificities of LFpen and FC in classifying teeth with D3 148
Table 34: Demographic profile of RA patients and mean value of core set of variables 150
Table 35: The Disease Activity Scores 151
Table 36: Table of various cut-offs of the disease activities scores 152
Table 37: Prior information on the sensitivity and specificity of DAS28-ESR$_4$, SDAI and CDAI 158
Table 38: Median, and quartiles values of sensitivity and specificity used to elicit the prior Beta distribution 160
Table 39: Number of participants classified as being in remission and non-remission in RABR 161
Table 40: Combination of DAS28-ESR$_4$, CDAI and SDAI responses using RABR dataset 161
Table 41: Estimated sensitivity and specificity of SDAI and CDAI in the RABR dataset assuming that the DAS28-ESR$_4$ is a gold standard 162
Table 42: Sensitivity and specificity of DAS28-ESR$_4$, CDAI and SDAI in the RABR dataset assuming that all scores are conditionally independent. 163
Table 43: Sensitivity and specificity of DAS28-ESR$_4$, CDAI and SDAI in the RABR dataset assuming that all scores are conditionally dependent. 164
Table 44: Sensitivity and specificity of DAS28-ESR$_4$, CDAI and SDAI in the RABR dataset assuming that all scores are conditionally dependent (sensitivity analysis). 170
Table 45: Methods employed for single binary index test 231
Table 46: Methods employed for multiple binary index tests 235
Table 47: Methods employed for single ordinal index test 240
Table 48: Methods employed for single continuous index test 242
Table 49: Methods employed for single continuous index test with focus on covariate-specific ROC 247
Table 50: Methods employed for multiple ordinal or continuous index tests 249
Table 51: Methods employed for single ordinal index test with ROC surface and VUS 250
Table 52: Methods employed for single continuous index test with ROC surface and VUS 251

Table 53: Methods employed for multiple binary index tests and categorical disease status.....253
Table 54: Methods employed to evaluate index test(s) when the sensitivity and specificity of the imperfect reference standard is known precisely. 254
Table 55: Methods employed to evaluate index test(s) when the sensitivity and specificity of the imperfect reference standard is unknown. 256
Table 56: Construction of reference standard. 261
Table 57: Table of other methods employed to evaluate medical test(s) 262

List of Figures

Figure 1: Diagnostic accuracy study with disease status of all participants known before the application of the index test.....5
Figure 2: Classical design of diagnostic accuracy study. All participants undertake both the index test and gold standard. 6
Figure 3: 2 by 2 contingency table of classification of binary test responses.....7
Figure 4: Example of receiving operating characteristic curves.....10
Figure 5: Diagnostic accuracy study with missing gold standard.....12
Figure 6: Differential verification with complete verification 13
Figure 7: The PRISMA flow-diagram of articles selected and included in the systematic review. 23
Figure 8: Imputation and bias-correction for partial verification methods with binary diagnostic outcome**.** 29
Figure 9: Imputation and bias-correction for partial verification methods in three class diagnostic outcomes where ROC and VUS are estimated 30
Figure 10: Guidance flowchart of methods employed to evaluate medical tests in missing and no gold standard scenarios. 34
Figure 11: The mean, standard error, mean square error and bias of the unadjusted and corrected sensitivity and specificity of the index test when the reference standard is prefect.....57
Figure 12: The mean, standard error, mean square error and bias of the unadjusted and corrected sensitivity and specificity of index test when the reference standard is imperfect and better than the index test. 61
Figure 13: The mean, standard error, mean square error and bias of the unadjusted and corrected sensitivity and specificity of index test when the reference standard is imperfect and worse than the index test. 65
Figure 14: The mean, standard error, mean square error and bias of the unadjusted and corrected sensitivity and specificity of index test when the reference standard is imperfect and have same sensitivity and specificity as the index test. 69
Figure 15: The unadjusted and corrected mean sensitivity and mean specificity of the index test when the sensitivity (or specificity) of the reference standard or index test is varied and the prevalence is fixed at 0.3. 71

Figure 16: Changes in prevalence largely impact the unadjusted and Brenner corrected sensitivity and specificity of the index test, unlike the Staquet et al correction method. 73
Figure 17: The unadjusted and corrected sensitivities and specificities of the index test under different variations of conditional dependence between the index test and the reference standard. 85
Figure 18: Trace plots of the sensitivity and specificity of the three tests and prevalence when all tests are conditionally independent. 123
Figure 19: Density plots of the sensitivity, specificity of the three tests and prevalence when all tests are conditionally independent. 124
Figure 20: Auto-correlation plot of the sensitivity and specificity of the three tests .124
Figure 21: Gelman diagnostic plots of the sensitivities and specificities of the three tests.....125
Figure 22: Scatterplots of DAS28-ESR_4, SDAI and CDAI.....153
Figure 23: Correlation matrix plot of DAS28-ESR_4, SDAI and CDAI 153
Figure 24: Histogram and density plot of SDAI, CDAI and DAS28-ESR_4.....154
Figure 25: Density plots of the prior and posterior distribution of the sensitivities of DAS-ESR, SDAI and CDAI using the FEM_W on the RABR dataset.....165
Figure 26: Density plots of the prior and posterior distribution of the specificities of DAS-ESR, SDAI and CDAI using the FEMW on the RABR dataset.....166
Figure 27: Estimated sensitivities and specificities of DAS28-ESR_4, SDAI and CDAI under the different assumptions (RABR dataset) 168
Figure 28: Density plots of the prior and posterior distribution of the sensitivities of DAS-ESR, SDAI and CDAI using the FEM_W on the RABR dataset.....170
Figure 29: Density plots of the prior and posterior distribution of the specificities of DAS-ESR, SDAI and CDAI using the FEMW on the RABR dataset.....171
Figure 30: Trace plots of the sensitivities and specificities of the three tests assuming that all tests are conditionally independent.....297
Figure 31: Density plots of the sensitivities and specificities of the three tests assuming that all tests are conditionally independent.....298
Figure 32: Autocorrelation plots of the sensitivities and specificities of the three tests assuming that all tests are conditionally independent. 298
Figure 33: Gelman diagnostic plots of the sensitivities and specificities of the three tests assuming that all tests are conditionally independent. 299

Figure 34: Auto-correlation plots of the prevalence, sensitivities and specificities of the tests assuming that all tests are conditionally dependent.....300
Figure 35: Trace plots for the sensitivities and specificities of test 1, test 2 and test 3, and prevalence assuming that all tests are conditionally dependent.....300
Figure 36: Density plots of the sensitivities and specificities of the three tests and the prevalence assuming that all tests are conditionally dependent.....301
Figure 37: Gelman diagnostic plots of prevalence, sensitivities and specificities of the three tests.....302
Figure 38: Trace plots of sensitivities and specificities of the three tests, and the prevalence assuming that all tests are conditionally dependent.....303
Figure 39: Density plots of sensitivities and specificities of the three tests, and the prevalence assuming that all tests are conditionally dependent.....305
Figure 40: Trace plots of the sensitivities and specificities of the three tests.....306
Figure 41: Autocorrelation of prevalence, sensitivities and specificities of the three tests.....306
Figure 42: Density plots of the sensitivities and specificities of the three tests.....307
Figure 43: Gelman diagnostic plots for the sensitivities and specificities of three tests 308
Figure 44: Trace plots of sensitivities and specificities of the three tests.....309
Figure 45: Density plots of sensitivities and specificities of the three tests.....309
Figure 46: Auto-correlation plots of sensitivities and specificities of the three tests310
Figure 47: Gelman Diagnostic plot of sensitivity and specificity of the three tests..310
Figure 48: Trace plots of sensitivities and specificities of three tests.....311
Figure 49: Density plots of the sensitivity and specificities of the three tests under the CD assumption (REML) using priors centred on the simulated truth.....312
Figure 50: Trace plots of sensitivities and specificities of the three tests, and the prevalence via the FEM_W model.....313
Figure 51: Density plots of the sensitivities and specificities of the three tests using the FEM_W model and assuming all tests are conditionally dependent.....313
Figure 52: Trace plots of sensitivities and specificities of the three tests 314
Figure 53: Density plots of sensitivities and specificities of the three tests, and the prevalence under the assumption of conditionally dependence (REML).....315
Figure 54: Density Diagnostic plots of 123CD under the assumption of conditional dependence using the FEM_W. 316

Figure 55: Trace Diagnostic plots of 123CD under the assumption of conditional dependence using the FEM_W. 317
Figure 56: Trace plots of sensitivities and specificities of DAS28-ESR_4, SDAI and CDAI 318
Figure 57: Auto-correlation plots of sensitivities and specificities of DAS28-ESR_4, SDAI and CDAI 319
Figure 58: Density plots of sensitivities and specificities of DAS28-ESR_4, SDAI and CDAI.....319
Figure 59: Gelman diagnostic plots of sensitivities and specificities of DAS28-ESR_4, SDAI and CDAI 320
Figure 60: Density plots of the prior and posterior distribution of the sensitivities of DAS28-ESR_4, SDAI and CDAI using the REML on the RABR dataset 321
Figure 61: Density plots of the prior and posterior distribution of the specificities of DAS28-ESR_4, SDAI and CDAI using the REML on the RABR dataset 321
Figure 62: Trace plots of sensitivities and specificities of DAS28-ESR_4, SDAI and CDAI (REML).....322
Figure 63: Density plots of sensitivities and specificities of DAS28-ESR_4, SDAI and CDAI (REML) 323
Figure 64: Auto-correlation plots of sensitivities and specificities of DAS28-ESR_4, SDAI and CDAI (REML) 323
Figure 65: Gelman diagnostic plots of sensitivities and specificities of DAS28-ESR_4, SDAI and CDAI (REML) 324
Figure 66: Trace plots of sensitivities and specificities of DAS28-ESR_4, SDAI and CDAI FEM_W.....325
Figure 67: Density plots of sensitivities and specificities of DAS28-ESR_4, SDAI and CDAI FEM_W.....326
Figure 68: Trace plots of the sensitivities and specificities of SDAI, CDAI and DAS28-ESR_4 (FEM_W). 327
Figure 69: Density plots of the sensitivities and specificities of SDAI, CDAI and DAS28-ESR_4 (FEM_W). 327

List of Abbreviations

Abbreviation	Meaning
ASC-H	Atypical squamous cell high grade
AUC	Area under the ROC curve
BBM	Beta – binomial model
BPS	Bladder pain syndrome
CD	Conditional dependence
CDAI	Clinical Disease Activity Index
CI	Conditional independence
CINAHL	Cumulative index to Nursing and Allied Health Literature
CRP	C-reactive protein level
CRS	Composite reference standard
CT	Computed tomography
CTC	Circulating tumour cell
D3	Dentine caries lesions
DAS28ESR	Disease Activity Score 28 joints
dCRS	Dual composite reference standard
DIC	Deviance information criteria
EBM	Evidence based medicine
EDGA	Evaluator determined disease activity
ELISA	Enzyme-linked immunosorbent assay
EMBASE	Excerpta Medica dataBASE
ESCC	Oesophageal squamous cell carcinoma
ESR	Erythrocyte sedimentation rate

FC	Florescence camera
FEM	Fixed effect latent class model
FEM_W	Fixed effect latent class model proposed by Wang et al
FMM	Finite mixture latent class model
FN	False negative
FNAB	Fine needle aspiration biopsy
FP	False positive
GRE	Gaussian random effect
HAVS	Hand and arm or full body vibration test
HDA	High disease activity
HER2	Human epidermal growth factor receptor 2
HIV	Human immune virus
HMC	Hamilton Monte Carlo
HRA	High resolution anoscopy
HSIL	High grade squamous intraepithelial lesion
HWCI	Hui and Walter conditional independence
ICDAS	International Caries Detection and Assessment System
LCA	Latent class analysis
LCM	Latent class model
LDA	Low disease activity
LFpen	Laser fluorescence pen
MAR	Missing at random
MATCH	Multidisciplinary Assessment of Technology for Healthcare
MCMC	Markov Chain Monte Carlo

MDA	Moderate disease activity
MEDLINE	Medical Literature Analysis and Retrieval System Online
MNAR	Missing not at random
MRI	Magnetic resonance imaging
MSE	Mean square error
NC	Non-cavitated caries lesions
NUTS	No-U-Turn sampler
PCR	Polymerase chain reaction
pD	Number of effective parameters
PDGA	Patient global disease activity
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PSYCINFO	Psychology information database. It is a database of abstract of literature in the field of psychology.
ptVAS	Patients global disease activity visual analogue scale
QoI	Quality of life
RA	Rheumatoid arthritis
RABR	Rheumatoid arthritis clinical dataset at baseline discriminating between participants in remission and non-remission
REM	Random effect latent class model
REML	Random effect model via logit link
REMP	Random effect model via probit link
rFN	Relative False negative
rFP	Relative False positive

ROC	Receiver operating characteristics
RS	Reference standard
rTN	Relative True negative
rTP	Relative True positive
SCOPUS	Source-neutral abstract and citation database
SDAI	Simplified Disease Activity Index
SE	Standard error
SHELF	Sheffield ELicitation Framework
SJC	Swollen joint count
Sn	Sensitivity
Sp	Specificity
TEN28	Tenderness upon touching 28 joints
TJC	Tender joint count
TLCM	Traditional latent class model
TN	True negative
TNM	Tumour node metastasis
TP	True positive
VUS	Volume under ROC surface

Chapter 1: General Introduction

1.1. Introduction

Prior to adopting a new medical test in a healthcare system, the test should be evaluated in terms of safety, performance and efficacy. As part of the evaluation of the test, a few core questions are considered, such as:

- Is the test safe?
- Does the test work under controlled laboratory conditions?
- Does the test work under real-life clinical conditions in the hands of the intended users?
- Do the long- and short-term benefits of doing the test outweighs any risks associated with it?

All these questions can be answered using robustly designed studies[1, 2].

Estimating the diagnostic accuracy of a medical test is just one part of the puzzle, but it is an important step in evaluating any new test[3]. Ultimately, clinicians, health practitioners and patients need assurance that a test has the ability to discriminate between patients who have the target condition that the test is designed to detect, and those who do not. Basic measures used to assess the accuracy of a medical test are "*sensitivity"* and "*specificity*". Studies that focus on estimating these measures are "*diagnostic accuracy studies*". These terms will be defined and discussed later in this chapter.

1.1.1. Medical test

Medical tests are employed to support clinical decisions about the management and care of individuals. According to the Cochrane Diagnostic Test Accuracy (DTA) portfolio[4], a medical test is defined as "*any observation, measurement made on an individual as well as more classic technology based test that involve medical and laboratory procedures on a person or on a sample or tissue or fluid from a person; and the key thing is that the observation is made to infer something about the health state of the individual*". Medical tests perform various roles such as screening, diagnosis, staging of disease, surveillance, and prognosis amongst others [4]. The descriptions of some of these roles are given in Table 1.

Table 1: Description of different roles performed by different medical tests.

Roles	Description
Screening	Medical tests employed for screening purposes (screening tests) are used to identify people with an increased risk of the target condition or disease or for early detection of disease. They are normally used on apparently healthy people who may have not shown signs or symptoms of the disease and the aim is to detect the target condition early [5, 6]. An example of a screening test is the cervical cancer screening test offered to women age 25 to 64 for the detection of cervical cancer [7].
Diagnosis	Medical tests employed for diagnostic purposes (diagnostic tests) are used to confirm the presence or absence of the target condition in people who may have shown signs or symptoms of the target condition or may have had positive screening test results [4, 8]. Examples include the fine needle aspiration biopsy (FNAB) used in the diagnosis of thyroid nodule disease [9].
Prognosis	Prognostic medical tests are used to "predict patients' likelihood of experiencing a medical event" in the future, such as developing a disease, recovery from a disease or death[10]. They can also be used to inform patients' treatment options. An example is the identification of prognostic markers or factors used when deciding on the treatment of breast cancer such as the tumour size, human epidermal growth factor receptor 2 (HER2) status, estrogen-receptor (ER) status[11] amongst others.

Table 1 cont.: Description of different roles performed by different medical test.

Roles	Description
Staging	Medical tests serving as staging tests are used to describe the progression of a disease. An example is the Tumour Node Metastasis (TNM) staging system which is a standard test used in describing the severity or growth of cancer [12]. Another example of tests used for staging purpose are the simplified disease activity index (SDAI), which is used to stage patients with rheumatoid arthritis as having mild, moderate or high disease activity[13].
Surveillance	Surveillance tests are employed to detect early signs of a target condition or illness among people who are exposed to agents causing the target condition[14] even before they show signs or symptoms. These are tests offered to people who are potentially exposed to hazardous substance or noise such as radiation or biological agents amongst others. An example is the Hand and arm or full body vibration test[15] (HAVS) employed to test for hand-arm vibration syndrome among workers exposed to machine-induced vibration while working.
Monitoring	Monitoring tests are used to monitor the progression of a disease or the response of patients to a treatment[16]. An example is the follow-up tests offered to cancer survivors such as imaging tests (mammogram, colonoscopy, bone scans and x-ray amongst others) and blood tests which measure the level of blood tumour markers in the blood[17-19]. In addition, the circulating tumour cell (CTC) is employed to monitor tumour progression in oesophageal squamous cell carcinoma (ESCC)[20].

A medical test can serve more than one purpose described in Table 1. For example the disease activity scores (DAS28) can be used to monitor disease activity in patients with rheumatoid arthritis and stratify them into remission, low, moderate or high disease activity[21]. It can also be used to guide rheumatologists on treatment options or doses

of treatment for the patients. Therefore, this test can be used for monitoring and as a guide for treatment decisions. Another example is the mammogram which can be offered as a screening test for early detection of breast cancer in women, and also as a monitoring test for women who have undergone breast cancer treatment.

Furthermore, some diagnostic or screening tests can be used to provide information about the spread of a target condition in a population and to monitor how effective an intervention or prevention has been over a time-period in that population. This process is also known as surveillance. A topical example would be the COVID-19 tests which are currently being used to monitor infection rates and how effectively some of the measures or policies put in place have prevented or affected the spread of COVID-19[22].

1.1.2. Diagnostic Accuracy

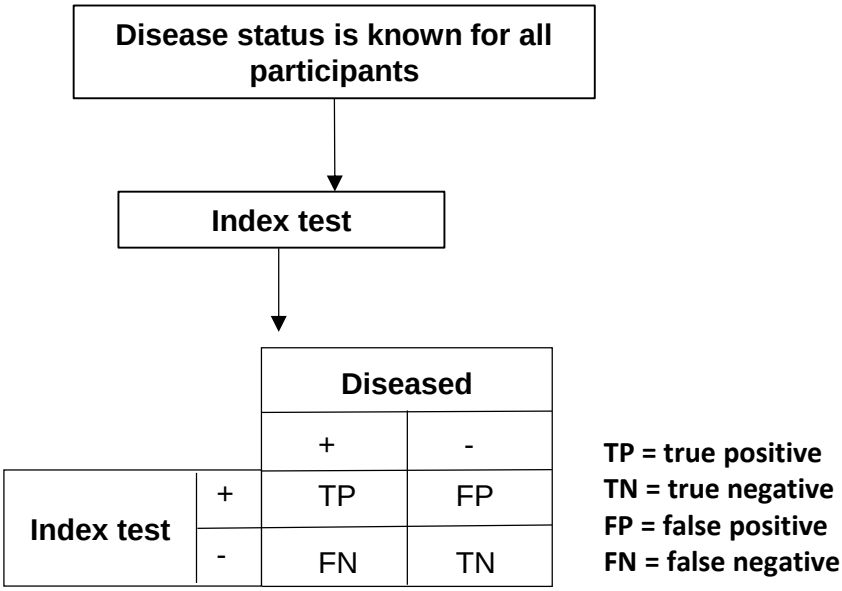
The diagnostic accuracy of a test is the ability of the test to appropriately discriminate between people with the target condition and people without the target condition[23, 24]. The target condition is often a disease. Studies that seek to evaluate the ability of a new (or index) test to differentiate participants with or without the target conditions are often referred to as diagnostic accuracy studies[25]. Throughout this thesis, I will use the terminology "***participants***" to describe people included in a diagnostic accuracy study rather than "***patients***" because those included in the study may be healthy people or those who may or may not have shown signs and symptoms of disease. The purpose of any diagnostic accuracy study is to estimate the ability of the index test to distinguish between participants with or without the target condition in a specific population of individuals. The test that is under evaluation in a diagnostic accuracy study is termed an '***index test'***. Aside from the various purposes of medical tests (Table 1); there are different roles the index test could take in the current diagnostic pathway. For instance, the index test may replace an existing medical test, or be a triage test (where the results inform decisions to conduct another test or not), or an add-on test[3] to the current tests undertaken. The potential role it will take should be taken into consideration when designing the diagnostic accuracy study to identify the most appropriate reference standards and estimate its accuracy within the diagnostic pathway.

To evaluate the diagnostic accuracy of an index test, the test is compared to existing test, which is currently used to diagnose the same target condition (disease) as the

index test. This latter test is referred to as a reference standard (RS) (see section **1.1.3** below).

This study is often referred to as a ***diagnostic accuracy study***[3]. Within the diagnostic accuracy study, it is also expected that the participants will undergo both the index test and the reference standard. However, if the disease status of the participants is already known before conducting the index test, the accuracy of the index test can be estimated without applying the reference standard[26]. The classification of the index test response with respect to the disease status is depicted in Figure 1 and the sensitivity and specificity of the evaluated test are derived using the formulas described in section **1.1.5**.

Figure 1: Diagnostic accuracy study with disease status of all participants known before the application of the index test



1.1.3. Reference Standard

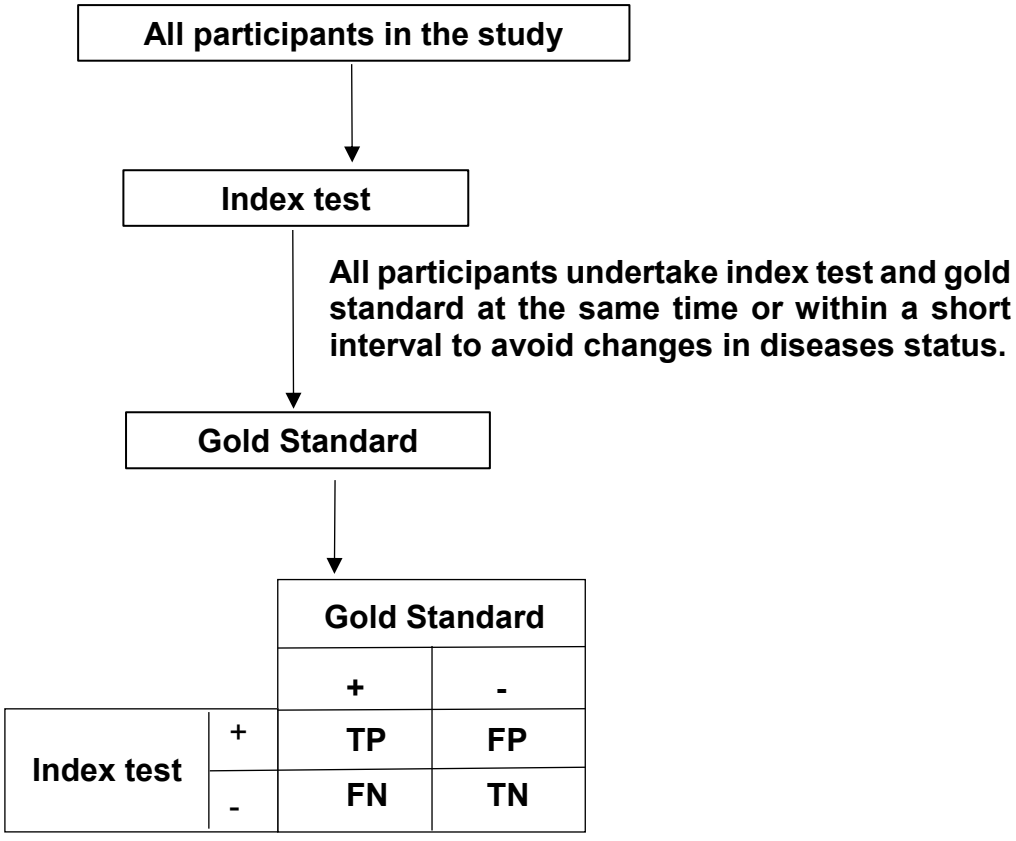
The reference standard is an existing test that is used as a benchmark to evaluate the index test[27, 28]. The preferred reference standard is often the best test used to diagnose the target condition of interest. Sometimes, the assumption is made that the reference standard perfectly discriminates between participants with or without the target condition. In this case, the reference standard is referred to as a gold standard.

1.1.4. Gold Standard

A gold standard as implied and defined by some researchers, is the "best available reference test"[29, 30] used as a benchmark to evaluate the index test. However, for other researchers, a gold standard is defined as a perfect reference test with 100% sensitivity and 100% specificity[31, 32]. With the former definition (or application), any imperfection in the presumed "gold standard" is not accounted or corrected for. In this thesis, the latter definition of a gold standard is employed.

The classical procedure of evaluating the index test using a gold standard is described in Figure 2 and the responses from both tests can be classified into a 2 – by – 2 contingency table given that the disease status of the participants can be classified into two states (diseased and non-diseased).

Figure 2: Classical design of diagnostic accuracy study. All participants undertake both the index test and gold standard.



This dichotomy is often artificial if test results are reported as a continuous outcome and hence these outcomes can be classified into a 2 x 2 table using cut-offs. Receiver Operating Characteristic curve (ROC) analysis is used to determine a threshold (test positive cut-off) above which the condition is judged to be present[33]. The test positive

cut-off is used to categorise the patients as having the target condition or not; and the 2-by-2 contingency table (*Figure 3*) can then be used to display the results.

The next section on diagnostic accuracy statistics assumes that the reference standard is a gold standard. However, if the reference standard is not a gold standard, estimating the sensitivity and specificity of the index test without considering the imperfection of the reference standard will yield biased estimates. This issue is discussed in more detail towards the end of this chapter (section **1.1**).

1.1.5. Diagnostic Accuracy Statistics

Diagnostic accuracy studies use accuracy statistics such as sensitivity, specificity, predictive values (positive and negative), area under the ROC curve, and likelihood ratios to convey the ability of an index test to discriminate between those with the disease (target condition) and those without the disease[24, 34]. The 2x2 table (Figure 3) displays concordance and discordance between the two tests (the index test and reference test). Disagreements between the two tests under comparison can be classified as false results (either as a false positive or false negative) for the index test being evaluated. With the assumption that the reference standard is a gold standard, the value of true positives (TPs), false positives (FPs), false negatives (FNs) and true negatives (TNs) are used to estimate the sensitivity and specificity of the index test.

Figure 3: 2 by 2 contingency table of classification of binary test responses

		Reference Test	
		Positive	Negative
Index Test	Positive	True Positives (**TP**)	False Positives (**FP**)
	Negative	False Negatives (**FN**)	True Negatives (**TN**)

The diagnostic accuracy statistics are defined below alongside how they are estimated or obtained using the quantities on Figure 3.

- **Sensitivity**

The sensitivity (Sn) of a test is the probability of that test to correctly classify participants with the target condition (or disease) as having the disease.

$$Sn = \frac{TP}{TP + FN}$$

- **Specificity**

The specificity (Sp) of a test is the probability of that test to correctly identify participants without the target condition as not having the disease.

$$Sp = \frac{TN}{TN + FP}$$

- **Positive Predictive Value**

The positive predictive value (PPV) is the probability that participants who have a positive index test results have the disease.

$$PPV = \frac{TP}{TP + FP}$$

- **Negative Predictive Value**

The negative predictive value is the probability that participants who have a negative index test results do not have the disease.

$$NPV = \frac{TN}{TN + FN}$$

- **Likelihood Ratio**

A positive likelihood ratio tells us the relative likelihood that a positive test result would be expected in a person with the disease compared to a person without the disease. The positive likelihood ratio is the ratio of sensitivity to the false positive rate (1 – Sp).

$$LR(+) = \frac{Sn}{(1 - Sp)}$$

A negative likelihood ratio tells us the relative likelihood that a person with the disease would have a negative test result compared to a person without the disease. The negative likelihood ratio is the ratio of the false negative rate (1 – Sn) to the specificity**.**

$$LR(-) = \frac{(1 - Sn)}{Sp}$$

- **Diagnostic Odds Ratio**

Diagnostic odds ratio (DOR) is the ratio of the odds of disease in individuals who test positive relative to the odds of disease in individuals who test negative. It measures the ability of a test to discriminate between the diseased and non-diseased participants correctly. It can be calculated using the positive likelihood ratio and the negative likelihood ratio, or the NPV and PPV, or sensitivity and specificity of a test.

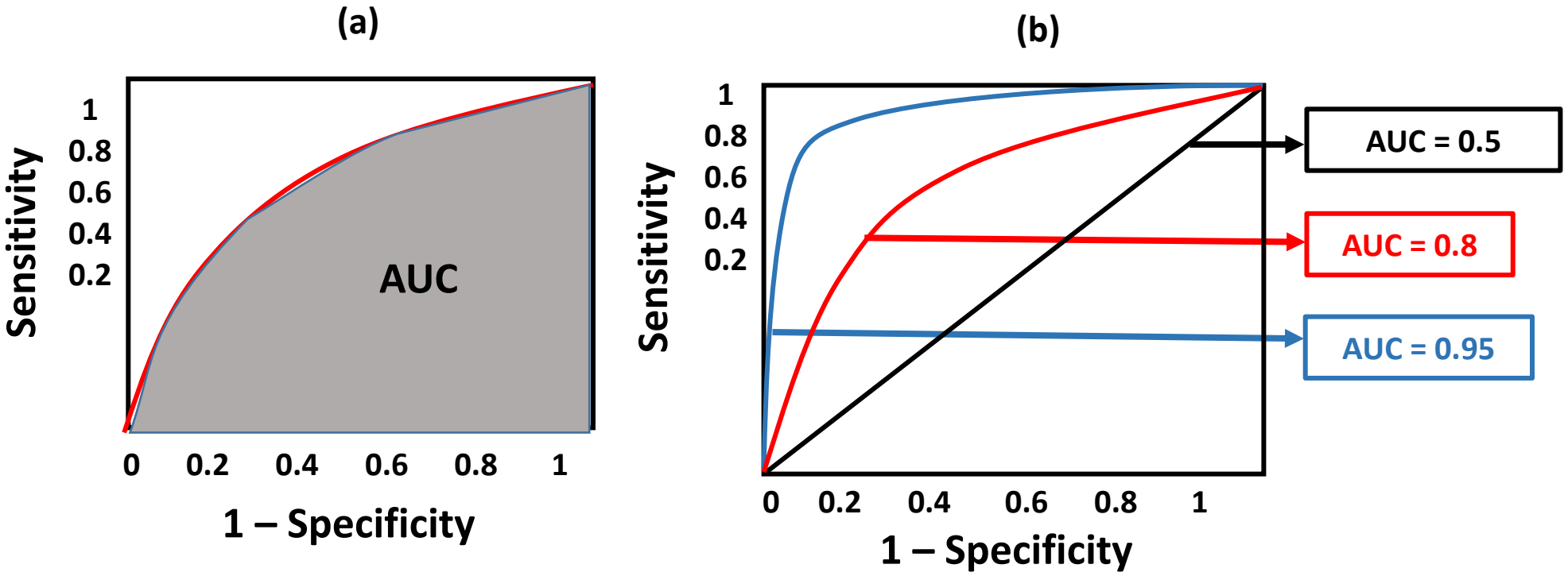
$$DOR = \frac{LR(+)}{LR(-)} = \frac{Sn \times Sp}{(1 - Sn) \times (1 - Sp)} = \frac{PPV \times NPV}{(1 - PPV) \times (1 - NPV)} = \frac{TP \times TN}{FN \times FP}$$

- **Receiver Operating Characteristic Curve**

The ROC is employed to display and measure the accuracy of a test with a continuous outcome at different cut-offs. The receiver operating characteristic (ROC) curve (

Figure 4) is a plot of true positive rate (sensitivity) against the false positive rate (1 – specificity) for the different cut-offs. Each point on a ROC curve is the sensitivity and false positive rate at a positivity threshold or cut-off employed to classify individuals into diseased or non-diseased. The area under the ROC curve (AUC) is the grey-coloured area under the ROC curve (see Figure 4(a)). The AUC measures the ability of the test to discriminate between participants with or without the target condition at different cut-offs. An AUC of 0.5 indicates a poor test as its ability to distinguish between those with and without the target condition is no better than chance. It is expected that an excellent test will have an AUC close to one. In Figure 4(b) the black diagonal line (also known as the reference line) indicates the ROC of a useless test which randomly classifies individuals. The red ROC curve displays a test with good discriminating ability and the blue ROC curve displays a test with much better ability to discriminate between participants with or without the disease. In the comparison of all the tests plotted on Figure 4(b) the test with the highest AUC (0.95) is preferred.

Figure 4: Examples of receiving operating characteristic curves



1.2. Statement of Problem

When the true disease status of the participants is unknown the index test is evaluated in evaluated against with the preferred reference standard. Some researchers may assume the reference standard to be a gold standard despite its misclassification errors[35-38]. This approach may be favoured because it is simple or because the error is assumed to be insignificant/trivial. However, there are many studies which take into consideration the imperfection of the reference standard or the missingness of the gold standard to accurately estimate the sensitivity and specificity of the index test and avoid bias. ***Bias*** is considered to be the difference between the true / actual value of a parameter and the estimate of that parameter[39]. The techniques or methods employed in evaluating the accuracy of an index test depend on:

- The accuracy of the reference standard used in the study
- The availability of the reference standard to all participants of the study.

The "availability of the reference standard" aims to answer the question, "*Will the reference standard be applied to all the participants in the study (complete verification) or will (or did) a sub-sample of the participants undergo the reference standard (partial verification)*?" These two components play a significant role in deciding which method will be employed to evaluate the index test. If all participants in the study underwent both the index test and the reference standard; and the reference standard is assumed to be a gold standard then it can be further assumed that it is appropriate to employ the classical design (Figure *2*) and the sensitivity and specificity of the index test can be estimated using the methods described in section **1.1.5**.

Consider a scenario where all the participants underwent the index test but only a sub-sample of the participants underwent the gold standard. This may occur for some reasons like the gold standard being invasive, unethical, expensive, time-consuming, or unavailable to all participants at the time of the study. By extension, the consequences of this is that the true disease status for some of the participants will be missing. When there is such missing information, evaluating the index test using only the complete cases introduces verification or workup bias[40-42]. ***Verification bias*** is a bias in the estimated diagnostic accuracy caused by ignoring information on the participants that were not verified using the chosen reference standard[43, 44]. To overcome this bias some researchers consider verifying the remaining participants with

an alternative reference standard. The alternative reference standard is often imperfect but may be less invasive or less costly compared to the gold standard. This method still introduces differential verification bias because the two tests can differ in quality (or accuracy measures) and in how they measure the target condition[45]. ***Differential verification bias*** is bias on the estimated diagnostic accuracy of the index test as a result of using different reference standards and combining the results as a single reference standard[46, 47]. Studies where the true disease status of a sub-sample of the participants is verified with the gold standard is depicted in Figure 5.

Figure 5: Diagnostic accuracy study with missing gold standard

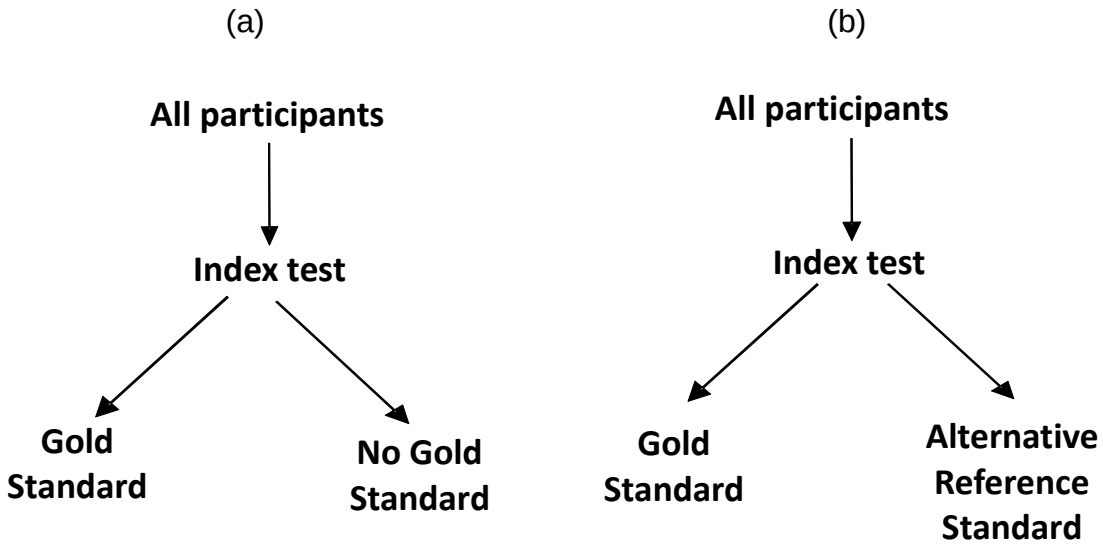
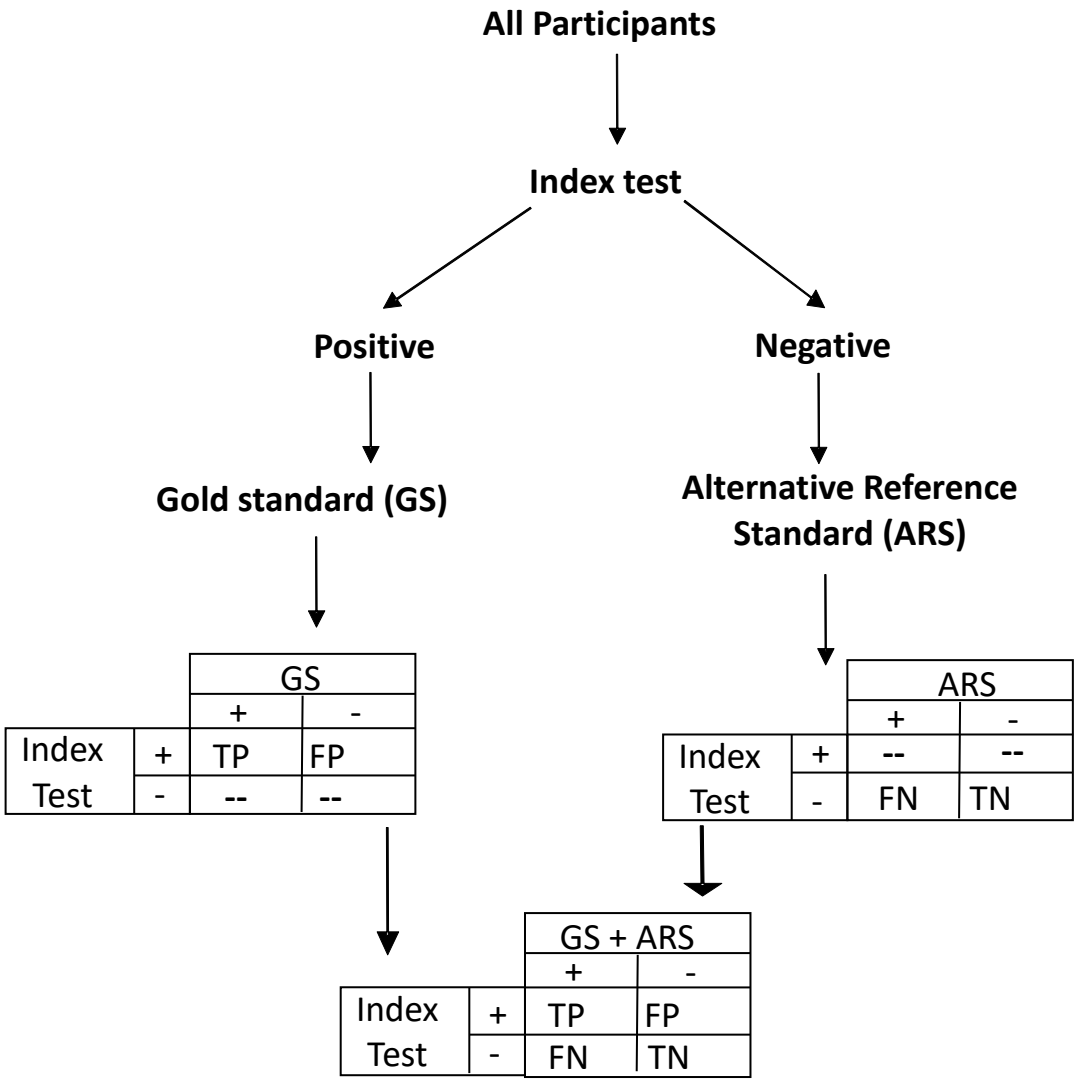


Figure 5(a) describes a study where the participants, whose true disease status was unverified by the gold standard, are assumed missing (partial verification study design). Figure 5(b) is a study where the participants whose true disease status was unverified by the gold standard underwent an alternative reference standard (differential verification study design).

An adapted version of the study design depicted in Figure 5(b) is the complete differential verification design[48] (Figure 6), because all participants with a positive index test results undertake the gold standard, and all participants with negative index test results undertake the alternative reference standard. This design may be used when the gold standard is invasive and so it may be unacceptable for all participants to undergo the gold standard. Ultimately, the results of both reference standards are pulled together to estimate the accuracy measures of the index test.

Figure 6: Differential verification with complete verification



Another potential scenario is where there is no accepted / consensus reference standard, because all tests available (validated or not) are unreliable. For example, in the diagnosis of bladder pain syndrome (BPS), there are some validated tests (questionnaires) developed based on expert opinion such as the O'Leary-Saint Interstitial Cystitis Symptom Index. However, there are no universally accepted reference / gold standard tests to diagnosis this target condition[49-51].

A further variant of the problem is where the best reference standard is not perfect because it still has some misclassification error**.** Some of the participants classified as having the target condition may not have the target condition and vice versa. Thus, using an imperfect reference standard imposes bias in the diagnostic accuracy

statistics of the index test, which is known as reference standard bias [52,53]. ***Reference standard bias*** is a bias on the estimated diagnostic accuracy of the index test as a result of using an imperfect reference standard [28].

Evaluating an index test in the aforementioned scenarios will potentially introduce bias. These biases can either overestimate or underestimate the sensitivity and specificity of the index test[54], if the classical approach of diagnostic accuracy study is followed. The consequences of this is that an index test may be adopted by practitioners because of the "false accurate" data. Similarly, a test may be disregarded/abandoned because of "false inaccurate" data. In either circumstance, patients may receive sub-optimal care and suffer poor outcomes.

In recent years, rapid advances in technology having led to an influx of sophisticated medical diagnostic technologies which seek to optimize disease detection and so, evaluating a medical test in the absence of a gold standard is a very pertinent issue with implications for medical test developers, evaluators, patients, commissioners and clinicians alike. They may all be faced with the question *"How can we ascertain that the new tests developed, accurately discriminate between those with or without the target condition?*

1.2.1. Review of types of methods

Various research has been undertaken to develop statistical methods to evaluate the sensitivity and specificity of the index test in the absence of a gold standard. A review of methods in context was carried out by Rutjes et al[41] in 2007. The identified methods were grouped into four broad groups which are:

- Impute or adjust for missing data on reference standard
- Correct imperfect reference standard
- Construct reference standard and
- Validate index test result.

A summary of the identified methods by Rutjes et al[41] is presented in Table 2. As part of my PhD research, a systematic review has been conducted as an update to the review by Rutjes et al[41]. The methodology and results of this review are reported in chapter two.

Table 2: Summary of the methods employed to evaluate medical tests in the absence of a gold standard.

Classification group	Characteristics
Impute or adjust for missing data on reference standard	This group presumes that there is a perfect reference standard. However, not all patients can be verified using the reference standard test. Evaluating the index test is done either by: A. Imputing outcomes for the missing data (reference standard) using a single or multiple imputation technique for patients who did not receive the reference standard. OR B. Adjust / correct the estimate of the sensitivity and specificity using information from patients who underwent both tests (index test and reference standard test) through statistical modelling.
Correct imperfect reference standard	This group accepts that all the reference standard tests do not perfectly discriminate between patients with the target condition and patients without the target condition. However, if the amount and the type of error associated with the reference standard test(s) is known, then estimates of the sensitivity and specificity of the index test can be corrected using some algebraic function.
Validate index test	In this group, the index test is evaluated based on what it is supposed to measure. It does not evaluate the index test against a reference standard rather it studies the index test results in relation to clinical characteristics associated to the target condition of interest.

Table 2 cont.: Summary of the methods employed to evaluate medical tests in the absence of a gold standard.

Classification group	Characteristics
Construct reference Standard	This group combines information from different tests to form a reference standard. There are two sub-groups within this group. The first sub-group includes methods where all patients receive the index test but different sets of reference tests (A & B) and the second sub-group is where all patients receive the same set of tests (index test and reference standard tests see C, D and E).
	A. Differential Verification where all patients receive the index test but a subgroup of patients receive a different reference standard test.
	B. Discrepancy analysis: All patients receive the index test and a reference standard test. However, patients with discordant results are subjected to another reference test known as the resolver test.
	C. Composite reference standard: Two or more imperfect reference standard tests are combined by a pre-specified rule to construct a reference standard that is used to evaluate the index test.
	D. Panel / Consensus Diagnosis: With this approach, a group of experts determine the presence or absence of the target condition in each patient using multiple sources information.
	E. Latent class analysis: This method uses a statistical model to combine information from all the tests to form a reference standard. The true diseases status is assumed to be unknown.

1.3. Scope of the research study

This research study focuses on methodologies or techniques used or proposed in evaluating the diagnostic accuracy of medical tests in scenarios where the gold standard is not applied to a sub-sample of the participants or the reference standard employed in the study is imperfect or there is no reference standard at all.

1.4. Aim of the PhD research study

Given the problems potentially encountered in evaluating the diagnostic accuracy of a medical test in the absence of a gold standard and building upon what other researchers have done: this study aims to investigate all techniques proposed to evaluate the sensitivity and specificity of medical tests in the absence of a gold standard, compare some methods through simulation studies, and apply some methods identified to a real-life clinical dataset.

1.4.1. Objectives

- To identify all proposed methods developed and employed to evaluate the medical tests in the absence of a gold standard.
- To explore and compare methods in order to select promising techniques
- To apply the appropriate proposed methods to analysis of the clinical dataset available for this study.

1.5. Research methodology

To achieve the stated objectives set out above: firstly, a systematic review was carried out to identify all existing methods which included clinical applications of these methods. Their underpinning assumptions, as well as the strengths and weakness of the methods were explored. The systematic review is an update to the review carried out by Rutjes et al[41]. The decision to update the review was based on the timeliness and the relevance of the review question[55, 56], and the emergence of databases not searched in the previous review, such as the Web of Science and Wiley Library. The emergence of these databases were expected to increase the number of eligible articles that will be retrieved for the review. The systematic review is reported in chapter two alongside the methodology underpinning the review and the results. Next, some of the methods identified from the systematic review were compared using simulation studies and clinical datasets. These comparisons were carried out because they have not been compared in previous studies. The comparison of the methods is reported in

chapter three. Finally, the latent class model was further explored and employed to analyse the clinical dataset under investigation in this research study. Latent class analysis was chosen as it was identified as the most appropriate and promising method for our dataset from the results of the systematic review. The investigation of the latent class model is reported in chapter four and the analysis of the real-life clinical dataset is reported in chapter five.

The thesis concludes in chapter six with a discussion on the key contributions of the research, the potential significance of this work and future avenues for research.

1.6. Significance of the research study

Conventionally, an index test is evaluated by comparing it with the best available reference standard when the true disease status of the participants is unknown. Often, the reference standards available are imperfect. Thus, evaluating the diagnostic accuracy of an index test without taking into consideration the imperfection of the reference standard leads to a biased estimation of the accuracy measures of the index test. This bias may lead to over or underestimates of the true sensitivity and specificity of the test under evaluation. The effect of ignoring this bias may mean that if the test is introduced into routine practice, it may have detrimental consequences to patients, clinicians and public health.

In the research reported in this thesis, the updated systematic review provides recommendations for the use of the most promising methodologies to improve the estimation of diagnostic accuracy when there is no gold standard. The research also compared statistical methods identified from the updated review. The focus of this comparison was on the methods used to correct the accuracy measures of the index test given that the accuracy measures of the imperfect reference standard are known. This element of the work informs test evaluators which methods to consider or discontinue when evaluating the accuracy of an index test given the accuracy measures of the imperfect reference standard are known. The research also investigated latent class models using both simulation methods and the analysis of the real-life dataset. This work provides information on which latent class model to employ when evaluating the accuracy of three conditionally dependent tests with higher order correlations.

Chapter Two: Systematic Review

2.1. Introduction

Numerous methods have been proposed and employed in evaluating the diagnostic accuracy of medical tests in the absence of a gold standard. Some of the methods were identified in a review undertaken by Rutjes et al[41]. In this work and as described in section **1.2.1**, the methods were classified into four groups which are:

- Impute or adjust for missing data on reference standard
- Correct imperfect reference standard
- Construct reference standard and
- Validate index test result.

Given the time which has elapsed since this review was undertaken and the increased research recently into diagnostic accuracy studies and their methodology, it is plausible that there have been methods proposed since the original review. Therefore, in order to meet the aims of this research on methods applied to evaluate medical tests in the absence of a gold standard, it is important to identify and explore all methods proposed or employed in diagnostic accuracy studies. These motivations are what inspired my quest to undertake a systematic review.

The term “no gold standard” is defined as scenarios where the reference standard is imperfect (known to have misclassification error), or there are no generally accepted reference standard(s), or there is a missing reference standard (partial and differential verification) for some participants of the study (Rutjes at al[41]).

This chapter has been published in PLOS ONE journal as a systematic review entitled “Diagnostic test evaluation methodology: A systematic review of methods employed to evaluate diagnostic tests in the absence of gold standard – An update”[57].

This chapter comprises of the following sections: the methods employed to carry out the review (section **2.2**); the results (section **2.3**); guidance developed from the results of the review to aid researchers on methods to choose from when evaluating the test performance of medical tests in the absence of a gold standard (section **2.4**); the discussion (section **2.5**) and conclusions of the review (section **2.6**).

2.2. Methodology

A systematic review was used to identify all methods that have been proposed or employed in evaluating medical tests when there is no gold standard. This methodology was chosen because it is the best approach to avoid subjective bias that could arise in selecting research studies or authors of one's own choice. In addition, this approach allows standardised searching across multiple databases for any article or research study that is related to the keywords or search terms used in the search procedure, cutting across different journals, countries, languages etc. Thus, expanding one's knowledge about the topic of interest in order to make a sound judgement or conclusion about the topic of interest.

On deciding to undertake this systematic review, I sought the consent of one of the key- authors of the previous review[41] (Professor Patrick Bossuyt) to update their review. After receiving the consent, a protocol was developed, peer-reviewed and registered on PROSPERO (the registration number is CRD42018089349).

2.2.1. Eligibility Criteria

The review includes methodological articles (that is papers that proposed or developed a method) and application articles (that is papers where any of the proposed methods were applied).

Inclusion criteria

- Articles published in English language in a peer-reviewed journal.
- Articles that focus on evaluating the diagnostic accuracy of new (index) test when there is a missing gold standard, no gold standard or an imperfect reference standard.

Exclusion criteria

- Articles that assumed that the reference standard was a gold standard and the gold standard was applied to all participants in the study. Books were excluded as they could contain information already published in a peer-review journal.
- Books, dissertations, theses, conference abstracts, and articles not published in a peer reviewed journal.
- Systematic reviews and meta-analyses of the diagnostic accuracy of medical test(s) for a target condition (disease) in the absence of gold standard for some or all of the participants. However, individual articles included in these reviews that met the inclusion criteria were included.

2.2.2. Search strategies and selection of articles

The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement[58] was used as a guideline when conducting this systematic review. PRISMA is "*an evidence-based minimum set of items for reporting in systematic reviews and meta-analyses*". The PRISMA checklist for this review is provided in the appendix (Appendix A.1). The following bibliographic databases were searched: EMBASE, MEDLINE, SCOPUS, WILEY online library (which includes Cochrane library, EBM), PSYCINFO, Web of Science, and CINAHL.

Search term

The keywords employed in the databases to search out articles included in this review are:

("No gold standard*" OR "without gold standard*" OR "missing gold standard*" OR "imperfect reference standard*" OR "no reference standard*" OR "missing reference standard*" OR "partial verification" OR "differential verification") AND ("medical test*" OR "new test*" OR "index test*" OR "diagnostic test*" OR "screening test*" OR "routine*). An example of the comprehensive format is reported in Appendix A.2.

Search date

The search dates were from January 2005 – February 2019. This is because, this review is an update of a review by Rutjes et al[41] who searched up to 2005. However, original methodological articles that proposed and described a method to evaluate medical tests when there is a missing or no gold standard published before 2005 were also included in the review. These original articles were identified by "snowballing"[59] from the references of articles identified by my review.

Selection procedure

All articles obtained from the electronic databases were imported to Endnote X8.0.2[60]. Duplicated articles were removed including books, dissertations, conference abstracts, and articles not published in a peer reviewed journal.

The selection of articles to be included in this review was made done by three people (Chinyereugo Umemneku-Chikere, A. Joy Allen, and Kevin Wilson). The sifting process was in two stages: by title and abstract and then by full text against the inclusion and exclusion criteria. Any discrepancies between reviewers were resolved in a group meeting.

2.2.3. Data synthesis

A data collection form was developed for this review which was piloted on seven studies and remodified to fit the purpose of this review. A copy of the data collection form is in Appendix A.3. Information extracted from the included articles was synthesized narratively. An example of a completed data extraction form is in Appendix A.4.

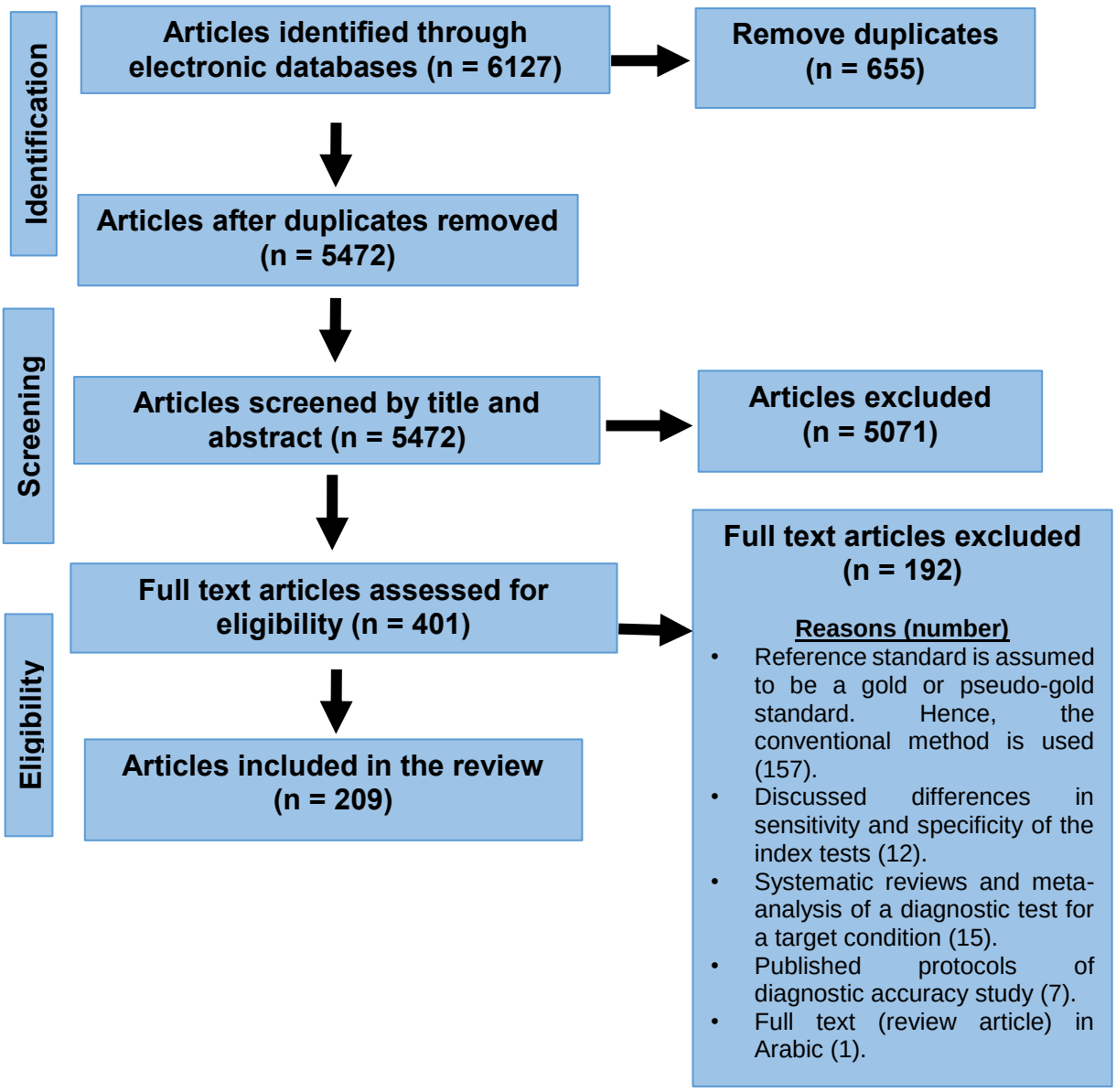
2.3. Results

A total of 6127 articles were identified; 5472 articles were left after removing the duplicated articles; 5071 articles were excluded after sifting by title and abstract; 401 articles went forward to full text assessment; and a total of 209 articles were included in the review. The search and selection procedure are depicted using the PRISMA [58] flow-diagram (Figure 7). The articles included in this review used a wide variety of different study designs including cross-sectional studies, retrospective studies, cohort studies, prospective studies and simulation studies.

One hundred and one papers, which developed a statistical method or model that could be employed to estimate the accuracy measures of the index test in the absence of a gold standard, were identified. These methods were categorised into four groups based on the availability and / or application of a gold standard to the participants in the study. These group are:

- Group 1: Methods employed when there is a missing gold standard.
- Group 2: Correction methods which adjust for using an imperfect reference standard whose diagnostic accuracy is known.
- Group 3: Methods employed when using multiple imperfect reference standards.
- Group 4: ***"other methods"***. This group includes methods such as study of agreement, test positivity rate, and considering alternative study design like validation.

Figure 7: The PRISMA flow-diagram of articles selected and included in the systematic review.



Methods in groups 2, 3 and 4 are employed when there is no gold standard to evaluate the diagnostic accuracy of the index test, while methods in group 1 are employed when there is a gold standard to evaluate the diagnostic accuracy of the index test. However, the gold standard is applied to only a sub-sample of the participants. A summary of all methods identified in the review, their key references and the clinical applications of these methods are reported in Table 3.

Table 3: Summary of classification of methods employed when there is missing or no gold standard.

Main Classification	Main Characteristics	Key references of paper that developed method (number)	Clinical Application of the methods
Group 1: **Method employed when there is missing gold standard:** • Imputation and bias-correction for partial verification methods • Differential verification	The true disease status is verified with the gold standard only in a subsample of the study participants. The methods are grouped into ***imputation and bias-correction for partial verification methods*** (Figure 8 and Figure 9) and the ***differential verification*** approach.	***Imputation and bias correction methods***[43, 61-108] (49) ***Differential verification***[39, 47, 109] (3)	***Imputation & Bias-correction methods***[110-114] ***Differential verification***[115; 116]
Group 2: **Correction methods**	The reference standard is imperfect. However, there is available information about the sensitivity and specificity of the reference standard which is used to correct or adjust the estimated sensitivity and specificity of the index test.	***Correction methods***[117-122] (6)	***Correction methods***[123-125]

Table 3 cont.: Summary of classification of methods employed when there is missing or no gold standard

Main Classification	Main Characteristics	Key references of paper that developed method (number)	Clinical Application of the methods
Group 3: Methods employed when using multiple imperfect reference standards or tests. • Discrepancy analysis • Latent class analysis (LCA) • Composite reference standard (CRS) • Expert or panel or consensus diagnosis	A gold standard or an imperfect reference standard with known diagnostic accuracy may not be available. Thus, multiple imperfect tests may be employed to evaluate the index test. Methods in this group include discrepancy analysis, latent class analysis (frequentist and Bayesian), composite reference standard (CRS), and panel or consensus diagnosis.	***Discrepancy analysis***[126, 127] (2) ***Latent class analysis:*** ***Frequentist LCA***: [128-139] (12) ***Bayesian LCA***: [140-147] (8) ***ROC (NGS)***:[148-158] (11) **CRS**[159-162] (4) **Panel or consensus diagnosis**[163] (1)	***Discrepancy analysis***[164-167] ***Latent class analysis:*** ***Frequentist LCA***[168-180] ***Bayesian LCA***[181-202]; [188, 203-221] ***ROC (NGS)***[222, 223] ***CRS***[224-232]; [233] ***Consensus diagnosis***[234-238]

LCA is latent class analysis; CRS is composite reference standard. ROC is receiver operating characteristics; NGS is no gold standard

Table 3 cont.: Summary of classification of methods employed when there is missing or no gold standard

Main Classification	Main Characteristics	Key references of paper that developed method (number)	Clinical Application of the methods
Group 4: **Other designs** • Considering an alternative study design like a validation study. • Study of agreement • Test positivity rate	***Analytic validation*** of a medical test is the process of verifying the test typically under controlled laboratory conditions which may not reflect the ultimate context of use. This also covers the technical performance of the tests. Case-control studies are common designs for these studies. ***Studies of agreement*** aim to investigate the concordance between two or more tests. ***Test positivity rate*** is the proportion of participants who have positive results on a test. This approach was used by Van Dyck et al [239] to reduce the number of people subjected to further evaluation.	***Validation***[240, 241] (2) ***Study of agreement***[242, 243] (2) ***Test positivity rate***[239] (1)	***Validation***:[244, 245] ***Study of agreement***[193, 246-250] ***Test positivity rate***[239, 243]

2.3.1. Methods employed when gold standard is missing

In some diagnostic accuracy studies, not all the participants get their disease status verified by a gold standard for reasons such as unavailability of the gold standard or medical reasons, amongst others. Such studies have partial verification as a subsample of the participants undergo the gold standard test and some will have their true disease status missing. Fifty-one papers were identified from the review that were developed to evaluate the diagnostic accuracy of index test(s) when the true disease status of some participants is not verified with the gold standard. These methods are divided into two subgroups:

- Imputation and bias-correction methods
- Differential verification methods

Imputation and bias-correction for partial verification methods

This includes methods to correct for verification bias while the disease-status of the unverified participants is left unverified. Forty-eight papers were developed in this group. These methods are further classified based on the result of the index test (binary, ordinal or continuous), the number of index tests evaluated (single or multiple), the assumptions made about verification (ignorable or missing at random – MAR) or non-ignorable or missing not at random – MNAR), and the classification of the diagnostic outcomes (disease-status). The identified methods in this subgroup are displayed in Figure 8 and Figure 9. There are two basic assumptions that are made regarding the missing disease status of the unverified participants, which are that they are missing at random or missing not at random[251-253].

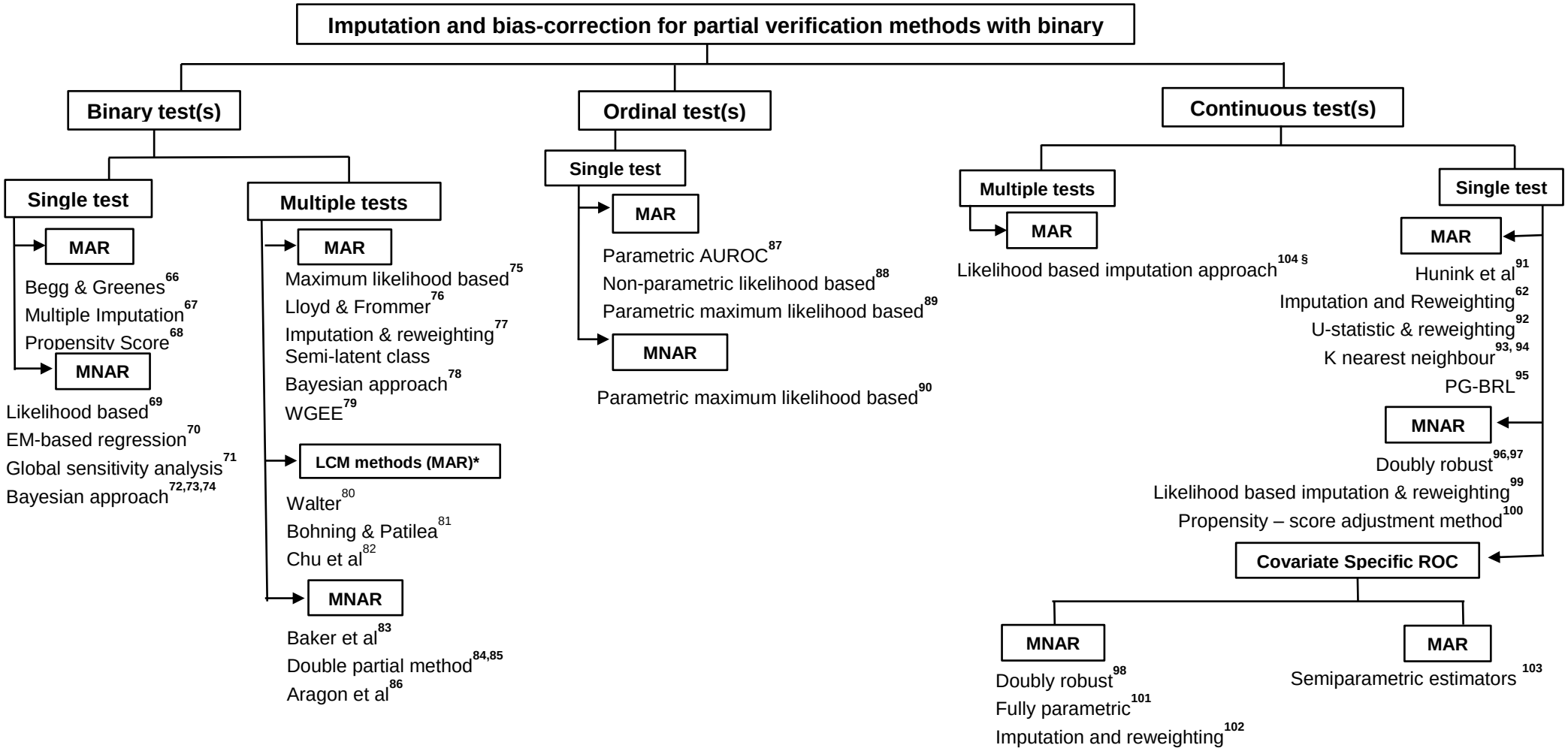
- **Missing at random (MAR)**: Missing at random implies that the missing disease status of the unverified participants can be fully accounted for or explained by the observed data (or variables in the dataset) such as the test results of another test etc.
- **Missing not at random (MNAR):** Missing not at random implies that the missing disease status of the unverified participants cannot be accounted for or explained by the observed data or variables that are related to the disease status.

In both assumptions, the missing disease status of the unverified participants are not missing randomly across the population; otherwise, the missingness is known as missing completely at random (MCAR).

Differential verification approach

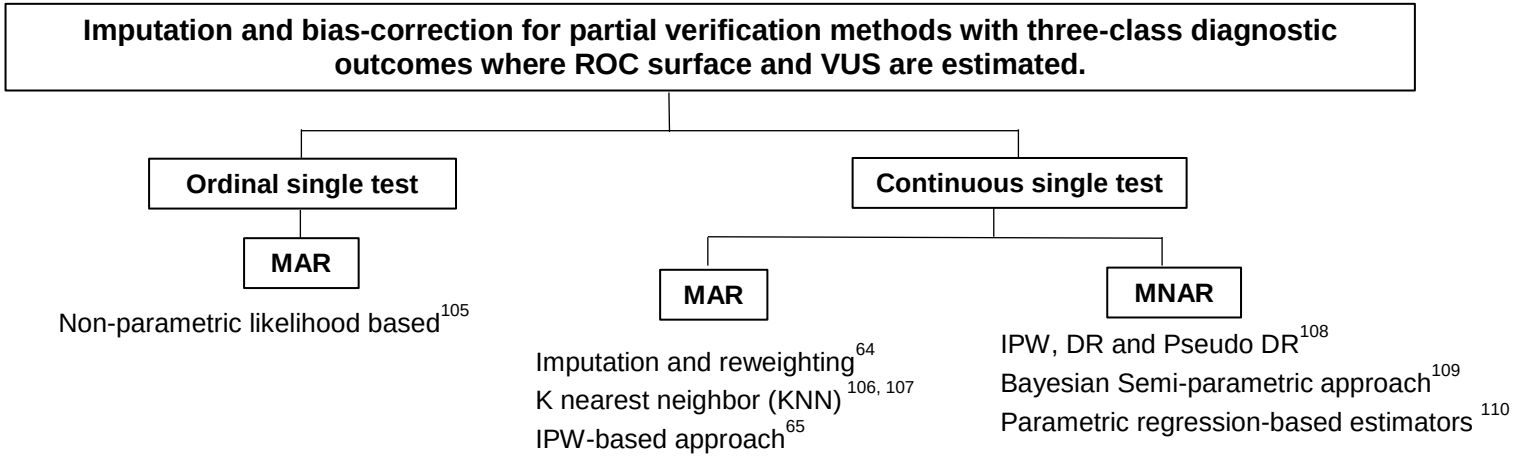
Participants whose disease status was not verified with the gold standard could undergo another reference standard (that is imperfect or less invasive than the gold standard[39]) to ascertain their disease status. This is known as ***differential verification***[48]. Differential verification has been explored by Alonzo et al, De Groot et al and Naaktgeboren et al[40, 48, 254]. They discussed the bias associated with differential verification, and how results using this approach could be presented. There are three identified statistical methods in this group. They are: a Bayesian latent class approach proposed by De Groot et al[47], a Bayesian method proposed by Lu et al[53] and a ROC approach proposed by Glueck et al[255]. These three methods aim to simultaneously adjust for differential verification bias and reference standard bias that arise from using an alternative reference standard (i.e. imperfect reference standard) for participants whose true disease status was not verified with the gold standard. Differential verification bias arises from using an alternative test to verify the missing disease status for those participants whose true disease status was not verified with the gold standard. Reference standard bias arises from using an imperfect reference standard.

Figure 8: Imputation and bias-correction for partial verification methods with binary diagnostic outcome.



*These LCM methods are employed when only participants with negative results in the index tests do not receive the gold standard.
§This method is applied to more than two-phase (multiphase) design.

Figure 9: Imputation and bias-correction for partial verification methods in three class diagnostic outcomes where ROC and VUS are estimated



Abbreviations used within the imputation and bias-correction figures (Figure 8 and 9).

LCM: latent class model
MAR: missing at random
MNAR: missing not at random
AUROC: Area under the ROC curve
ROC: Receiver operating characteristic
IPW: inverse probability weighting
WGEE: weighted generalised estimating equation
PG-BRL: partial gold Bayesian rank likelihood
DR: doubly robust
VUS: volume under the surface
EM: Expectation maximization
KNN: K nearest neighbour

2.3.2. Correction methods

This group includes algebraic methods developed to correct the estimated sensitivity and specificity of the index test when the sensitivity and specificity of the imperfect reference standard are known. There were seven statistical methods identified in this group, described in five different articles[117-121]. The method by Emerson et al[121] does not estimate a single value for sensitivity or specificity, unlike the other correction methods[117-120] but produces an upper bound value and a lower bound value for the sensitivity and specificity of the index test. These bounded values are used to explain the uncertainty around the estimated sensitivity and specificity of the index test.

2.3.3. Methods with multiple imperfect reference standards

A gold standard result or accurate information about the diagnostic accuracy of the imperfect reference standard are often not available to evaluate the index test. In these situations, multiple imperfect reference standards can be employed to evaluate the index test. Methods in this group include:

Discrepancy analysis

This compares the index test with an imperfect reference standard. Participants with discordant results undergo another imperfect test, called the resolver test, to ascertain their disease status. Discrepancy analysis is typically not recommended because it produces biased estimates[126, 256]. Modifications of this approach have been proposed[127, 164, 239]. In these, some of the participants with concordant responses (true positives and true negatives) are sampled to undertake the resolver test alongside participants with discordant responses (false negative – FN and false positive – FP). However, further research is needed to explore if these modified approaches are adequate to remove or reduce the potential bias.

Latent class analysis

The test performance of all the tests employed in the study are evaluated simultaneously using probabilistic models with the basic assumption that the true disease status of the participants is latent or unobserved. There are frequentist LCAs and Bayesian LCAs. The frequentist LCAs use only the data from the participants in the study to estimate the sensitivity and specificity of the tests; while the Bayesian LCAs employ external information (e.g. expert opinion or estimates from a previous research study) on the sensitivity and specificity of the tests evaluated in addition to the empirical data obtained from participants within the study. The LCAs assume that

the tests (index test and reference standards) are either conditionally independent given the true disease status or the tests are conditionally dependent. To model the conditional dependence among the tests, various latent class models (LCMs) with different dependence structures have been developed such as the Log-linear LCM[129], Probit LCM[130], extended log-linear and Probit LCMs[135], Gaussian Random Effect LCM[132] and two-crossed random effect LCM[134]. However, some studies[32],[257] have shown that different latent class models with different conditional dependence structures produce different estimates of sensitivities and specificities even though the posterior distributions of the estimated parameters in each model converge, showing that each model has a good fit of the data. This is plausible because of the non-identifiability of the model and the impact of the priors on the posterior distributions.

Construct composite reference standard

This method combines results from multiple imperfect tests (excluding the index test) with a predetermined rule to construct a reference standard that is used to evaluate the index test. By excluding the index test as part of the composite reference standard, incorporation bias can be avoided[159]. A novel method identified under the composite reference standard is the "dual composite reference standard (dCRS)" proposed by Tang et al[162]. The dCRS aims to find a composite reference standard (CRS) with both high sensitivity and high specificity to reduce the errors in the estimated sensitivity and specificity of the index test. Thus, the sensitivity and specificity of the index test is evaluated with the same composite reference standard but with different combination rules. The sensitivity of the index test is evaluated using the "all positive combination rule" of the combined reference standards and the specificity of the index test is evaluated using the "any positive combination rule" of the combined reference standards.

Panel or consensus diagnosis

This method uses the decision from a panel of experts to ascertain the disease status of each participant, which is then used to evaluate the index test.

2.3.4. Other designs

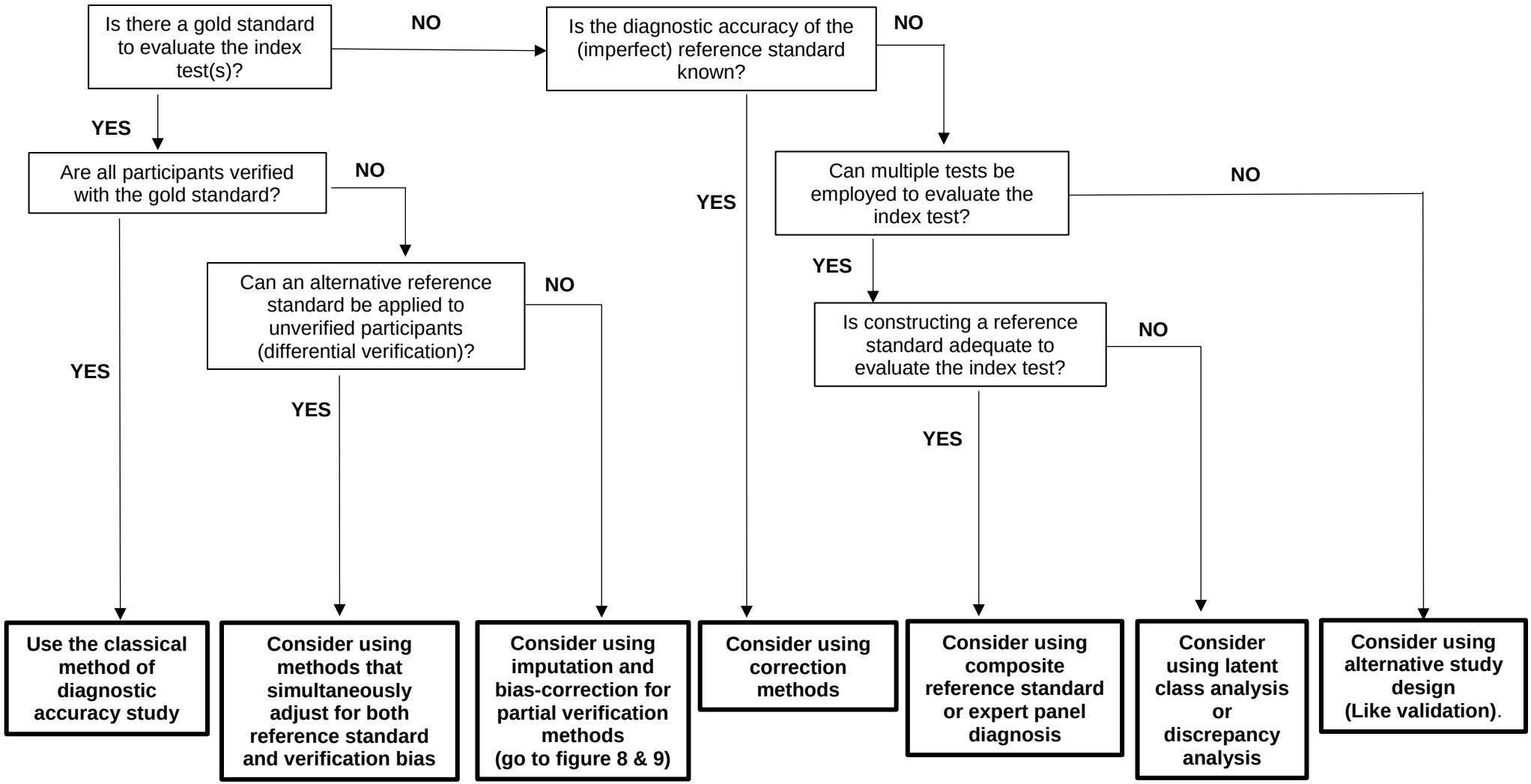
This group includes methods that fit the inclusion criteria but could not be placed into the other three groups. They include study of agreement, test positivity rate and the use of an alternative study design such as analytic validation**.** Study of agreement and test positivity rate are best used as exploratory tools alongside other methods[180, 225]

because they are not robust enough to assess the diagnostic ability of the medical test. Validation of a medical test cuts across different disciplines in medicine such as psychology and laboratory and experimental medicine. With this approach, the medical test is assessed based on what it is designed to do[241], that is if the test is able to diagnose or detect the target condition for which it is designed. Other designs include case-control designs (where the participants are known to have or not have the target condition) [258, 259], laboratory based studies or experimental studies which are undertaken with the aim to evaluate the analytical sensitivity and specificity of the index test[240, 260, 261].

2.4. Guidance to researchers

The guidance flowchart (Figure 10) constructed, is a modification and extension of the guidance for researchers flow-diagram developed by Reitsma et al[262]. Since, evaluating the accuracy measures of the index test is the focus of any diagnostic accuracy study, the flowchart starts with asking the first question “Is there a gold standard to evaluate the index test?” Following the responses from each question box (not bold); methods are suggested (bold boxes at the bottom of the flowchart) to guide clinical researchers, test evaluators, and researchers as to the different methods to consider.

Figure 10: Guidance flowchart of methods employed to evaluate medical tests in missing and no gold standard scenarios.



2.5. Discussion

This review sought to identify and review new and existing methods employed to evaluate the diagnostic accuracy of a medical test in the absence of a gold standard. The identified methods are classified into four main groups based on the availability and/or the application of the gold standard on the participants in the study. The four groups are: methods employed when only a sub-sample of the participants have their disease status verified with the gold standard (group 1); correction methods (group 2); methods using multiple imperfect reference standards (group 3) and other methods (group 4) such as study of agreement, test positivity rate and alternative study designs like validation.

In this review additional statistical methods have been identified that were not included in the earlier reviews on this topic by Reitsma et al [262] and Alonzo [251]. A list of all the methods identified in this review are presented in the Appendix (Appendix A.5 – supplementary information). This includes a brief description of the all the methods identified alongside their strengths and weaknesses and published articles of where the methods have been clinically applied. Most of the methods developed to evaluate the index test when the gold standard is missing for some of the participants are scarcely applied clinically [66, 90]. This may be due to the complexity of these methods (in terms of application and interpretation of results), and/or a disconnection between the fields of expertise of those who develop (e.g. statisticians) and those who employ the methods (e.g. clinical researchers). For example, the publication of such methods in specialist statistical journals may not be readily accessible to those developing clinical research studies. In order to close this gap, two flow-diagrams (Figure 8 and Figure 9) were constructed in addition to the modified guidance flowchart (Figure 10) as guidance tools to aid clinical researchers and test evaluators in the choice of methods to consider when evaluating medical tests in the absence of a gold standard. In addition, an R package (*bcROCsurface*)[105] and an interactive web application (Shiny app) that estimates the ROC surface and VUS in the presence of verification bias have been developed by To Duc [105].

One of the issues not addressed in this current review was methods that evaluate the differences in diagnostic accuracy of two or more tests in the presence of verification bias. Some published articles that consider this issue are Nofuentes and Del Castillo [263-267], Marin-Jimenez and Nofuentes [268], Harel and Zhou [269] and Zhou and Castelluccio [270]. This review also did not consider methods employed to estimate the

time-variant sensitivity and specificity of diagnostic tests in the absence of a gold standard. This issue has recently been addressed by Wang et al [271]. These methods were outside the scope of my research.

In terms of the methodology, a limitation of this review is the exclusion of books, dissertations, theses, conference abstracts and articles not published in the English language (such as the review by Masaebi et al [272] which was published in 2019), which could imply that there could still be some methods not identified by this review. A difference in the search strategy between the review reported in this chapter and the previous review by Rutjes et al[41], is that experts were not contacted for unpublished papers in their archives which are related to methods used to evaluate the diagnostic accuracy of an index test in the absence of a gold standard.

Regarding the methods identified in this review, further research could be carried out to explore the different modification to the discrepancy analysis approaches to understand if these modifications reduce or remove the potential bias. In addition, further research is needed to determine if the identified methods, developed to evaluate an index test in the absence of a gold standard are robust. Therefore, in line with this suggested area of further research, chapter three statistically compares the correction methods identified from the review via simulation approach to ascertain which developed method is robust.

Given the large numbers of statistical methods that have been developed, especially to evaluate medical tests when there is a missing gold standard, and the complexity of some of these methods, more interactive web applications (e.g. a Shiny package in R [273]) could be developed to implement these methods in addition those developed by To Duc [105] and Lim et al [274]. The development of such interactive web tools with clear instructions for use will aid the applications of these methods and help bridge the gap between the method developers and the clinical researchers or tests evaluators who are the end users of these methods. The review was updated in December 2020 before the submission of this thesis. Twenty-two articles were identified. These papers were clinical application papers. The results is reported in the Appendix ().

2.6. Conclusion

Various methods have been proposed and applied in the evaluation of medical tests when there is a missing gold standard result for some participants, or no gold standard at all. These methods depend on the availability of the gold standard, its application to

all or a subsample of participants in the study, the availability of alternative reference standard(s), and underlying assumption(s) made with respect to the index test(s) and / or participants in the study.

Knowing the appropriate method to employ when analysing the data from participants in a diagnostic accuracy study in the absence of a gold standard can helps to make statistically robust inference on the accuracy of the index test. This evidence, in addition to data on cost-effectiveness, utility and usability of the test will support clinicians, policy makers and stake holders to decide on whether to adopt the index test into their practice.

In the next chapter (chapter three), the correction methods will be explored and investigated to understand how they perform (in terms of their statistical properties) under different simulated scenarios.

Chapter Three: Comparison of Correction Methods

3.1. Introduction

From the systematic review (chapter two), several methods were identified that are employed to evaluate a medical test in the absence of a gold standard. The identified methods were grouped into four groups, which are:

- Group 1: Methods employed when there is a missing gold standard.
- Group 2: Correction methods which adjust for using an imperfect reference standard whose diagnostic accuracy is known.
- Group 3: Methods employed when using multiple imperfect tests.
- Group 4: ***"other methods"***. This group includes methods such as study of agreement, test positivity rate, and considering alternative study designs such as validation studies.

In this chapter, simulation studies are employed to investigate and compare the correction methods identified from the review. I have purposely not investigated all the correction methods identified, rather I have focused on those methods where there was a lack of comparisons in the literature. Simulations were used to create different scenarios (various "truths") and to use the results from the simulations to help understand the applicability of these methods under certain circumstances. Understanding how these methods perform in different simulated scenarios will help us to make appropriate choices in diagnostic accuracy studies.

Out of the seven correction methods identified in the systematic review; three correction methods are employed to simultaneously evaluate the diagnostic accuracy of multiple index tests. The three approaches take into consideration the conditional dependence of the index tests, given the true disease status of the participants and the known diagnostic accuracy of the imperfect reference test. Three of the seven approaches are based on different conditional dependence structures; which are Gaussian Random Effects (GRE) - initially proposed by Qu et al [132], the Beta-Binomial (BB), and the Finite Mixture (FM) approaches initially proposed by Albert et al [32, 133]. These three correction methods have been explored and compared in a previous study by Albert [120] and all methods estimate the parameters of interest provided the model is specified correctly.

Four further approaches are employed to evaluate a single diagnostic test with a dichotomous outcome using an imperfect reference standard whose sensitivity and

specificity are known (either from previous validation studies, experimental or field studies or case-control studies). In addition, the reference standard and the index test are assumed to be conditionally independent given the true disease status. Two or more tests are assumed to be conditionally dependent given the true disease status if the tests diagnose the same target condition (disease) using a related or the same biological component. For example, magnetic resonance imaging (MRI) and computed tomography (CT) could be considered as conditionally independent (although they are both imaging tests) because they use different biological components[275-278] – to diagnose the target condition. The CT scan uses the radiation opacity of tissue while a MRI scan uses magnetism to excite water molecules to produce images. An example of conditionally dependent tests would be where some enzyme-linked immunosorbent assay (ELISA) or polymerase chain reaction (PCR) tests [167, 198, 279] are used to diagnose a target condition as these enzyme-linked immunosorbent assay or polymerase chain reaction tests use the same biological components.

Mathematically, two tests (say T_1 and T_2) are assumed to be conditionally independent if:

$$\Pr(T_1 = 1, T_2 = 1|D = 1) = \Pr(T_1 = 1|D = 1) \times \Pr(T_2 = 1|\, D = 1)$$

$$\Pr(T_1 = 0, T_2 = 0|D = 0) = \Pr(T_1 = 0|D = 0) \times \Pr(T_2 = 0|\, D = 0)$$

Where D denotes the diseases status; D equal to 1 indicates the presence of disease and D equal to 0 indicates the absence of disease. T_1 equal to 1 indicates a positive test result from T_1 and T_1 equal to 0 indicates a negative result from T_1 (similarly for T_2).

The four correction methods employed to evaluate a single index test with a binary outcome are:

1. The Gart and Buck [118] correction method
2. The Staquet et al [119] correction method
3. The Brenner [117] correction method
4. The Emerson et al [121] correction method

These four correction methods are described in more detail below. Two of these correction methods, the Brenner correction method and the Staquet et al correction method, were compared with the classical method employed to estimate the sensitivity and specificity of the index test assuming the reference standard is a gold standard.

The estimated sensitivity and specificity from the classical method are referred to as "*unadjusted sensitivity*" and *unadjusted specificity*" respectively in this chapter. They are called unadjusted because if the reference standard is not a gold standard, the estimates obtained are biased and need to be corrected. From *Table 4*, the unadjusted sensitivity and specificity of the index test and the sample prevalence of the target condition are:

$$Sn_T = \frac{a}{e}; \qquad Sp_T = \frac{d}{f}; \qquad Prr = \frac{e}{N}$$

Table 4: 2 by 2 contingency table of the index test and imperfect reference standard

	Reference standard		
Index test	Positive = 1	Negative = 0	Total
Positive = 1	a	b	a + b = g
Negative = 0	c	d	c + d = h
	a + c = e	b + d = f	a + b + c + d = N

The confidence intervals for sensitivity and specificity can be obtained using five different approaches. These are the Wald confidence interval[280], the Wilson Score interval[281], the Clopper-Pearson[280] interval, and Agresti – Coull interval[282]. In this thesis the Wilson Score interval is used to obtain the 95% confidence interval for the estimated sensitivity and specificity of the index test explored in the clinical dataset. The Wilson score interval is calculated as:

$$(LL_{\hat{\theta}}, UL_{\hat{\theta}}) = \frac{1}{1+\frac{z^2}{n_*}}\left(\hat{\theta} + \frac{z^2}{2n_*}\right) \pm \frac{z}{1+\frac{z^2}{n_*}}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n_*} + \frac{z^2}{4n_*^2}}$$

where $\boldsymbol{n_*}$ is not the total number of participants in the study but rather it is the total number of participants with positive (negative) test results on the reference standard when calculating the confidence interval of the estimated sensitivity (specificity). $\hat{\theta}$ is the estimated sensitivity (or specificity) and z is the standard normal distribution value at a given percentage of confidence interval.

The correction methods were compared via simulation studies. More details of the simulation process and the different scenarios explored is discussed in detail in section **3.3**.

Basic notation

Let D be the true disease status with value 0 and 1 (0 represents non-diseased and 1 represents diseased). R is the reference standard and T is the index test. The known sensitivity and specificity of the reference standard are denoted as Sn_R and Sp_R respectively. The estimated sensitivity and specificity of the index test are denoted as Sn_T and Sp_T respectively. As described above, they are also known as the unadjusted sensitivity and specificity of the index test.

3.1.1. Gart & Buck Correction method

Gart and Buck [118] proposed a pair of estimators to correct for the estimated sensitivity and specificity of the index test when the diagnostic accuracy of the reference test is known and the index and reference tests are conditionally independent given the true disease status.

The Gart and Buck estimators are:

$$Sn_{cor}^{GB} = \frac{Sp_R \times Prr \times Sn_T + (1 - Sp_R)(1 - Prr) \times Sp_T - (1 - Sp_R)(Sp_R - \hat{P}J)}{\hat{P}J} \quad (1)$$

$$Sp_{cor}^{GB} = \frac{Sn_R * (1 - Prr) \times Sp_T + (1 - Sn_R)Prr \times Sn_T - (1 - Sn_R)(1 - Sp_R + \hat{P}J)}{J(1 - \hat{P})} \quad (2)$$

where the corrected sensitivity and specificity of the index test are denoted as Sn_{cor} and Sp_{cor} respectively. The estimated population prevalence is denoted as $\hat{P}$ and calculated as:

$$\hat{P} = \frac{Prr + Sp_R - 1}{J}$$

Where the sample prevalence of the target condition is denoted as Prr and J is the Youden index for the diagnostic test which is calculated as:

$$J = Sn_R + Sp_R - 1$$

The standard errors of the estimators are obtained via the delta method (a non-parametric approach of obtaining standard errors and confidence intervals)[283, 284].

3.1.2. Staquet et al correction method

Staquet et al[119] proposed a pair of estimators to correct for the sensitivity and specificity of the index test when the index test and reference standard are conditionally independent given the true disease status, and the sensitivity and specificity of the reference standard are known though imperfect.

The **Staquet at al estimators** are:

$$Sn_{cor}^{sq} = \frac{gSp_R - b}{N(Sp_R - 1) + e}; \qquad Sp_{cor}^{sq} = \frac{hSn_R - c}{NSn_R - e}; \qquad \hat{P} = \frac{N(Sp_R - 1) + e}{N(Sn_R + Sp_R - 1)} \tag{3}$$

where $\hat{P}$ is the estimated population prevalence and the values of N, g, b, h and e are expressed in Table 4.

The Staquet et al[119] estimators are equivalent to the Gart and Buck [118] method. This is shown using algebraic expression below.

Simplification of the Gart and Buck estimators to obtain the Staquet et al estimators

$$Sn_{cor}^{GB} = \frac{Sp_{RS} \times Prr \times Sn_{IT} + (1 - Sp_{RS})(1 - Prr) \times Sp_{IT} - (1 - Sp_{RS})(Sp_{RS} - \hat{P}J)}{\hat{P}J}$$

$$= \frac{Sp_{RS} \times \frac{e}{N} \times \frac{a}{e} + (1 - Sp_{RS})\left(\frac{f}{N} \times \frac{d}{f}\right) - (1 - Sp_{RS})(Sp_{RS} - Prr - Sp_{RS} + 1)}{J(\hat{P})}$$

$$= \frac{\frac{a}{N}(Sp_{RS}) + \frac{d}{N} - \frac{d}{N}(Sp_{RS}) - (1 - Sp_{RS})(1 - Prr)}{J\hat{P}}$$

$$= \frac{\frac{a}{N}(Sp_{RS}) + \frac{d}{N} - \frac{d}{N}(Sp_{RS}) - \frac{f}{N}(1 - Sp_{RS})}{J\hat{P}}$$

$$= \frac{\frac{a}{N}(Sp_{RS}) + \frac{d}{N} - \frac{d}{N}(Sp_{RS}) - \frac{f}{N} + \frac{f}{N}(Sp_{RS})}{J\hat{P}}$$

$$= \frac{\frac{a}{N}(Sp_{RS}) - \frac{d}{N}(Sp_{RS}) + \frac{f}{N}(Sp_{RS}) - \frac{f}{N} + \frac{d}{N}}{J\hat{P}}$$

$$= \frac{\frac{a-d+f}{N}(Sp_{RS}) + \frac{d}{N} - \frac{f}{N}}{J\hat{P}}$$

$$= \frac{\frac{a+b}{N}(Sp_{RS}) - \frac{b}{N}}{J\hat{P}}$$

$$= \frac{g(Sp_{RS}) - b}{N(Prr + Sp_{RS} - 1)}$$

$$= \frac{g(Sp_{RS}) - b}{N(Prr) + N(Sp_{RS} - 1)}$$

$$Sn_{cor}^{sq} = \frac{g(Sp_{RS}) - b}{N(Sp_{RS} - 1) + e}$$

$$Sp_{cor}^{GB} = \frac{Sn_{RS} \times (1 - Prr) \times Sp_{IT} + (1 - Sn_{RS}) Prr \times Sn_{IT} - (1 - Sn_{RS})\left(1 - Sp_{RS} + \hat{P}J\right)}{J\left(1 - \hat{P}\right)}$$

$$= \frac{Sn_{RS} \times \frac{f}{N} \times \frac{d}{f} + (1 - Sn_{RS}) \times \frac{e}{N} \times \frac{a}{e} - (1 - Sn_{RS})(1 - Sp_{RS} + Sp_{RS} + Prr - 1)}{J\left(1 - \hat{P}\right)}$$

$$= \frac{\frac{d}{N}(Sn_{RS}) + \frac{a}{N}(1 - Sn_{RS}) - (1 - Sn_{RS})(Prr)}{J\left(1 - \hat{P}\right)}$$

$$= \frac{\frac{d}{N}(Sn_{RS}) + \frac{a}{N} - \frac{a}{N}(Sn_{RS}) - \frac{e}{N}(1 - Sn_{RS})}{J\left(1 - \frac{Prr + Sp_{RS} - 1}{J}\right)}$$

$$= \frac{\frac{d}{N}(Sn_{RS}) + \frac{a}{N} - \frac{a}{N}(Sn_{RS}) - \frac{e}{N} + \frac{e}{N}(Sn_{RS})}{\frac{J(J - Prr - Sp_{RS} + 1)}{J}}$$

$$= \frac{\frac{d-a+e}{N}(Sn_{RS}) + \frac{a}{N} - \frac{e}{N}}{J\left(\frac{Sp_{RS} + Sn_{RS} - 1 - Prr - Sp_{RS} + 1}{J}\right)}$$

$$= \frac{\frac{h}{N}(Sn_{RS}) - \frac{c}{N}}{Sn_{RS} - Prr}$$

$$= \frac{h(Sn_{RS}) - c}{N(Sn_{RS} - Prr)}$$

$$Sp_{Cor}^{sq} = \frac{h(Sn_{RS}) - c}{NSn_{RS} - e}$$

Hence, the Gart and Buck correction method is not explored in the simulation study

3.1.3. Brenner correction method

Brenner [117] proposed two pairs of estimators to correct the estimated sensitivity and specificity of an index test when using an imperfect reference standard. The first pair of estimators proposed by Brenner [117] assumes that the index test and the reference standard test are conditionally independent given the true disease status.

The estimators are defined as:

$$Sn_{cor}^{B1} = \frac{Prr \times Sn_R \times Sn_T + (1 - Prr)(1 - Sp_R)(1 - Sp_T)}{Prr \times Sn_R + (1 - Prr)(1 - Sp_R)} \quad (4)$$

$$Sp_{cor}^{B1} = \frac{Prr \times (1 - Sn_R)(1 - Sn_T) + (1 - Prr) \times Sp_R \times Sp_T}{Prr \times (1 - Sn_R) + (1 - Prr) \times Sp_R} \quad (5)$$

The second pair of estimators proposed by Brenner [117], assumes that there is a positive correlation between the index test and the reference standard. Thus, the estimators are developed to adjust for positive correlation between the classification error of the index test and reference standard.

The estimators are defined as:

$$Sn_{cor}^{B2} = \frac{Prr \times Sn_T + (1 - Prr) \times (1 - Sp_R)}{Prr \times Sn_R + (1 - Prr) \times (1 - Sp_R)} \quad (6)$$

$$Sp_{cor}^{B2} = \frac{Prr \times (1 - Sn_R) + (1 - Prr) \times Sp_T}{Prr \times (1 - Sn_R) + (1 - Prr) \times Sp_R} \quad (7)$$

Both pairs of estimators will be considered in the simulation studies.

3.1.4. Emerson et al correction method

The Emerson et al correction method[121] does not estimate a single value for the sensitivity or specificity of the index test. Rather, it estimates upper and lower bound values for the sensitivity and specificity of the index test as well as the prevalence of the target condition. The estimators employed to obtain the upper bounds of the sensitivity and specificity of the index test are:

$$Sn_{max} = \frac{Sn_T \times Prr + (1 - \eta_R)(1 - Prr)}{\hat{P}} \quad (8)$$

$$Sp_{max} = \frac{(1 - \psi_R) \times Prr + Sp_T(1 - Prr)}{1 - \hat{P}} \quad (9)$$

where:

$$\eta_R = \frac{Sp_R(1 - \hat{P})}{1 - Prr}; \qquad \hat{P} = \frac{Prr + Sp_R - 1}{Sn_R + Sp_R - 1}; \qquad \psi_{RS} = \frac{Sn_R \times \hat{P}}{Prr}$$

The estimators employed to obtain the lower bounds of the sensitivity and specificity of the index test are:

$$Sn_{min} = \frac{(\psi_R + Sn_T - 1)Prr + 0 \times (1 - Prr)}{\hat{P}} = \frac{(\psi_{RS} + Sn_T - 1)Prr}{\hat{P}} \quad (10)$$

$$Sp_{min} = \frac{0 \times Prr + (\eta_R + Sp_T - 1) \times (1 - Prr)}{1 - \hat{P}} = \frac{(\eta_R + Sp_T - 1) \times (1 - Prr)}{1 - \hat{P}} \quad (11)$$

The estimators developed by Emerson et al[121] are not compared with the Staquet et al and Brenner correction methods in this simulation study because they do not produce single value estimates like the Staquet et al and Brenner correction methods but rather bounded value estimates. These bounded values explains the confidence of the

estimated sensitivity and specificity of the index test[121]. The narrower the interval or bounds, the more confident the estimate and the wider the interval the less confident the estimate obtained.

3.2. Aims of the simulation study

The aims of this simulation study are to:

- Explore the statistical properties of the correction methods: Brenner, and Staquet et al[117, 119].
- Compare two correction methods, the Brenner and Staquet et al correction methods, to understand how robust these methods are.

3.3. Methodology

Simulation studies were undertaken following the suggested guidelines by Morris, White [285]. This includes: **P**lanning for the simulation study, **C**oding and execution, **A**nalysis and **R**eporting the simulation study appropriately.

1. **Planning for the simulation study**: Planning for a simulation study entails stating out clearly the **A**ims (objectives) of the study, the **D**ata-generating mechanism, the **E**stimands, the **M**ethods and the **P**erformance measures (**ADEMP**).
   - The **aims** of this simulation study are defined in section **3.2.**
   - **Data generating mechanism:** The data-generating mechanism describes how the dataset is generated or simulated. The simulated variables are the discordant and concordant values (relative true positive – rTP, relative true negative – rTN, relative false positive – rFP, and relative false negative – rFN) of a dichotomous index and reference test. They are called *relative* because they are obtained in comparison to a reference standard which is not a gold standard (so the true disease status of the participants are unknown). These variables were simulated using the multinomial distribution. The prevalence of the target condition as well as the sensitivities and specificities of the index and reference tests were assumed to be known. The probability of each cell (rTP, rTN, rFP and rFN) is obtained using the relationship between the prevalence of the target condition, the sensitivity and specificity of the index and reference tests, and the covariance term or correlation between the two tests among the disease group and non-diseased groups. The probability values estimated for each cell are multiplied by the total number of participants to report the number of participants in each cell. These values are assumed to be known during the

simulation process. Thus, when the two tests are conditionally independent given the true the disease status, the covariance terms among the diseased and non-diseased group across both tests are zero. Further, if the two tests are conditionally dependent given the true disease status, then the conditional dependence across the two tests is modelled using the fixed effect method[286]. Finally, if the reference standard is a gold standard (perfect test), then the assumed (known) sensitivity and specificity of the reference test are both 1. In this case, the rTP, rTN, rFN, rFP are error free, so the estimated sensitivity and specificity of the index test are unbiased.

- **Estimands:** The estimands are the random values obtained from using the estimators of interest. In this simulation study, the estimators of interest are the different correction methods employed to estimate the sensitivity and specificity of the index test. The estimands are the different sensitivities and specificities of the index test obtained by applying the estimators of interest on various simulated samples.
- **Methods**: The methods are the estimators of interest, used to derive the estimands. These are:
  a) Unadjusted sensitivity and specificity: This is the classical method employed to estimate sensitivity and specificity of the index test assuming that the reference test is perfect (see Table 4).
  b) Brenner corrected sensitivity and specificity
  c) Staquet et al corrected sensitivity and specificity
- **Performance measures:** The performance measures employed to assess the correction methods are well-used properties of a good estimator. The properties chosen are bias, mean square error (MSE) and consistency. These measures were chosen as they are well-known measures employed to assess a good estimator. These measures were used to assess the estimators of interest under the conditional independence and conditional dependence assumption given the true disease status.

  a) **Bias**: An estimator is statistically unbiased if the difference between the true value of the parameter (θ) and the expected (mean) value of the estimator ($E(\hat{\theta})$) is equal to zero[287, 288]. That is:

$$\theta - E(\hat{\theta}) = 0$$

If the true value is not equal to the expected value, then the estimator is biased.

$$Bias = E(\hat{\theta}) - \theta$$

b) **Mean square error (MSE)**: The mean square error is the mean or average of the squared difference between the estimator and the true parameter[288]. That is

$$E\left[(\theta - \hat{\theta})^2\right] = \frac{1}{n}\sum_{i=1}^{n}(\theta_i - \hat{\theta}_i)^2$$

c) **Consistency**: An estimator is consistent if, as the sample size increases ($n \to \infty$) the mean value of the estimates obtained using the given estimator approaches the true value[288, 289] and the variance decreases to zero.

$$E(\hat{\theta}) \to 0 \ \ as \ \ n \ \to \infty; \quad Var(\hat{\theta}) \to 0 \ \ as \ \ n \to \infty$$

An estimator can be unbiased and consistent, it can be unbiased and not consistent, and it can be biased (for finite sample sizes) and consistent.

2. **Coding and Execution**: Functions used to simulate the random samples (from the multinomial distribution) and to calculate the estimands were coded and executed using the R statistical software (R Studio)[290]. The script is attached as an appendix (Appendix B.1).

3. **Analysis**: This section is divided into two cases (section **3.3.1** and section **3.3.2**). The first case is where the index test and the reference standard (RS) are conditionally independent given the true disease status (under the assumption that the RS is perfect or imperfect) – section **3.3.1**. The second case is where the index test and the reference standard are conditionally dependent given the true disease status (under the assumption that the RS is perfect or imperfect) – section **3.3.2**.

4. **Reporting**: At the end of the analysis in each section (sections **3.3.1** and **3.3.2**) the results are presented with tables and graphs, alongside some observations and discussion.

3.3.1. Comparison of correction methods – conditional independence

Let D be the two-class true disease status of the participants which takes the value 0 or 1. D equal to 0 indicates that the participant does not have the disease (or D-) and

D equal to 1 indicates the participant has the disease (or D+). Let the prevalence of the disease be denoted as p.

$$\Pr(D+) \;=\; \Pr(D=1) = \; p\,; \qquad 0 \le p \le 1.$$

The diseased and non-diseased groups are mutually exclusive because a participant cannot be classified as diseased and non-diseased at the same time. Thus, the joint probability of diseased and non-diseased is 1. Let T denote the index test and R denote the reference standard. T and R have dichotomised results - negative and positive results – which implies that each participant is classified as having the disease (positive) or not having the diseased (negative) by each test. Under conditional independence, having a positive (or negative) result with the reference test does not affect the likelihood of having a positive (or negative) result with the index test. However, under conditional dependence, having a positive (negative) result with the reference test can affect the likelihood of having a positive (negative) result on the index test, as both tests measure the same biological component.

Therefore, statistically R and T are conditionally independent given the true disease status if and only if:

$$\Pr(T=1, R=1|D=1) = \Pr(T=1|D=1) \times \Pr(R=1|\,D=1)$$

and

$$\Pr(T=1, R=1|D=0) = \Pr(T=1|D=0) \times \Pr(R=1|\,D=0)$$

This implies that:

$$\Pr(T=0, R=0|D=0) = \Pr(T=0|D=0) \times \Pr(R=0|\,D=0)$$

$$\Pr(T=0, R=0|D=1) = \Pr(T=0|D=1) \times \Pr(R=0|\,D=1)$$

$$\Pr(T=1, R=0|D=1) = \Pr(T=1|D=1) \times \Pr(R=0|\,D=1)$$

$$\Pr(T=1, R=0|D=0) = \Pr(T=1|D=0) \times \Pr(R=0|\,D=0)$$

$$\Pr(T=0, R=1|D=1) = \Pr(T=0|D=1) \times \Pr(R=1|\,D=1)$$

$$\Pr(T=0, R=1|D=0) = \Pr(T=0|D=0) \times \Pr(R=1|\,D=0)$$

The sensitivity and specificity of the reference standard are $Sn_R = Pr(R = 1|\, D = 1)$ and $Sp_R = \Pr(R = 0|\, D = 0)$ respectively. The sensitivity and specificity of the index test are $Sn_T = Pr(T = 1|\, D = 1)$ and $Sp_T = \Pr(T = 0|\, D = 0)$ respectively.

The covariance between two tests among the diseased group (participants with the target condition) can be expressed as:

$$Cov_d(R = 1, T = 1|D = 1) = \Pr(R = 1, T = 1|\, D = 1) - (Sn_R \times Sn_T)$$

$$= \rho_d \sqrt{Sn_R \times Sn_T \times (1 - Sn_R) \times (1 - Sn_T)}$$

The covariance between the two tests among the non – diseased group can be expressed as:

$$Cov_{nd}(R = 0, T = 0|D = 0) = \Pr(R = 0, T = 0|\, D = 0) - (Sp_R \times Sp_T)$$

$$= \rho_{nd} \sqrt{Sp_R \times Sp_T \times (1 - Sp_R) \times (1 - Sp_T)}$$

The correlation between T and R among the disease group is denoted by ρ_d, and the correlation between T and R among the non-disease group is denoted by ρ_{nd}. R and T are conditionally independent when ρ_d and ρ_{nd} are equal to zero. It is logical that the sensitivities and specificities of the tests are between 0 and 1; and the correlation coefficient is between 0 and the absolute value of 1.

If the reference standard and the index test are conditionally independent given the true disease status, then the covariances defined above are zero; the same applies to their correlation coefficients. The cell probabilities of classifying participants under the assumption that the index test and the reference standard are conditionally independent given the true disease status are depicted in the 4 x 2 table and 2 x 2 table displayed as Table 5.

Table 5: Cell probabilities of the 4 x 2 and 2 x 2 classification of participants

	Diseased (+)		**Diseased (-)**	
	RS +	**RS –**	**RS +**	**RS –**
T +	$p_1(Sn_R \times Sn_T)$	$p_1(1 - Sn_R)Sn_T$	$p_0(1 - Sp_R)(1 - Sp_T)$	$p_0 \times Sp_R(1 - Sp_T)$
T –	$p_1 \times Sn_R(1 - Sn_T)$	$p_1(1 - Sn_R)(1 - Sn_T)$	$p_0 \times Sp_T(1 - Sp_R)$	$p_0 \times Sp_R \times Sp_T$
OR				

	Reference standard (RS)	
	Positive (+)	**Negative (-)**
T +	$p_1(Sn_R \times Sn_T) + \left(p_0(1-Sp_R)(1-Sp_T)\right)$	$p_1(1-Sn_R)Sn_T + (p_0 \times Sp_R(1-Sp_T))$
T –	$p_1 \times Sn_R(1-Sn_T) + (p_0 \times Sp_T(1-Sp_R))$	$p_1(1-Sn_R)(1-Sn_T) + (p_0 \times Sp_R \times Sp_T)$

Sn_R is sensitivity of reference standard; Sn_T is sensitivity of index test; Sp_R is specificity of the reference standard; Sp_T is specificity of the index test; RS is reference standard; T+ is index test positive; T- is index test negative, p_1 is the prevalence of the target condition (diseased); p_0 is 1 – p_1 (which is the prevalence of non-diseased).

In this section the correction methods (Brenner and Staquet et al) and the classical method are explored and compared under the assumption that the reference standard and the index test are conditionally independent given the true disease status. That is, there is no covariance terms between the two tests (reference standard and index test) among the diseased and non-diseased groups. This comparison helps us to understand how these estimators perform in the simulated scenarios. The different scenarios explored in this section are:

- When the reference standard is perfect.
- When the reference standard is imperfect and better than the index test. That is the sensitivity and specificity of the reference standard are higher than the sensitivity and specificity of the index test.
- When the reference standard is imperfect and worse than the index test. That is the sensitivity and specificity of the reference standard are lower than the sensitivity and specificity of the index test.
- When the accuracy measures of the reference standard are imperfect and the same as the index test. That is the sensitivity and specificity of the reference standard are the same as the sensitivity and specificity of the index test.

In all scenarios, 200 random samples of sizes between 50 and 1000 were simulated using the multinomial distribution. Varying sample sizes were chosen to aid the understanding of how the sample size impacts the estimates of interest. The sensitivities and specificities of the index test and reference standard were varied to reflect the different scenarios simulated. The choice of parameters employed in the simulation study was based upon clinical case studies identified from the systematic

review reported in Chapter two. For example, the clinical datasets used in this chapter have prevelances varying between 0.1, 0.3 and 0.9. The sensitivity and specificity of the RS employed in the study by Mathew et al[125] dataset were 0.74 and 0.91 respectively, the sensitivities and specificities of the RS used in the datasets reported in Matos et al[124] were 0.8 both for the NC detection and 0.79 and 0.99 respectively for the D3 detection. Therefore, using a fixed prevalence of 0.3 (and varying the prevalences from zero to one), alongside setting the sensitivity and specificity of the reference standard to be either 0.8 and/or 0.9 in the simulation study reflects possible clinical cases identified from the review.

Algebraic relationship between classical, Brenner and Staquet et al correction method

In this section, the different estimators are explored to understand how they are mathematically related. Mathematically, the Brenner estimators can be reduced to:

$$Sn_{cor}^{B1} = \frac{Prr \times Sn_R \times Sn_T + (1-Prr)(1-Sp_{RS})(1-Sp_T)}{Prr \times Sn_R + (1-Prr)(1-Sp_R)}$$

$$= \frac{\left(\frac{e}{N} \times Sn_R \times \frac{a}{e}\right) + \left(\frac{f}{N} \times (1-Sp_R) \times \frac{b}{f}\right)}{\frac{eSn_R}{N} + \frac{f(1-Sp_R)}{N}}$$

$$= \frac{\frac{1}{N}(aSn_R) + \frac{1}{N}\left(b(1-Sp_R)\right)}{\frac{1}{N}\left(eSn_R + f(1-Sp_R)\right)}$$

$$\boldsymbol{Sn_{cor} = \frac{aSn_R + b(1-Sp_R)}{eSn_R + f(1-Sp_R)}}$$

$$Sp_{cor}^{B1} = \frac{Prr \times (1-Sn_R)(1-Sn_T) + (1-Prr)Sp_R \times Sp_T}{Prr(1-Sn_R) + (1-Prr)Sp_R}$$

$$= \frac{\left(\frac{e}{N}(1-Sn_R) \times \frac{c}{e}\right) + \left(\frac{f}{N} \times Sp_R \times \frac{d}{f}\right)}{\frac{e(1-Sn_R)}{N} + \frac{fSp_R}{N}}$$

$$= \frac{\frac{1}{N}(c(1-Sn_R)) + \frac{1}{N}(dSp_R)}{\frac{1}{N}(e(1-Sn_R) + fSp_R)}$$

$$\boldsymbol{Sp_{cor} = \frac{c(1 - Sn_R) + dSp_R}{e(1 - Sn_R) + fSp_R}}$$

As a reminder the second pair of estimators is:

$$Sn_{cor}^{B2} = \frac{Prr \times Sn_T + (1 - Prr) \times (1 - Sp_R)}{Prr \times Sn_R + (1 - Prr) \times (1 - Sp_R)}$$

$$Sp_{cor}^{B2} = \frac{Prr \times (1 - Sn_R) + (1 - Prr) \times Sp_T}{Prr \times (1 - Sn_R) + (1 - Prr) \times Sp_R}$$

If the reference standard is perfect ($Sn_R = Sp_R = 1$) the Staquet et al and Brenner corrected estimators for sensitivity and specificity reduces to the classical estimator for sensitivity (Sn_T) and specificity (Sp_T).

For the Staquet et al estimators:

$$Sn_{cor}^{sq} = \frac{gSp_R - b}{N(Sp_R - 1) + e} = \frac{g - b}{e} = \frac{a + b - b}{e} = \frac{a}{e} = Sn_T$$

$$Sp_{cor}^{sq} = \frac{hSn_R - c}{NSn_R - e} = \frac{h - c}{N - e} = \frac{c + d - c}{e + f - e} = \frac{d}{f} = Sp_T$$

For the Brenner estimators:

$$Sn_{cor}^{B1} = \frac{aSn_R + b(1 - Sp_R)}{eSn_R + f(1 - Sp_R)} = \frac{a}{e} = Sn_T$$

$$Sp_{cor}^{B1} = \frac{c(1 - Sn_R) + dSp_R}{e(1 - Sn_R) + fSp_R} = \frac{d}{f} = Sp_T$$

$$Sn_{cor}^{B2} = \frac{Prr \times Sn_T + (1 - Prr) \times (1 - Sp_R)}{Prr \times Sn_R + (1 - Prr) \times (1 - Sp_R)} = \frac{Prr \times Sn_{IT}}{Prr} = Sn_T$$

$$Sp_{cor}^{B2} = \frac{Prr \times (1 - Sn_R) + (1 - Prr) \times Sp_T}{Prr \times (1 - Sn_R) + (1 - Prr) \times Sp_R} = \frac{(1 - Prr) \times Sp_T}{(1 - Prr)} = Sp_T$$

where a, b, c, d, e, f, g, h and N are described in Table 4.

Therefore, when the reference standard is perfect:

$$Sn_{cor}^{sq} = Sn_{cor}^{B1} = Sn_{cor}^{B2} = Sn_T \quad and \quad Sp_{cor}^{sq} = Sp_{cor}^{B1} = Sp_{cor}^{B2} = Sp_T$$

However, if the reference standard is not perfect, the Staquet et al corrected sensitivity is a function of the known specificity of the reference standard and the observed data and the Staquet et al corrected specificity is a function of the known sensitivity and the observed data.

That is, the Staquet et al estimators can be written as:

$$Sn_{cor}^{sq} = \frac{gSp_R - b}{N(Sp_R - 1) + e} = \frac{(a+b)Sp_R - b}{(e+f)(Sp_R - 1) + e} = \frac{aSp_R + bSp_R - b}{eSp_R + fSp_R - e - f + e}$$
$$= \frac{\boldsymbol{aSp_R + b(Sp_R - 1)}}{\boldsymbol{eSp_R + f(Sp_R - 1)}}$$

$$Sp_{Cor}^{sq} = \frac{hSn_R - c}{NSn_R - e} = \frac{(c+d)Sn_R - c}{(e+f)Sn_R - e} = \frac{cSn_R + dSn_R - c}{eSn_R + fSn_R - e} = \frac{\boldsymbol{dSn_R + c(Sn_R - 1)}}{\boldsymbol{fSn_R + e(Sn_R - 1)}}$$

The Brenner corrected sensitivity and specificity are function of both the sensitivity and specificity of the reference standard as well as the observed data, i.e.

$$Sn_{cor}^{B1} = \frac{\boldsymbol{aSn_R + b(1 - Sp_R)}}{\boldsymbol{eSn_R + f(1 - Sp_R)}}; \quad Sp_{cor}^{B1} = \frac{\boldsymbol{dSp_R + c(1 - Sn_R)}}{\boldsymbol{fSp_R + e(1 - Sn_R)}}$$

The letters in red are common in both methods, and the variations or differences are displayed using the green colour.

Perfect reference standard

Initially, the sensitivity and specificity of the reference standard were set to 1 (indicating a perfect test) and the sensitivity and specificity of the index test were 0.8 and 0.7 respectively. The prevalence of disease was 0.3. The mean values of the unadjusted and corrected sensitivities and specificities of the index test were obtained alongside the standard error, the mean squared error (MSE) and empirically assessed bias. These were reported in Table 6 and Figure 11.

The results in Table 6 and Figure 11 showed that if the reference standard is perfect, the mean sensitivities and specificities (unadjusted and corrected) are very similar and are approximately equal to the simulated truth. The estimators are unbiased as their bias and MSE is approximately zero. The standard error decreases as the sample size increases. In fact, the standard error tends to zero as the sample size increases.

Table 6: Estimates from unadjusted and corrected sensitivities and specificities of the index test under the conditional independence assumption when the reference standard is perfect

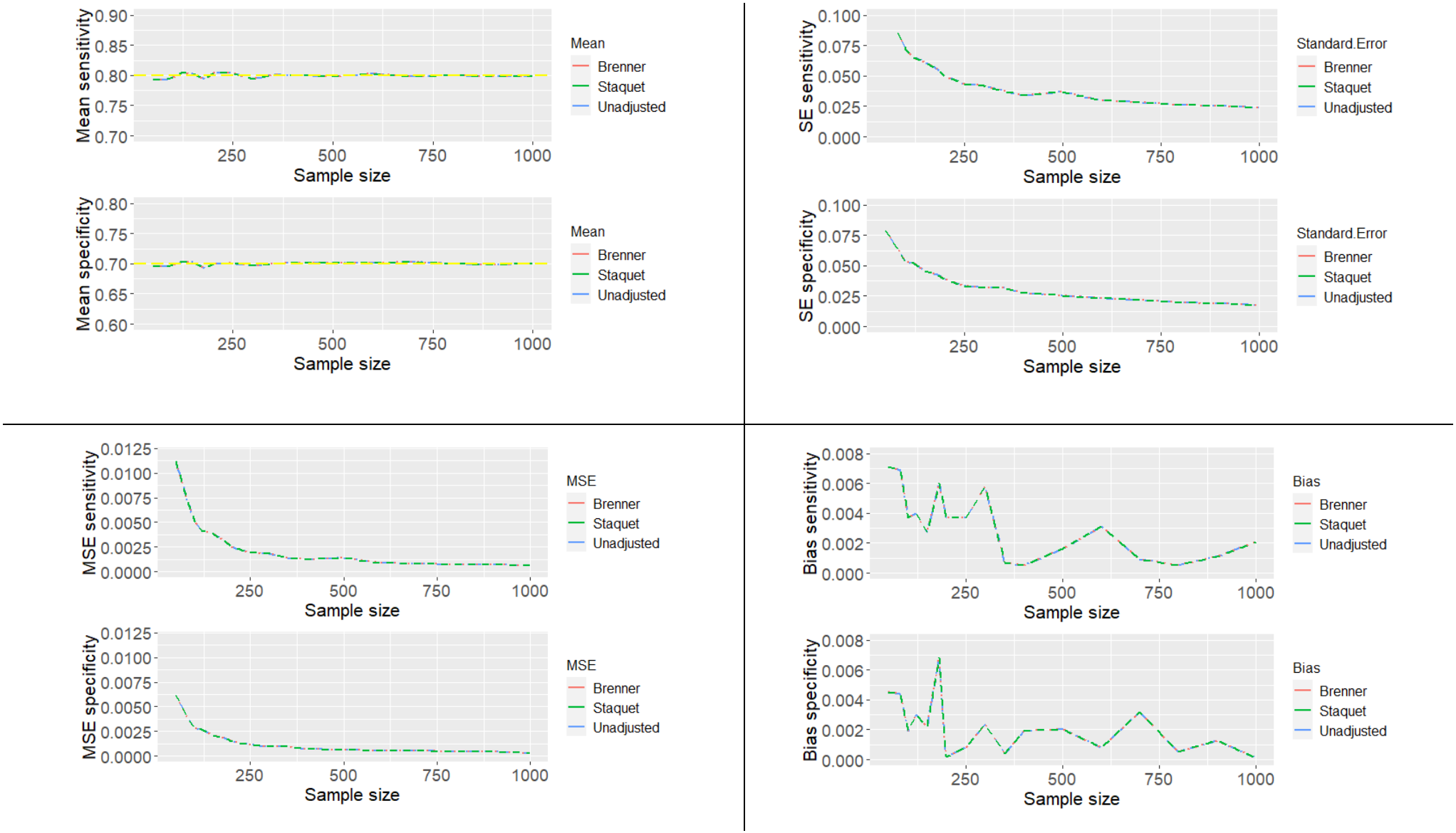
Sample size		Methods					
		Unadjusted Sensitivity	**Unadjusted Specificity**	**Brenner Sensitivity**	**Brenner Specificity**	**Staquet et al Sensitivity**	**Staquet et al specificity**
	Properties measured	**ST = 0.8**	**ST = 0.7**	**ST = 0.8**	**ST = 0.7**	**ST = 0.8**	**ST = 0.7**
50	Mean	0.793	0.696	0.793	0.696	0.793	0.696
	SE	0.106	0.079	0.106	0.079	0.106	0.079
	MSE	0.011	0.006	0.011	0.006	0.011	0.006
	Bias	0.007	0.005	0.007	0.005	0.007	0.005
100	Mean	0.796	0.698	0.796	0.698	0.796	0.698
	SE	0.071	0.053	0.071	0.053	0.071	0.053
	MSE	0.005	0.003	0.005	0.003	0.005	0.003
	Bias	0.004	0.002	0.004	0.002	0.004	0.002
200	Mean	0.804	0.70	0.804	0.70	0.804	0.70
	SE	0.05	0.039	0.05	0.039	0.05	0.039
	MSE	0.003	0.002	0.003	0.002	0.003	0.002
	Bias	0.004	0.000	0.004	0.000	0.004	0.000

Table 6 cont.: Estimates from unadjusted and corrected sensitivities and specificities of the index test under the conditional independence assumption when the reference standard is perfect

Sample size	Methods						
		Unadjusted Sensitivity	Unadjusted Specificity	Brenner Sensitivity	Brenner Specificity	Staquet et al Sensitivity	Staquet et al specificity
	Properties measured	ST = 0.8	ST = 0.7	ST = 0.8	ST = 0.7	ST = 0.8	ST = 0.7
500	Mean	0.798	0.702	0.798	0.702	0.798	0.702
	SE	0.037	0.025	0.037	0.025	0.037	0.025
	MSE	0.001	0.001	0.001	0.001	0.001	0.001
	Bias	0.002	0.002	0.002	0.002	0.002	0.002
1000	Mean	0.798	0.700	0.798	0.700	0.798	0.700
	SE	0.024	0.017	0.024	0.017	0.024	0.017
	MSE	0.001	0.000	0.001	0.000	0.001	0.003
	Bias	0.002	0.000	0.002	0.000	0.002	0.000

MSE is mean squared error; SE is standard error; ST is simulated truth.

Figure 11: The mean, standard error, mean square error and bias of the unadjusted and corrected sensitivity and specificity of the index test when the reference standard is prefect.



In Figure 11(a), the dashed yellow line represents the simulated truth, 0.8 for the sensitivity of the index test and 0.7 for the specificity of the index test. This yellow line is aligned to the red, blue and green lines which represent the Brenner corrected sensitivity (or specificity), unadjusted sensitivity (or specificity), and Staquet et al corrected sensitivity (specificity) respectively. The yellow line is the simulated true values of the sensitivity and the specificity of the index test.

Imperfect reference standard

In this section, the reference standard is assumed to be imperfect. The imperfection of the reference standard is varied to reflect different scenarios that will be explored. Multiple (200) random samples of sizes 50 to 1000 were simulated using the multinomial distribution. The values employed for the simulations are as follows: the sensitivity and specificity of the reference standard are both 0.9 and the sensitivity and specificity of the index test are 0.8 and 0.7 respectively. The prevalence of the target condition is 0.3. The estimated values are presented in Table 7 and Figure 12.

From Table 7 and Figure 12, the unadjusted and Brenner corrected sensitivities and specificities of the index test are poorly estimated irrespective of the sample size. Their mean sensitivities and specificities are lower than the simulated truth. Thus, they are biased estimators for the sensitivity and specificity of the index test when the reference standard is imperfect. The Staquet et al corrected sensitivity and specificity appear to be approximately unbiased. The yellow line is the simulated true values of the sensitivity and the specificity of the index test

Table 7: Unadjusted and corrected sensitivities and specificities of the index test when the reference standard is imperfect and better than the index test

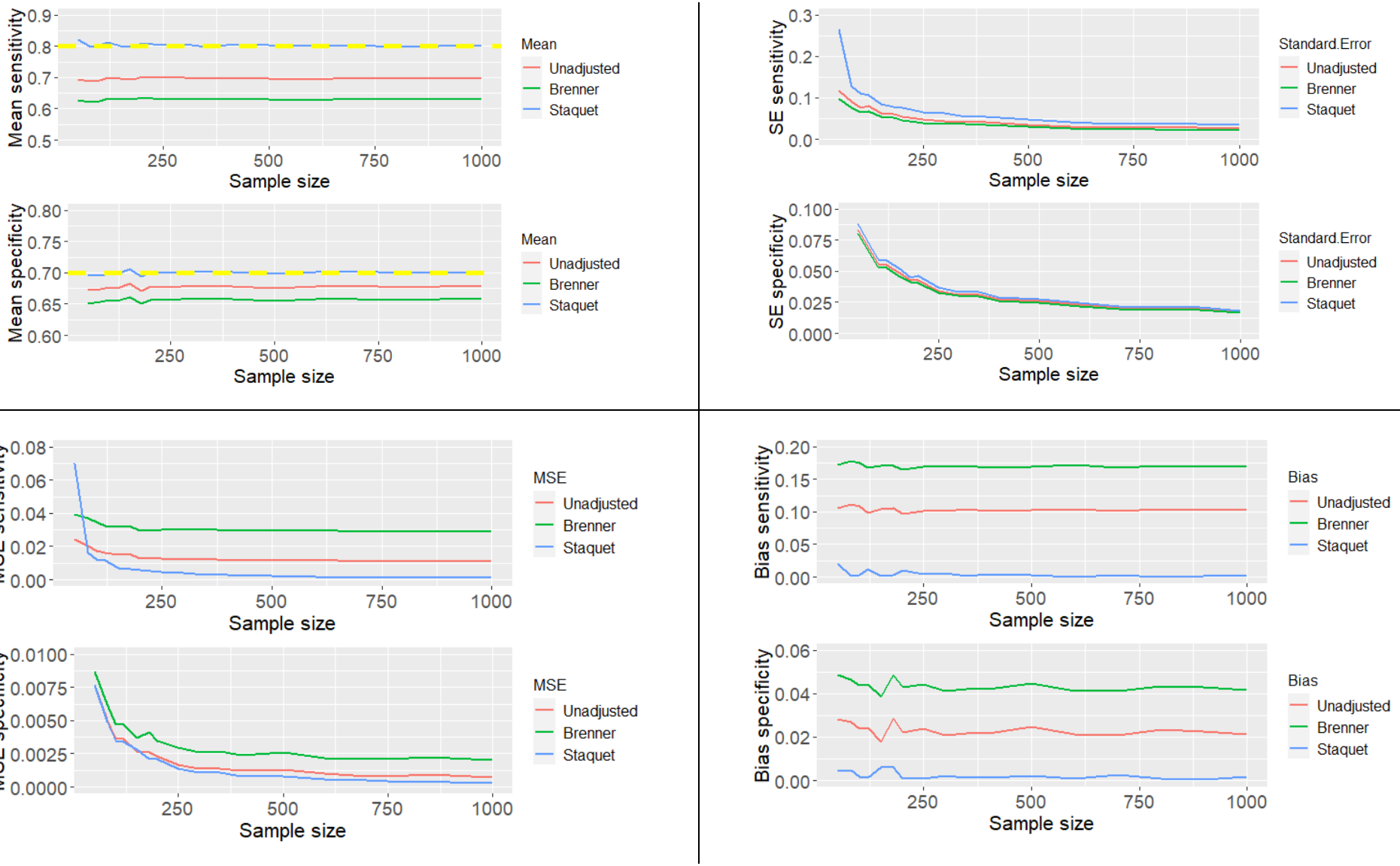
Sample size		Methods					
		Unadjusted Sensitivity	**Unadjusted Specificity**	**Brenner Sensitivity**	**Brenner Specificity**	**Staquet et al Sensitivity**	**Staquet et al specificity**
	Properties measured	**ST = 0.8**	**ST = 0.7**	**ST = 0.8**	**ST = 0.7**	**ST = 0.8**	**ST = 0.7**
50	Mean	0.694	0.672	0.627	0.651	0.820	0.696
	SE	0.116	0.083	0.099	0.08	0.264	0.088
	MSE	0.025	0.008	0.04	0.009	0.080	0.008
	Bias	0.106	0.028	0.173	0.049	0.020	0.004
100	Mean	.692	0.676	0.624	0.656	0.804	0.698
	SE	0.077	0.056	0.065	0.053	0.11	0.06
	MSE	0.018	0.004	0.035	0.005	0.012	0.004
	Bias	0.108	0.024	0.176	0.044	0.004	0.002
200	Mean	0.703	0.678	0.635	0.657	0.81	0.701
	SE	0.055	0.043	0.047	0.041	0.075	0.043
	MSE	0.013	0.002	0.03	0.004	0.006	0.002
	Bias	0.100	0.022	0.165	0.043	0.010	0.001

Table 7 cont.: Unadjusted and corrected sensitivities and specificities of the index test when the reference standard is imperfect and better than the index test

Sample size	Methods						
		Unadjusted Sensitivity	Unadjusted Specificity	Brenner Sensitivity	Brenner Specificity	Staquet et al Sensitivity	Staquet et al specificity
	Properties measured	ST = 0.8	ST = 0.7	ST = 0.8	ST = 0.7	ST = 0.8	ST = 0.7
500	Mean	0.698	0.675	0.631	0.655	0.803	0.698
	SE	0.036	0.026	0.031	0.025	0.048	0.028
	MSE	0.012	0.001	0.03	0.003	0.002	0.001
	Bias	0.0102	0.025	0.169	0.045	0.003	0.002
1000	Mean	0.697	0.679	0.630	0.658	0.801	0.702
	SE	0.026	0.017	0.022	0.017	0.035	0.018
	MSE	0.011	0.001	0.03	0.002	0.001	0.000
	Bias	0.103	0.021	0.17	0.042	0.001	0.002

MSE is mean squared error; SE is standard error; ST is simulated truth.

Figure 12: The mean, standard error, mean square error and bias of the unadjusted and corrected sensitivity and specificity of index test when the reference standard is imperfect and better than the index test.



The second scenario explored assumes that the reference standard is worse than the index tests. The sensitivity and specificity of the index test are both 0.9 and the sensitivity and specificity of the reference standard are 0.8 and 0.7 respectively. The prevalence of the target condition is 0.3. The estimated values are presented in Table 8 and Figure 13.

From Table 8 and Figure 13, it is clear that the unadjusted and Brenner corrected sensitivities and specificities of the index test are poorly estimated irrespective of the sample size. Their means are less than the simulated truth. The Staquet et al corrected sensitivity and specificity are estimated with less bias. However, at small sample sizes, say below 200, there are illogical results. Illogical results imply having sensitivities, specificities or prevalence that are above one or below zero. Illogical results are discussed later in this section. The yellow line is the simulated true values of the sensitivity and the specificity of the index test

Table 8: Unadjusted and corrected sensitivities and specificities of the index test when the reference standard is imperfect and the index test is better than the reference standard

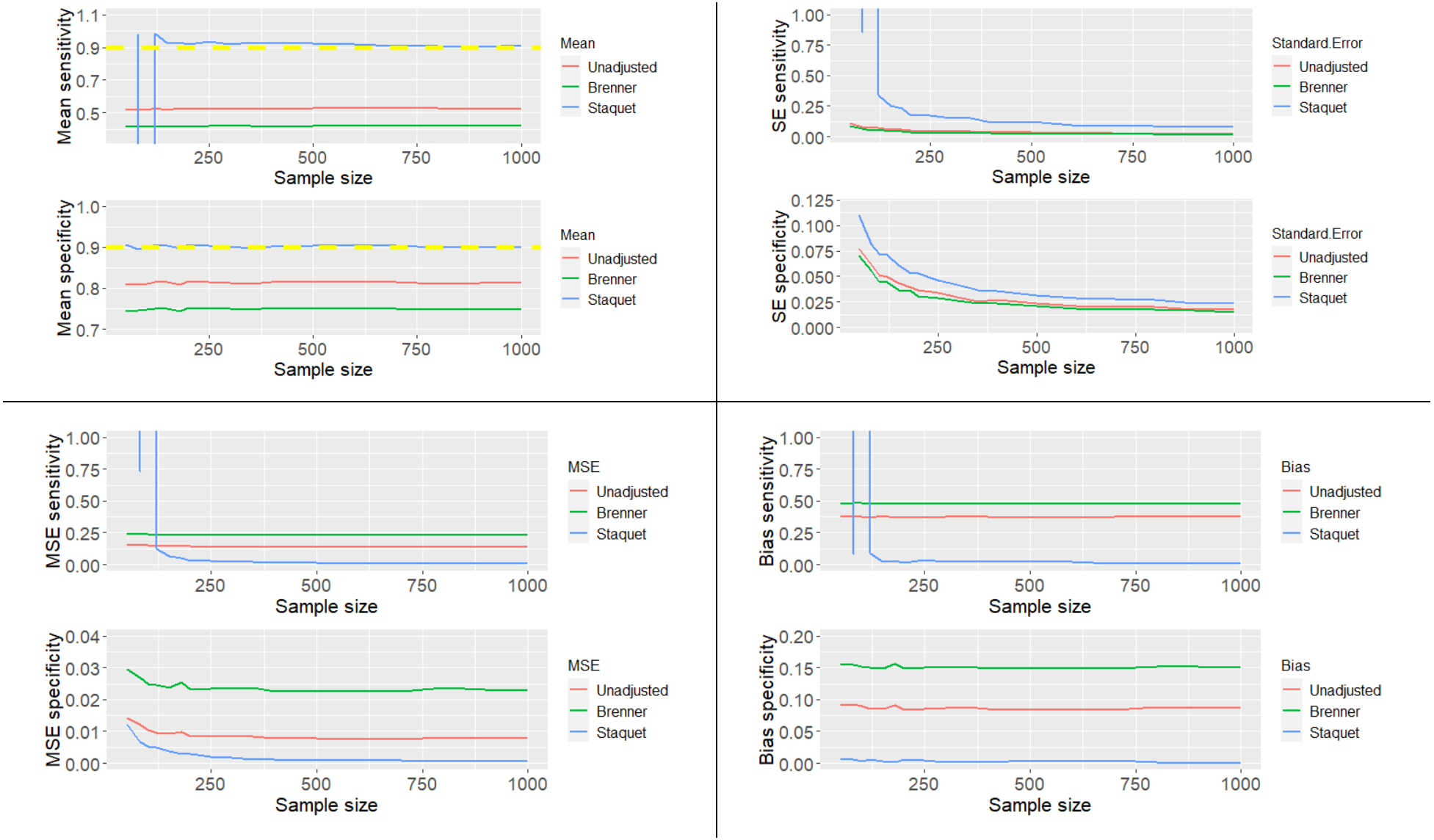
Sample size		Methods					
		Unadjusted Sensitivity	**Unadjusted Specificity**	**Brenner Sensitivity**	**Brenner Specificity**	**Staquet et al Sensitivity**	**Staquet et al specificity**
	Properties measured	**ST = 0.9**	**ST = 0.9**	**ST = 0.9**	**ST = 0.9**	**ST = 0.9**	**ST = 0.9**
50	Mean	0.524	0.809	0.42	0.743	-2×10^{13}	0.906
	SE	0.107	0.078	0.085	0.07	2.04×10^{14}	0.110
	MSE	0.153	0.014	0.238	0.03	4.19×10^{28}	0.012
	Bias	0.376	0.091	0.48	0.157	2.03×10^{13}	0.006
100	Mean	0.520	0.811	0.414	0.784	-1.9×10^{13}	0.898
	SE	0.073	0.051	0.055	0.045	1.94×10^{14}	0.072
	MSE	0.149	0.011	0.239	0.025	3.77×10^{28}	0.005
	Bias	0.38	0.089	0.486	0.152	1.93×10^{13}	0.002
200	Mean	0.523	0.816	0.417	0.751	0.917	0.907
	SE	0.053	0.036	0.039	0.031	0.176	0.053
	MSE	0.145	0.008	0.235	0.023	0.031	0.003
	Bias	0.37	0.084	0.483	0.149	0.017	0.007

Table 8 cont.: Unadjusted and corrected sensitivities and specificities of the index test when the reference standard is imperfect and the index test is better than the reference standard

Sample size	Methods						
		Unadjusted Sensitivity	Unadjusted Specificity	Brenner Sensitivity	Brenner Specificity	Staquet et al Sensitivity	Staquet et al specificity
	Properties measured	ST = 0.9	ST = 0.9	ST = 0.9	ST = 0.9	ST = 0.9	ST = 0.9
500	Mean	0.53	0.815	0.421	0.75	0.923	0.905
	SE	0.033	0.023	0.025	0.021	0.121	0.031
	MSE	0.138	0.008	0.23	0.023	0.015	0.001
	Bias	0.370	0.085	0.479	0.150	0.023	0.005
1000	Mean	0.526	0.814	0.418	0.75	0.913	0.900
	SE	0.024	0.017	0.018	0.015	0.080	0.023
	MSE	0.141	0.008	0.232	0.023	0.006	0.001
	Bias	0.374	0.086	0.482	0.150	0.013	0.000

MSE is mean squared error; SE is standard error; ST is simulated truth.

Figure 13: The mean, standard error, mean square error and bias of the unadjusted and corrected sensitivity and specificity of index test when the reference standard is imperfect and worse than the index test.



The third scenario explored assumes that the reference standard and the index test have the same sensitivity and specificity. The simulated true values of the sensitivities and specificities of the reference standard and index test are 0.9. The prevalence of the target condition is 0.3. The estimated values are presented in Table 9 and Figure 14.

From Table 9 and Figure 14, the unadjusted and Brenner corrected sensitivities and specificities of the index test are poorly estimated irrespective of the sample size as their means are less than the simulated truth. The estimates obtained from the Brenner correction method are usually worse than the unadjusted estimates. The Staquet et al corrected sensitivity and specificity are estimated with less bias. The yellow line is the simulated true values of the sensitivity and the specificity of the index test

Table 9: Unadjusted and corrected sensitivities and specificities of the index test when the reference standard is imperfect and has same sensitivity and specificity as the index test.

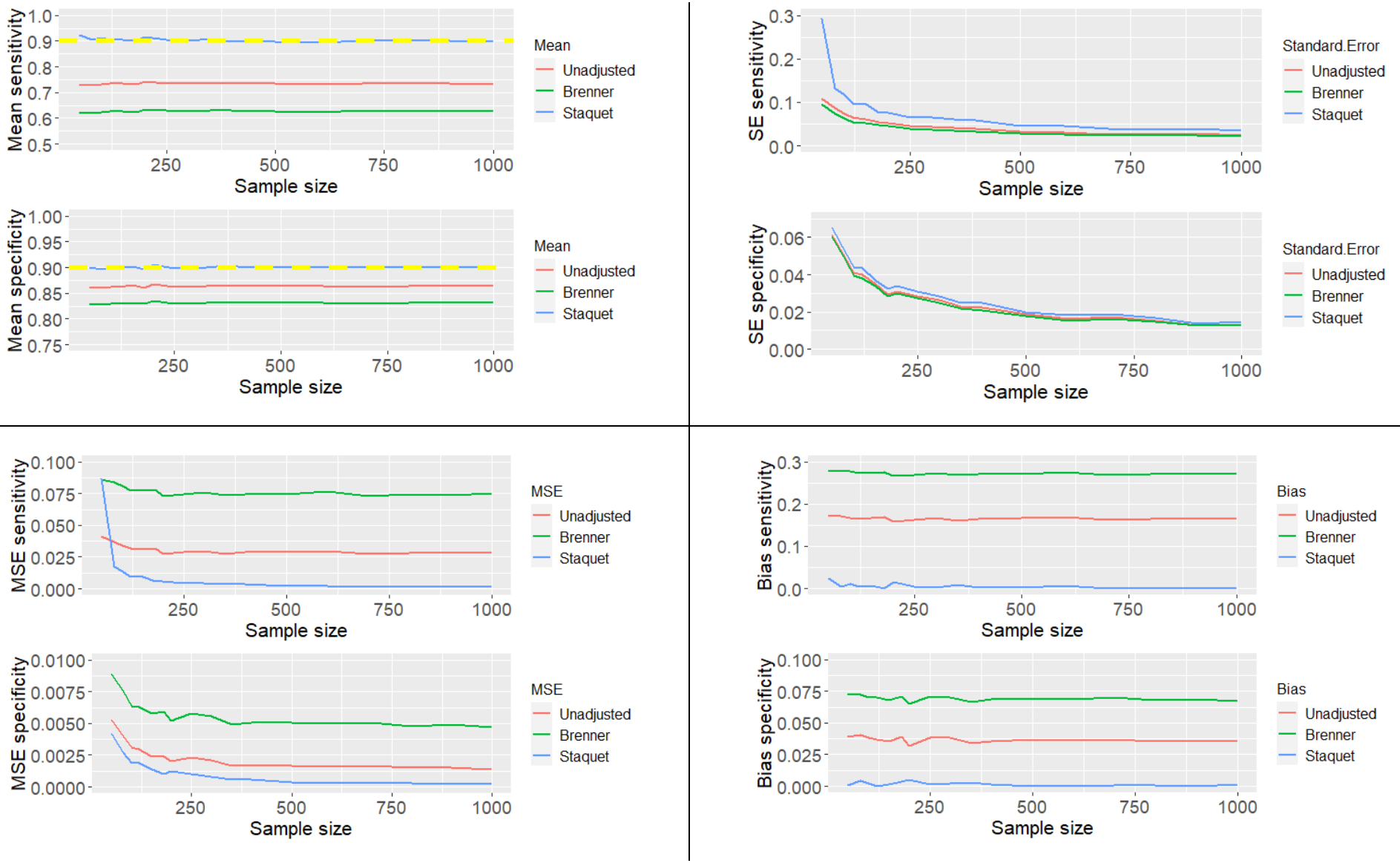
Sample size		Methods					
		Unadjusted Sensitivity	**Unadjusted Specificity**	**Brenner Sensitivity**	**Brenner Specificity**	**Staquet et al Sensitivity**	**Staquet et al specificity**
	Properties measured	**ST = 0.9**	**ST = 0.9**	**ST = 0.9**	**ST = 0.9**	**ST = 0.9**	**ST = 0.9**
50	Mean	0.729	0.861	0.623	0.827	0.924	0.0899
	SE	0.108	0.061	0.095	0.06	0.294	0.065
	MSE	0.041	0.005	0.086	0.009	0.087	0.004
	Bias	0.171	0.039	0.277	0.073	0.024	0.001
100	Mean	0.732	0.862	0.623	0.830	0.912	0.898
	SE	0.074	0.041	0.064	0.04	0.117	0.044
	MSE	0.034	0.003	0.081	0.006	0.014	0.002
	Bias	0.168	0.038	0.277	0.07	0.012	0.002
200	Mean	0.743	0.868	0.633	0.835	0.915	0.905
	SE	0.053	0.031	0.046	0.030	0.076	0.034
	MSE	0.028	0.002	0.073	0.005	0.006	0.001
	Bias	0.158	0.032	0.267	0.065	0.015	0.005

Table 9 cont.: Unadjusted and corrected sensitivities and specificities of the index test when the reference standard is imperfect has same sensitivity and specificity as the index test

Sample size	Methods						
		Unadjusted Sensitivity	Unadjusted Specificity	Brenner Sensitivity	Brenner Specificity	Staquet et al Sensitivity	Staquet et al specificity
	Properties measured	ST = 0.9	ST = 0.9	ST = 0.9	ST = 0.9	ST = 0.9	ST = 0.9
500	Mean	0.734	0.863	0.628	0.831	0.899	0.900
	SE	0.032	0.019	0.028	0.018	0.045	0.020
	MSE	0.029	0.002	0.075	0.005	0.002	0.000
	Bias	0.166	0.037	0.272	0.069	0.001	0.000
1000	Mean	0.733	0.865	0.627	0.832	0.898	0.901
	SE	0.025	0.013	0.021	0.013	0.035	0.015
	MSE	0.029	0.001	0.075	0.005	0.001	0.000
	Bias	0.167	0.035	0.273	0.068	0.002	0.001

MSE is mean squared error; SE is standard error; ST is simulated truth.

Figure 14: The mean, standard error, mean square error and bias of the unadjusted and corrected sensitivity and specificity of index test when the reference standard is imperfect and have same sensitivity and specificity as the index test.



Other combinations

In this section, four scenarios were explored where the sensitivity (or specificity) of the RS or IT varied from zero to one (with an increments of 0.01).

a) **Scenario one**: The sensitivity of the RS was varied from 0 to 1 as the specificity of RS and IT, and the sensitivity of IT were fixed at 0.9, 0.8 and 0.8 respectively.

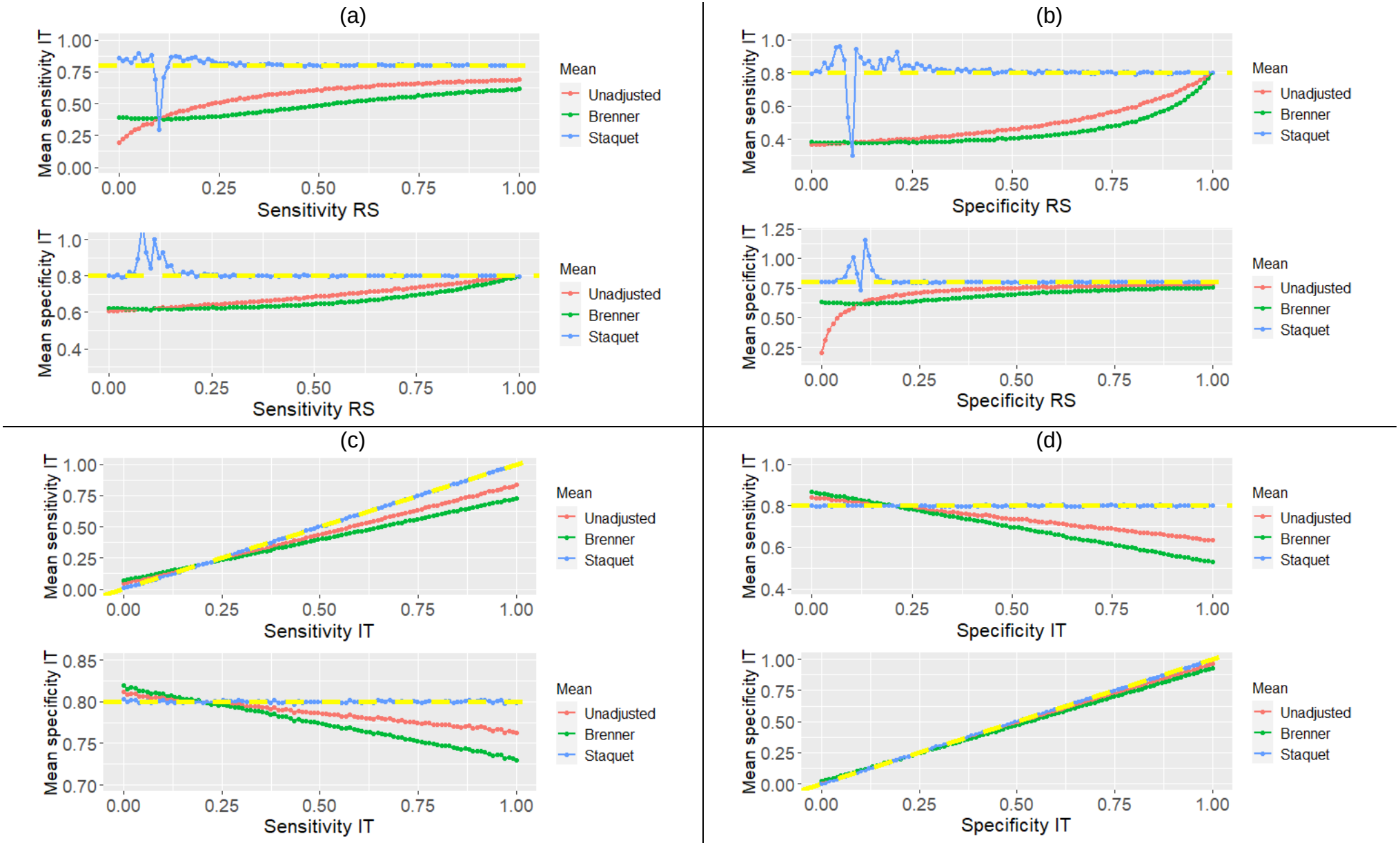
b) **Scenario two**: The specificity of the RS was varied from 0 to 1 as the sensitivity of RS and IT, and the specificity of IT were fixed at 0.9, 0.8 and 0.8 respectively.

c) **Scenario three**: The sensitivity of the IT was varied from 0 to 1 as the specificity of RS and IT, and the sensitivity of RS were fixed at 0.9, 0.8 and 0.9 respectively.

d) **Scenario three**: The specificity of the IT was varied from 0 to 1 as the sensitivity of RS and IT, and the specificity of RS were fixed at 0.9, 0.8 and 0.9 respectively.

The mean sensitivity and mean specificity of the IT in the four scenarios are reported in Figure 15 as (a), (b), (c), and (d) respectively. From Figure 15, the estimates obtained from the Staquet et al approach are approximately equivalent to the simulated true values for the index test. However, when the sensitivity (or specificity) of the RS is low (< 0.3), the estimates obtained using the Staquet et al correction method could be inaccurate. Conventionally, the reference standard in clinical case studies do not have low sensitivity or specificity. The sensitivity and specificity of a reference standard are often above 0.5 (readers can explore further combination of sensitivity and specificity using the R-Code in the Appendix **B.1**).

Figure 15: The unadjusted and corrected mean sensitivity and mean specificity of the index test when the sensitivity (or specificity) of the reference standard or index test is varied and the prevalence is fixed at 0.3.

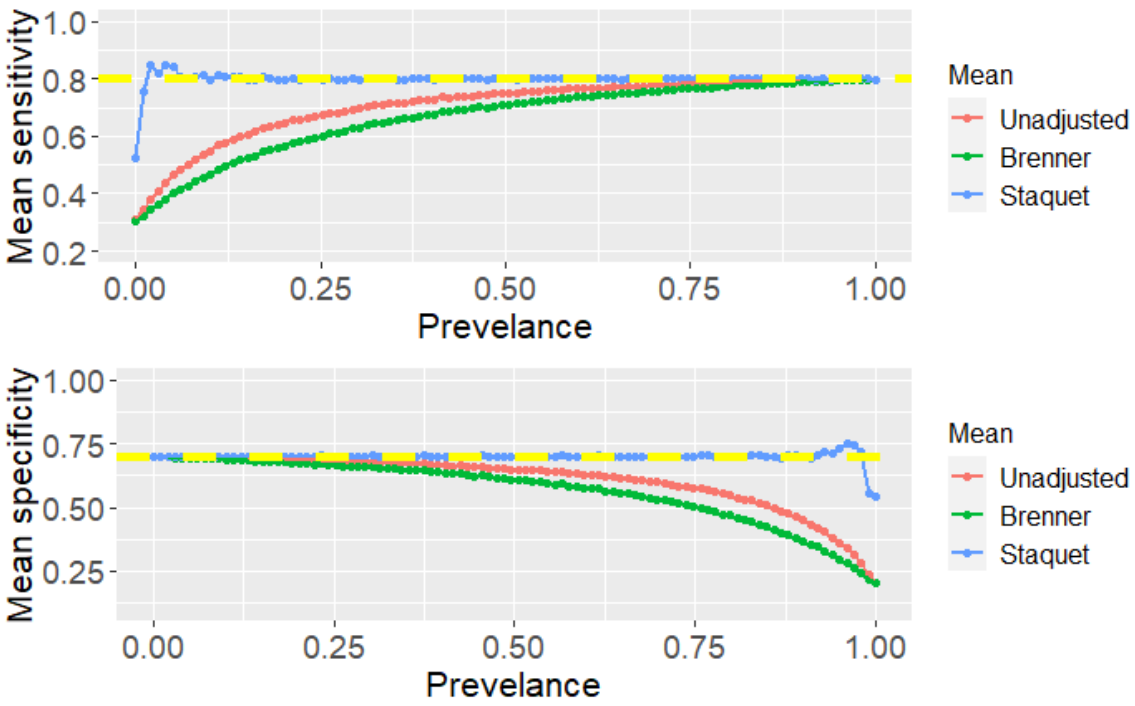


Impact of varying the prevalence on unadjusted and correction methods

In this section, I wish to explore the impact of varying the prevalence of the target condition given that the sensitivity and specificity of the index test and reference standard are fixed across simulated samples. This investigation will help to understand which of the methods are the least impacted by changes in prevalence and which methods uphold the assumption of constant sensitivity and specificity across populations of differing prevalence.

Multiple (200) samples of 1000 participants were simulated at 100 different prevalence rates (varying from 0 to 1). The mean unadjusted and corrected mean sensitivities and mean specificities of the index test are displayed Figure 16. Changes in the prevalence of the disease have a large impact on the unadjusted and Brenner corrected sensitivities and specificities of the index test compared to the Staquet et al corrected sensitivity and specificity. The unadjusted and Brenner corrected sensitivities increase towards the simulated truth as the prevalence of the target condition increases. The Brenner and unadjusted specificities tend toward the simulated truth as the prevalence decreases. It could be that the estimates obtained via the classical approach and Brenner correction method are impacted by the sample prevalence. With the Staquet et al correction method, irrespective of the varying prevalences, the estimated sensitivity and specificity of the index test are the same as the simulated truth and are constant. This indicates that the Staquet et al correction method upholds the assumption of constant sensitivity and specificity across all populations (with different prevalences). However, there are some illogical sensitivities produced when the prevalence is very low.

Figure 16: Changes in prevalence largely impact the unadjusted and Brenner corrected sensitivity and specificity of the index test, unlike the Staquet et al correction method.



Illogical results

Illogical results for sensitivities and specificities occur when their estimated values are above or below their possible range, which in each case is zero to one inclusive. Undefined sensitivities or specificities occur when the estimated specificities or sensitivities are mathematically undefined. In our results (Table 8 and Figure 13), illogical and undefined estimates were obtained for the Staquet et al correction method but not the Brenner correction method or the classical method. There are various reasons for obtaining illogical or undefined estimates such as high or low prevalence and small sample sizes. When the sample size is small (for example 50), there is the likelihood of having undefined specificities when the prevalence is high (for example 0.85). When the estimated prevalence is 1, illogical results are always obtained in the Staquet et al method. Although, this case would not be of interest in practice. A consequence of undefined specificities is that; it is impossible to obtain a mean specificity. Therefore, they are excluded when computing the mean specificities, standard deviation, bias and MSE from the Staquet et al estimands. Illogical

sensitivities (specificities) occurs when the sample size is small, and the prevalence is very low (high).

Table 10 reports the number of illogical and undefined estimates produced out of 200 random simulated samples at various sample sizes and prevalences using the Staquet et al correction method. When the prevalence is low (e.g. 0.1) or very high (e.g. 0.95), there is a large number of illogical estimates, even with large sample sizes.

Table 10: Number of illogical and undefined results obtained at various sample sizes and prevalences

Sample size	Number of samples out of 200 producing illogical sensitivity and/or specificity at different prevalence values (number of samples with undefined results)					
	Prevalence					
	0.1	**0.2**	**0.3**	**0.5**	**0.85**	**0.95**
50	64 (0)	33 (0)	0 (0)	3 (0)	42 (6)	57 (23)
80	55 (0)	26 (0)	0 (0)	0 (0)	27 (2)	65 (16)
100	49(0)	23 (0)	0 (0)	0 (0)	18 (0)	53 (15)
120	48 (0)	13 (0)	0 (0)	0 (0)	12 (0)	64 (8)
150	52 (0)	11(0)	0 (0)	0 (0)	6 (0)	54 (6)
200	44 (0)	13 (0)	0(0)	0 (0)	4 (0)	58 (3)
250	36 (0)	10 (0)	0 (0)	0 (0)	3 (0)	42 (4)
300	27 (0)	6 (0)	0 (0)	0 (0)	3 (0)	46 (0)
350	22 (0)	0 (0)	0 (0)	0 (0)	3 (0)	46 (2)
400	22 (0)	2 (0)	0 (0)	0 (0)	0 (0)	46 (1)
500	15 (0)	0 (0)	0 (0)	0 (0)	0 (0)	38 (0)
700	14 (0)	0 (0)	0 (0)	0 (0)	0 (0)	26 (1)
1000	2 (0)	0 (0)	0 (0)	0 (0)	0 (0)	11 (0)

Conditions for obtaining illogical estimates using Staquet et al[119] approach

The Staquet et al[119] correction method is explored algebraically to understand the conditions for obtaining illogical estimates. Illogical estimates are estimates (sensitivity, specificity and prevalence) that are outside [0, 1].

Illogical estimates for prevalence

Algebraically, illogical estimate (greater than one) is obtained for the estimated prevalence if:

$$N(Sp_{RS} - 1) + e > N(Sn_{RS} + Sp_{RS} - 1)$$

$$NSp_{RS} - N + e > NSn_{RS} + NSp_{RS} - N$$

$$e > NSn_{RS}$$

$$Sn_{RS} < \frac{e}{N} = Prr \qquad Condition\ (1)$$

Illogical estimates for sensitivity and specificity

Algebraically, illogical estimates are obtained for the sensitivity of the IT via the Staquet et al[119] approach if:

$$Sp_{RS} < \frac{d}{h} = rNPV \qquad Condition\ (2)$$

$$Sp_{RS} < \frac{b}{g} = rPPV' \quad \textbf{and} \quad Sp_{RS} > \frac{f}{N} = 1 - Prr = Prr' \qquad Condition\ (3a)$$

$$Sp_{RS} > \frac{b}{g} = rPPV' \quad \textbf{and} \quad Sp_{RS} < \frac{f}{N} = 1 - Prr = Prr' \qquad Condition\ (3b)$$

Condition (2) produces an estimated corrected sensitivity whose absolute value is greater than one and condition (3) produces a negative estimate that is estimate less than zero.

Similarly, illogical estimates are obtained for the specificity of IT if:

$$Sn_{RS} < \frac{a}{g} = rPPV \qquad Condition\ (4)$$

$$Sn_{RS} < \frac{c}{h} = rNPV' \quad \text{and} \quad Sn_{RS} > \frac{e}{N} = Prr \qquad Condition\ (5a)$$

$$Sn_{RS} > \frac{c}{h} = rNPV' \quad \text{and} \quad Sn_{R} < \frac{e}{N} = Prr \qquad Condition\ (5b)$$

Condition (4) produces an estimated corrected specificity whose absolute value is greater than one, and condition (5) produces negative estimates.

The $\boldsymbol{rNPV}$ refers to the ***"relative negative predictive value"***. It is the proportion of participants with negative results in both the IT and RS divided by total number of participants with a negative IT result. It is termed ***relative*** because it is obtained in relation to the RS which is imperfect. If the RS was a gold standard, it would be called the negative predictive value (NPV). Therefore, the complement of $rNPV$ ($rNPV'$), is estimated as:

$$1 - rNPV = rNPV' = 1 - \frac{d}{h} = \frac{c}{h}$$

The ***"relative positive predictive value (rPPV)"*** is the proportion of participants with positive test results in both the IT and RS divided by the total number of participants with a positive IT result.

Key conclusions from conditional independence assumption.

Following the simulation studies in this section, firstly, when the reference standard is perfect, the classical method, Brenner corrected method and Staquet et al corrected method will all accurately estimates the sensitivity and specificity of the index test irrespective of the prevalence in the population and sample size. Secondly, when the reference standard is imperfect and conditionally independent of the index test given the true disease status, it does not matter whether the sensitivity and specificity of the reference standard is better than the index test, or worse or they are the same, the Staquet et al method is an unbiased and consistent estimator for sensitivities and

specificities. It outperformed the Brenner corrected and classical methods irrespective of the prevalence rates. Thirdly, estimates obtained using the classical and Brenner methods are always affected by the prevalence of the target condition. If the prevalence is high, the sensitivity is more likely to be estimated correctly and the specificity of the index test is underestimated, and if the prevalence is low the sensitivity is underestimated, and the specificity is more likely to be estimated correctly. Finally, the classical and Brenner methods do not produce illogical results (that is estimates outside 0 and 1) irrespective of the sample size or prevalence of the target condition. However, illogical estimates can be obtained with the Staquet et al method especially if the prevalence of the diseases is very low or high. Thus, having illogical results is not a sufficient indication of conditional dependence between the two tests. It could be that the sample size is small and the distribution of participants across the cells in the 2 x 2 contingency table is the reason for obtaining the illogical results.

3.3.2. Comparison of correction methods – conditional dependence

Two medical tests (say T_1 and T_2) are assumed to be ***statistically*** conditionally dependent given the true disease status (D) if:

$$\Pr(T_1 = 1, T_2 = 1|D = d) \neq \Pr(T_1 = 1|D = d) \times \Pr(T_2 = 1|\, D = d); \qquad d = 0{,}1$$

In this case, the correlation coefficient (ρ) between the two tests given the true disease status (the diseased or non-diseased group) is not equal to zero ($\rho \neq 0$). The correlation between the two tests, can be expressed using the covariance between the two tests results among the diseased and the non-diseased groups. Therefore, to introduce conditional dependence between the two tests (index test and reference standard) in our simulation, the fixed effects modelling approach[54, 122, 286, 291, 292] will be employed. The joint probability of any two tests (say T_1 and T_2) given the true disease status can be expressed as[54, 286]:

$$P(T_1 = t_1, T_2 = t_2|D = d) = \prod_{i=1}^{2} P(T_i = t_i|D = d) + \varphi_{t_1,t_2|d}$$

where $\varphi_{t_1,t_2|d}$ is the conditional dependence term among the diseased / non-diseased group. The joint probability of the two tests is expressed as:

$$P(T_1 = t_1, T_2 = t_2) =$$

$$p\prod_{i=1}^{2}P(T_i = t_i|D = 1) + \varphi_{t_1,t_2|1} + (1 - p)\prod_{i=1}^{2}P(T_i = t_i|D = 0) + \varphi_{t_1,t_2|0}$$

The two tests explored in the simulation study reported in this chapter are the index test denoted by T and the reference test denoted by R. Constraints are required on the covariance parameters to ensure that the correlations remain in the interval of -1 and 1. The inequality[54] and equality constraints of the conditional dependence terms[286] using the fixed effects model are expressed as follows:

The inequality constraints are:

$$-Sn_R \times Sn_T + \max(0, Sn_R + Sn_T - 1) \leq \varphi_{1,1|1} \leq \min(Sn_R, Sn_T) - (Sn_R \times Sn_T)$$

$$-Sp_R \times Sp_T + \max(0, Sp_R + Sp_T - 1) \leq \varphi_{0,0|0} \leq \min(Sp_R, Sp_T) - (Sp_R \times Sp_T)$$

where, $\varphi_{1,1|1}$ is the covariance term among the diseased participants with a positive response on both the index test and reference standard and $\varphi_{0,0|0}$ is the covariance term among the non-diseased participants with a negative response on both the index test and reference standard. The maximum and minimum values of the covariance terms depend on the sensitivity and specificity of the index test and reference standard.

The equality constraints are:

$$\varphi_{0,0|d} + \varphi_{0,1|d} = 0 \quad or \quad \varphi_{0,1|d} + \varphi_{1,1|d} = 0\,;$$

$$\varphi_{0,0|d} + \varphi_{1,0|d} = 0 \quad or \quad \varphi_{1,0|d} + \varphi_{1,1|d} = 0;$$

$$d = 0,1$$

and

$$\varphi_{0,0|d} + \varphi_{0,1|d} + \varphi_{1,0|d} + \varphi_{1,1|d} = 0$$

Following the inequality and equality constraints, there are seven possible combinations of covariances that can exist between the two tests given the true disease status. A dependence on the sensitivities of both tests (correlation between the two tests in the diseased group) does not imply dependence on the specificities of the two tests (correlation between the two tests in the non-diseased group)[292]. These seven possible combinations are:

- **Case 1**: The covariances between the two tests in the diseased and non-diseased groups are both positive.

- **Case 2**: The covariances between the two tests in the diseased and non-diseased groups are both negative.
- **Case 3**: The covariance between the two tests in the diseased group is positive and the covariance in the non- diseased group is negative.
- **Case 4**: The covariance between the two tests in the diseased group is negative and the covariance in the non- diseased group is positive.
- **Case 5**: The covariance between the two tests in the diseased group is either positive or negative and the covariance in the non-diseased group is 0.
- **Case 6**: The covariance between the two tests in the non – diseased group is either positive or negative and the covariance in the diseased group is 0.
- **Case 7**: The covariance between the two tests in the diseased and non-diseased groups are zero. This is a special case of when the two tests are conditionally independent given the true disease status.

Cases 1 – 6 are known as pairwise correlation[54, 286] because the two tests' responses are correlated. Cases 1 – 6 will have different variations because changing the values of the covariance terms among diseased and / or non – diseased groups will always yield different cell probabilities, different cell frequencies (see Table 11) and ultimately different estimates of the sensitivity and specificity of the index test if the conditional dependence between the two tests is not taken into consideration. It is important to note that if the value of the covariance term is very low (that is close to zero) then it is likely that it will have no significant impact on the estimated sensitivity and / or specificity of the index test. The probability of each cell in the 4 x 2 and 2 x 2 cell probabilities tables given that the reference standard and the index test are conditionally dependent given the true disease status is depicted in Table 11.

Table 11: 4 x 2 and 2 x 2 tables of cell probabilities classified by the true disease, reference standard and index tests results

	Diseased (+)		**Diseased (-)**	
	RS +	**RS –**	**RS +**	**RS –**
T +	$p_1(Sn_R \times Sn_T + \varphi_{11\|1})$	$p_1\left((1-Sn_R)Sn_T + \varphi_{01\|1}\right)$	$p_0((1-Sp_R)(1-Sp_T) + \varphi_{11\|0})$	$p_0(Sp_R(1-Sp_T) + \varphi_{01\|0})$
T –	$p_1\left(Sn_R(1-Sn_T) + \varphi_{10\|1}\right)$	$p_1((1-Sn_R)(1-Sn_T) + \varphi_{00\|1})$	$p_0\left(Sp_T(1-Sp_R) + \varphi_{10\|0}\right)$	$p_0\left(Sp_R \times Sp_T + \varphi_{00\|0}\right)$
OR				
	Reference standard (RS)			
	Positive (+)		**Negative (-)**	
T +	$p_1\left(Sn_R \times Sn_T + \varphi_{11\|1}\right) + p_0((1-Sp_R)(1-Sp_T) + \varphi_{11\|0})$		$p_1\left((1-Sn_R)Sn_T + \varphi_{01\|1}\right) + p_0(Sp_R(1-Sp_T) + \varphi_{01\|0})$	
T –	$p_1\left(Sn_R(1-Sn_T) + \varphi_{10\|1}\right) + p_0\left(Sp_T(1-Sp_R) + \varphi_{10\|0}\right)$		$p_1\left((1-Sn_R)(1-Sn_T) + \varphi_{00\|1}\right) + p_0\left(Sp_R \times Sp_T + \varphi_{00\|0}\right)$	

Sn_R is sensitivity of reference standard; Sn_T is sensitivity of index test; Sp_R is specificity of the reference standard; Sp_T is specificity of the index test; RS is reference standard; T+ is index test positive; T- is index test negative, p_1 is the prevalence of the target condition; p_0 is 1 – p_1. $\varphi_{00|0}$, $\varphi_{11|0}, \varphi_{01|0}, \varphi_{10|0}, \varphi_{00|1}$, $\varphi_{11|1}, \varphi_{01|1}, \varphi_{10|1}$ are covariance terms.

Perfect reference standard

Taking the inequality constraints [54, 286], we can see that the covariance term among the diseased group is zero.

$$-Sn_R \times Sn_T + \max(0, Sn_R + Sn_T - 1) \leq \varphi_{1,1|1} \leq \min(Sn_R, Sn_T) - (Sn_R \times Sn_T)$$

$$\Rightarrow -1 \times Sn_T + \max(0, 1 + Sn_T - 1) \leq \varphi_{1,1|1} \leq \min(1, Sn_T) - (1 \times Sn_T)$$

$$\Rightarrow -Sn_T + Sn_T \leq \varphi_{1,1|1} \leq Sn_T - Sn_T$$

$$\Rightarrow 0 \leq \varphi_{1,1|1} \leq 0$$

The same applies to the non-diseased group. This implies that the two tests are conditionally independent. Thus, provided that the reference standard employed in the study is a perfect test, there is no need to adjust for any conditional dependence as it does not exist. This is simple to understand as the dependence is between the errors made by the tests, and a perfect reference standard would not make any errors. No simulation pertaining to a perfect reference standard is performed in this section as this will be a replica of the simulation performed in section **3.3.1** while assuming the reference standard is perfect.

Imperfect reference standard

In this section, the reference standard is not perfect, and it is correlated with the index test. The simulated true values for the sensitivity and specificity of the reference standard are both 0.9. The sensitivity and specificity of the index test are both 0.8 and the prevalence of the target condition is 0.3. Following the inequality constraints, the bounded value of the covariance terms among the diseased group is:

$$-Sn_R \times Sn_T + \max(0, Sn_R + Sn_T - 1) \leq \varphi_{1,1|1} \leq \min(Sn_R, Sn_T) - (Sn_R \times Sn_T)$$

$$\Rightarrow -0.02 \leq \varphi_{1,1|1} \leq 0.08$$

and among the non-diseased group is:

$$-Sp_R \times Sp_T + \max(0, Sp_R + Sp_T - 1) \leq \varphi_{0,0|0} \leq \min(Sp_R, Sp_T) - (Sp_R \times Sp_T)$$

$$\Rightarrow -0.02 \leq \varphi_{0,0|0} \leq 0.08$$

In addition to the Brenner correction method and the Staquet et al correction method, the Brenner correction method for two positively correlated tests (Equations 6 and 7) were also investigated in this section (section **3.3.2**). To investigate the performance

of these correction methods assuming that the tests are conditionally dependent, 100 samples of 1000 participants were simulated at 100 different prevalence values (from 0 to 1) using the multinomial distribution. The different scenarios of conditional dependence investigated were cases 1 – 6. Case seven which is a case of conditional independence, has already being explored in section **3.3.1**. The estimated unadjusted and corrected sensitivities and specificities of the index test under different variation of conditional dependence between the index test and the reference standard are displayed in Figure 17 (plots a – h). In Figure 17, the Brennerpos represents estimates obtained from the second pair of estimators proposed by Brenner, which is employed to correct for the sensitivity and specificity of the index test given that the index test and reference standard are positively correlated. Each plot displayed in Figure 17 labelled a – h is:

a. **Case 1**: $cov_d = cov_{nd} = 0.08.$
b. **Case 2**: $cov_d = -0.02;\ cov_{nd} = -0.01.$
c. **Case 3**: $cov_d = 0.07;\ cov_{nd} = -0.01.$
d. **Case 4**: $cov_d = -0.02;\ cov_{nd} = 0.08.$
e. **Case 5a**: $cov_d = 0.08;\ cov_{nd} = 0.$
f. **Case 5b**: $cov_d = -0.02;\ cov_{nd} = 0.$
g. **Case 6a**: $(cov_d = 0;\ cov_{nd} = 0.08.$
h. **Case 6b**: $cov_d = 0;\ cov_{nd} = -0.01.$

From the plots, all the methods perform poorly in estimating the sensitivity and specificity as they are either over estimated or underestimated.

Key conclusions from the conditional dependence assumption.

Following the simulation study, when the reference standard and the index test are conditionally dependent, all estimators (classical, Brenner and Staquet et al correction methods) performs poorly as they either underestimate or overestimate the accuracy measures of the index test. This is expected for the explored methods except the Brenner correction method for positively correlated tests in case 1; because the Brenner correction method for positively correlated tests was developed to be implemented when the two tests are conditionally dependent given the true disease

status. In addition, the sensitivity and specificity estimated from the Staquet et al approach is not constant across different populations unlike when the reference standard and index tests are conditionally independent. Thus, the estimates change with the prevalence of the target condition. Furthermore, there are still illogical results produced when using the Staquet et al methods, especially at very high or very low prevalence while the Brenner and Classical methods do not produce illogical estimates.

Figure 17: The unadjusted and corrected sensitivities and specificities of the index test under different variations of conditional dependence between the index test and the reference standard.

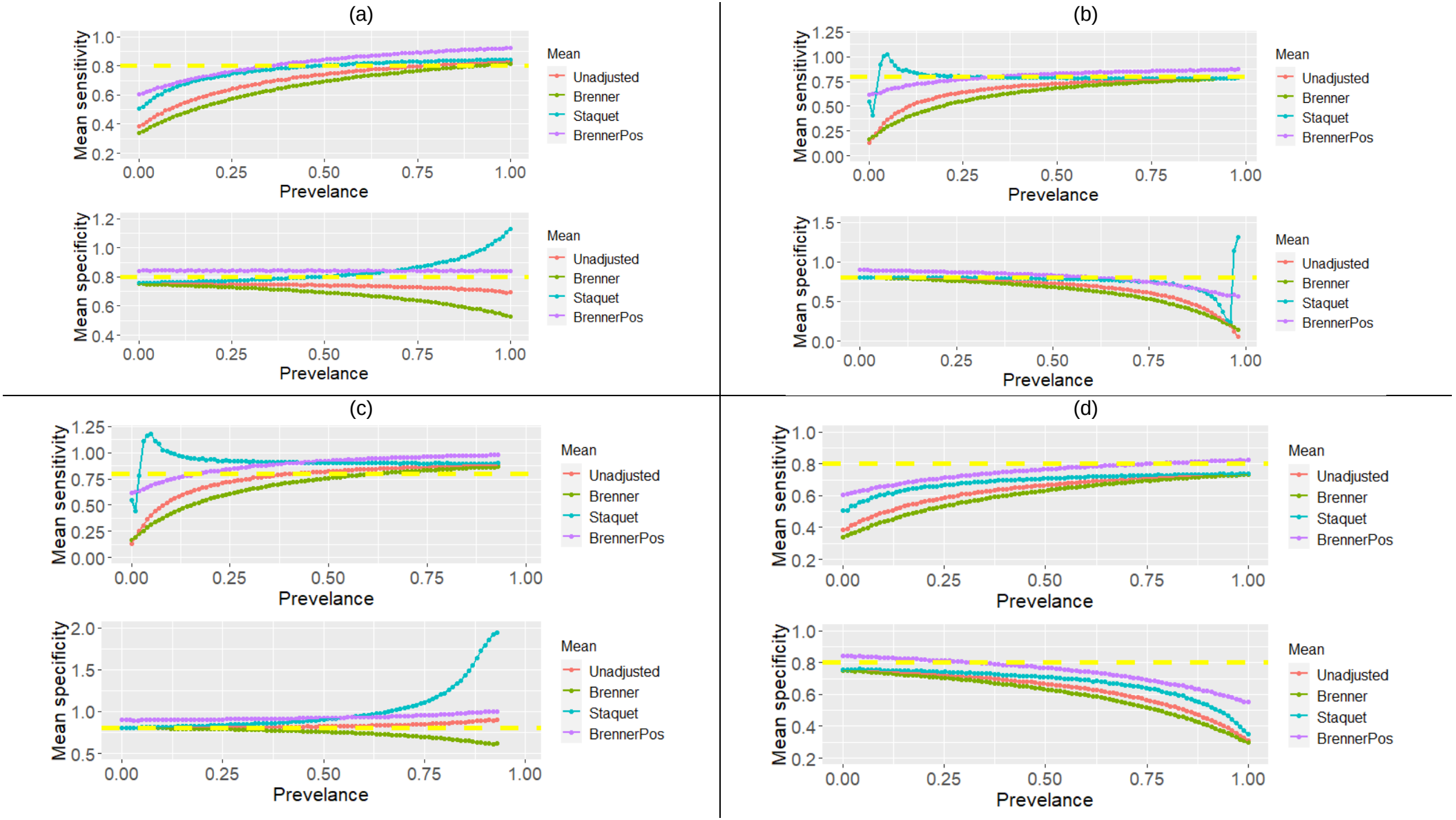
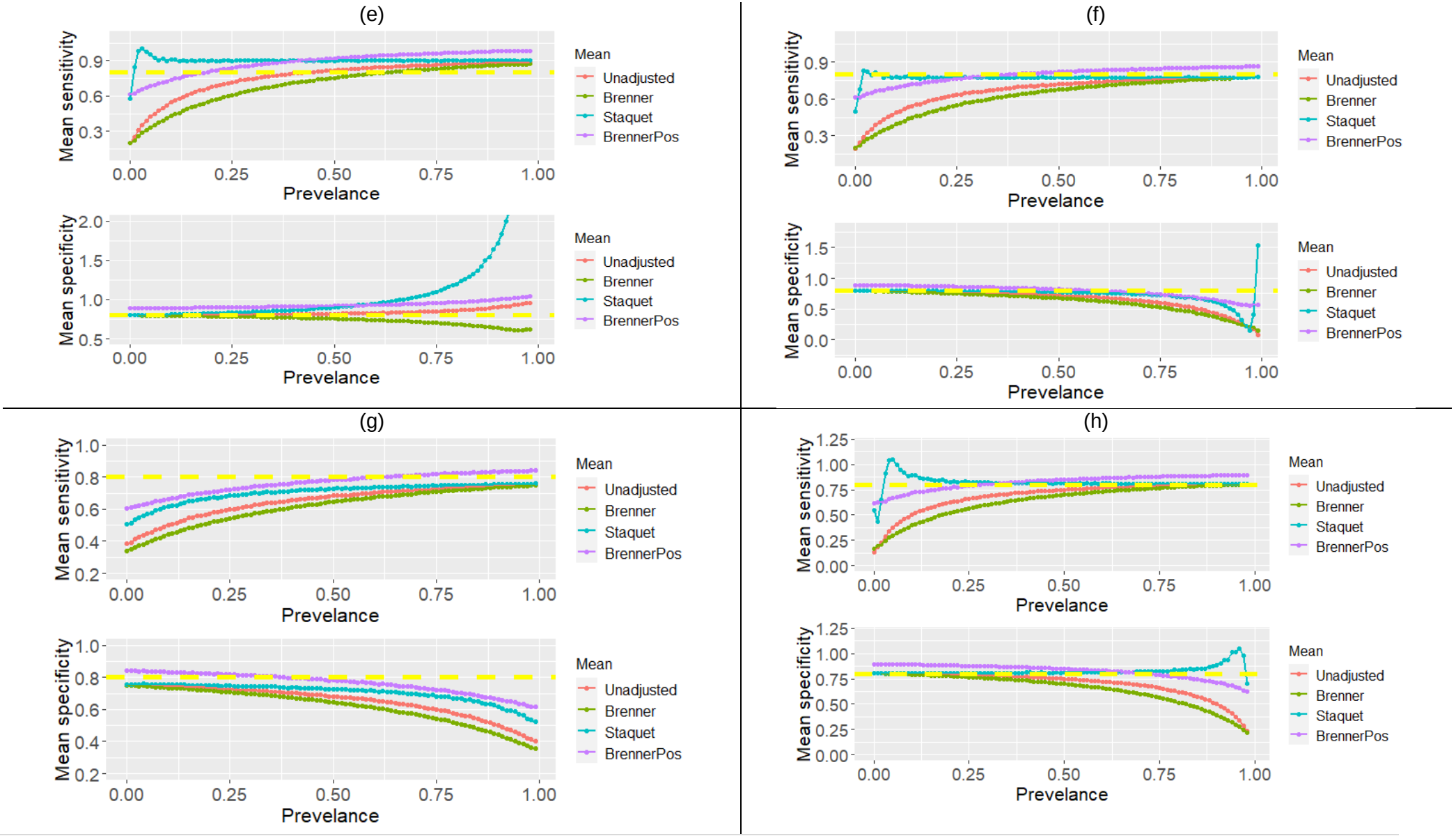


Figure 17 cont.: The unadjusted and corrected sensitivities and specificities of the index test under different variations of conditional dependence between the index test and reference standard



3.4. Application of methods to a clinical dataset

In this section three clinical datasets from two published articles (Mathews et al[125] and Matos et al[124]) were reanalysed to explore how the different methods (classical, Brenner and Staquet et al correction methods) employed can affect the estimates obtained and in turn potentially affect the adoption and usage of the evaluated test in clinical practice. In addition, exploration of these clinical datasets could support the findings observed in the simulation studies.

Reanalysis of the clinical dataset by Mathews et al

The extracted clinical dataset from Mathews et al[125] (Table 12) aims to evaluated the sensitivity and specificity of high resolution anoscopy (HRA) cytology in discriminating HIV patients into high grade squamous intraepithelial lesion (HSIL) and atypical squamous cells high grade (ASC-H) or not.

Table 12: Results of HRA cytology and punch biopsy in classifying patients into high grade and non-high grade squamous intraepithelial lesion

	Biopsy $\geq AIN2$	**Biopsy $< AIN2$**	**Total**
Cytology HSIL or ASC-H	40	22	62
Cytology < HSIL	22	177	199
	62	199	261

HSIL: high grade squamous intraepithelial lesion; ASC-H: atypical squamous cells; AIN: anal intraepithelial neoplasia. Biopsy ≥ AIN2 indicate positive results and Biopsy < AIN2 indicate negative result.

The punch biopsy was employed as the reference standard, which is assumed to be imperfect. According to Mathews et al[125], the sensitivity and specificity of punch biopsy were extracted from Byrom et al[125, 293] which are 0.74 and 0.91 respectively. The study by Mathews et al[125] employed the Staquet et al approach to correct for the sensitivity and specificity of HRA cytology given that the accuracy measures of punch biopsy are known and assuming both tests (index and reference standard) are conditionally independent. In this research, this dataset (Table 12) was reanalysed using the Brenner correction method.

The estimated population prevalence ($\hat{P}$) and sample prevalence (Prr) are 0.27 and 0.23 respectively (≅0.3), indicating a very low likelihood, based on the simulation study, of obtaining an illogical result when using the Staquet et al approach.

The corrected and unadjusted sensitivity and specificity of HRA cytology are presented in Table 13.

Table 13: Unadjusted and corrected sensitivities and specificities of HRA cytology

Accuracy measures	Methods		
	Unadjusted (95% CI)	**Brenner (95% CI)**	**Staquet et al (95% CI)**
Sensitivity	0.65 (0.52, 0.75)	0.50(0.38, 0.62)	0.89 (0.79, 0.95)
Specificity	0.89 (0.84, 0.93)	0.85 (0.79, 0.89)	0.96 (0.92, 0.98)

CI is confidence interval

Following the results obtained (Table 13), firstly, the estimates from the Staquet et al approach are not illogical. Secondly, not correcting for the imperfection of the reference standard (that is using the classical method), underestimates the sensitivity and specificity of HRA cytology when compared to the estimates obtained via the Staquet et al approach. In addition, correcting for the imperfection of the reference standard using the Brenner correction method further underestimates the sensitivity and specificity of HRA cytology. Therefore, discouraging the use of HRA cytology to rule out the diagnosis of HSIL because the sensitivity of HRA cytology via the Brenner correction method is 0.5. Furthermore, correcting for the imperfection of the reference standard (biopsy) using the Staquet et al correction method, adjusts the values of the sensitivity and specificity of HRA cytology. Thus, having an excellent sensitivity of 0.9 and an excellent specificity that is close to one.

Reanalysis of the clinical dataset by Matos et al

The dataset considered the detection of occlusal caries lesions in primary teeth. The Matos et al [124] study used the Brenner[117] correction method to estimate the sensitivities and specificities of two fluorescence – based devices (Fluorescence camera (FC) and DIAGNOdent – a pen type laser fluorescence (LFpen)) used in detecting occlusal caries lesions, under the assumption that the sensitivity and specificity of the reference standard – International Caries Detection and Assessment System (ICDAS) criteria) –

are known. The fluorescence devices (FC and LFpen) were assumed to be conditionally independent of the reference standard. In this research study, the dataset was reanalysed using the classical method, the Brenner correction method, and the Staquet et al correction method.

The two target conditions are non-cavitated caries lesions (NC) and Dentine caries lesions (D3). D3 stands for Development Dental Defects. The total number of teeth considered for detecting noncavitated lesions (NC) were 383, of which 91.6% (351/383) had NC and 8.4% (32/383) were sound using visual inspection as the reference standard. The total number of teeth considered for D3 were 407, of which 94.8% (386/407) were sound and 5.2% (21/407) had dentine caries lesions using operative intervention as the reference standard. The teeth were treated independently.

According to Matos et al[124], the sensitivity and specificity of the reference standard were obtained from previous studies[294-298]. For NC detection, the sensitivity and specificity of visual inspection (reference standard) are 0.796 and 0.799 respectively. For the D3, the sensitivity and specificity of the reference standard are 0.786 and 0.995 respectively.

The 2 x 2 classification of the patients' teeth stratified by two different examiners and the two fluorescence devices (FC and LFpen) are reported in Table 14 and Table 16 (reproduced from the result tables in the Matos et al [124] study). I will reanalyse the dataset using the Matos et al[124] stated accuracy measures for the reference standards. Reanalysing the dataset in Table 14, the results of the unadjusted, Brenner corrected and the Staquet et al corrected sensitivity and specificity are reported in Table 15. The sample prevalence is 0.916.

Table 14: Results of the visual inspection (reference standard) and fluorescence - based devices (LFpen and FC) by two separate examiners

Visual inspection for NC (Examiner 1)		
Index test	**Positive (+)**	**Negative (-)**
LFpen positive	**241**	**6**
LFpen negative	**110**	**26**
Total	**351**	**32**

Visual inspection for NC (Examiner 1)		
Index test	**Positive (+)**	**Negative (-)**
FC positive	**156**	**3**
FC negative	**195**	**29**
Total	**351**	**32**

Visual inspection for NC (Examiner 2)		
Index test	**Positive (+)**	**Negative (-)**
LFpen positive	**237**	**6**
LFpen negative	**114**	**26**
Total	**351**	**32**

Visual inspection for NC (Examiner 2)		
Index test	**Positive (+)**	**Negative (-)**
FC positive	**158**	**4**
FC negative	**193**	**28**
Total	**351**	**32**

Table 15: Sensitivity and Specificity of LFpen and FC stratified by examiner 1 and 2 with the NC detection.

Non-cavitated caries lesion (NC)			
Methods – LFpen (Examiner 1)			
Accuracy measures	**Unadjusted (95% CI)**	**Brenner (95% CI)**	**Staquet et al (95% CI)**
Sensitivity	0.69 (0.64, 0.73)	0.68 (0.63, 0.73)	0.70 (0.65, 0.75)
Specificity	0.81 (0.65, 0.91)	0.44 (0.28, 0.61)	0.04 (0.01, 0.17)
Methods – LFpen (Examiner 2)			
Accuracy measures	**Unadjusted (95% CI)**	**Brenner (95% CI)**	**Staquet et al (95% CI)**
Sensitivity	0.68 (0.63, 0.73)	0.66 (0.61, 0.71)	0.69 (0.64, 0.74)
Specificity	0.81 (0.64, 0.91)	0.45 (0.29, 0.62)	0.06 (0.02, 0.20)
Methods – FC (Examiner 1)			
Accuracy measures	**Unadjusted (95% CI)**	**Brenner (95% CI)**	**Staquet et al (95% CI)**
Sensitivity	0.44 (0.39, 0.50)	0.44 (0.39, 0.49)	0.45 (0.40, 0.50)
Specificity	0.91 (0.76, 0.97)	0.65 (0.48, 0.79)	0.36 (0.22, 0.53)
Methods – FC (Examiner 1)			
Accuracy measures	**Unadjusted (95% CI)**	**Brenner (95% CI)**	**Staquet et al (95% CI)**
Sensitivity	0.45 (0.40, 0.50)	0.44 (0.39, 0.49)	0.46 (0.41, 0.51)
Specificity	0.88 (0.73, 0.95)	0.64 (0.47, 0.78)	0.37 (0.23, 0.54)
CI is confidence interval; LFpen: laser florescence pen; FC: fluorescence camera			

Following the estimated sensitivities and specificities of LFpen and FC (Table 15) in discriminating between participants with non-cavitated caries (NC); firstly, the results appear to be consistent across different examiners. Secondly, the sensitivity of LFpen and FC are consistent across all methods confirming the observation from the simulation that if the prevalence of the target condition is high, then the sensitivity is likely to estimated correctly by all methods. Furthermore, the Staquet at al corrected specificity is lower than the Brenner corrected and unadjusted specificities. Therefore, in clinical application, if the LFpen and the reference standard are conditionally independent (as is assumed), then the Staquet et al corrected sensitivity and specificity for the LFpen indicates that the LFpen has reasonable sensitivity (≅0.7) but very poor specificity (≅0.05) in diagnosing non-cavitated caries. In addition, if the FC device is conditionally independent of the reference standard (as is assumed), then the FC device has poor sensitivity and poor specificity in diagnosing non-cavitated caries and is unlikely to be recommended for use in practice for ruling in the diagnosis of NC based on the specificity' values.

Reanalysing the dataset in Table 16, the prevalence of the D3 is 0.052. Having a low prevalence, indicates that the specificity of LFpen and FC are likely to be correctly estimated by all methods. However, the sensitivity could be poorly estimated. The unadjusted, Brenner corrected and the Staquet et al corrected sensitivity and specificity are reported in Table 17.

Table 16: Results of the operative intervention (reference standard) and fluorescence - based devices (LF and FC) classified by examiners

Operative intervention for D3 (Examiner 1)		
Index test	Positive (+)	Negative (-)
LFpen positive	20	45
LFpen negative	1	341
Total	21	386

Operative intervention for D3 (Examiner 1)		
Index test	Positive (+)	Negative (-)
FC positive	21	38
FC negative	0	348
Total	21	386

Operative intervention for D3 (Examiner 2)		
Index test	Positive (+)	Negative (-)
LFpen positive	21	54
LFpen negative	0	332
Total	21	386

Operative intervention for D3 (Examiner 2)		
Index test	Positive (+)	Negative (-)
FC positive	19	46
FC negative	2	340
Total	21	386

Table 17: Sensitivity and Specificity of LFpen and FC stratified by examiner 1 and 2 with the dentine caries lesions detection.

Dentine caries lesion (D3)			
Methods – LFpen (Examiner 1)			
Accuracy measures	**Unadjusted (95% CI)**	**Brenner (95% CI)**	**Staquet et al (95% CI)**
Sensitivity	0.95 (0.77, 0.99)	0.86 (0.66, 0.95)	1.04 (NaN)
Specificity	0.88 (0.85, 0.91)	0.87 (0.83, 0.90)	0.90 (0.87, 0.93)
Methods – LFpen (Examiner 2)			
Accuracy measures	**Unadjusted (95% CI)**	**Brenner (95% CI)**	**Staquet et al (95% CI)**
Sensitivity	1 (0.85, 1)	0.91 (0.72, 0.98)	1.09 (NaN)
Specificity	0.86 (0.82, 0.89)	0.85 (0.81, 0.88)	0.87 (0.83, 0.90)
Dentine caries lesion (D3)			
Methods – FC (Examiner 1)			
Accuracy measures	**Unadjusted (95% CI)**	**Brenner (95% CI)**	**Staquet et al (95% CI)**
Sensitivity	1.00 (0.85, 1.00)	0.91 (0.72, 0.98)	1.09 (NaN)
Specificity	0.90 (0.87, 0.93)	0.89 (0.86, 0.92)	0.92 (0.89, 0.94)
Methods – FC (Examiner 2)			
Accuracy measures	**Unadjusted (95% CI)**	**Brenner (95% CI)**	**Staquet et al (95% CI)**
Sensitivity	0.91 (0.72, 0.98)	0.82 (0.61, 0.93)	0.99 (0.83, 1.00)
Specificity	0.88 (0.84, 0.91)	0.87 (0.83, 0.90)	0.89 (0.85, 0.92)
CI is confidence interval; LFpen: laser florescence pen; FC: fluorescence camera; NaN is not available or cannot be estimated			

From the estimates in Table 17, the results are consistent across all examiners and the specificities of LFpen and FC are consistent across all methods, as observed in the simulation, that, at low prevalence, the specificity of the test is likely to be estimated correctly by all methods. Thus, the LFpen and FC have an excellent ability to rule in (specificity $\cong 0.9$ to one decimal place) the diagnosis of dentine caries lesion (D3). However, the sensitivities of the index tests are inconsistent across all the methods, with Staquet et al having an illogical sensitivity (> 1).

Following the simulation study, the classical method outperforms the Brenner correction methods and the Staquet et al method outperforms both the classical and Brenner corrected methods. However, at very high or low prevalence, the Staquet et al method could provide illogical results when both tests are conditionally independent despite the large sample sizes available in the study. This is what has occurred in the analysis of this clinical dataset. Hence, to overcome the challenges of illogical results, considering another statistical method – a latent class model – could be considered to estimate the sensitivities and specificities of the LFpen and FC in detecting D3. Latent class models are discussed in detail in chapter four, and the reanalysis of this clinical dataset using a latent class model is revisited in this chapter (section **4.8**).

Assessing the clinical datasets for the possibility of obtaining illogical estimates

The Mathews et al[125] dataset was assessed to ascertain if illogical estimates could be obtained via the Staquet et al[119] approach, the statistics below were estimated:

$$rPPV = 0.645; \quad rNPV = 0.889; \quad rPPV' = 0.355; \quad rNPV' = 0.111; \quad Prr = 0.23$$

The sensitivity of the RS (0.74) is greater than the sample prevalence (0.23), hence, obtaining illogical prevalence is unlikely. In addition, the sensitivity of RS is greater than the rPPV (0.645), the $rNPV'$ (0.111) and the sample prevalence (0.23); therefore, obtaining an illogical sensitivity via Staquet et al[119] approach is unlikely. The specificity of the RS (0.91) is greater than the rNPV (0.889), $rPPV'$ (0.355) and Prr' (0.77). Thus, an illogical specificity estimate will not be obtained using the Staquet et al approach. In summary, none of the conditions for obtaining illogical estimates were fulfilled in this dataset.

The first clinical dataset from Matos et al[124] was assessed for the possibility of obtaining illogical estimates and the following statistics were calculated:

NC, LFpen Examiner 1

$$rPPV = 0.975\,; \quad rPPV' = 0.025; \quad rNPV' = 0.809; \quad rNPV = 0.191; \quad Prr = 0.916$$

NC, FC Examiner 1

$$rPPV = 0.981\,; \quad rPPV' = 0.019; \quad rNPV' = 0.871; \quad rNPV = 0.129; \quad Prr = 0.916$$

The sensitivity of the RS (0.796) is less than the sample prevalence (0.92), hence, there is a likelihood of obtaining illogical estimated prevalence. The estimated prevalence is 1.2 (which is illogical). The specificity of visual inspection (0.799) is greater than the rNPV (0.191 or 0.129). It is also greater than the complement of the rPPV ($rPPV' = 0.025$) and the complement of the sample prevalence ($Prr' = 1 - Prr = 0.004$). Thus, obtaining illogical sensitivity for the index tests (LFpen and FC) are unlikely. The sensitivity of visual inspection (0.796) is less than the rPPV (0.975 or 0.981) indicating the likelihood of obtaining illogical estimates for the specificities of FC and LFpen whose absolute value is greater than one. The sensitivity of the RS is also less than the sample prevalence (0.916) and less than the complement of the relative NPV (0.871 for FC, and 0.809 for LFpen). In summary, condition (1) and condition (3a) were fulfilled in this dataset. Illogical estimated prevalence was obtained but the estimated specificities are logical (that is within [0, 1]). Therefore, this result needs to be treated with scepticism.

The second clinical dataset from Matos et al[124] was assessed to ascertain of obtaining illogical results and the following statistics are calculated.

D3, LFpen Examiner 1

$$rPPV = 0.308; \quad rPPV' = 0.692; \quad rNPV' = 0.003; \quad rNPV = 0.997; \quad Prr = 0.052$$

D3, FC Examiner 1

$$rPPV = 0.356; \quad rPPV' = 0.644; \quad rNPV' = 0; \quad rNPV = 1; \quad Prr = 0.052$$

The sensitivity of the RS (0.786) is greater than the sample prevalence (0.052). Hence, obtaining illogical estimated prevalence is unlikely. The sensitivity of the RS (0.786) is also greater than the rPPV (0.31 or 0.36), and the complement of the rNPV (0). Therefore, the likelihood of obtaining illogical specificities for LFpen and FC are unlikely. The specificity of visual inspection (0.995) is less than the rNPV (1 for FC and 0.997 for LFpen). It is also greater than the complement of the rPPV ($rPPV' = 0.025$) and the complement of the prevalence ($Prr' = 0.95$). Thus, obtaining illogical

sensitivity estimates for the index tests (LFpen and FC) is likely as the condition (2) is met. In summary, illogical estimated sensitivity was obtained for the index tests (1.04 and 1.09).

3.5. Summary

In this chapter, I used simulation studies to compare methods employed to correct for an imperfect reference standard to estimate the sensitivity and specificity of a binary response index test given that the accuracy measures (sensitivity and specificity) of the imperfect reference standard are known. The methods compared were the Brenner and Staquet et al correction methods. Both methods assume that the index test and the reference standard are conditionally independent given the true disease status.

Different scenarios under the assumption of conditional dependence and conditional independence were explored to understand the statistical properties (bias, consistency and MSE) of the classical and two correction methods under investigation, and how they perform to estimate the accuracy measures of the index test. For example, under the assumption of conditional independence, I looked at the statistical properties of the correction and classical methods when the sensitivity and specificity of the reference standard are known and better than the index test, worse than the index test and the same as the index test. Coverage probability[299] (which is the proportion of confidence intervals calculated in a particular way that contains the true value of the parameter) was not explored as it is a property of the confidence interval procedure which is not within the scope of my research.

The results from the simulation studies indicate that at low prevalences (0 – 0.5) the sensitivity is underestimated using imperfect reference standards, and at high prevalences (0.5 – 1) the specificity of the index test is underestimated. The findings support the findings from other research that also suggest the use of a population with a high prevalence of disease to estimate the sensitivity and a low prevalence of diseases for the specificity[118, 128] if this is possible.

Adjusting for the error in the imperfect reference standard is essential to correctly estimate the accuracy measures of the index test. Exploring the correction methods, the Staquet et al correction method outperforms the Brenner correction method provided the index test and reference standard are conditionally independent irrespective of whether the reference standard is better than the index test or not. However, at very low or high prevalence there is the likelihood of having illogical results

(that is estimates of sensitivity or specificity that are greater than 1). The Staquet et al method was further explored to understand the conditions for illogical estimates.

The knowledge gained from the simulation study was applied to reanalyse clinical datasets by Mathews et al[125] and Matos et al [124].

Mathews et al[125], employed the Staquet et al correction method to correct for the imperfection of the reference standard (Biopsy) in order to estimate the sensitivity and specificity of the index test (HRA cytology). I reanalysed the dataset using the Brenner correction method in addition to the Staquet et al method. I found that the Brenner correction method underestimated the sensitivity and specificity of HRA cytology in comparison to the Staquet et al method. Therefore, if the Brenner estimates were employed to make a clinical decision on HRA cytology, HRA cytology would be considered a poor test in ruling out the diagnosis of HSIL as it had a sensitivity of 0.5 and a very good test in ruling in the diagnosis of HSIL as it had a specificity of 0.85. Using the Staquet et al approach to correct for the imperfection of the reference standard, the sensitivity and specificity of HRA cytology were 0.89 and 0.96 respectively, making it an excellent test to rule in and rule out the diagnosis of HSIL among HIV patients.

Matos et al [124] used the Brenner correction method to adjust for the imperfection in the reference standard. Applying the Staquet et al correction method (which has shown to outperform Brenner correction method), showed that the specificities of LFpen and FC in detecting non-cavitated caries (NC) were significantly lower than those obtained when using the Brenner correction method. Hence, if the estimates from the Staquet et al correction method were used then FC and LFpen may not reach the threshold for changing clinical decision making and therefore would be less likely to be recommended in practice to rule in the diagnosis of NC.

When the index test and the reference standard are conditionally dependent given the true disease status, using the Staquet et al or Brenner correction methods is not recommended, even if the diagnostic accuracy of the reference standard is known. Using a method that can account or correct for this dependence is essential to correctly estimate the sensitivity and specificity of the index test is recommended. Bayesian latent class models could be considered for this (these methods are investigated in chapter four) because the knowledge of the accuracy measures of the reference standard can be employed as prior information to make the latent class model

identifiable and therefore better able to accurately estimate the sensitivity and specificity of the index test.

Chapter Four: Latent Class Models

4.1. Introduction

Latent class models (LCMs) have been suggested by different researchers[128, 139, 291, 300-302] to evaluate the diagnostic accuracy of multiple tests simultaneously in the absence of a gold standard. With the LCM, none of the tests under investigation are used as a benchmark to determine the presence or absence of disease. Therefore, the true disease status of all the participants is unobserved (latent). The prevalence of the disease is estimated alongside other sensitivity and specificity of interest such as the sensitivities and specificities of the tests. There are different types of latent class models that have been proposed and have been employed in diagnostic accuracy studies. These range from the traditional frequentist, Hui and Walter[128] latent class model to more complex latent class models which consider the conditional dependence among the tests being evaluated.

This chapter discusses the different LCMs employed in diagnostic accuracy studies with specific focus on the possible choices of LCMs to consider for the analysis of the clinical dataset explored in chapter five. Briefly, the clinical dataset consists of test responses from three scores that are conditionally dependent given the true disease status. The clinical dataset is discussed comprehensively in chapter five.

The aim of the investigation in this current chapter is to explore how the different LCMs perform under varying assumptions (for example, conditional independence and dependence assumptions) and to inform which LCMs will be most appropriate to analyse the clinical dataset explored in chapter five. This chapter outlines various LCMs and simulation studies carried out to explore these models. A generated dataset from the simulation reflects the characteristics of the clinical dataset; thus, providing a guide of what to expect when analysing the clinical dataset. Finally, this chapter discusses the performance of the latent class models explored within the different simulated scenarios.

The notation used throughout this chapter is described in the next section (section **4.2**).

4.2. Basic notation

The notations stated here are employed in this chapter. Some are similar to those used in chapter three and some have changed to provide more details relevant for this chapter.

Let D be the true latent disease status of a participants with two classes – diseased and non-diseased. Let J be the number of tests that are evaluated and N ($i = 1, 2, 3, \dots, N$) be the total number of participants in the study. Each test has a binary outcome (0 for participants classified as non-diseased by a test and 1 for participants classified as diseased by a test). A test response can be continuous or have more than two categories. However, in this research, all test responses are dichotomised to have a binary response. Thus, for a test with a continuous response, cut-offs will be employed to dichotomise the test responses into two classes. Let t_{ij} $(0\ or\ 1)$ be the test response of the j[th] test for the i[th] participant. Let $p_1 = Pr\ (D = 1)$ denote the prevalence of the diseased in the population and $p_0 = Pr\ (D = 0)$ denote the probability of no disease in the population.

$$p_d = \Pr(D = d)\,; \qquad p_0 = 1 - p_1; \qquad d = 0, 1$$

Furthermore, it is expected that the i[th] participant will undergo all the tests employed in the diagnostic accuracy study, and the set of test responses from each individual is independent. Given the J tests, there will be 2^J combinations of the test responses. So, let K be the total number of combinations of the test responses and n_k $(k = 1, 2, \dots K)$ be the number of participants in each of the k tests' combinations. For example, if there are three index tests, each with a dichotomised response, then there will be eight (K = 8) possible combinations of the test responses and n_k is the number of participants in each combination, or the cell frequency of each combination.

A randomly selected participant is considered, hence the subscript i is dropped. Let $T_j = t_j$ be the binary result of the j[th] test and the vector $\boldsymbol{T}$ denote the combination of all the test results. For example, when J = 3, the test results are $T_1 = t_1, T_2 = t_2, T_3 = t_3$, where $t_j \in \{0, 1\}$ and the vector $\boldsymbol{T} = (t_1, t_2, t_3)$. Let the probability of the test combination be denoted as $\Pr(\boldsymbol{T})$). Let Sn denote the sensitivity of a test and Sp denote the specificity of a test. Then the sensitivity and specificity of the j[th] test is presented as:

$$Sn_j = \Pr\big(T_j = 1 \,\big|D = 1\big)\,; \qquad Sp_j = \Pr\big(T_j = 0 \big| D = 0\big)\,; \quad j = 1, 2, \dots, J$$

4.3. Latent class model

LCMs are a statistical technique which uses observed information (responses) from the participants in a study to group the participants into unobserved (latent) classes[303]. The observed information takes the form of categorical variables and the unobserved

variable is also a categorical variable[304]. For example, in a diagnostic accuracy study, given that there are multiple tests to evaluate, the test responses obtained from the participants are the observed information and the true disease status of each participant (diseased or non-diseased) is the latent variable.

Employing latent class modelling in diagnostic accuracy studies originated with the Hui and Walter traditional latent class model (TLCM)[128]. The TLCM was proposed with the simplifying assumption that the tests evaluated in the diagnostic accuracy study are conditionally independent given the true disease status of the participants. This assumption is violated when the tests under evaluation are conditionally dependent given the true disease status. Since the development of the TLCM, LCMs with different conditional dependence structures have been proposed. The applications of these LCMs range from frequentist approaches (where only the observed data is employed to estimate the parameters of interest) to Bayesian approaches (where existing data or expert opinion is combined with the observed data to estimate the parameters of interest). Employing the TLCM[128, 305] to estimate the sensitivity and specificity of tests that are conditionally dependent will result in biased estimates. To eliminate the bias that arises as a result of assuming conditional independence, LCMs which capture the conditional dependence of the tests can be employed. These LCMs include the fixed effect latent class model (FEM)[54, 286, 291], the random effect latent class model (REM)[132], the finite mixture latent class model (FMM)[32, 120, 133], the Beta-Binomial latent class model (BBM)[120], and the grade of membership latent class model[306] amongst others. This chapter will focus on describing the key concepts behind the Hui and Walter LCM, the FEM, REM and FMM, and simulation studies are employed to explore these models. The choice of methods discussed is motivated by the clinical dataset which will be the focus of further data analysis in chapter five.

4.3.1. Traditional latent class model

A basic assumption underlying the TLCM is that the tests under evaluation are conditionally independent. In a one-prevalence study, a minimum of three tests need to be employed to make the model identifiable[140], because the degrees of freedom $(2^J - 1)$ is equal to or greater than the number of parameters to estimate $(2 \times J + 1)$. A model is said to be identifiable if there exists a unique solution for every unknown parameter in the model[307].

The TLCM uses the multinomial distribution to model the relationship between the parameters of interest and the observed data. Thus, using the notations introduced in section **4.2**, the likelihood of the observed data is defined as:

$$L(n_1, n_2, \dots, n_k) \sim Multinomial\ (\boldsymbol{Pr}(\boldsymbol{T})\,, N) \tag{12}$$

Where N is the total number of participants in the study and $\boldsymbol{n} = (n_1, n_2, \dots, n_k)$ is the number of participants in each k[th] test combination. Both N and n are observed data, while $\boldsymbol{Pr}(\boldsymbol{T})$ is a vector of the joint probability of the test responses ($\Pr(\boldsymbol{T})$) of each k[th] test combination.

$$\Pr(\boldsymbol{T}) \;\; = \sum_{d=0}^{1} \Pr(\boldsymbol{T}|D) \times \Pr(D = d)$$

$$= \sum_{d=0}^{1} \Pr\big(T_1 = t_1, T_2 = t_2, \dots, T_J = t_j \big| D = d\big) \times \Pr(D = d)$$

$$= \sum_{d=0}^{1} p_d \times \Pr(\boldsymbol{T}|D = d)$$

$$= p_1\{\Pr(\boldsymbol{T}|D = 1)\} + p_0\{\Pr(\boldsymbol{T}|D = 0)\} \tag{13}$$

The probability of the test responses within the diseased group is denoted by $\Pr(\boldsymbol{T}|D = 1)$ and $\Pr(\boldsymbol{T}|D = 0)$ is the probability of the test responses within the non-diseased group. The probabilities of the test responses within the diseased group and non-diseased group can be expressed in terms of the sensitivities and the specificities of the index tests[199]:

$$\Pr\big(\boldsymbol{T_j}|D = 1\big) = \prod_{j=1}^{J} Sn_j^{t_j}\ \big(1 - Sn_j\big)^{(1-t_j)};\ \ \Pr\big(\boldsymbol{T_j}|D = 0\big) = \prod_{j=1}^{J} Sp_j^{(1-\,t_j)}\ \big(1 - Sp_j\big)^{t_j} \tag{14}$$

Equations (13) and (14) are used to estimate the probability of each test result combination ($\Pr(\boldsymbol{T})$) which are substituted into the likelihood in Equation (12) to estimate the parameters of interest.

For example, let's assume that there are three tests (t_1, t_2, t_3) to evaluate in a study, none of the three tests is a gold standard and all three tests are conditionally independent. I wish to estimate the sensitivities and specificities of the three tests using the TLCM approach.

The probability of the test responses is:

$$\Pr(\boldsymbol{T}) = p_1\{\Pr(\boldsymbol{T}|D=1)\} + p_0\{\Pr(\boldsymbol{T}|D=0)\}$$

where

$$\Pr(\boldsymbol{T}|D=1) = Sn_1^{t_1} Sn_2^{t_2} Sn_3^{t_3} (1-Sn_1)^{1-t_1} (1-Sn_2)^{1-t_2} (1-Sn_3)^{1-t_3}$$

$$\Pr(\boldsymbol{T}|D=0) = Sp_1^{1-t_1} Sp_2^{1-t_2} Sp_3^{1-t_3} (1-Sp_2)^{t_2} (1-Sp_3)^{t_3} (1-Sp_1)^{t_1}$$

4.3.2. Fixed effect latent class model

The fixed effect latent class model (FEM) assumes that the evaluated tests could be conditionally dependent given the true disease status of the participants. Hence, the FEM models the conditional dependence among the tests using covariance terms within the diseased and non-diseased groups, and these covariance terms are fixed across all participants. Covariance measures the joint variability of two variables[308]. As explained in section **3.3.2**, the pairwise covariance terms between two tests can be expressed using the correlations between the two tests and the sensitivities or specificities of the two tests. The FEM uses the multinomial distribution, like the TLCM to model the relationship between the observed data and the parameters of interest (see Equation (12)). However, the joint probability of the test responses ($\Pr(\boldsymbol{T})$) is redefined to include the covariance term(s). Hence, the FEM is adapted from the TLCM with additional terms, the covariance terms, employed to model the conditional dependence among multiple tests, including pairwise conditional dependence between two tests and higher order conditional dependence among multiple tests.

Let the covariance term among the diseased group be $\varphi_{\boldsymbol{t}|1}$ and the non-diseased group be $\varphi_{\boldsymbol{t}|0}$. The joint probability of the test responses is described as[199, 274, 309]:

$$\begin{aligned}\Pr(\boldsymbol{T}=\boldsymbol{t}) &= \sum_{d=0}^{1} p_d \times \Pr(\boldsymbol{T}|D=d) + \varphi_{\boldsymbol{t}|d}; \quad d = 0,1 \\ &= p_1\{\Pr(\boldsymbol{T}|D=1) + \varphi_{\boldsymbol{t}|1}\} + p_0\{\Pr(\boldsymbol{T}|D=0) + \varphi_{\boldsymbol{t}|0}\} \qquad (15)\end{aligned}$$

The probability of the test responses within the diseased group is denoted as $\Pr(\boldsymbol{T}|D=1) + \varphi_{\boldsymbol{t}|1}$ and $\Pr(\boldsymbol{T}|D=0) + \varphi_{\boldsymbol{t}|0}$ is the probability of the test responses within the non-diseased group. The FEM often uses pairwise covariance terms because of the challenges in modelling higher order correlations. However, with the approach developed by Wang et al[286], third-order correlation can now be modelled using the

FEM. A third-order correlation infers correlation among three tests. For example, let us assume that there are three tests under evaluation (t_1, t_2, t_3), and all three tests are correlated. That is, all three tests have some degree of pairwise correlation between them. So, test 1 and test 2 are correlated, test 2 and test 3 are correlated, and test 1 and test 3 are correlated. The correlation that exist among the three tests is called a third-order correlation.

With the introduction of the covariance terms, the total number of parameters to estimate is now $2^{J+1} + 2J + 1$. This includes J sensitivities, J specificities, 1 prevalence (for a single population study), and 2^{J+1} covariance terms (both for the diseased and non-diseased groups); while the degrees of freedom is $2^J - 1$. So, this makes the FEM non-identifiable because the number of parameters to estimate is larger than the degrees of freedom[140, 307].

In order to overcome this restriction, equality constraints[54, 286] are imposed on the covariance terms in a FEM (there are $J + 1$ constraints). This reduces the number of covariance terms to be estimated to $2^{J+1} - (2J + 2)$. Nevertheless, on its own, this does not make the model identifiable. Hence, prior information (deterministic or probabilistic)[310] is used to model some of the parameters in the FEM. In addition to the equality constraints, the FEM also uses some inequality constraints [54, 291] on the covariance terms which allows the covariance terms to be bound by the marginal probabilities of the tests, so that the joint probabilities of the tests do not exceed one and that negative probabilities cannot be estimated.

Let us assume that there are two tests (T_1 and T_2) to evaluate and each test has a binary outcome ($t_j = 0$ or 1). There are seven parameters to estimate under the conditional dependence assumption. These are two sensitivities, two specificities of the two tests, the prevalence and two covariance terms (one for the diseased group – $covs_{12}$ and the other is the non-diseased group – $covc_{12}$). There are four possible combinations of the test responses which are:

- An all-positive response (11)
- Two one-positive responses (10, 01)
- An all-negative response (00)

The number of participants in each test combination is represented as $n_k,\ k = 1, 2, 3, 4$.

The likelihood of each combination is modelled using the multinomial distribution.

$$L(n_1, n_2, \ldots, n_4) \sim Multinomial\ (\Pr(\boldsymbol{T}), N)$$

The joint probability of each test responses is:

$$\Pr(T_1 = 1, T_2 = 1) = p_1(Sn_1 \times Sn_2 + covs_{12}) + p_0((1 - Sp_1) \times (1 - Sp_2) + covc_{12})$$

$$\Pr(T_1 = 0, T_2 = 1) = p_1\big((1 - Sn_1) \times Sn_2 - covs_{12}\big) + p_0(Sp_1 \times (1 - Sp_2) - covc_{12})$$

$$\Pr(T_1 = 1, T_2 = 0) = p_1(Sn_1 \times (1 - Sn_2) - covs_{12}) + p_0((1 - Sp_1) \times Sp_2 - covc_{12})$$

$$\Pr(T_1 = 0, T_2 = 0) = p_1((1 - Sn_1) \times (1 - Sn_2) + covs_{12}) + p_0(Sp_1 \times Sp_2 + covc_{12})$$

Since, there are only 3 ($2^J - 1$) degrees of freedom, the model is non-identifiable. Hence, some constraints (either deterministic or probabilistic) need to be placed on some of the parameters to make the model identifiable so that the model provides a unique solution.

Inequality constraints are used to provide the range of values (i.e. the upper and lower bound values) that the covariance terms can take, and this is determined by the sensitivities (specificities) of the tests. The covariance terms can be positive or negative and the model can have dependence among the diseased group only, among the non-diseased group only, or in both groups. The inequality constraints are[54]:

$$-Sn_1 \times Sn_2 + \max(0, Sn_1 + Sn_2 - 1) \leq \varphi_{1,1|1} \leq \min(Sn_1, Sn_2) - (Sn_1 \times Sn_2)$$

$$-Sp_1 \times Sp_2 + \max(0, Sp_1 + Sp_2 - 1) \leq \varphi_{0,0|0} \leq \min(Sp_1, Sp_2) - (Sp_1 \times Sp_2)$$

Now let us suppose there are three tests to evaluate ($\boldsymbol{T} = T_1, T_2, T_3$), and each test has a binary outcome ($t_j = 0$ or 1). This implies that there are 8 (2^3) possible combinations of the test responses, which are:

- An all-positive response (111)
- Three two-positive responses (110, 101, 011)
- Three one-positive responses (001, 010, 100)
- An all-negative response (000)

Firstly, let us assume that two (Test 1 and Test 2) out of the three tests are conditionally dependent among the diseased group and both tests are conditionally independent of the third test (Test 3).

Thus, using Equation (15), the probabilities of the test responses are[54]:

$$\Pr(\boldsymbol{T}) = p_1\{\Pr(\boldsymbol{T}|D = 1) + \varphi_{t|1}\} + p_0\{\Pr(\boldsymbol{T}|D = 0)\}$$

$$\Pr(\boldsymbol{T}|D=1)+\varphi_{\boldsymbol{t}|1}=\left(Sn_1^{t_1}Sn_2^{t_2}(1-Sn_1)^{1-t_1}(1-Sn_2)^{1-t_2}+(-1)^{(t_1-t_2)}\varphi_{12|1}\right)\times Sn_3^{t_3}(1-Sn_3)^{1-t_3}$$

$$\Pr(\boldsymbol{T}|D=0)=Sp_1^{1-t_1}Sp_2^{1-t_2}Sp_3^{1-t_3}(1-Sp_1)^{t_1}(1-Sp_2)^{t_2}(1-Sp_3)^{t_3}$$

Secondly, let us assume that the three tests are correlated among the diseased and non-disease group. Hence, the joint probability of the test responses taking into consideration this third order correlation is described by Wang et al[286]:

$$\Pr(\boldsymbol{T}=111)=p_1(Sn_1\times Sn_2\times Sn_3+covs_{111})+p_0((1-Sp_1)\times(1-Sp_2)\times(1-Sp_3)+covc_{111})$$

$$\Pr(\boldsymbol{T}=110)=p_1(Sn_1\times Sn_2\times(1-Sn_3)+covs_{110})+p_0((1-Sp_1)\times(1-Sp_2)\times Sp_3+covc_{110})$$

$$\Pr(\boldsymbol{T}=101)=p_1(Sn_1\times(1-Sn_2)\times Sn_3+covs_{101})+p_0((1-Sp_1)\times Sp_2\times(1-Sp_3)+covc_{101})$$

$$\Pr(\boldsymbol{T}=100)=p_1(Sn_1\times(1-Sn_2)\times(1-Sn_3)+covs_{100})+p_0((1-Sp_1)\times Sp_2\times Sp_3+covc_{100})$$

$$\Pr(\boldsymbol{T}=011)=p_1((1-Sn_1)\times Sn_2\times Sn_3+covs_{011})+p_0(Sp_1\times(1-Sp_2)\times(1-Sp_3)+covc_{011})$$

$$\Pr(\boldsymbol{T}=010)=p_1((1-Sn_1)\times Sn_2\times(1-Sn_3)+covs_{010})+p_0(Sp_1\times(1-Sp_2)\times Sp_3+covc_{010})$$

$$\Pr(\boldsymbol{T}=001)=p_1((1-Sn_1)\times(1-Sn_2)\times Sn_3+covs_{001})+p_0(Sp_1\times Sp_2\times(1-Sp_3)+covc_{001})$$

$$\Pr(\boldsymbol{T}=000)=p_1((1-Sn_1)\times(1-Sn_2)\times(1-Sn_3)+covs_{000})+p_0(Sp_1\times Sp_2\times Sp_3+covc_{000})$$

The FEM model by Wang et al[286] allows researchers to study the pairwise and higher order conditional dependences that can exist between or among the tests evaluated.

Equality constraints on the covariance terms reduces the number of covariance terms to estimate[286]. Hence, instead of estimating 16 covariance terms, 8 covariance terms will be estimated (four for the diseased group and four for the non-diseased group) and the remaining 8 covariance terms can be estimated using the 8 estimated covariance terms through the set of equations expressed below[286].

So, for the diseased group the covariance terms are:

$$covs_{111}, covs_{000}, covs_{001}, covs_{011}, covs_{100}, covs_{101}, covs_{110}\ and\ covs_{010}$$

$covs_{111},\ covs_{000},\ covs_{001}\ and\ covs_{011}$ are used to calculate the remaining covariance terms

$$covs_{100}=covs_{001}+\ covs_{111}-covs_{000}$$

$$covs_{101}=-(covs_{001}+\ covs_{011}+covs_{111})$$

$$covs_{110}=covs_{001}-\ covs_{111}+covs_{000}$$

$$covs_{010}=-(covs_{001}+\ covs_{011}+covs_{000})$$

For the non-diseased group, the covariance terms are:

$$covc_{111}, covc_{000}, covc_{001}, covc_{011}, covc_{100}, covc_{101}, covc_{110}\ and\ covc_{010}$$

$covc_{111},\ covc_{000},\ covc_{001}\ and\ covc_{011}$ are used to calculate the remaining covariance terms

$$covc_{100} = covc_{001} + \ covc_{111} - covc_{000}$$

$$covc_{101} = -(covc_{001} + \ covc_{011} + covc_{111})$$

$$covc_{110} = covc_{001} - \ covc_{111} + covc_{000}$$

$$covc_{010} = -(covc_{001} + \ covc_{011} + covc_{000})$$

The pairwise conditional dependence terms between any two of the three tests are as follows:

For the disease group:

$$covst_{12} = covs_{111} + covs_{110}$$

$$covst_{13} = covs_{111} + covs_{101}$$

$$covst_{23} = covs_{111} + covs_{011}$$

The pairwise covariance between test 1 and test 2 is $covst_{12}$, the covariance term between test 2 and test 3 is $covst_{23}$, and the pairwise covariance term between test 1 and test 3 is $covst_{13}$.

For the non-diseased group:

$$covct_{12} = covc_{000} + covc_{001}$$

$$covct_{13} = covc_{000} + covc_{010}$$

$$covct_{23} = covc_{000} + covc_{100}$$

The pairwise covariance between test 1 and test 2 is $covct_{12}$, the covariance term between test 2 and test 3 is $covct_{23}$, and the pairwise covariance term between test 1 and test 3 is $covct_{13}$.

There are twenty-three parameters to estimate for a diagnostic accuracy study evaluating three tests, if the three tests are assumed to be conditionally dependent both among the diseased and non-diseased groups. This number is greater than the

degrees of freedom which is seven. The equality constraints reduce the number of conditional dependence parameters to estimate to eight (four for each disease group). So, the total parameters to estimate is now fifteen. However, this is still greater than the degrees of freedom (seven); therefore, a minimum of eight (2^J) informative priors are needed to make the model identifiable.

4.3.3. Random effect latent class model

The REM assumes that the causes of the correlations between the tests are unobserved and they are subject-specific[132, 199] (i.e. it could differ for each participant in the study), unlike the FEM which assumes that the conditional dependencies between the tests are fixed across all participants in the study. Hence, the REM does not use covariance terms like FEM; rather the REM models the conditional dependence of the tests using a continuous latent variable. Let that continuous latent variable be defined as $\boldsymbol{Z}$ ($\boldsymbol{Z} = Z_d; d = 0, 1$). The REM assumes that the test response for each participant depends not only on the unobserved disease status (because of the absence of a gold standard) but also on the unobserved continuous variable[311, 312]. The continuous latent variable is assumed to follow the multivariate standard normal distribution ($\boldsymbol{Z_d} \sim N(0,1)$). The number of random variables in $\boldsymbol{Z}$ depends on the number of disease classes. For instance, if there are two latent disease classes (diseased and non-diseased), then there are two continuous random variables in $\boldsymbol{Z}$, one for each disease class ($\boldsymbol{Z} = (Z_{d=1}, Z_{d=0})$). Moreover, the disease status and the continuous latent variable are assumed to be independent.

The test response of the i[th] participant on the j[th] test is denoted as $t_{ij} = 0 \; or \; 1$**.** Zero indicates a negative test response and one indicates a positive test response. D is the true disease status of a participant, which is latent. With the REM, the latent disease status of each participant (D_i) is modelled using a Bernoulli distribution with the probability equal to the prevalence of the disease (p_1) in the population and the test response (t_{ij}) of a participant is modelled using the Bernoulli distribution with a probability related to each participant (p_{ij}). That is:

$$D_i \sim Bernoulli\ (p_1);$$

where D_i is the disease status of the i[th] individual, and the test response of each participant is:

$$t_{ij} \sim Bernoulli\ (p_{ij})$$

A random variable (say X) follows a Bernoulli distribution if X takes only two values (0 and 1) such that the probability of X = 1 is denoted as p and the probability of X = 0 is denoted as 1 - p.

The probability that a test response is positive given the latent disease status (D) and the latent continuous random variable (Z) is modelled using a regression equation[132, 199]:

$$p_{ij} = \Pr\left(t_{ij} = 1 \middle| D_i = d_i, \boldsymbol{Z} = \boldsymbol{z}\right) = \eta^{-1}\left(a_{jd} + \boldsymbol{b}_{jd} z\right) \quad (16)$$

where η is a link function, a_{jd} is the intercept and $\boldsymbol{b_{jd}}$ is the coefficients vector

The link functions often employed in a diagnostic accuracy study are the probit link function[132, 313] or the logit link function[312]. The coefficient vector models the dependence in the tests induced by the random effects among the diseased or non-diseased groups; however, this does not translate to the correlation between the tests in the same way as the covariance terms in the FEM. The coefficient vector is also called the variance parameter[132] for a probit link model.

In addition to the prevalence, sensitivity and specificity of the tests, other parameters to be estimated using the REM are the coefficients of the random effects (b_{jd}), which are called the variance parameters[132] and the intercept parameter, a_{jd}. Constraints can be placed on the variance parameters to reduce the number of parameters to estimate and make the REM identifiable.

Firstly, the variance parameter can be set to zero (i.e. $b_{jd} = 0;\ for\ d = 0, 1$). In this scenario, the tests under evaluation are assumed to be conditionally independent given the disease status (D) and the latent variable (Z). Hence, the probability of having a positive test response given disease status is:

$$P\left(t_{ij} = 1 \middle| D\right) = \eta^{-1}(a_{jd})$$

Alternatively, one can assume that the variance parameters are the same for all the tests within the diseased and non-diseased groups ($b_{jd} = b_d \neq 0;\ for\ d = 0, 1$). In this case, the coefficient of the latent variable Z is a constant (b_d). That is:

$$P\left(t_{ij} = 1 \middle| D, \boldsymbol{Z}\right) = \eta^{-1}(a_{jd} + b_d z)$$

Finally, the variance parameter (b_{jd}) can be employed to model the dependence (direct effect[132]) of two correlated tests. For example, if there are three tests under evaluation,

test 1 and test 2 can be correlated and both tests (test 1 and test 2) can be independent of test 3; thus, the variance parameter is used to model the dependence between test 1 and test 2 only.

The estimated sensitivities and specificities of the tests are the average of the estimated sensitivities and specificities from each participant in the study. This is because the REM assumes that the test responses are conditional on the disease status (D) and the latent variables (Z) are specific to each participant. So, the probability of the test responses given the disease status and latent continuous variable is calculated for each participant in the study. Hence, taking the average of the estimated sensitivities and specificities is required.

The random effect latent class model is more computationally burdensome than the fixed effect latent class model because the conditional dependence of the tests evaluated is based on each participant, whereas for the FEM it is evaluated collectively on participants with same test response combination.

4.3.4. Finite mixture latent class model

The FMM models assumes that the conditional dependence is unobserved, like the REM, and heterogeneous among the participants in the study[120]. However, it does not model this latent variable with the Gaussian distribution like the REM[75, 314, 315]. The FMM models this dependence by using mixtures of distributions that are asymmetrical. The FMM assumes that there will be participants who will be correctly classed as diseased or non-diseased by the tests and some will be misclassified. Hence, the FMM tries to take this into consideration when modelling the conditional dependence of the tests[306, 314] by using an indicator variable, l_{id}, to label participants as correctly or wrongly diagnosed.

Therefore, the probability of having a positive test response given the true disease status and the latent variable is described below[133]:

$$\Pr\left(t_{ij}=1 \middle| D=d, L=l_{id_i}\right)=\begin{cases} l_{i1}+(1-l_{i1})w_j(1) & if\ d_i=1 \\ 1-\left(l_{i0}+(1-l_{i0})\left[1-\ w_j(0)\right]\right) & if\ d_i=0 \end{cases}$$

where l_{id_i} is an indicator variable which takes the value 0 or 1.

The probability of a positive test response given the true disease status and the latent variable is further described as a four-class latent class model [32, 306]:

$$\Pr\left(t_{ij}=1 \middle| D=d, L=l_{id_i}\right)=\begin{cases}1, & if\ d_i=1\ \ and\ \ l_{i1}=1\\ 0, & if\ d_i=0\ \ and\ \ l_{i0}=1\\ w_j(1), & if\ d_i=1\ \ and\ \ l_{i1}=0\\ 1-\ w_j(0), & if\ d_i=0\ \ and\ \ l_{i0}=0\end{cases}$$

The term $w_j(0)$ is the probability test j correctly classifies a patient as non-diseased and $w_j(1)$ is the probability test j correctly classifies a patient as diseased.

4.4. Bayesian approach

Under the conditional independence assumption, the four latent class models described above – TLCM, FEM, REM and FMM – are identifiable in a single population study when there are at least three tests[32, 128]. However, when the three tests are assumed to be conditionally dependent given the true disease status, estimating the diagnostic accuracy of the tests using only the observed data when there is no gold standard is challenging as the models are non-identifiable. Thus, to make the models identifiable, the Bayesian approach has been recommended[140, 291, 316-318].

The Bayesian approach combines the observed data and prior information on the parameters of interest to obtain posterior distributions for the parameter[317]. The parameters of interest in this research study are the sensitivities and specificities of the tests being evaluated and the prevalence of the target condition (disease). Prior information is obtained from expert opinions, previous pilot or experimental studies, or literature related to the parameters of interest[319]. The Bayesian approach combines the observed data and prior information via the continuous form of Bayes' Theorem:

$$Pr(\boldsymbol{\theta}|data)=\frac{\Pr(data|\boldsymbol{\theta})\Pr(\boldsymbol{\theta})}{\Pr(data)} \qquad (17)$$

where $\boldsymbol{\theta}$ is a vector of the unknown parameters, $\Pr(\boldsymbol{\theta}|data)$ is the posterior distribution of the parameters, $\Pr(data|\boldsymbol{\theta})$ is the likelihood function, $\Pr(data)$ is the probability of the observed data and $\Pr(\boldsymbol{\theta})$ is the prior distribution.

The likelihood function describes the relationship between the parameters of interest and the observed data. The prior distribution describes the prior information for the parameters in a density function[320, 321]. The posterior distribution is the probability distribution obtained after the observed data and the prior information are combined typically via a sampling technique such as Markov Chain Monte Carlo (MCMC)[322]. Inference about a parameter is made using its posterior distribution (via the mean, median, mode or quartiles). Bayesian inference is accomplished using Gibbs Sampling

(BUGS) - WINBUGS[323] or openBUGS[324] software can be used to perform Bayesian analysis amongst others. There are different types of models employed in Bayesian analysis. These range from a simple model with only one parameter to complex - hierarchical models which need various hyper-parameters. Hyper-parameters are parameters employed to model the prior distribution of the parameter of interest[325]. They are related to the parameters of interest via some function such as the likelihood function. They could be a single value or a distribution. For example, the REM model (see section **4.3.3**) uses some hyper-parameters (the intercepts and variance parameters) to estimate the parameters of interest (sensitivity, specificity and prevalence). The intercepts and the variance parameters are called "hyper-parameters" because they are not the parameters of interest in the research; however, they play a very significant role in accurately estimating the parameters of interest and reflect some of the assumptions underlying a model. These hyper-parameters are part of the model. They are often used to constrain the parameters of interest. Hence, attention is needed when specifying these hyper-parameters so that the parameters of interest are accurately estimated.

4.4.1. Specification of prior information

Prior information about a parameter could be probabilistic (i.e. a probability distribution) or deterministic (i.e. a single value)[140]. Prior information is used to express a belief about a parameter in a model before evidence (such as the observed data) is taken into consideration in the analysis[316, 317]. The belief could be subjective if it is the opinion of an individual or group of individuals or it could be non-subjective if it is obtained from previous research studies such as pilot experiments or a published research study which employed real-life data. There are different measures that are used to elicit prior information about a parameter. Such measures include the mean, median, mode, probability values, uncertainty intervals or quartiles, proportions, and plots or graphs[326, 327]. The measures are also called Quantities of Interest (QoI)[328]. These measures are used to form a probability distribution that reflects the parameter of interest which is to be estimated. For example, in diagnostic accuracy studies, parameters such as prevalence, sensitivity and specificity, whose values range from 0 to 1, typically use the Beta distribution with hyper-parameters α and β as their prior distribution, because they are the conjugate prior to a Bernoulli distribution. Conjugate priors are prior distributions that are of the same family of probability distributions as the posterior distribution[329]. The hyper-parameters of this distribution can be obtained

via the mean if the information on the parameter is a single value estimate. For example, if the prior information on the sensitivity of a test 1 is 0.9. This value (0.9) can be taken to be the mean value of the Beta distribution. Hence, using the mean function of a Beta distribution which is:

$$E(X) = \frac{\alpha}{\alpha + \beta}$$

Possible combinations of hyper-parameters for the Beta distribution in this case are $\alpha = 6$ and $\beta = 0.667$ or $\alpha = 4.05$ and $\beta = 0.45$, since:

$$\frac{\alpha}{\alpha + \beta} = \frac{6}{6 + 0.667} = 0.9; \qquad \frac{4.05}{4 + 0.45} = 0.9$$

The probability intervals or quantiles or variance of the parameter can also be used to obtain the hyper-parameters of the prior distribution. There are some web-based applications such as the SHeffield ELicitation Framework (SHELF)[328, 330] and MATCH (Multidisciplinary Assessment of Technology for Healthcare)[331] that have been developed to aid in the determination of the prior probability distributions for the parameters of interest using the measures inputted by the user.

There are different ways to obtain information on the measures of interest from experts, which include face-to-face interviews, group workshops and probabilistic Delphi panel exercises[319, 328, 332] amongst others. Elicitation of priors from experts is thought to be more rigorous if there are multiple experts, and especially if there are many measures to obtain from the experts. It is an approach worth taking if there are no published articles or existing information about the measures of interest, or there are discrepancies in existing information about the measures of interest. Whichever approach is used to obtain prior information, it is necessary that the information is from a credible source and it is elicited accurately so that the posterior distributions obtained are not biased.

In Bayesian analysis, informative or non-informative priors can be used. A non-informative prior expresses no specific information about the parameter to be estimated [333-335]. It is often flat and covers a large range of plausible values for the parameter of interest. An example of a non-informative prior employed in diagnostic accuracy studies is Beta (1,1). It is expected that the posterior distribution of the parameter (using non-informative priors) will be influenced only by the observed data. Informative priors provide some information about the parameter to be estimated. They

do not cover a large range of values, so the parameter may be constrained. They can be classified as weak or strong priors. Weakly informative priors are used to express partial (weak) information about the parameter and strong priors are used to express strong or very specific information about the parameter[333-335]. Using a very strong prior could impact the posterior distribution significantly. However, if the observed dataset is large and the model's degrees of freedom are equal to or greater than the number of parameters to estimate then the posterior distribution is typically minimally impacted by the prior distribution whether the prior information is informative or non-informative. Furthermore, using a deterministic prior (i.e. an exact value rather than probabilistic prior) indicates very strong knowledge about the parameter. Parameters with deterministic priors are not estimated within the model.

4.4.2. Inference using the posterior distribution

Once the prior probabilistic information is obtained, the choice of model (see section 4.3) is ascertained, and the observed dataset is formatted appropriately, the parameters of interest are estimated using a numerical technique such as Markov Chain Monte Carlo (MCMC)[322] via statistical software such as WINBUGS[323]. The posterior distribution is obtained, which is used to make inferences about the parameter(s) of interest. The mean or median of the posterior distribution of the parameters of interest is often reported alongside the standard deviation or 95% credible interval. Different model diagnostics have been proposed such as the deviance information criterion (DIC), Gelman-Rubin diagnostic measures, autocorrelation plots and trace plots[336], to assess the posterior distributions of the parameters. These diagnostics are often used to "assess the convergence" of the posterior distributions of the parameters or to "select a model" (that fits best) among various models that may have been employed to analyse the observed dataset in a study.

Assessing convergence

Assessing the convergence of a parameter in a model is to ascertain that the samples generated from the algorithm (via MCMC) are from the posterior distribution of the parameter. A basic measure used to assess convergence is the *potential scale reduction factor* (denoted as $\hat{R}$). It is expected that at convergence the value of the potential scale reduction factor is equal to one (i.e. $\hat{R} = 1$). The Gelman and Rubin's potential scale reduction factor is estimated when multiple chains are employed to run

the Bayesian analysis. It is calculated based on the variance of the estimated parameter between chains, and the variance within a chain[317]. Plots employed to assess convergence are trace plots, autocorrelation plots and the Gelman-Rubin diagnostic plot[337]. It is expected that at convergence:

- The trace plots will be caterpillar-shaped, and where multiple chains are employed; all the chains will overlap each other. This is also to assess mixing.
- The auto correlation from the samples will tend to zero as the number of samples generated increases.
- The shrinkage factor in the Gelman – Rubin diagnostic plot will tend to one as number of the samples generated increases. The $\hat{R}$ uses estimates of the variance within and between the chains to monitor convergence[338].

Selection of a model

The deviance information criterion (DIC) [336] can be used to compare the different models employed to analyse a dataset in a study. DIC is a model fit criterion which is particularly suited to Bayesian models, and penalises models for increasing numbers of parameters. The DIC is a function of the posterior mean of the Bayesian deviance and the effective number of parameters (pD).

$$Deviance = -2\log p(y|\boldsymbol{\theta})$$

Where y is the observed data, $\boldsymbol{\theta}$ is the vector of parameters to estimate, and $p(y|\theta)$ is the likelihood of the observed data given the parameters of interest. The effective number of parameters (pD) is calculated as:

$$pD = \frac{Var(Deviance)}{2};\ \text{for complex models} \quad OR \quad pD = \bar{D} - \hat{D};$$

An example of a complex model is the REM. The posterior mean of the deviance is $\bar{D}$ and $\hat{D}$ is the deviance of the mean.

$$pD = E[-2\log(p(y|\theta)] + 2\log(p\left(y\middle|\bar{\theta}(y)\right)$$

where $\bar{\theta}(y) = E(\bar{\theta}|y)$ is the posterior mean of the parameters.

Hence, the DIC is calculated as: $DIC = \bar{D} + pD$

The model with the smallest DIC is chosen to be the model that best fits the data. Using the DIC as a criterion for model selection alone is not always encouraged[336], especially when the knowledge of the relationship between the parameters of interest and the observed data is well-known. This knowledge should be applied alongside the DIC in the selection of a model.

4.4.3. Advantages and disadvantage of Bayesian approach over frequentist approach

Advantage of Bayesian approach

- The Bayesian approach incorporates prior information about the unknown parameters and uses the observed data to estimate the parameters of interest[146] which could be more reliable if the prior is accurate and elicited carefully.
- The Bayesian approach aids in solving non-identifiable models[291]. Under the conditional dependence assumption, most latent class models cannot obtain a unique solution for the parameters of interest especially when the degrees of freedom are smaller than the number of parameters to estimate. Thus, using prior information about some of the parameters in the model helps to make the model identifiable and so a unique solution is provided.
- The Bayesian approach can be employed to solve the small sample – data problem[339]. Some research studies may lack sufficient data to reach a valid inference for some parameters of interest. However, if there are reliable opinions (expert opinions) about these parameters, these opinions can be employed as a prior to infer the parameters of interest. That is, the data analyst does not have to rely on large sample asymptotic theory[339].

Disadvantages of the Bayesian approach

- The Bayesian approach requires knowledge of the probability distribution of the parameters of interest for reliable inference. It is important to understand which distributional form to use for the parameters as this can impact the posterior distribution obtained. Several web-applications like MATCH[331] have been developed to help non-statisticians in eliciting prior information.
- The Bayesian approach employs prior information which could be based on expert opinions which are subjective. If the expert opinions are inaccurate, biased estimates are obtained. This is especially true if the study's observed

data sample size is small and the model is non-identifiable with the observed data alone.

- Bayesian analysis can be computationally complex and resource heavy.

4.5. Simulation

In this section, three datasets are generated to show the possible associations that can exist among three medical tests. The first dataset assumes that the three tests are conditionally independent (section **4.5.2**), the second dataset assumes that only two of the three tests are conditionally dependent (and both tests are conditionally independent of the third test), and the last dataset assumes that all the three tests are conditionally dependent among the diseased group (section **4.5.3**).

The datasets were simulated in R studio[290, 340, 341]. The simulated datasets represent possible test responses from the three tests under two basic assumptions – a conditional independence assumption and a conditional dependence assumption among the tests given the true disease status. The openBUGS[324, 342-345] software was employed to analyse the simulated datasets to estimate the sensitivities and specificities of the tests as well as the prevalence of the disease via the REM, FMM, TLCM and FEM. The RStan[346] package via R studio was used to analyse the generated dataset using the FEM approach by Wang et al[286] (FEM_W). RStan[346] uses the No-U-Turn sampler (NUTS)[347] an extension of the Hamiltonian Monte Carlo (HMC) algorithm, which is a sampling algorithm that does not employ the random walk behaviour that Gibbs sampling uses. In addition, the NUTS is less sensitive to correlated parameters, so this sampling technique allows for higher order correlations among three or more tests. The R-code written to simulate the different datasets is presented in the appendix (Appendix C.1), the WINBUGS code and RStan code used to analyse the datasets is also reported in the appendix (Appendix C.2). Diagnostic plots were employed to assess the convergence of the posterior distribution of the parameters.

4.5.1. Simulated values

Let us assume that there are three tests (T_1, T_2, T_3) to evaluate, and the response of each test is either 0 or 1. Here, 0 indicates the absence of the target condition and 1 indicates the presence of the target condition. The prevalence of the disease is 0.3, and the sensitivities and specificities of the three tests are:

$$Sn_1 = 0.9, \quad Sn_2 = 0.8, \quad Sn_3 = 0.7, \quad Sp_1 = 0.9, \quad Sp_2 = 0.8, \quad Sp_3 = 0.9$$

The choice of prevalence, sensitivities and specificities are arbitrary, but these values were chosen to represent values which are plausible for a diagnostic accuracy study. The fixed effect model by Wang et al[286] was employed to simulate the datasets. Inequality constraints[54] are employed to calculate the possible bound values that the covariance terms can take given the sensitivity and specificity of the tests. However, under the conditional independence assumption, all the covariance terms are zero. That is the covariance terms among the diseased group ($covs$) and the covariance terms among the non-diseased group ($covc$) between all tests are zero.

$$covs_{12} = covs_{13} = covs_{23} = covc_{12} = covc_{13} = covc_{23} = 0$$

With the FEM approach, the number of participants (or cell frequency) of each test combination $(111, 101, 110, 100, 000, 001, 011, 010)$ is simulated using a multinomial distribution. Therefore, using the information on the sensitivity and specificity of each test, alongside the prevalence and covariance terms, 100 different datasets of 500 participants were simulated using the "rmultinom" function in R. The choice of using 500 participants (rather than a small sample size like 50 or 100) was made to ensure that the inferences drawn about the models are consistent. This is because of the possibility that the sample size could impact the choice of one method over another. Studies of much larger sample size (10,000 for example) are possible but comparatively much rarer. I took the average of the 100 simulated datasets as my dataset for each scenario investigated. Using this approach reduces the random variability that could arise from using only one simulated sample (of 500 participants). Another option could have been to use a very large dataset, for example, 50,000 participants; however, using such a large sample size would be computationally intensive and there would be the risk of the sampling iterations failing (terminating) especially when using the REM and FEM. The simulated dataset is employed to explore how the different latent class models described in section **4.3** recover the simulated truth. In order to estimate the parameters of interest, the Bayesian approach was employed. This is because of the identifiability issues discussed earlier (section **4.3.2**).

4.5.2. Conditional independence assumption

Under the conditional independence assumption, all the covariance terms equate to zero. The simulated dataset is reported in Table 18 and the estimated parameters of interest are reported in Table 19.

Table 18: Simulated dataset of 500 participants assuming conditional independence

Test 1 result	Test 2 result	Test 3 result	Mean frequency of test results
1	1	1	**77**
1	1	0	**39**
1	0	1	**21**
1	0	0	**34**
0	1	1	**14**
0	1	0	**60**
0	0	1	**27**
0	0	0	**228**

The estimated sensitivities and specificities of the three tests as well as the prevalence from all the LCMs explored are approximately the same as the simulated truth (rounding to one decimal place – Table 19). The estimated sensitivities, specificities and prevalence from the TLCM and FEM models are identical because both models have the same likelihood function. The random effect and finite mixture models require a good choice of priors for the hyper-parameters to accurately reflect the underlying assumption of the analysis and to estimate the parameters of interest.

Therefore, given that the three tests are conditionally independent, the dependence (or variance) parameters of the REM, FMM and FEM were equated to zero. For the REM, the intercept parameters were modelled using weakly informed priors (via a truncated normal distribution). The intercept for the diseased (d = 1) and non-diseased group (d = 0) are stated as:

$$a_{j1} \sim N(0, 0.1)I(-1,) \quad and \quad a_{j0} \sim N(0, 0.1)I(, 1);$$

$I(-1,)\ and\ I(,1)$ are used to truncate the normal distribution to constrain the intercept parameter. $I(-1,)$ implies values from -1 to infinity and $I(,1)$ implies values from minus infinity to 1.

For the FMM model the indicator variables (l_{d_i}) used non-informative priors ($Beta(1,1)$), and the probability of each test to correctly classify participants as diseased or non-diseased ($w_j(d)$) used a weakly informed prior as described below.

$$W_j(d) \sim Beta\ (0.5, 0.5); \qquad d = 0, 1$$

The prevalence in all the models was modelled using a non-informative prior ($Beta\ (1,1)$).

Table 19: Estimated prevalence, sensitivities and specificities of the three tests from the different LCMs under the conditional independence assumption.

Parameters	Simulated truth	TLCM Mean (SD)	FEM Mean (SD)	REM (Probit) Mean (SD)	REM (Logit) Mean (SD)	FMM Mean (SD)
Sn_1	0.9	0.90 (0.05)	0.90 (0.05)	0.94 (0.06)	0.92 (0.05)	0.91 (0.05)
Sn_2	0.8	0.80 (0.05)	0.80 (0.05)	0.81 (0.05)	0.80 (0.05)	0.80 (0.05)
Sn_3	0.7	0.69 (0.05)	0.70 (0.05)	0.70 (0.05)	0.70 (0.05)	0.70 (0.06)
Sp_1	0.9	0.90 (0.03)	0.90 (0.03)	0.90 (0.03)	0.90 (0.03)	0.90 (0.03)
Sp_2	0.8	0.80 (0.03)	0.80 (0.03)	0.79 (0.03)	0.80 (0.03)	0.80 (0.03)
Sp_3	0.9	0.90 (0.02)	0.90 (0.03)	0.89 (0.02)	0.90 (0.02)	0.90 (0.02)
Prevalence	0.3	0.30 (0.03)	0.30 (0.02)	0.29 (0.04)	0.29 (0.04)	0.30 (0.04)

SD is standard deviation; TLCM is traditional latent class model; FEM is fixed effect model; REM is random effect model; FMM is finite mixture model; $Sn_1, Sn_2, Sn_3,$ are the sensitivities of test 1, test 2 and test 3 and Sp_1, Sp_2, Sp_3 are specificities of test 1, test 2 and test 3.

The trace plots, density plots, auto-correlation and Gelman diagnostic plots of the estimated parameters are reported in Figure 18 to Figure 21. The trace plots are caterpillar-shaped indicating that the model has converged to the target posterior distribution. The density plots of the parameters are bell-shaped and unimodal. The autocorrelations of the parameters are approximately zero. The Gelman diagnostic plots show that the shrinkage factor ($\hat{R}$) for all the parameters is 1, indicating the model has converged to the posterior distributions.

There are three different colours on the diagnostic plots which represent the three chains employed to run the analysis.

Figure 18: Trace plots of the sensitivity and specificity of the three tests and prevalence when all tests are conditionally independent.

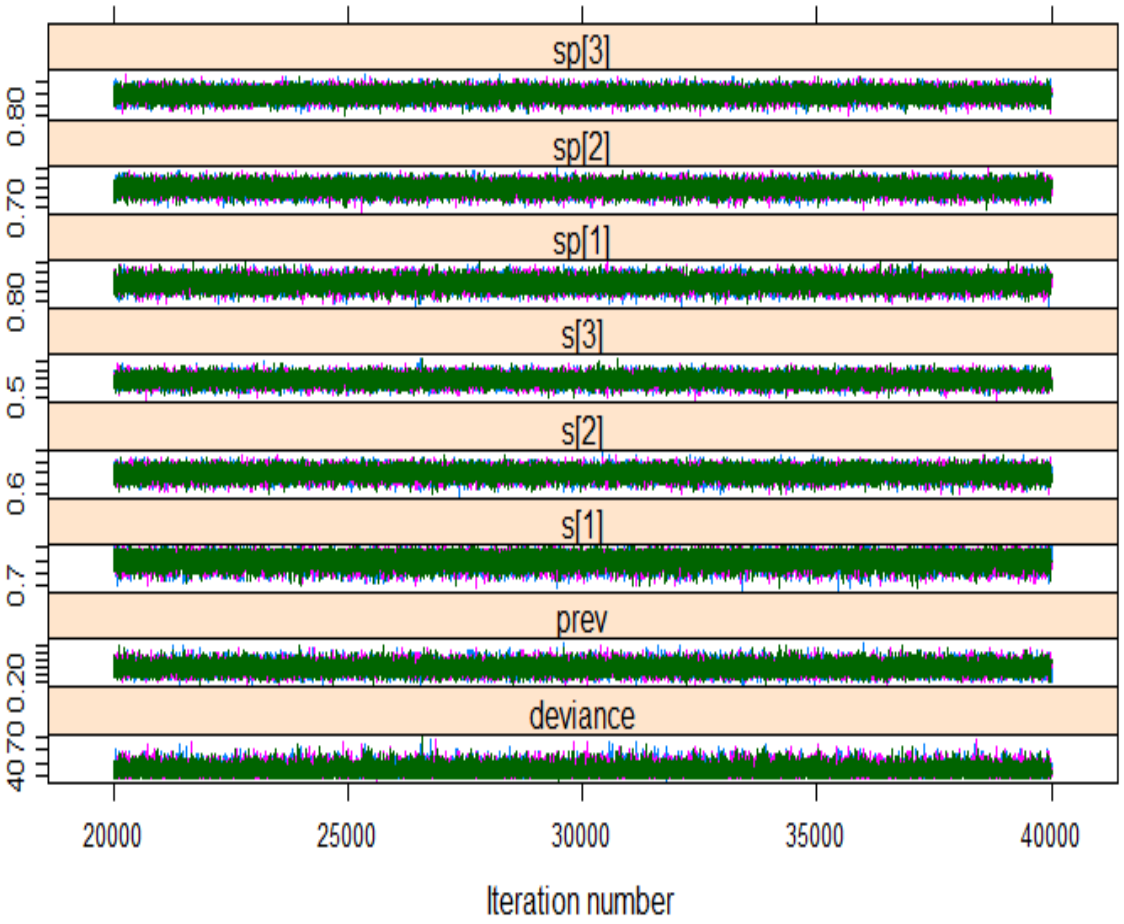


Figure 19: Density plots of the sensitivity, specificity of the three tests and prevalence when all tests are conditionally independent.

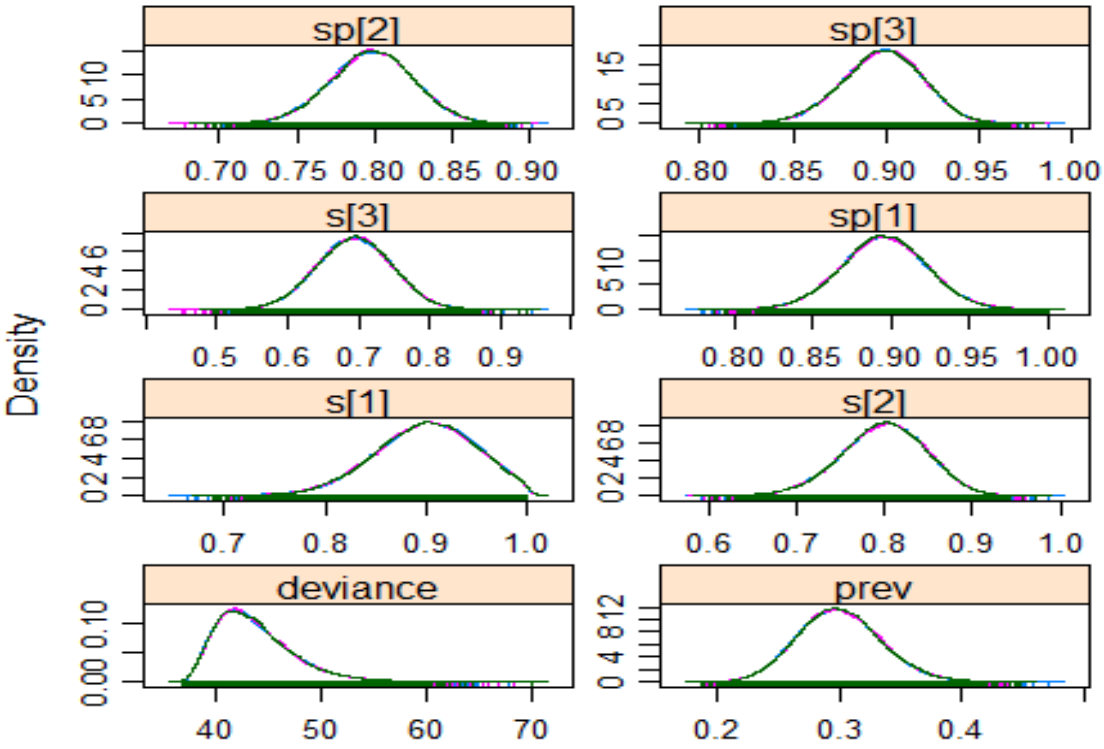


Figure 20: Auto-correlation plot of the sensitivity and specificity of the three tests

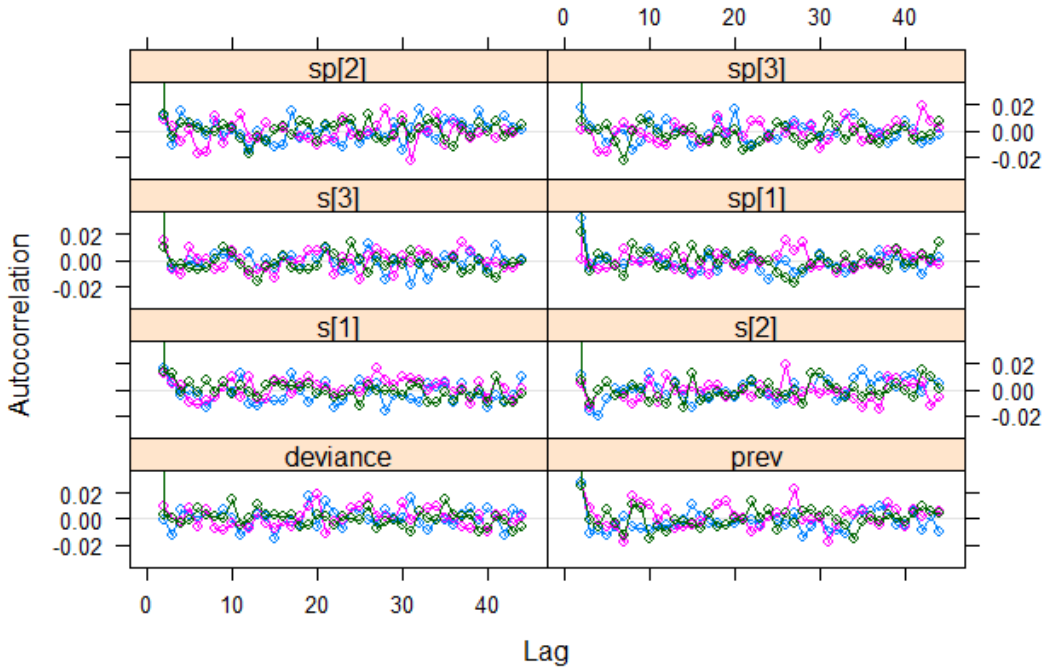
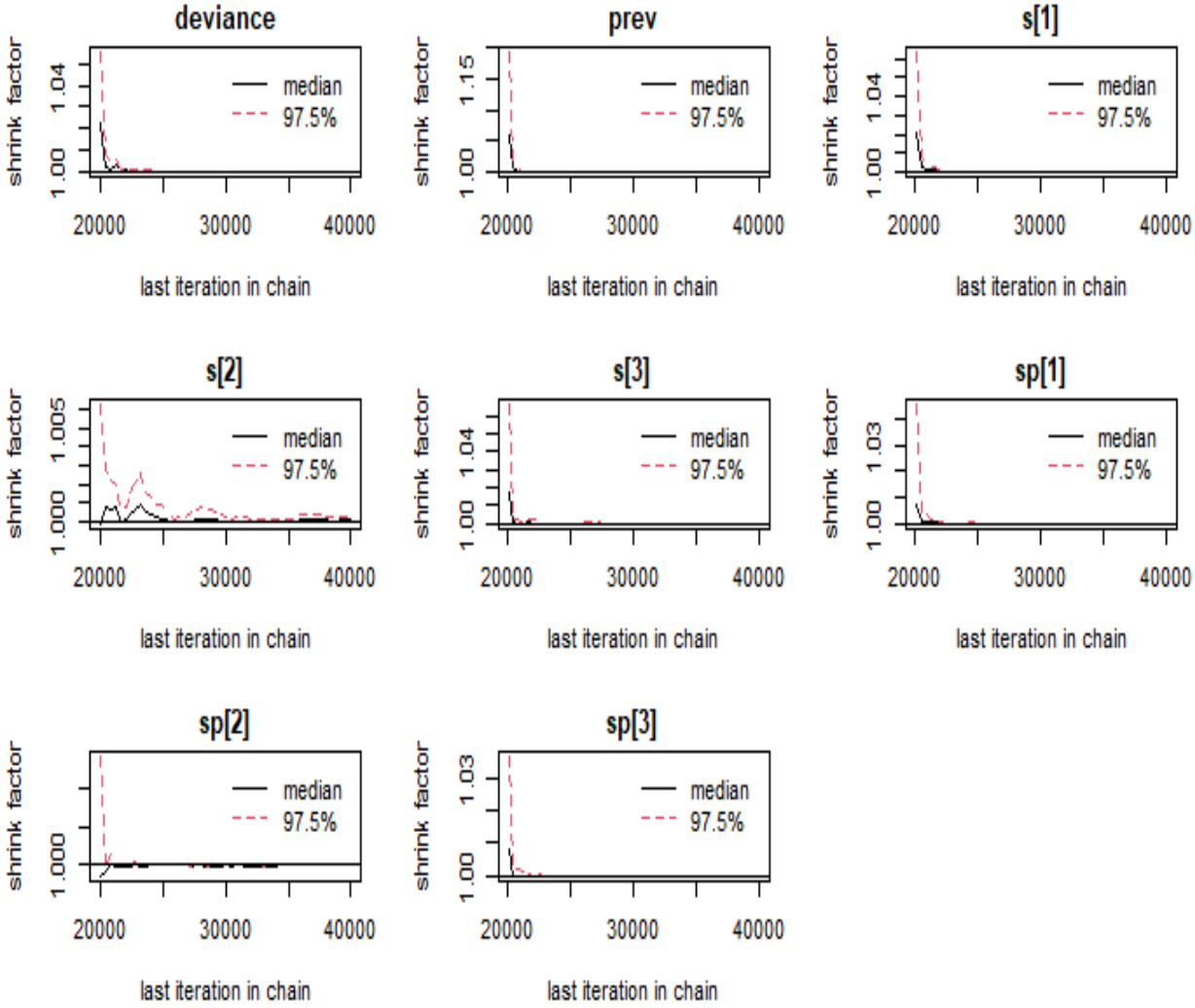


Figure 21: Gelman diagnostic plots of the sensitivities and specificities of the three tests



4.5.3. Conditional dependence assumption

To explore the LCMs under the conditional dependence assumption, two scenarios were simulated. Firstly, datasets were simulated assuming that two of the three tests were conditionally dependent given the true disease status. Secondly, all the three tests were assumed to be conditionally dependent given the true disease status. As discussed in section **4.3.2**, the number of parameters to estimate increases under the conditional dependence assumption and is often larger than the degrees of freedom in the model which makes the model non-identifiable. Therefore, to make the model identifiable, informative priors that constrain some or all of the parameters of interest are employed. Using the simulated truth in section **4.5.1**, the plausible informative prior distributions for the sensitivities and specificities for the three tests are:

- $Sn_1 \sim Beta\ (6,\ 0.667)$
- $Sn_2 \sim Beta(4, 1)$
- $Sn_3 \sim Beta\ (5.95, 2.55)$
- $Sp_1 \sim Beta(4.05, 0.45)$
- $Sp_2 \sim Beta(4, 1)$
- $Sp_3 \sim Beta(4.05, 0.45)$

The above informative priors are centred on the truth. The pairwise covariance terms of the fixed effects model employed a Uniform prior distribution that is bounded as follows:

- $Cov_{23} \sim Unif(0, ub23);$ $\qquad ub23 = \min(Sn_2, Sn_3) - Sn_2 \times Sn_3$

The variance parameter is modelled using a Gamma distribution for the FMM and truncated normal distribution for the REM:

- $b_1 \sim Gamma(1,1)$
- $b_1 \sim Normal\ (0, 0.1)I(0,)$

The prevalence was modelled using a flat beta prior:

- $p_d \sim Beta\ (1,1)$

SCENARIO ONE

The pairwise covariance term between the two tests is 0.11. The choice of the covariance term is based on the inequality constraints of the FEM approach. In addition, I wanted to use the strongest possible positive correlation between test 2 and test 3. The simulated true values are:

$$p_d = 0.3; \quad Sn_1 = 0.9, \quad Sn_2 = 0.8, \quad Sn_3 = 0.7, \quad Sp_1 = 0.9, \quad Sp_2 = 0.8,$$
$$Sp_3 = 0.9, \quad covs_{12} = 0; \quad covs_{13} = 0; \quad covs_{23} = 0.11$$

The dataset generated is reported in Table 20. Firstly, the dataset was analysed assuming that all tests were conditionally independent. This will help understand what happens to the estimates when the underlying assumption is misspecified. The estimated sensitivities and specificities of the tests are presented in Table 21.

Table 20: Simulated dataset of 500 participants assuming conditional dependence between two tests (test 2 and test 3).

Test 1 result	Test 2 result	Test 3 result	Mean frequency of test results
1	1	1	**92**
1	1	0	**23**
1	0	1	**5**
1	0	0	**50**
0	1	1	**15**
0	1	0	**61**
0	0	1	**28**
0	0	0	**226**

Table 21: Estimated prevalence, and sensitivities and specificities of the three tests from the different LCMs under the conditional independence assumption.

Parameters	Simulated truth	FEM Mean (SD)	REM (Probit) Mean (SD)	REM (Logit) Mean (SD)	FMM Mean (SD)
Sn_1	0.9	0.91 (0.04)	0.94 (0.05)	0.93 (0.04)	0.93 (0.05)
Sn_2	0.8	0.97 (0.02)	1.00 (0.01)	0.99 (0.02)	0.98 (0.03)
Sn_3	0.7	0.89 (0.05)	0.94 (0.06)	0.91 (0.05)	0.91 (0.06)
Sp_1	0.9	0.82 (0.02)	0.82 (0.03)	0.82 (0.02)	0.82 (0.02)
Sp_2	0.8	0.79 (0.02)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)
Sp_3	0.9	0.89 (0.02)	0.89 (0.02)	0.89 (0.02)	0.89 (0.02)
Prevalence	0.3	0.22 (0.02)	0.21 (0.02)	0.22 (0.02)	0.22 (0.03)

SD is standard deviation; TLCM is traditional latent class model; FEM is fixed effect model; REM is random effect model; $Sn_1, Sn_2, Sn_3,$ are the sensitivities of test 1, test 2 and test 3 and Sp_1, Sp_2, Sp_3 are specificities of test 1, test 2 and test 3.

From Table 21, the sensitivities of test 2 and test 3 are overestimated across all the LCMs. This is in line with the research study by Vacek [54], that if two tests are positively correlated among the diseased group, their sensitivities are overestimated if the conditional dependence between the two tests is not accounted for. The specificity of test 1 and the prevalence are underestimated and the specificities of test 2 and test 3 are estimated accurately (rounding to one decimal place). The diagnostic plots are reported in Appendix C.3.

Secondly, the dataset in Table 20 was analysed using the correct underlying assumption (conditional dependence between test 2 and test 3) and informative priors for some or all the parameters of interest. In analysing this dataset using the appropriate assumption, two versions of fixed effect latent class models were employed. The first version is the fixed effect model introduced by Wang et al[286] (FEM_W), which allows for higher order conditional dependence to be estimated as this approach was employed to simulate the datasets. Another version of the FEM is the well-known FEM[54] employed to model only the pairwise conditional dependence between two tests. The FMM was not used because of the complexity of the FMM in evaluating pair-wise correlated tests simultaneously with tests that they are conditionally independent.

Reanalysing the dataset (Table 20) while taking into consideration the conditional dependence between the two tests (test 2 and test 3), informative priors centred on the simulated true values (see below) were employed to analyse the dataset in Table 20. Generally, non-informative prior was used for the prevalence. The choice of using a non-informative prior for the prevalence is intentional as I want to find out if the latent class models can recover the true prevalence.

- $Sn_1 \sim Beta\ (6, 0.667)$
- $Sn_2 \sim Beta(4,1)$
- $Sn_3 \sim Beta\ (3.5, 1.5)$
- $Sp_1 \sim Beta(4.05, 0.45)$
- $Sp_2 \sim Beta(4, 1)$
- $Sp_3 \sim Beta(4.05, 0.45)$
- $p_d \sim Beta(1, 1)$

Table 22 shows the results obtained when the right model specification (with conditional dependence between test 2 and test 3 among the diseased group) and

informative priors (for some or all the parameters of interest) centred on the truth are used. The estimated values from the random effect model via probit link (REMP) is the same as the simulated truth, though it has larger uncertainty (SD) compared to the FEMs and random effect model via the logit link (REML). This could be an indication that the choice of prior affects the posterior distributions of the parameters. However, this observation is investigated later (within this section) by using different priors not centred on the simulated truth.

Table 22: Estimated prevalence, sensitivities and specificities of the three tests from the different LCMs assuming conditional dependence of two tests (test 2 and test 3)

Parameters	Truth	FEM Mean (SD)	FEM_W Mean (SD)	REMP Mean (SD)	REML Mean (SD)
Sn_1	0.9	0.93 (0.04)	0.73 (0.09)	0.90 (0.11)	0.93 (0.04)
Sn_2	0.8	0.86 (0.10)	0.72 (0.09)	0.80 (0.16)	0.92 (0.04)
Sn_3	0.7	0.76 (0.11)	0.65 (0.10)	0.70 (0.19)	0.75 (0.11)
Sp_1	0.9	0.88 (0.06)	0.90 (0.06)	0.90 (0.13)	0.96 (0.05)
Sp_2	0.8	0.79 (0.02)	0.83 (0.05)	0.80 (0.16)	0.79 (0.02)
Sp_3	0.9	0.89 (0.02)	0.94 (0.04)	0.90 (0.13)	0.89 (0.02)
Prevalence	0.3	0.27 (0.05)	0.38 (0.08)	0.30 (0.05)	0.34 (0.04)
$Covs_{23}$	0.11	0.08 (0.05)	0.08 (0.04)	NA	NA

SD is standard deviation; NA means not applicable; The FEM_W is the FEM by Wang et al and the FEM is the traditional latent class model that takes into consideration pairwise correlation between two tests (test 2 and 3). REMP is random effect model via probit link and REML is the random effect model via logit link. Covs is the pairwise covariance term between the two tests subscripted.

The FEM and REML overestimate the sensitivities of test 2 and test 3 and the REML overestimates the specificity of test 1. In addition, the covariance between test 2 and test 3 was underestimated, which could be the reason why the sensitivities of test 2 and test 3 were not estimated correctly. The FEM_W underestimates the prevalence, and sensitivities of test 1 and test 2. This posed questions because the dataset was simulated using the FEM_W approach (section **4.5.1**). On investigating the simulated datasets and the clinical case studies reported by Wang et al[286], I noticed that conditional independence among the tests in the non-diseased group means that the specificities of tests are very close to one, such that the possible covariance terms among the non-diseased group are insignificant or close to zero[286]. Therefore, another dataset was simulated using the values below:

$$p_d = 0.3; \quad Sn_1 = 0.9, \quad Sn_2 = 0.8, \quad Sn_3 = 0.7, \quad Sp_1 = 0.99, \quad Sp_2 = 0.99,$$
$$Sp_3 = 0.99, \quad covs_{12} = 0, \quad covs_{13} = 0; \quad covs_{23} = 0.11$$

The generated dataset is presented in Table 23.

Table 23: Simulated dataset of 500 participants assuming conditional dependence between two tests (test 2 and test 3) with specificities close to one.

Test 1 result	Test 2 result	Test 3 result	Mean frequency of test result
1	1	1	**92**
1	1	0	**19**
1	0	1	**5**
1	0	0	**25**
0	1	1	**12**
0	1	0	**4**
0	0	1	**3**
0	0	0	**340**

Table 23 was analysed taking into consideration the conditional dependence between test 2 and test 3. With the changes in the specificities of the three tests, the informative priors for the specificities of the three tests centred on the truth are Beta (113.45, 0.419) [286]. The estimated mean and standard deviation of the parameters of interest are presented in Table 24. The FEM estimates the parameters of interest accurately including the covariance term between test 2 and test 3. The FEMw underestimates the sensitivities of test 1 and test 2. The REML overestimates the sensitivities of test 2 and test 3 despite using informative priors. This could be because of the conditional dependence that exists between the two tests. The estimated sensitivities and specificities from the REMP and FMM are the same as the simulated truth. However, it has larger uncertainty (SD) compared to other estimates obtained from the FEMs and REML. This could be an indication that the choice of prior has a large impact on the posterior distributions in the REMP model. The diagnostic plots from the FEM and REML model are displayed in Appendix C.4 and Appendix C.5 respectively. The trace plots are caterpillar – shaped indicating that the model has converges to the target posterior distribution. The density plots of the parameters are bell-shaped and

unimodal. The Gelman diagnostic plots shows that the shrinkage factor ($\hat{R}$) for all the parameters is 1, indicating the model has converged to the posterior distributions.

Table 24: Estimated prevalence, sensitivities and specificities of the three tests from the different LCMs assuming conditional dependence of two tests (test 2 and test 3).

Parameters	Truth	FEM Mean (SD)	FEM_W Mean (SD)	REMP Mean (SD)	REML Mean (SD)
Sn_1	0.9	0.89 (0.03)	0.85 (0.05)	0.90 (0.11)	0.87 (0.03)
Sn_2	0.8	0.80 (0.04)	0.75 (0.05)	0.80 (0.16)	0.94 (0.03)
Sn_3	0.7	0.70 (0.04)	0.67 (0.05)	0.70 (0.19)	0.84 (0.06)
Sp_1	0.99	0.99 (0.01)	1.00 (0.01)	1.00 (0.01)	0.99 (0.01)
Sp_2	0.99	0.99 (0.01)	0.99 (0.01)	1.00 (0.01)	1.00 (0.01)
Sp_3	0.99	0.99 (0.01)	0.99 (0.01)	1.00 (0.01)	1.00 (0.01)
Prevalence	0.3	0.31 (0.02)	0.32 (0.03)	0.32 (0.03)	0.32 (0.02)
$Covs_{23}$	0.11	0.10 (0.03)	0.11 (0.03)	NA	NA

SD is standard deviation; NA means not applicable; The FEM_W is the FEM by Wang et al and the FEM is the traditional latent class model that takes into consideration pairwise correlation between two tests (test 2 and 3). REMP is random effect model via probit link and REML is the random effect model via logit link. Covs is the pairwise covariance term between the two tests subscripted.

To explore the impact of the priors on the posterior distributions, alternative priors not centred on the simulated true values were employed to reanalyse the dataset in Table 23. The informative priors (not centred on the simulated true values) are:

- $Sn_1 \sim Beta\ (4,1)$ centred on 0.8
- $Sn_2 \sim Beta(3.5, 1.5)$ centred on of 0.7
- $Sn_3 \sim Beta\ (6, 0.667)$ centred on 0.9
- $Sp_1 \sim Beta(3.5, 1.5)$ centred on 0.7
- $Sp_2 \sim Beta(4.05, 0.45)$ centred on 0.9
- $Sp_3 \sim Beta(4,1)$ centred on 0.8

The estimated parameters of interest are presented in Table 25. The diagnostic plots are presented in Appendix C.6.

Table 25: Estimated prevalence, sensitivities and specificities of the three tests from the different LCMs assuming conditional dependence of two tests (test 2 and test 3).

Parameters	Truth	Centred priors	FEM Mean (SD)	FEM_W Mean (SD)	REMP Mean (SD)	REML Mean (SD)
Sn_1	0.9	0.8	0.88 (0.03)	0.81 (0.06)	0.80 (0.16)	0.88 (0.03)
Sn_2	0.8	0.7	0.90 (0.05)	0.83 (0.07)	0.70 (0.19)	0.95 (0.03)
Sn_3	0.7	0.9	0.80 (0.05)	0.74 (0.06)	0.90 (0.11)	0.87 (0.05)
Sp_1	0.99	0.7	0.95 (0.02)	0.95 (0.02)	0.70 (0.19)	0.95 (0.03)
Sp_2	0.99	0.9	0.99 (0.01)	0.99 (0.01)	0.90 (0.13)	0.99 (0.01)
Sp_3	0.99	0.8	0.99 (0.01)	0.99 (0.01)	0.80 (0.16)	0.99 (0.01)
Prevalence	0.3	NA	0.28 (0.03)	0.30 (0.03)	0.32 (0.03)	0.28 (0.03)
$Covs_{23}$	0.11	NA	0.03 (0.03)	0.06 (0.04)	NA	NA

SD is standard deviation; NA means not applicable; The FEM_W is the FEM by Wang et al and the FEM is the traditional latent class model that takes into consideration pairwise correlation between two tests (test 2 and 3). REMP is random effect model via probit link and REML is the random effect model via logit link. Covs is the pairwise covariance term between the two tests subscripted.

The results presented in Table 25 indicate that the choice of priors employed had a large impact on the REMP, as the estimated values are centred on the priors and the variances of the estimated parameters are very large compared to those from the FEMs and REML. The results from FEMs and REML are not largely impacted by the choice of priors as the estimated values are not centred on the priors. However, the choice of priors affects the estimates obtained.

SCENARIO TWO

In scenario two, all the tests were assumed to be correlated among the diseased group. This implies that all the tests have pairwise conditional dependence and some third order correlation in the diseased group. This simulated scenario replicates the conditions in the clinical dataset which is analysed in chapter five (where all the three tests being evaluated have pair-wise covariance terms). The simulated truth for scenario two is:

$$p_d = 0.3; \quad Sn_1 = 0.9, \quad Sn_2 = 0.8, \quad Sn_3 = 0.7, \quad Sp_1 = 0.9, \quad Sp_2 = 0.8, \quad Sp_3 = 0.9$$

$$covs_{12} = 0.02; \quad covs_{13} = 0.06; \quad covs_{23} = 0.12$$

The equality and inequality constraints[286] were applied so that the sensitivities and specificities of the tests do not exceed their boundaries (0 and 1). The simulated dataset is reported in Table 26.

Table 26: Simulated dataset of 500 participants assuming conditional dependence among all tests.

Test 1 result	Test 2 result	Test 3 result	Mean frequency of test result
1	1	1	**100**
1	1	0	**20**
1	0	1	**8**
1	0	0	**45**
0	1	1	**11**
0	1	0	**62**
0	0	1	**23**
0	0	0	**231**

Assuming that the all the tests are conditionally independent given the true disease status overestimates the sensitivities of all the tests and underestimates the specificities of test 1. The estimated sensitivities and specificities of the tests are reported in Table 27 and the diagnostic plots are in Appendix C.7.

Table 27: Estimated sensitivities and specificities of the three tests from the different LCMs under the conditional independence assumption.

Parameters	Simulated truth	FEM Mean (SD)	REM (Probit) Mean (SD)	REM (Logit) Mean (SD)	FMM Mean (SD)
Sn_1	0.9	0.94 (0.03)	0.98 (0.03)	0.96 (0.03)	0.96 (0.04)
Sn_2	0.8	0.95 (0.03)	0.98 (0.02)	0.97 (0.03)	0.96 (0.04)
Sn_3	0.7	0.92 (0.04)	0.96 (0.04)	0.94 (0.04)	0.93 (0.05)
Sp_1	0.9	0.84 (0.02)	0.82 (0.02)	0.83 (0.02)	0.83 (0.02)
Sp_2	0.8	0.79 (0.02)	0.77 (0.03)	0.78 (0.02)	0.78 (0.03)
Sp_3	0.9	0.90 (0.02)	0.90 (0.02)	0.91 (0.02)	0.90 (0.02)
Prevalence	0.3	0.24 (0.02)	0.22 (0.02)	0.23 (0.02)	0.23 (0.02)

SD is standard deviation; TLCM is traditional latent class model; FEM is fixed effect model; REM is random effect model; $Sn_1, Sn_2, Sn_3,$ are the sensitivities of test 1, test 2 and test 3 and Sp_1, Sp_2, Sp_3 are specificities of test 1, test 2 and test 3.

Reanalysing the dataset described in Table 26, taking into consideration the conditional dependence among the three tests; informative priors centred on the simulated truth were employed for all the parameters of interest except the prevalence (as noted earlier this allows the model to be identifiable). The estimated sensitivities and specificities of test 1, test 2 and test 3 are reported in Table 28. The diagnostic plots of the result presented in Table 28 is reported in Appendix C.8.

Table 28: Estimated prevalence, and the sensitivities and specificities of the three tests from the different LCMs assuming conditional dependence among all three tests.

Parameters	Simulated truth	FEM_W Mean (SD)	REMP Mean (SD)	REML Mean (SD)	FMM Mean (SD)
Sn_1	0.9	0.76 (0.09)	0.9 (0.11)	0.88 (0.10)	0.9 (0.11)
Sn_2	0.8	0.73 (0.09)	0.8 (0.16)	0.83 (0.12)	0.8 (0.16)
Sn_3	0.7	0.68 (0.10)	0.7 (0.18)	0.73 (0.15)	0.7 (0.19)
Sp_1	0.9	0.91 (0.06)	0.9 (0.13)	0.86 (0.05)	0.9 (0.13)
Sp_2	0.8	0.83 (0.05)	0.8 (0.17)	0.77 (0.04)	0.8 (0.16)
Sp_3	0.9	0.95 (0.03)	0.9 (0.13)	0.9 (0.02)	0.9 (0.13)
Prevalence	0.3	0.38 (0.08)	0.4 (0.06)	0.38 (0.06)	0.39 (0.22)
$Covs_{12}$	0.02	0.05 (0.04)	NA	NA	NA
$Covs_{13}$	0.06	0.05 (0.04)			
$Covs_{23}$	0.12	0.08 (0.04)			

SD is standard deviation; NA means not applicable; The FEM_W is the FEM by Wang et al. REMP is random effect model via probit link and REML is the random effect model via logit link. Covs is the pairwise covariance term between the two tests subscripted.

From Table 28, the estimated sensitivities and specificities of the three tests from all the REMs and FMM are the same as the simulated truth except for the prevalence. However, the variances in the REMP and FMM are large compared to the REML and FEM_W. The sensitivities of test 1 and test 2 obtained from the FEM_W model are underestimated and the specificity of test 3 is overestimated. This is similar to what was observed in scenario one. When simulating datasets using the fixed effect approach by Wang et al[286], conditional independence between two tests among the diseased (non-diseased) group implied that the sensitivities (specificities) of the tests have to be close to one. Hence, another dataset was generated using the information below assuming that the specificities of all the tests are close to one:

$$p_d = 0.3; \quad Sn_1 = 0.9, \quad Sn_2 = 0.8, \quad Sn_3 = 0.7, \quad Sp_1 = 0.99, \quad Sp_2 = 0.99, \quad Sp_3 = 0.99$$

$$covs_{12} = 0.02; \quad covs_{13} = 0.06; \quad covs_{23} = 0.12$$

The generated dataset is presented in Table 29. This dataset was analysed using the informative priors centred on the simulated truth and the results obtained are reported in Table 30.

Table 29: Simulated dataset of 500 participants assuming conditional dependence among all tests and the specificities of all the tests are close to one.

Test 1 result	Test 2 result	Test 3 result	Mean frequency of test result
1	1	1	**101**
1	1	0	**13**
1	0	1	**5**
1	0	0	**22**
0	1	1	**4**
0	1	0	**8**
0	0	1	**1**
0	0	0	**345**

Table 30: Estimated prevalence, and the sensitivities and specificities of the three tests from the different LCMs assuming conditional dependence among all three tests.

Parameters	Simulated truth	FEM_W Mean (SD)	REMP Mean (SD)	REML Mean (SD)	FMM Mean (SD)
Sn_1	0.9	0.86 (0.05)	0.90 (0.11)	0.92 (0.06)	0.90 (0.11)
Sn_2	0.8	0.77 (0.05)	0.80 (0.16)	0.81 (0.1)	0.80 (0.16)
Sn_3	0.7	0.69 (0.05)	0.70 (0.19)	0.65 (0.14)	0.70 (0.19)
Sp_1	0.99	1.00 (0.01)	1.00 (0.01)	0.99 (0.01)	1.00 (0.01)
Sp_2	0.99	1.00 (0.01)	1.00 (0.01)	1.00 (0.01)	1.00 (0.01)
Sp_3	0.99	1.00 (0.00)	1.00 (0.01)	1.00 (0.00)	1.00 (0.01)
Prevalence	0.3	0.32 (0.03)	0.4 (0.08)	0.4 (0.05)	0.43 (0.23)
$Covs_{12}$	0.02	0.03 (0.03)	NA	NA	NA
$Covs_{13}$	0.06	0.05 (0.03)			
$Covs_{23}$	0.12	0.11 (0.03)			

SD is standard deviation; NA means not applicable; The FEM_W is the FEM by Wang et al. REMP is random effect model via probit link and REML is the random effect model via logit link. Covs is the pairwise covariance term between the two tests subscripted.

From Table 30 the estimated sensitivities and specificities are approximately the same as the simulated true values (rounding to one decimal place) except the prevalence from the REMs and FMM. This could be because no informative prior was employed for the prevalence parameter. The diagnostic plots from the FEM_W and REML are reported in Appendix C.9 and Appendix C.10 respectively. The trace plots are caterpillar – shaped indicating that the model has converges to the target posterior distribution. The density plots of the parameters are bell-shaped and unimodal. The Gelman diagnostic plots shows that the shrinkage factor ($\hat{R}$) for all the parameters is 1, indicating the model has converged to the posterior distributions.

Informative priors, not centred on the simulated truth were employed to reanalyse the dataset reported in Table 26. The estimates obtained are reported in Table 31.

Table 31: Estimated prevalence, and the sensitivities and specificities of the three tests from the different LCMs assuming conditional dependence among all three tests.

Parameters	Simulated truth	Centred priors	FEM Mean (SD)	REMP Mean (SD)	REML Mean (SD)	FMM Mean (SD)
Sn_1	0.9	0.8	0.85 (0.06)	0.80 (0.16)	0.92 (0.06)	0.80 (0.16)
Sn_2	0.8	0.7	0.84 (0.07)	0.70 (0.19)	0.90 (0.07)	0.70 (0.19)
Sn_3	0.7	0.9	0.76 (0.07)	0.90 (0.11)	0.80 (0.12)	0.90 (0.11)
Sp_1	0.99	0.7	0.95 (0.02)	0.70 (0.19)	0.95 (0.02)	0.70 (0.19)
Sp_2	0.99	0.9	0.99 (0.01)	0.90 (0.13)	0.99 (0.01)	0.90 (0.13)
Sp_3	0.99	0.8	1.00 (0.00)	0.80 (0.16)	0.99 (0.01)	0.80 (0.16)
Prevalence	0.3	NA	0.29 (0.03)	0.41 (0.08)	0.34 (0.07)	0.64 (0.27)
$Covs_{12}$	0.02	NA	0.05 (0.04)	NA	NA	NA
$Covs_{13}$	0.06		0.07 (0.04)			
$Covs_{23}$	0.12		0.07 (0.04)			

SD is standard deviation; NA means not applicable; The FEM_W is the FEM by Wang et al. REMP is random effect model via probit link and REML is the random effect model via logit link. Covs is the pairwise covariance term between the two tests subscripted.

From Table 31 the FMM and REMP are largely impacted by the choice of priors as their values are centred on the priors and they have large uncertainties compared to estimates from FEM_W and REML. Moreover, using priors not centred on the true values can affect some of the estimated parameters, for example the sensitivities of test 2 and test 3 in the REML model and the sensitivity of test 3 in the FEM_W model were overestimated compared to the estimated sensitivities when informative priors centred on the truth were employed (see Table 30). The diagnostic plots of estimates obtained using priors not centred on the simulated true values are reported in appendix (Appendix C.11).

4.6. Limitations

Firstly, the dataset was simulated using the FEM approach proposed by Wang et al[286] (FEM_W) which allows pairwise and third order correlations between and among the tests to be modelled. This approach could have impacted the estimates obtained from other types of LCMs such as the REMs and FMM especially when the conditional dependence terms are not estimated accurately. However, in real – life scenarios, the

observed dataset cannot be proven by any statistical test to have a fixed or random effect conditional dependence structure. These are assumptions made by the researchers based on the knowledge of the observed dataset to justify the choice of LCM chosen. Secondly, simulated random samples may be affected by sampling variation, although this variation was minimized by the simulation of 100 samples of 500 participants and using the mean values rather than generating a single sample of 500 participants. Although, effort has been made to ensure that the results produced are reproducible by setting the seed to produce the simulated datasets in R. However, if a different seed is used, a different dataset will be generated, and slightly different estimates could be obtained; but the difference in the estimates obtained may not be significant. In addition, the scenarios explored in this chapter are limited to three imperfect tests that are correlated among the diseased group. There are other possible scenarios where the values for the sensitivity and specificity of the threes tests and the covariance terms are different from what is employed in this chapter. Thus, the comparison and inferences made in this chapter are by necessity limited. Furthermore, this simulation study is limited to three imperfect tests in a population where there is no gold standard; diagnostic accuracy studies can evaluate more or fewer than three imperfect tests in a population. However, the advantage of having more than three tests makes the models more identifiable without the need for informative priors, provided that the tests with conditional dependence are few.

4.7. Summary

4.7.1. Conditional Independence

Under the conditional independence assumption, the REMs, FEM, TLCM and FMM are all identifiable with a minimum number of three tests in a population because the number of parameters to estimate is at least equal to the number of degrees of freedom. However, the REM and FMM require a good choice of hyper-priors (weakly informed hyper-priors) for the hyper-parameters, to estimate the parameters of interest accurately. Based on the simulation study, the TLCM is recommended when all the tests being evaluated are conditionally independent because:

- They are simple models that do not have hyper-parameters; hence they do not require specification of prior distribution for any hyper-parameters.

4.7.2. Conditional dependence

Estimating the sensitivity and specificity of three conditionally dependent imperfect tests in a population requires that informative priors are used to make the LCMs identifiable and to estimate the parameters of interest accurately. The different LCMs explored in this chapter are FEMs, REMs and FMM and all these LCMs have different conditional dependence structures.

Fixed effect models

From the simulation study, when two tests are conditionally dependent and both tests are conditionally independent of the third test, then using the basic FEM (with pairwise conditional dependence) would be recommended as opposed to the fixed effect model proposed by Wang et al[286]. This is because there is no higher order correlation that exist among the tests. However, when the three tests have pairwise conditional dependence among themselves (that is test 1 and test 2 are correlated, test 2 and test 3 are correlated and test 1 and test 3 are correlated), then using the FEM by Wang et al[286] (FEM_W) should be considered, because of the third order correlations that exist among the tests.

A disadvantage with the FEM_W is that conditional independence among the diseased (non-diseased) group implies that the evaluated tests are expected to have sensitivities (specificities) that are close to one so that the covariances terms of the tests among the non-diseased group are insignificant or close to zero. This is a case of conditional independence as observed in section **3.3.2** because tests with approximately 100% sensitivities (specificities) are conditionally independent among the diseased (non-diseased) group. However, as observed in the simulation studies, conditional independence between two tests among either the diseased (or non-diseased) group does not always imply that the sensitivities (or specificities) of the tests must be close to one. It implies that the covariance terms between the two tests among the disease groups (diseased and non-diseased) are zero. Therefore, the FEM_W approach could be employed if the expected specificities (sensitivities) of all the tests employed in the diagnostic accuracy study are approximately one, in order to ensure conditional independence among the non-diseased (diseased) group. For example, Table 20 and Table 26, showed two datasets simulated when the specificities and sensitivities of the imperfect tests are not approximately equal to 1. However, the three tests are conditionally independent among the non-diseased group, because the covariance terms among the non-diseased group were zero. Analysing these datasets with the

FEM_W produced inaccurate estimates (Table 22 and Table 28) for the prevalence and sensitivities of the tests despite generating the datasets using the FEM_W approach and using informative priors centred on the simulated true values. Therefore, if the three tests are conditionally dependent among the diseased (non-diseased) group, and the specificities (sensitivities) of all the tests are expected to be approximately one then using the FEM_W approach should be considered.

Random effect models

Using informative priors directly for the parameters of interest in the REML has less impact on the posterior distributions of the parameters but using informative priors directly for the parameters of interest in the REMP has a significant impact on the posterior distributions of the parameters. A possible way to reduce this impact could be to transform the informative priors to informative hyper-priors. An example is the study by Dendunkuri et al[348], where a mathematical approach (bisection method[349, 350]) was employed to estimate the hyper-parameters of the probit link function using the prior information on the sensitivity and specificity of the tests being evaluated. Transforming informative priors of the parameters of interest into hyper-priors of hyper-parameters could be complex. This would be the case when the number of tests to evaluate is more than two. This could be the reason why most Bayesian diagnostic accuracy studies[70, 140, 309] employ the random effect model via the logit link.

From the simulation, the REMP is largely impacted by the choice of priors. The REML is less impacted by the choice of priors. However, specifying the correct priors will help the REML to estimate the sensitivities and specificities of the tests accurately.

Finite mixture model

From the simulation study, the choice of priors employed has a large impact on the posterior distribution of the model. To reduce this impact, researchers could consider transforming the informative priors to hyper-priors, however these can be very complex because of the model structure.

Comparing all latent class models based on observations from the simulation

If there are three imperfect tests to evaluate in a population and only two of the tests are conditionally dependent either among the diseased or non-diseased groups (i.e. scenario one); the FEM and REML models are preferred over the FMM and REMP models as they are less impacted by the choice of priors. The FEM may also be

preferred over the REML because it is computationally faster and can analyse very large datasets (for example, 50,000 participants).

Generally, the choice of LCM, can depend on the prior knowledge about the tests being evaluated and some underlying conditions on the participants in the datasets. The REMs can be chosen if there is a suspicion that the disease status of the participants has some other underlying factor such as age, which can contribute to the results obtained by the tests. The FMM can be chosen if there is a degree of certainty that some percentage of participants are correctly classified as diseased and non-diseased, because this information will help to specify the correct hyper-priors. The FEM can be employed if there is no prior knowledge of any variable that can influence the test responses of each participant or there is prior knowledge that all the tests evaluated have pairwise correlation between themselves and some higher order correlation and these pair-wise conditional dependencies could be of importance to the researcher. These comparison are shown in table (Table 32).

Following the observations from the simulation studies on the various latent class models, the FEM_W^{286} and the REML will be employed to analyse the clinical dataset in chapter five. This is because, both latent class models are less influenced by the choice of priors.

Table 32: Comparison of the different LCMs explored in this chapter

Comparison of LCMs based on the three tests scenarios explored in this chapter				
	LCMs			
Scenario	**FEM**	**FEM$_W$**	**REM**	**FMM**
Scenario One	Model is identifiable so model can recover the simulated true values.	Model is identifiable so model can recover the simulated true values.	Model is identifiable so model can recover the simulated true values.	Model is identifiable so model can recover the simulated true values.
Scenario Two	Model is non-identifiable and will require informative priors.	Model is non-identifiable and will require informative priors.	Model is non-identifiable and will require informative priors.	Model is non-identifiable and will require informative priors.
	Does not require hyper-priors	Does not require hyper-priors	Require informed hyper-priors.	Require informed hyper-priors
	Encourage the use of FEM over the FEM$_W$ in scenario two, because the evaluated tests are not expected to have sensitivities or specificities close to one.	Encourage its use if the specificities of the evaluated tests are expected to be close to one indicating that the tests are conditionally independent among the non-diseased group.	REM do not expect the sensitivities or specificities of the evaluated tests to be close to one. In addition, due to the complexity of specifying the hyper-priors of the probit link REM, the logit link is well-employed.	Complexity in specifying the right hyper-priors makes the choice of priors directly employed on the parameters of interest to impact largely the posterior distribution of the parameters of interest.

Scenario one: All the three tests are conditionally independent among the diseased and non-diseased groups

Scenario two: Two of the three tests are conditionally dependent among the diseased group only and both tests are conditionally independent of the third test among the diseased and non-diseased groups. LCM is latent class model; FEM is the classical fixed effect model with pairwise covariance structure; FEM$_W$ is the FEM by Wang et al with third-order covariance structure; REM is random effect model; FMM is Finite mixture model

Table 32 cont. Comparison of the different LCMs explored in this chapter

Comparison of LCMs based on the three tests scenarios explored in this chapter				
	LCMs			
Scenarios	**FEM**	**FEM$_W$**	**REM**	**FMM**
Scenario Three	Cannot be used for cases with higher order covariance terms than two.	Model not identifiable so informative priors are required.	Model not identifiable so informative priors are required.	Model not identifiable so informative priors are required.
		Model is encouraged to be used if the specificities of the evaluated tests are expected to be close to one indicating conditional independence among the non-diseased group. Otherwise, consider the REM.	REM does not expect the sensitivities or specificities of the evaluated tests to be close to one. In addition, due to the complexity of specifying the hyper-priors of the probit link, the logit link is employed.	Complexity in specifying the right hyper-priors makes the choice of priors directly employed on the parameters of interest impact largely the posterior distribution of the parameters of interest.

Scenario three: All three tests are conditionally dependent among the diseased group.

LCM is latent class model; FEM is the classical fixed effect model with pairwise covariance structure; FEM$_W$ is the FEM by Wang et al with third-order covariance structure; REM is random effect model; FMM is Finite mixture model

Table 32 cont. Comparison of the different LCMs explored in this chapter

General comparison of LCMs explored in this chapter				
	LCMs			
Issues	**FEM**	**FEM$_W$**	**REM**	**FMM**
Sampling algorithm	Model uses the MCMC sampling algorithm which can be implemented on Openbugs directly and R-Openbugs package.	This model uses the No-U-Turn sampling algorithm which is implemented in the R-Stan package because of the correlation that exits among the tests evaluated.	Model uses the MCMC sampling algorithm which can be implemented on Openbugs directly and R-Openbugs package.	Model uses the MCMC sampling algorithm which can be implemented on Openbugs directly and R-Openbugs package.
Variables that could affect disease status among participants	Encourage its use over the REM and FMM if there is no indication that the disease status of the participants could be affected by subject specific variables like age.	Encourage its use over the REM and FMM if there is no indication that the disease status of the participants could be affected by subject specific variables like age.	Encourage its use over FEM and FMM if there is an indication that the disease status of the participants are subject-specific that is affected by age, race or sex or other variables.	Encourage its use over FEM and REM if the probabilities that the evaluated tests correctly classify the participants into diseased and non-diseased are known. However, this will require using informative hyper-priors to represent such information.

LCM is latent class model; FEM is the classical fixed effect model with pairwise covariance structure; FEM$_W$ is the FEM by Wang et al with third-order covariance structure; REM is random effect model; FMM is Finite mixture model

4.8. Revisiting the clinical dataset from chapter three

In this section, the Matos et al[124] clinical dataset discussed in chapter three is reanalysed using the TLCM. This example investigated the clinical performance of LFpen and FC to classify teeth with or without Dentine Lesions using operative intervention as the reference standard (Table 16). In chapter three, illogical results were obtained from using the Staquet et al correction method (see Table 17). The results obtained when reanalysing the dataset with TLCM are presented in Table 33. Using the TLCM, the known sensitivity and specificity of the reference standard are employed as deterministic priors. Deterministic priors involve setting the sensitivity and specificity of the reference standard within the TLCM to an exact value, rather than converting them to a probability distribution. Deterministic priors were employed to make the TLCM comparable to Brenner and Staquet et al correction methods, therefore, the estimates obtained via TLCM could be compared to the estimates obtained from the Brenner and Staquet et al correction methods which do not use probabilistic priors.

From Table 33, the estimates from the TLCM are logical, as they are not above one unlike the estimated sensitivities from the Staquet et al correction method. The specificities of LFpen and FC are consistent across all methods (≅0.9). The estimated sensitivities from the TLCM is approximately the same as the unadjusted estimates (≅1). Thus, in the case where illogical estimates are obtain via the Staquet et al approach, using the TLCM is recommended if the two tests are conditionally independent.

Table 33: Estimated sensitivities and specificities of LFpen and FC in classifying teeth with D3

Dentine caries lesion (D3)								
	LFpen				FC			
Accuracy measures	**Unadjusted (SD)**	**Brenner (SD)**	**Staquet et al (SE)**	**TLCM Mean (SD)**	**Unadjusted (SD)**	**Brenner (SD)**	**Staquet et al (SD)**	**TLCM Mean (SD)**
Sensitivity	0.95 (0.05)	0.87 (0.07)	1.04 (NaN)	0.94 (0.05)	1(0.00)	0.91 (0.06)	1.09 (NaN)	0.96 (0.04)
Specificity	0.89 (0.02)	0.87 (0.02)	0.90 (0.02)	0.89 (0.02)	0.90 (0.02)	0.89 (0.02)	0.92 (0.01)	0.91 (0.02)
Prevalence of D3 via TLCM is 0.07 (0.01)								

LFpen: laser florescence pen; FC: fluorescence camera; NaN means not a number; SD is standard deviation; TLCM is traditional latent class model.

Chapter Five: Clinical Application of Latent Class Model

In chapter two, methods were identified by the systematic review that can be employed to evaluate the diagnostic accuracy of a medical test in the absence of a gold standard. Following the findings of this review[57], the latent class model (LCM) was identified as being the best approach to use when there are multiple imperfect tests to evaluate and none of the tests are considered as a gold standard. In the LCM, the sensitivity and specificity of all the tests are evaluated simultaneously. An advantage of LCMs is that conditional dependence (or independence) that exists among the tests under evaluation can be taken into consideration. Various LCMs were identified by the review, some of which were explored in chapter four. Two LCMs (FEM_W[286] and REML) were considered appropriate to analyse the clinical dataset used as a case study in this chapter because the scores evaluated in the clinical datasets are correlated, as explored in section **5.2** of this chapter. The clinical dataset is described in section **5.1** and **5.2**. The aims of the analysis presented in section **5.3**. The methods employed to analyse the clinical dataset under the different model assumptions are presented in section **5.4**. The results of the analysis of the clinical data are reported in section **5.5**, and this chapter concludes with a discussion of the findings and limitations of the case study (section **5.6**).

5.1. Description of the clinical data

The dataset explored in this chapter was obtained from the RA-MAP Consortium, Newcastle University. The RA-MAP Consortium is a multi-partner organisation of more than 140 individuals affiliated with 21 academic and industry organisations that are focused on making genomic medicine in rheumatoid arthritis a reality. They have established large a cohort dataset of patients with Rheumatoid Arthritis (RA) in 28 centres in the UK. The RA-MAP Consortium data comprises of 317 participants and there are 70 variables recorded for each participant. The variables include clinical information to enable the calculation of three specific scores which are: the Disease Activity Score of 28 joints – Erythrocyte Sedimentation Rate (DAS28-ESR), the Simplified Disease Activity Index (SDAI) and the Clinical Diseases Activity Index (CDAI). These scores measure the disease activity of patients suspected and diagnosed with rheumatoid arthritis. Demographic variables such as age, sex, race, height and weight, etc. were also collected. These variables were collected both at baseline and at six months (follow-up period). In this research, I will use the information

collected from the RA patients at baseline. The demographic characteristics of the participants in the cohort study (at baseline) are provided in Table 34

Table 34. The baseline data are newly diagnosed patients with seropositive RA. Further details on the characteristics of the participants, ethical approval and protocol for data collection is available in Tom et al[351].

Table 34: Demographic profile of RA patients and mean value of core set of variables

Variables	Statistics	Baseline dataset		
		Total	**Female**	**Male**
Age	Min.	20	20	22
	Max.	84	84	84
	Mean (SD)	53.12 (15.27)	51.25 (15.08)	58.01 (14.77)
DAS28-ESR$_4$	Min.	1.89	2.07	0.89
	Max.	8.68	8.68	8.14
	Mean (SD)	5.19 (1.38)	5.310 (1.37)	4.89 (1.38)
SDAI	Min.	2.22	2.22	2.53
	Max.	78.60	78.60	71.30
	Mean (SD)	28.80 (14.29)	29.80 (14.53)	26.19 (13.39)
CDAI	Min.	1.70	1.70	1.90
	Max.	66.10	66.10	65.60
	Mean (SD)	27.08 (13.47)	23.9 (12.49)	28.30 (13.67)

SD mean standard deviation; Min is minimum; Max is maximum; DAS28-ESR$_4$ is disease activity score of 28 joints including CRP (four variables); SDAI is simplified disease activity index; CDAI is clinical disease activity index.

RA is a disease of the joints and muscles[21] and disease activity is measured in RA patients to enable rheumatologists offer personal-based treatment to improve quality of life. Various scores and medical tools have been employed to measure the disease activity in RA patients, such as the DAS28 (CRP and ESR), SDAI, CDAI, X-ray, and Doppler signals amongst others. There are different variations of DAS28 including the DAS28-CRP and the DAS28-ESR[352-355]. However, I will be focusing on the DAS28-ESR$_4$ (which includes four variables) because it is a well-established and widely used measure among rheumatologists[21]. In addition, for most research studies on the

validity or accuracy of newly developed or proposed scores for the diagnosis of RA, DAS28-ESR$_4$ has been widely considered as a gold standard[356-358]. In addition to DAS28-ESR$_4$, I will evaluate SDAI and CDAI. The formulas used to calculate these scores are expressed in Table 35. These formulas were used to calculate the baseline score for each patient in the RA-MAP dataset.

Table 35: The Disease Activity Scores

Scores	Formula
$DAS28-ESR_4$	$\left(0.56 * \sqrt{TEN28} + 0.28 * \sqrt{SJC28} + 0.70 * In\,(ESR)\right) + 0.014 * ptGVAS$
$SDAI$	$TJC + SJC + PDGA + EDGA + CRP$
$CDAI$	$TJC + SJC + PDGA + EDGA$

Key: TEN28 is tenderness upon touching the 28 joints; SJC28 is swollen joint count in 28 joints; ESR is Erythrocyte sedimentation rate which is the rate at which red blood cells sediment in a period of one hour; and ptGVAS is the patient's global disease activity visual analogue scale (VAS). Tender joint count (TJC); the swollen joint count (SJC); patient global disease activities (PDGA), and the evaluator determined global disease activity (EDGA); DAS28-ESR$_4$ is disease activity score of 28 joints including ESR (four variables); SDAI is simplified disease activity index; CDAI is clinical disease activity index.

The values obtained from these scores are used to classify the patients into one of four disease activity groups; remission, low disease activity (LDA), moderate disease activity (MDA) and high disease activity (HDA)[21, 354]. The responses from these three scores are continuous, so cut-offs are employed to classify patients into the different levels of disease activity. The standard cut-off for the different scores in classifying the disease activities of the patients are[353, 359, 360] presented in Table 36.

5.1.1. Missing data

Out of 371 participants in the cohort, there were 51 participants whose baseline DAS28-ESR$_4$ score was missing. There were a further 2 participants whose baseline scores for SDAI and CDAI were also missing. Therefore, there are 264 participants with complete information on the three scores at baseline. Since the number of participants with incomplete information across the three scores (53 at baseline) can be considered relatively small compared to participants with complete information on the three scores, removing these participants from the final analysis seems logical. This will also circumvent the complexity that could arise from making assumptions

about the missingness mechanism (missing at random or missing not at random) of the data.

Table 36: Table of various cut-offs of the disease activities scores

Scores	Disease activity state	Accepted cut-off
SDAI[359]	Remission	$x \leq 3.3$
	Low disease activity	$3.3 < x \leq 11$
	Moderate disease activity	$11 < x \leq 26$
	High disease activity	$x > 26$
CDAI[359, 361]	Remission	$x \leq 2.8$
	Low disease activity	$2.8 < x \leq 10$
	Moderate disease activity	$10 < x \leq 22$
	High disease activity	$x > 22$
DAS28-ESR$_4$[353]	Remission	$x \leq 2.6$
	Low disease activity	$2.6 < x \leq 3.2$
	Moderate disease activity	$3.2 < x \leq 5.1$
	High disease activity	$x > 5.1$

DAS28-ESR$_4$ is disease activity score of 28 joints including ESR (four variables); SDAI is simplified disease activity index; CDAI is clinical disease activity index.

5.2. Exploration of clinical dataset

In this section, the scores are presented using scatterplots (Figure 22), which indicate a positive relationship between the scores. The correlation between the scores is reported using a correlation matrix plot (Figure 23) and the distributions of the variables are explored using histograms and density plots (Figure 24). The correlation between the responses from the SDAI and CDAI is 0.9, the correlation between the scores from SDAI and DAS28-ESR$_4$ is 0.9 and correlation between the scores from DAS28-ESR$_4$ and CDAI is 0.8 (Figure 23). These values indicate a strong correlation between the scores.

From the histograms and the density plots (Figure 24), the response from the scores are unimodal, steeped bell-shaped.

Figure 22: Scatterplots of DAS28-ESR4, SDAI and CDAI

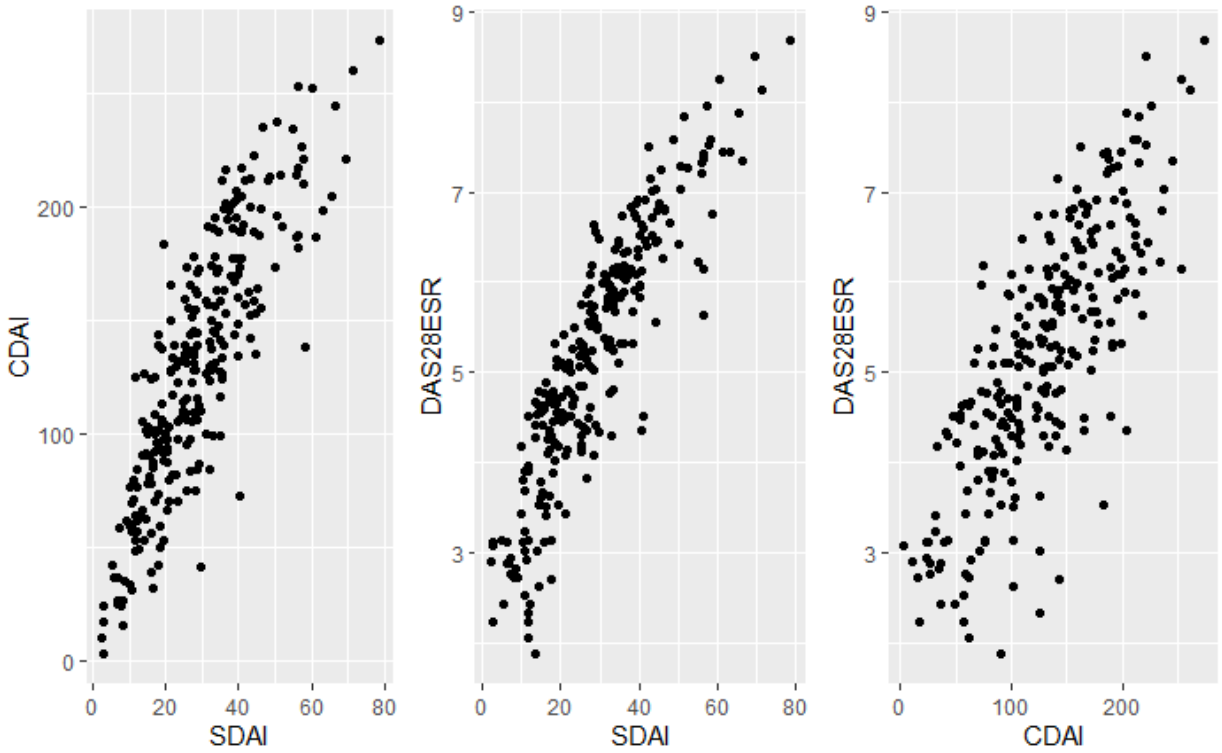


Figure 23: Correlation matrix plot of DAS28-ESR4, SDAI and CDAI

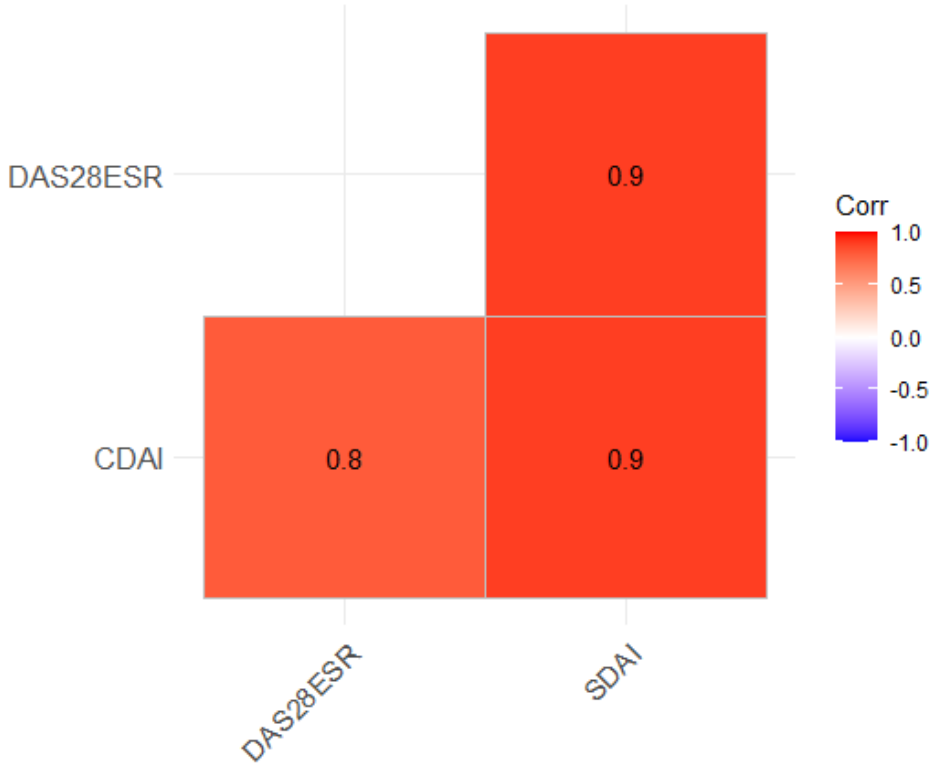
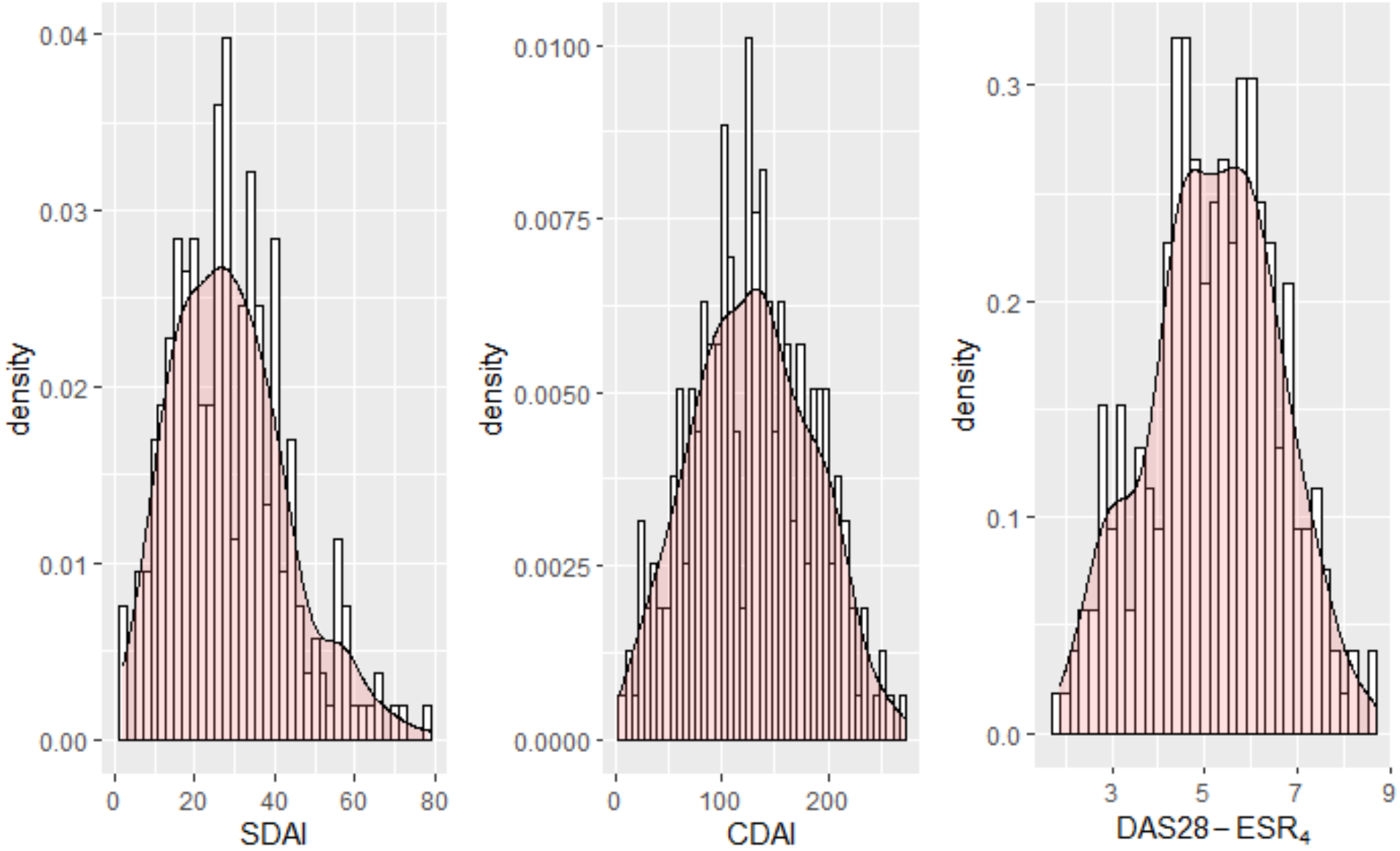


Figure 24: Histogram and density plot of SDAI, CDAI and DAS28-ESR$_4$



5.3. Aims of the clinical dataset analysis

To estimate the sensitivity and specificity of DAS28-ESR$_4$, CDAI and SDAI in discriminating between RA patients who are in remission and non-remission (which include low, moderate and high disease activity group) using the baseline data only. In estimating these accuracy measures, different model assumptions were employed to help understand the potential bias that the variations in the model assumptions may have on how the scores should be used in practice. The different model assumptions are:

- Estimating the sensitivity and specificity of SDAI and CDAI under the assumption that DAS28-ESR$_4$ is a gold standard.
- Estimate the sensitivity and specificity of SDAI and CDAI under the assumption that DAS28-ESR$_4$, SDAI and CDAI are conditionally independent and none of the scores are a gold standard.
- Estimate the sensitivity and specificity of DAS28-ESR, SDAI and CDAI under the assumption that the three tests are conditionally dependent and none of the scores are a gold standard.

5.4. Methodology for analysing the clinical datasets

Firstly, under the assumption that DAS28-ESR$_4$ is a gold standard, the classical approach is employed to estimate the sensitivity and specificity of SDAI and CDAI in discriminating between RA patients in remission and non-remission. This is the standard assumption often presumed in some literatures where SDAI and CDAI have been evaluated[356, 357, 362, 363]; hence this assumption is explored in this analysis.

In order to build on the findings of previous research studies, the three scores – DAS28-ESR$_4$, CDAI and SDAI – are considered to be correlated, and DAS28-ESR$_4$ is not a gold standard when evaluated with other devices such as x-ray [21, 354, 362-364]. Thus, the accuracy measures of SDAI, CDAI and DAS28-ESR$_4$ were estimated assuming that none of these scores is a gold standard while taking into consideration the correlations that exist among the scores.

To account for the correlations that exist among the scores, given that none of the scores are a gold standard, the Bayesian LCM was employed. Therefore, Bayesian LCM [140, 197] is employed to analyse the clinical dataset to overcome the non-identifiability problem which comes about because of the correlation among the scores.

The priors employed in this study were extracted from previous published research studies (see section **5.4.1**).

The MCMC[322] method (via ROpenBUGS[324, 344, 345]) was used to obtain the posterior distributions of the sensitivity and specificity of the individual scores. These estimates were then used to make inference on the sensitivity and specificity of the scores at baseline. The models were assessed to ensure that the posterior distributions of the parameters converged and were compared using the DIC. For the purpose of our study both informative and non-informative priors were employed. To elicit these priors, the SHELF[328, 330] was used.

5.4.1. Prior information on the disease activities scores (SDAI, CDAI and DAS28-ESR$_4$)

Prior information on the sensitivity and specificity of SDAI, CDAI and DAS28-ESR$_4$ were obtained from previous research studies. A literature review was undertaken for the purpose of this thesis to identify published articles with sensitivity and specificity of the DAS28-ESR$_4$, SDAI and CDAI. This review considered studies published up until January 2020. The key criterion for selection of articles from the searched articles was the study population; as studies whose study population was homogenous to the RA-MAP baseline clinical data were selected. The sensitivity and specificity of the evaluated scores (DAS28-ESR$_4$, SDAI, CDAI) to discriminate between RA patients at remission and non-remission were extracted from the selected articles. The sensitivity and specificity of the scores from selected published articles are reported in Table 37.

Limitations of the prior information obtained via literature review

Outlined below are the limitations experienced in the selection of articles used to specify the priors employed in analysing the clinical data using the Bayesian LCM approach.

- Different reference standards were used in different research studies to evaluate the test performance of these scores. Most studies used DAS28-ESR$_4$ as a reference standard to evaluate SDAI and CDAI[356, 357, 362, 363, 365, 366] while others used expert's opinions[359, 367] or other medical devices like ultrasound[368], x-ray[369] or Boolean scores[370] to evaluate SDAI, CDAI and DAS28-ESR$_4$. This meant that if there are two identified studies with the same population as the RA-MAP dataset and both studies used different reference standard, the datasets from the two identified studies cannot be pooled together. Therefore, the study with the larger

sample size is chosen. Choosing a study with the larger sample size reduces the uncertainty around the estimates obtained. For example, Ben Abdelghani et al[369] and Legrand et al[368] were two research studies identified from the review whose populations are homogenous to the RA-MAP baseline data. However, Ben Abdelghani et al[369] and Legrand et al[368] used ultrasound and x-ray stability as their reference standard respectively. The datasets from both studies cannot be combined because of the different reference standards used. Hence, the Legrand et al[368] study was chosen because it had a larger sample size (n = 133) than Ben Abdelghani et al[369] (n = 62).

Table 37: Prior information on the sensitivity and specificity of DAS28-ESR$_4$, SDAI and CDAI

Authors	Population (disease level)	Sample size	Index test (cut-off)	Reference standard	Sensitivity (95% CI)	Specificity (95% CI)
Legrand et al[368]	Newly recruited RA patients (Rem.)	133	SDAI (3.3)	X-ray stability	0.39 (0.28, 0.50)*	0.86 (0.75, 0.97)*
			CDAI (2.8)	X-ray stability	0.4 (0.29, 0.51)*	0.84 (0.72, 0.99)*
Ben Abdelghani et al[369]	Newly recruited RA patients (Rem.)	62	DAS28-ESR$_4$ (2.6)	Ultrasound	0.81 (0.67, 0.96)*	0.63 (0.50, 0.76)*

*Some articles did not publish their 95% confidence interval (CI) for their sensitivity and specificity obtained. Hence, the 95% CIs were estimated for those articles using the simple asymptotic method[371] and the information within the articles. * is used to depict the 95% CI which I calculated. Rem. implies remission. SDAI is Simplified Disease Activity Index, CDAI is Clinical Disease Activity Index and DAS28-ESR is the Disease Activity Score for 28 joints estimated using ESR (Erythrocyte sedimentation rate). HDA is high disease activity. RA is rheumatoid arthritis.

Specification of prior information

In this section, the specification of the prior information (probabilistic prior information) on the parameters of interest, which are the sensitivities and specificities of the DAS28-ESR$_4$, SDAI and CDAI, is detailed.

- The information reported in Table 37 was used to obtain the prior distributions for the sensitivity and specificity of the DAS28-ESR, SDAI and CDAI using the quartile method[372] in the SHELF[319, 328, 330].
- The estimated sensitivity and specificity ($\hat{\theta}$) of the scores from the identified articles served as the mean values and were used to estimate the standard error of the estimates using the normal approximation to the binomial distribution[373, 374].

$$SE(\hat{\theta}) = \sqrt{\frac{\hat{\theta}\,(1-\hat{\theta})}{n}};\ ;$$

Here n can be either the total number of positives as determined using the reference standard when the sensitivity is estimated or the total number of negatives as determined using the reference standard when specificity is estimated. The lower quartile (25% or Q1) and upper quartile (75% or Q3) were estimated using the formula below[375, 376]:

$$Q_1 = SE * Z_1 + \hat{\theta} \quad and \quad Q_3 = SE * Z_3 + \hat{\theta};\quad Z_1 = -0.67 \quad and \quad Z_3 = 0.67$$

The Z values within this formula are from the standard normal distribution for 25% and 75% quartiles.

- The estimated sensitivity and specificity from the identified articles were used as the median value (50%), and their estimated quartiles ($Q_1 and\ Q_3$) were used as the lower and upper quartiles in the SHELF elicitation application to elicit their Beta distributions (see Table 38).

Table 38: Median, and quartiles values of sensitivity and specificity used to elicit the prior Beta distribution

Variables	Q_1	Median	Q_3	Elicited Beta distribution
Baseline Remission				
Sens. DAS28-ESR$_4$	0.7531	0.813	0.8729	$Beta\ (15.2, 3.68)$
Spec. DAS28-ESR$_4$	0.5817	0.631	0.6803	$Beta\ (27.6, 16.20)$
Sensitivity SDAI	0.3565	0.391	0.4254	$Beta\ (35.8, 55.70)$
Specificity SDAI	0.8210	0.857	0.8928	$Beta\ (36.3, 6.26)$
Sensitivity CDAI	0.3654	0.4	0.4346	$Beta\ (36.6, 54.80)$
Specificity CDAI	0.7993	0.837	0.8747	$Beta\ (35.9, 7.20)$

SDAI is Simplified Disease Activity Index; CDAI is Clinical Disease Activity Index and DAS28-ESR is the Disease Activity Score for 28 joints estimated using ESR (Erythrocyte sedimentation rate); Q_1 is lower quartile and Q_3 is upper quartile.

5.5. Analysis of the RA-MAP baseline clinical data

In this section, the baseline data are analysed under the three assumptions described in section **5.3**. The baseline data are the score responses collected on the newly recruited RA patients. The score responses were used to classify participants into remission and non-remission. The number of participants classified into remission and non-remission groups using the standard cut-off for remission on DAS28-ESR$_4$, SDAI and CDAI (in Table 36) is reported in Table 39. This dataset will be referred to as **RABR** (rheumatoid arthritis baseline dataset to classify RA patients into remission or non-remission). In the baseline, there are extremely high numbers of RA patients whose disease activity is above remission, i.e. there is a low prevalence of remission among the participants.

Table 39: Number of participants classified as being in remission and non-remission in RABR

Scores	Cut-off Remission	Number of patients in remission	Number of patients in non-remission
DAS28-ESR$_4$	2.6	8	256
SDAI	3.3	4	260
CDAI	2.8	4	260

SDAI is Simplified Disease Activity Index; CDAI is Clinical Disease Activity Index and DAS28-ESR is the Disease Activity Score for 28 joints estimated using ESR (Erythrocyte sedimentation rate).

Furthermore, in the RABR dataset, let 1 denote a test response that is below the cut-off (remission) and 0 denote a test response that is above the cut-off (non-remission). Therefore, the number of RA patients based on the combination of the test responses from the baseline data is presented in Table 40 (RABR). The RABR table shows that 253 participants have been classified as "non-remission" by DAS28-ESR, SDAI and CDAI, and one participant has been classified as remission by all three tests.

Table 40: Combination of DAS28-ESR$_4$, CDAI and SDAI responses using RABR dataset

DAS28-ESR$_4$	SDAI	CDAI	Observed Frequency
1	1	1	1
1	1	0	0
1	0	1	0
1	0	0	7
0	1	1	3
0	1	0	0
0	0	1	0
0	0	0	253

SDAI is Simplified Disease Activity Index; CDAI is Clinical Disease Activity Index and DAS28-ESR is the Disease Activity Score for 28 joints estimated using ESR (Erythrocyte sedimentation rate).

5.5.1. Analysis of the baseline data assuming DAS28-ESR$_4$ is a gold standard

Firstly, the sensitivity and specificity of SDAI and CDAI was estimated using DAS28-ESR$_4$ as the reference standard (and a gold standard). The results obtained are presented in Table 41. The prevalence of RA patients in remission is 0.03.

Table 41: Estimated sensitivity and specificity of SDAI and CDAI in the RABR dataset assuming that the DAS28-ESR$_4$ is a gold standard

Diagnostic accuracy measures	SDAI Estimates (95% CI)	CDAI Estimates (95% CI)
Sensitivity	0.125 (0, 0.354)	0.125 (0, 0.354)
Specificity	0.988 (0.975, 1)	0.988 (0.975, 1)

SDAI is Simplified Disease Activity Index; CDAI is Clinical Disease Activity Index; CI means confidence interval

Assuming that DAS28-ESR$_4$ is a gold standard in classifying which patients are in and out of remission, the SDAI and CDAI have excellent specificity close to one (0.99) and very poor sensitivity (0.13). Both results are identical because both tests are highly correlated with correlation value as 0.9 (see Figure 23). The strong correlation between the two scores is in line with other researchers' findings about the scores[13, 356, 364, 377, 378].

5.5.2. Analysis of the baseline data assuming SDAI, CDAI and DAS28-ESR$_4$ are conditionally independent and none of the scores is a gold standard

Secondly, the sensitivity and specificity of SDAI, CDAI and DAS28-ESR4 were estimated under the assumption that none of the scores are a gold standard and all scores are conditionally independent given the true disease status using the traditional LCM (TLCM). The TLCM does not need informative priors as the model is identifiable when the number of tests to evaluate is more than two[120, 169]. Non-informative (flat Beta distribution – Beta (1, 1)) priors were employed to obtain the posterior distributions of the parameters of interest. The estimated sensitivities and specificities of the three scores are presented in Table 42.

Table 42: Sensitivity and specificity of DAS28-ESR$_4$, CDAI and SDAI in the RABR dataset assuming that all scores are conditionally independent.

Diagnostic accuracy measures	Mean (Standard deviation)
Sensitivity DAS28 – ESR$_4$	0.333 (0.178)
Specificity DAS28 – ESR$_4$	0.97 (0.011)
Sensitivity SDAI	0.808 (0.158)
Specificity SDAI	0.996 (0.004)
Sensitivity of CDAI	0.808 (0.158)
Specificity of CDAI	0.996 (0.004)
Prevalence	0.02 (0.009)
DIC	21.71

SDAI is Simplified Disease Activity Index; CDAI is Clinical Disease Activity Index; DAS28-ESR is the Disease Activity Score for 28 joints estimated using ESR (Erythrocyte sedimentation rate); DIC is deviance information criterion.

Under the conditional independence assumption, the results (Table 42) showed that CDAI and SDAI have sensitivities that are as good as (0.81) and better than the sensitivity of DAS28-ESR (0.33). The specificity of all the scores are high (i.e. greater than 0.9) when discriminating between patients in remission.

Diagnostics of TLCM

Multiple MCMC chains (three) were used to run the analysis, and a total of 40000 iterations were run for each chain. The number of thinning interval was three. This is to ensure that the posterior distribution of the parameters converged. To check for convergence of the posterior distributions, firstly, the trace plot for each parameter was assessed to ensure that the chains overlapped and that they were caterpillar-shaped. Secondly, the autocorrelation plot was used to ascertain that the sampling autocorrelation is around zero. Finally, the Gelman and Rubin shrinkage factors ($\hat{R}$) of each parameter was assessed to ensure that they were less than 1.01[379]. The shrinkage factors of all the parameters is one. The diagnostic plots of each analysis is in the appendix section of the thesis (Appendix D.1).

5.5.3. Analysis of the baseline data assuming SDAI, CDAI and DAS28-ESR$_4$ are conditionally dependent and none of the scores are a gold standard

The RABR dataset was analysed under the assumption that none of the scores are a gold standard while taking into account the conditional dependencies among the scores using the Bayesian LCM (fixed effect LCM by Wang et al (FEM$_W$) and random effect LCM via logit link – REML). In this case, informative priors reported in Table 38 were employed to analyse the dataset. The informative priors are provided again below, and the results from this analysis are reported in Table 43.

Sensitivity

- $DAS28 - ESR_4 \sim Beta\ (15.2, 3.68)$
- $SDAI \sim Beta(35.8, 55.70)$
- $CDAI \sim Beta(36.6, 54.80)$

Specificity:

- $DAS28 - ESR_4 \sim \text{Beta}\ (27.6, 16.20)$
- $SDAI \sim Beta(36.3, 6.26)$
- $CDAI \sim Beta(35.9, 7.20)$

Table 43: Sensitivity and specificity of DAS28-ESR$_4$, CDAI and SDAI in the RABR dataset assuming that all scores are conditionally dependent.

Diagnostic accuracy measures	FEM$_W$ Mean (SD)	REML Mean (SD)
Sensitivity DAS28 – ESR$_4$	0.68 (0.09)	0.76 (0.10)
Specificity DAS28 – ESR$_4$	0.93 (0.02)	0.92 (0.02)
Sensitivity SDAI	0.40 (0.05)	0.40(0.05)
Specificity SDAI	0.98 (0.01)	0.97 (0.01)
Sensitivity of CDAI	0.41 (0.05)	0.40 (0.05)
Specificity of CDAI	0.97 (0.01)	0.97 (0.01)
Prevalence	0.02 (0.01)	0.02 (0.01)
Deviance	56.40 (9.45)	141.31 (5.81)
pD	1.86	16.86
DIC	58.26	158.17

SDAI is Simplified Disease Activity Index; CDAI is Clinical Disease Activity Index; DAS28-ESR is the Disease Activity Score for 28 joints estimated using ESR (Erythrocyte sedimentation rate); DIC is deviance information criterion.

From Table 43, the estimated specificities of the three tests are above 0.9, showing excellent ability to rule in RA patients in remission or non-remission. The sensitivities of SDAI and CDAI are poor (< 0.5) and the sensitivity of DAS28 – ESR_4 is within the range of 0.7 and 0.8, which is quite good. Hence, making DAS28-ESR_4 a preferred choice of score among the three evaluated scores via these methods. The density plots of the priors and posterior distribution of the sensitivities and specificities of the three scores estimated from the REM_W are displayed in Figure 25 and Figure 26 respectively.

Figure 25: Density plots of the prior and posterior distribution of the sensitivities of DAS-ESR, SDAI and CDAI using the FEM_W on the RABR dataset

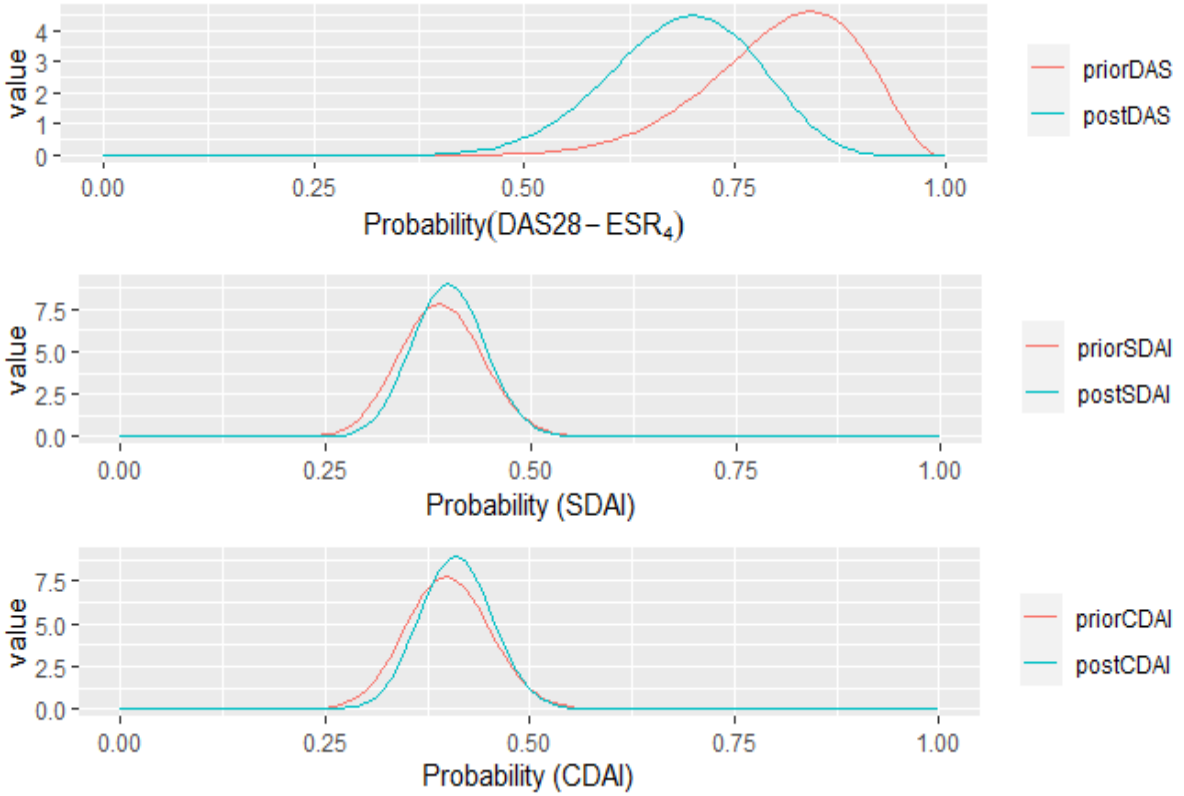
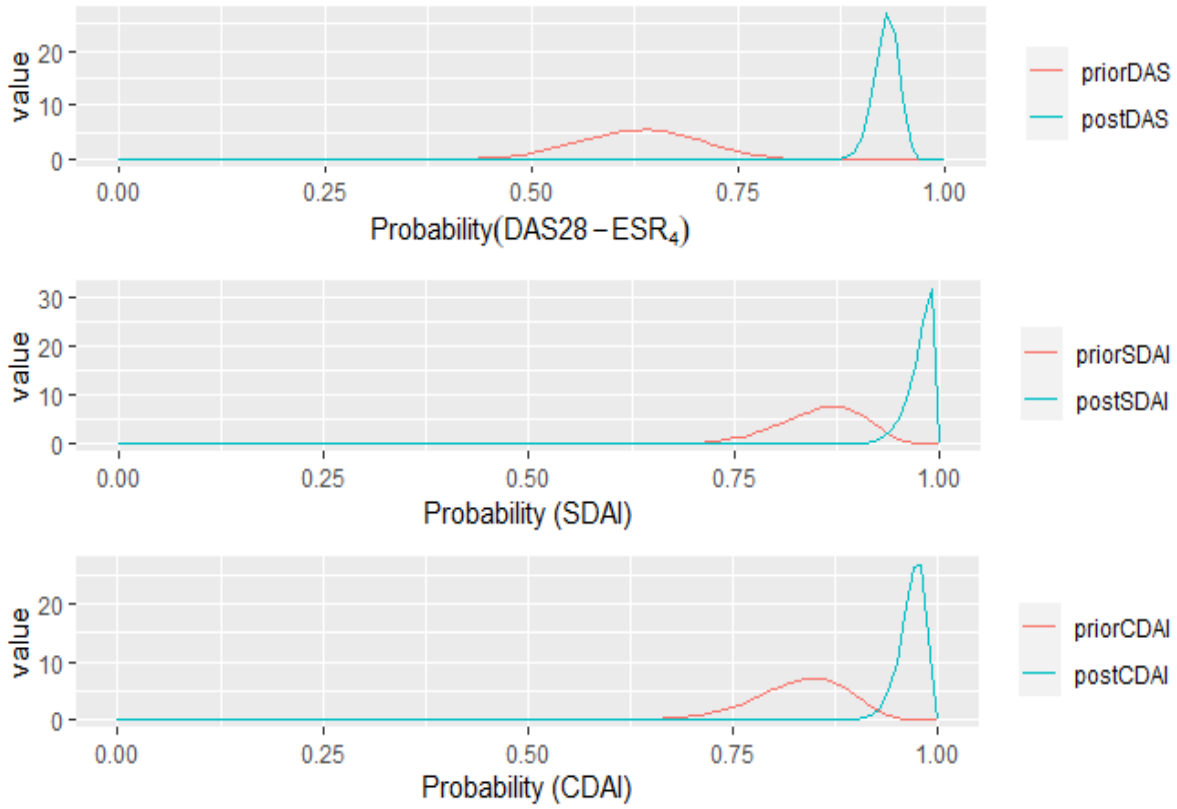


Figure 26: Density plots of the prior and posterior distribution of the specificities of DAS-ESR, SDAI and CDAI using the FEMW on the RABR dataset



The density plots of the priors and posterior distribution of the sensitivities and specificities of the three scores estimated via the REML is similar to FEM_W, and they are displayed in Appendix D.2. From the plots, the specificities of all the scores are not centred on the choice of priors.

Diagnostics of the FEM_W and REML

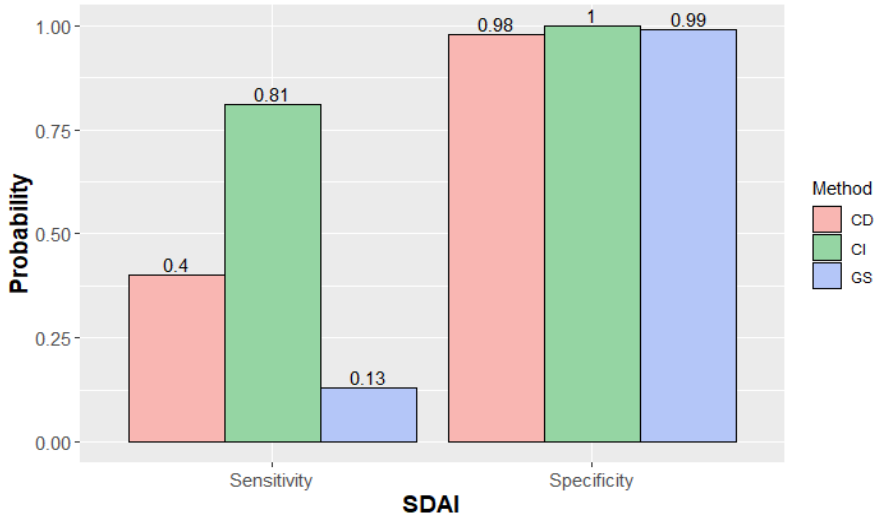
The FEM_W was implemented via the RStan program. 2,000,000 iterations were run for three chains each. 2000 iterations was used as the warm-up. The REML was analysed using Openbugs. 40,000 iterations were run for three chains each. All the parameters in the models converged to their posterior distributions. The convergence of the model was assessed using the trace plots, autocorrelation plots and the Gelman and Rubin diagnostic plots. These plots are presented in Appendix D.3, and the Gelman and Rubin shrinkage factor ($\hat{R}$) for each parameter is one, indicating a convergence of the model.

Comparison of estimates obtained from various assumptions

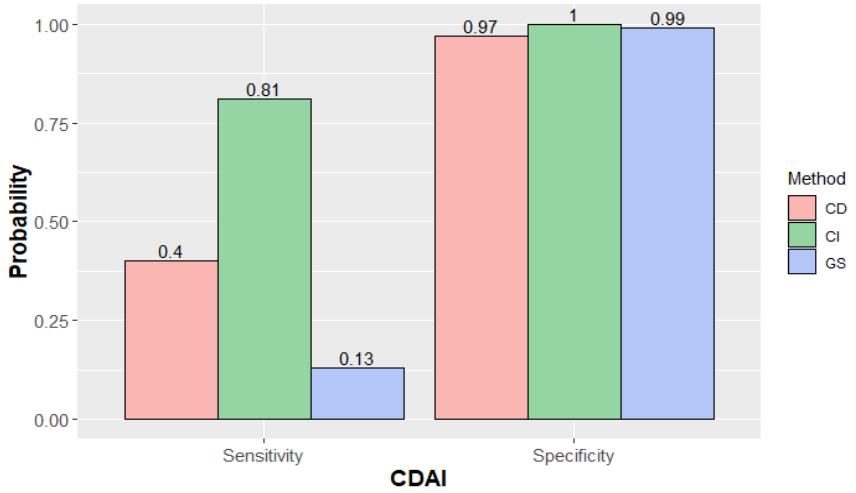
Analysing the RABR dataset under different assumptions produced different estimates. These different estimates are displayed in Figure 27. There are differences in the estimates obtained due to the different assumptions used in the analysis of the dataset. This is more evident among the estimated sensitivities of the scores. This could be because of the low prevalence of the RA patients in remission in the RABR dataset, which is 0.03. Hence, the specificities seem to be quite consistent across all assumptions (> 0.90). The estimates of SDAI and CDAI (across all assumptions) are similar because as mentioned earlier, previous research studies have shown that both scores are highly correlated. The estimated sensitivity of SDAI and CDAI (0.81) under the conditional independence assumption is larger than the estimated sensitivity of SDAI and CDAI (0.40) under the conditional dependence assumption, and the estimated sensitivity of DAS28-ESR_4 (0.33) under the conditional independence assumption is smaller than its estimated sensitivity ($\cong 0.7$) under the conditional dependence assumption. This could be because of the conditional dependence that exists among the scores. Therefore, not accounting for the conditional dependence between SDAI and CDAI overestimates the sensitivities of the two scores and underestimates the sensitivity of DAS28-ESR_4. However, taking into consideration the conditional dependence between the scores reduces the sensitivities of SDAI and CDAI and increases the sensitivity of DAS28-ESR_4. Therefore, if the assumptions used here hold, DAS28-ESR_4 is a preferred test among the three tests to classify newly diagnosed RA patients into remission or non-remission. However, this analysis and comparison highlights the importance of exploring the conditional dependence between tests rigorously to inform the choice of method used to evaluate the sensitivity and specificity of multiple tests in the absence of a gold standard.

Figure 27: Estimated sensitivities and specificities of DAS28-ESR$_4$, SDAI and CDAI under the different assumptions (RABR dataset)

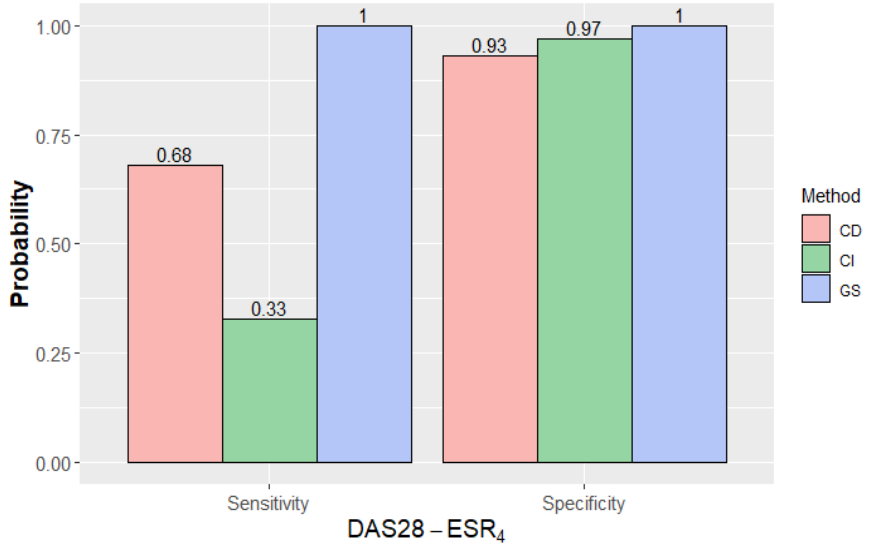
(a)



(b)



(c)



The blue bar is the estimated sensitivity and specificity when DAS28-ESR$_4$ is employed as a gold standard (GS). The green bar is the sensitivity and specificity when all the tests evaluated are assumed to be conditionally independent (CI) and the pink bar is the sensitivity and specificity when all the tests are assumed to be conditionally dependent (CD) and none of the scores is a gold standard.

Sensitivity analysis

As part of Bayesian analysis, sensitivity analysis is often carried out to check how changes in the model inputs affect the posterior inferences made. These include changes in the prior information, or the sampling distribution used. To check how sensitive the FEM_W and REML are in estimating the sensitivity and specificity of SDAI, CDAI and DAS28-ESR_4, different priors for the sensitivities of the three scores were employed. The prior distributions used for the sensitivities of the three tests were centred on 0.99, and the prior distributions of the specificities of the scores were centred on 0.4. The beta distributions are:

Sensitivity:

- $DAS28-ESR_4 \sim Beta(113.45, 0.419)$
- $SDAI \sim Beta(113.45, 0.419)$
- $CDAI \sim Beta(113.45, 0.419)$

Specificity:

- $DAS28-ESR_4 \sim Beta\ (2, 3)$
- $SDAI \sim Beta\ (2\ , 3)$
- $CDAI \sim Beta\ (2\ , 3)$

The choice of these prior distributions is arbitrary. The estimated sensitivity and specificity of SDAI, CDAI and DAS28-ESR_4 via the FEM_W and REML models are presented in Table 44. The density plots of the prior and posterior distributions of the sensitivities and specificities are presented in Figure 28 and Figure 29 respectively. From Table 44, and Figure 28, the choice of priors impacted the estimated sensitivities of the three scores, which indicates that the FEM_W and REML model are sensitive to change. This also implies that the posterior inference about the sensitivities of the scores is not robust[316, 380]. The specificities of the scores are unaffected by the changes (Figure 29) as their posterior distributions are not impacted by the choice of priors. The REML and FEM_W were assessed for convergence and their diagnostic plots are reported in Appendix D.4. In addition, the Gelman and Rubin shrinkage factor for all the parameters is one.

Table 44: Sensitivity and specificity of DAS28-ESR$_4$, CDAI and SDAI in the RABR dataset assuming that all scores are conditionally dependent (sensitivity analysis).

Diagnostic accuracy measures	FEM$_W$ Mean (SD)	REML Mean (SD)
Sensitivity DAS28 – ESR$_4$	0.98 (0.00)	0.99 (0.01)
Specificity DAS28 – ESR$_4$	0.96 (0.00)	0.96 (0.06)
Sensitivity SDAI	0.98 (0.00)	1 (0.00)
Specificity SDAI	0.98 (0.00)	0.98 (0.04)
Sensitivity of CDAI	0.98 (0.00)	1 (0.00)
Specificity of CDAI	0.98 (0.00)	0.98 (0.05)
Prevalence	0.01 (0.00)	0.03 (0.09)

SDAI is Simplified Disease Activity Index; CDAI is Clinical Disease Activity Index; DAS28-ESR is the Disease Activity Score for 28 joints estimated using ESR (Erythrocyte sedimentation rate).

Figure 28: Density plots of the prior and posterior distribution of the sensitivities of DAS-ESR, SDAI and CDAI using the FEM$_W$ on the RABR dataset

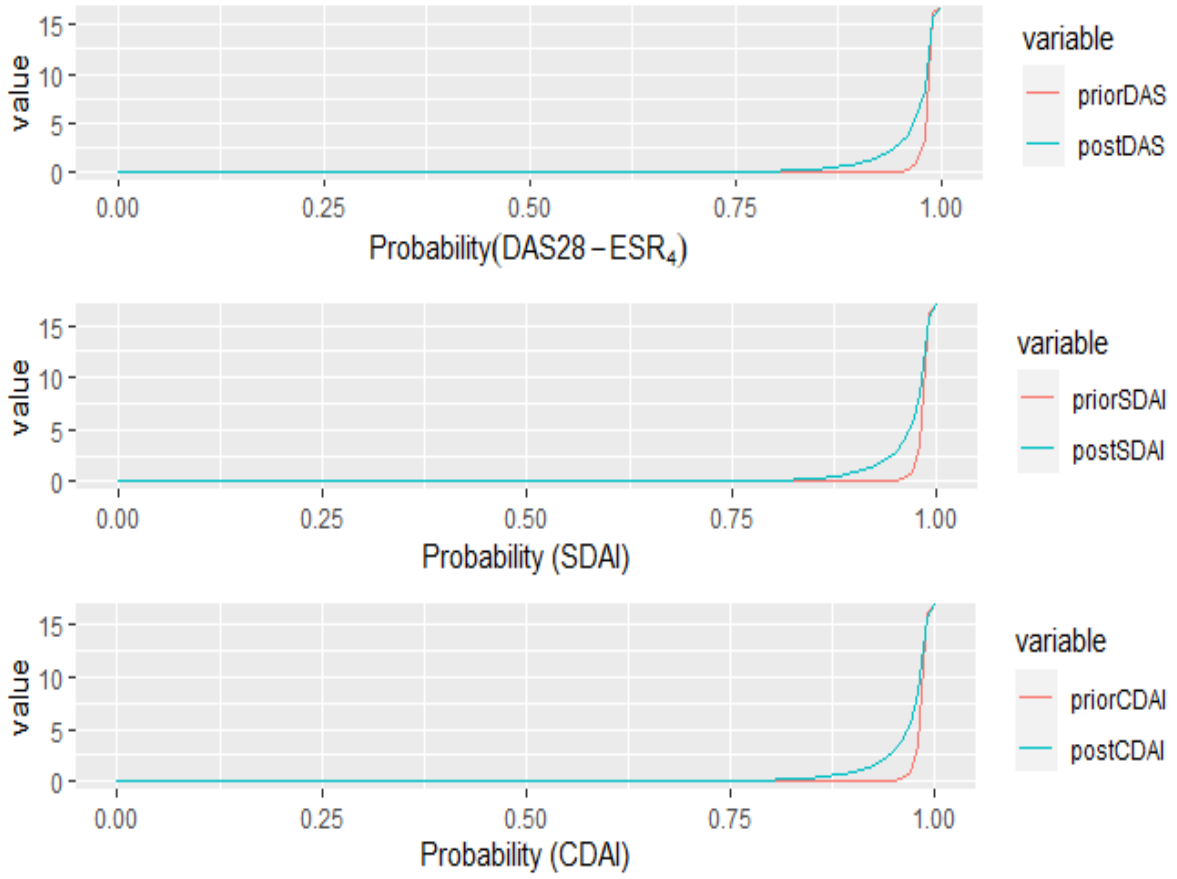
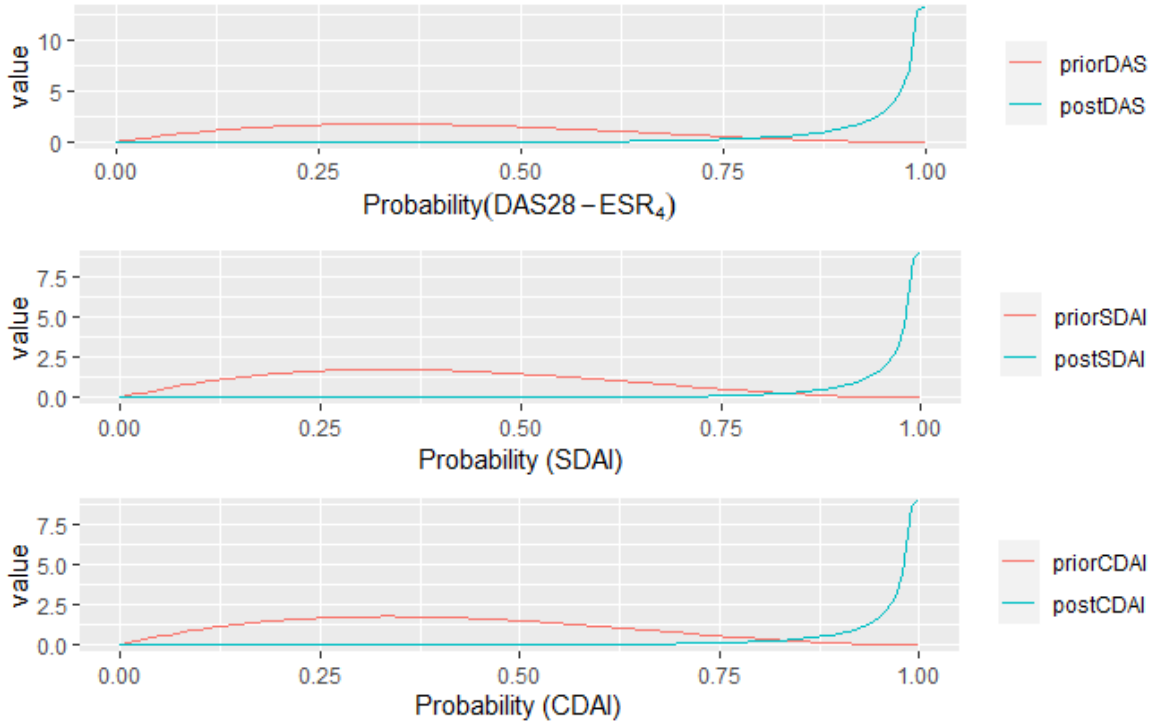


Figure 29: Density plots of the prior and posterior distribution of the specificities of DAS-ESR, SDAI and CDAI using the FEMW on the RABR dataset



5.6. Discussion

From the analysis of the clinical dataset under varying assumptions, it is obvious that various assumptions made about diagnostic tests employed in a diagnostic accuracy study affect the estimated accuracy measures obtained, like the assumptions of conditional dependence or independence or the assumption that one of the tests employed in the study is a gold standard. When the DAS28-ESR is employed as a gold standard to evaluate the diagnostic accuracy of CDAI and SDAI (which is an assumption made by some of the literature[356, 357, 362, 363, 365, 366]) the estimated sensitivity of SDAI and CDAI are very low. This indicates that neither measure is good in discriminating between rheumatoid arthritis patients in remission or non-remission at baseline compared to when the three tests are assumed to be imperfect (either conditionally independent or not). The high sensitivities of DAS28-ESR, SDAI and CDAI under the assumption of conditional independence could be an indication that the three tests are conditionally dependent among the diseased group. Hence, not accounting for the conditional dependence among the tests overestimates the sensitivity of the three scores, which is in line with our findings in chapter four and the study by Vacek [54]. Furthermore, the estimated specificities of the scores are robust.

A limitation is that there are only three tests to evaluate and these tests are conditionally dependent given the true disease status. Thus, the number of parameters to estimate is more than the degrees of freedom. Hence, the model is non-identifiable. Using informative priors is essential to make the model identifiable and the choice of informative priors could impact the posterior distributions. Consequently, increasing the number of tests to evaluate could reduce the problem of model identifiability, especially if included tests are conditionally independent of the other tests considered.

Another limitation is that the number of participants classified to be in remission by all scores at baseline is very small (there was only one participant in this category). This is expected because the baseline dataset is made up of newly diagnosed RA patients. Hence, the sensitivities of the scores are poor, volatile and highly sensitive to the choice of priors. Moreover, this could limit the utility of the results in clinical practice to populations homogenous to this case-study (RA-MAP baseline dataset).

Finally, the estimates obtained may not be transferable to other clinical datasets and the values could change if data from other samples from another population was used.

Comparison of latent class models

The TLCM, which models the three scores (DAS28-ESR, SDAI and CDAI) under the assumption that all the scores are conditionally independent given the true disease status were compared to the FEM_W and REML, which assumed that the scores are conditionally dependent given the true disease status. The deviance and DICs were used to compare the LCMs. The DIC and deviance from TLCM was smaller than the deviance and DIC from the conditional dependence models. However, following the background knowledge of the tests and the recommendation by Spiegelhalter et al [336] that "the DIC should not be used as a strict criterion for model choice", the conditional dependence model was chosen. Comparing the conditional dependence LCMs, the deviance and the DIC from the FEM_W model is smaller than the REML. This is because the FEM_W is considered as a simpler model than the REML which is a complex hierarchical model. The R code employed to analyse the dataset is reported in Appendix D.5.

In conclusion, different assumptions are made when evaluating medical tests in diagnostic accuracy studies; some of these assumptions are not testable statistically such as conditional dependence when the true disease status of the participants is unknown or assuming the reference standard employed is a gold standard. However,

these are assumptions that are based on the biological knowledge of the medical tests and they are employed to help in estimating the sensitivity and specificity accurately. Therefore, to carry out a diagnostic accuracy study, it is important to explore and research to understand the possible correlation that could exist among the medical tests being evaluated especially if there is no gold standard. It is also important to, where possible, specify the conditional dependence accurately if any exists. The clinical analysis carried out in this chapter demonstrates the effect of different assumptions on the estimations of the diagnostic accuracy of tests. These estimates can impact on the utility of a test in practice and where the test will be used to rule in or rule out diagnosis. Therefore, it is important to ensure the assumptions are valid.

Chapter Six: General Discussion

Conventionally, the sensitivity and specificity of an index test are evaluated by comparing the index test with the best available reference standard when the true disease status of each participant is unknown, and the reference standard is often assumed to be perfect (i.e. having 100% sensitivity and specificity). However, in reality the reference standard could be imperfect. Hence, evaluating the diagnostic accuracy of an index test without taking into consideration the imperfection of the reference standard leads to biased estimates of the sensitivity and specificity of the index test. This bias may overestimate the true sensitivity and specificity of the test under evaluation. Ignoring this bias may mean that the test is introduced into routine practice under the false assumption that it can accurately rule in or rule out disease. These assumptions can adversely affect a patient's health.

From the test developer's perspective, a miscalculation of the diagnostic accuracy of the test may result in developmental revenue being misdirected to or away from the test in question, representing a waste of scarce research resources. From the patients' and clinician's perspective, this could mean a misdiagnosis which could potentially lead to use of treatments that are not effective and may even be harmful or conversely, it could mean patients are falsely reassured as not having the disease, leading to potentially harmful delays in treatment. From the government perspective, it could lead to misappropriation of public funds in funding or buying an inaccurate medical intervention while ignoring an effective medical intervention. Topically, from a public health perspective, misestimating the accuracy of an index test for an infection could lead to failure in controlling outbreaks of infectious disease and inaccurate measures of the incidence rate.

6.1. Summary of the research study

Firstly, this work reviewed previous research studies that proposed methods and diagnostic accuracy studies to evaluate the test performance of an index test in the absence of a gold standard (summarised in chapter two). Several methods were identified from the review which were classified in four groups which are: methods employed when only a sub-sample of the participants have their disease status verified with the gold standard (group 1); correction methods (group 2); methods using multiple imperfect reference standards (group 3) and other methods (group 4) such as study of agreement, test positivity rate and alternative study designs like validation.

Following the results from the review, some identified methods which are commonly used in diagnostic accuracy studies and have not been compared in previous research studies were compared. The compared methods were the corrections methods by Brenner[117] and Staquet et al[119]. Both methods are employed to correct the sensitivity and specificity of an index test provided that the reference standard and the index test are conditionally independent, and the accuracy of the reference standard are known. These methods were compared via simulation studies and reanalysis of clinical datasets (chapter three). Combining the observations and results from the simulations studies and the clinical datasets, the Staquet et al correction method outperforms the Brenner correction method. Based on the findings from comparing the correction methods, and the systematic review carried out in chapter two, alongside the nature of the RA-MAP clinical dataset, there was a need to explore and investigate LCMs with various conditional dependence structures. Therefore, in chapter four, various LCMs with different conditional dependence structures were investigated. The investigation of these LCMs led to the selection of LCMs deemed best to analyse the RA-MAP clinical dataset. Finally, in chapter five, the RA-MAP clinical dataset was analysed using the some of the LCMs explored in chapter four.

6.2. Contributions of the research study

The miscalculation of the sensitivity and specificity of an index test by test evaluators or researchers could affect the adoption of an index test into routine practice or the continued use of an existing test. In addition, it could be both harmful to patients directly and indirectly (through the wasted use of scarce health care resources). My research has addressed this issue and made the following contributions as outlined in the following paragraphs:

Firstly, the systematic review conducted provided a list of methods developed to evaluate the accuracy measures of the index test in the absence of a gold standard. Clinical application studies where the identified methods have been applied were also cited as real-life examples to aid researchers and test evaluators in understanding how identified methods are applied. The strengths and weakness of the identified methods were also highlighted in this review to aid researchers and test evaluators in the choice of appropriate method for their research question. Based on the findings of the review, a guidance flowchart of the choice of methods to consider was constructed to guide test evaluators and researchers on available methods to consider when evaluating an index test in the absence of a gold standard.

This systematic review was an update of the review published in 2007 by Rutjes et al[381] (whose search was conducted up to 2005). New methods were identified by my updated review such as the dual composite reference standard by Tang et al[162]. More examples of studies describing clinical applications were also identified, and the guidance flowchart was reconstructed and expanded to include the extra methods identified in the review. There could be a variety of methods to consider when evaluating an index test in the absence of a gold standard. Thus, choosing the best method among the available methods is another question test evaluators and researchers need to consider when estimating the sensitivity and specificity of an index test. Hence, the guidance flowchart reported in chapter two will help guide researchers in the choice of which method is applicable to their scenario.

Chapter three compared some of the correction methods identified from the systematic review. These correction methods (Brenner[117] and Staquet et al[119]) are commonly used in diagnostic accuracy studies and no previous research was identified that directly compared the statistical performance of these methods. Both correction methods (Brenner [117] and Staquet et al[119] correction methods) evaluate the diagnostic accuracy of a test given that the accuracy measures of the reference standard is known and the reference standard and index test are conditionally independent. These methods were investigated to identify how they perform at estimating the sensitivity and specificity of the index test accurately irrespective of the prevalence of the target condition. This is the first study that I am aware of that has compared these methods. From this comparison, an algebraic expression was used to show that the estimated sensitivity or specificity from the corrected and unadjusted methods are similar when the reference standard is perfect. This algebraic expression showed that no matter the method employed, the estimates obtained are unbiased and the same when the reference standard is perfect or a gold standard. From this work, it was observed that the Staquet et al method maintains the assumption of constant test characteristics (sensitivity and specificity) across populations with different prevalences of disease. Implying that, irrespective of the prevalence in that target condition, the Staquet et al method is more robust in estimating the accuracies of the index test than the unadjusted and Brenner correction method. The exception to this, is that the Staquet et al method might provide illogical estimates when the prevalence is very high or very low. Therefore, based on these findings, the use of Brenner correction method in

correcting for the sensitivity and specificity of a medical test in diagnostic accuracy study should be discontinued.

In chapter four, various LCMs with differing conditional dependence structure (i.e. FEM, REM and FMM) were investigated using simulation studies. The purpose of this was to identify which LCM to consider when there are three tests to evaluate and all three tests are conditionally dependent given the true disease status. From the simulation studies, the REM via logit link (REML) and the FEM by Wang et al[286] (FEM_W), are less affected by the choice of prior distributions used. This makes them good options for LCMs to consider. However, it was observed from this work that conditional independence among the diseased (non-diseased) group in the FEM_W implies that the sensitivities (specificities) of the evaluated tests need to be close to one such that the covariances between the tests are insignificant or approximately zero else the FEM_W may not estimate the sensitivity and specificity accurately.

Finally, chapter five provides an example of a clinical application of the FEM proposed by Wang et al[286] (FEM_W). In addition, this is the first study, that I am aware of, that estimated the sensitivity and specificity of the DAS28-ESR_4, SDAI and CDAI under the assumption that all scores are correlated and that none of the scores are a gold standard. No published research study that I have been able to identify has estimated the accuracy measures of these scores taking into consideration the correlation that exist among them and some studies have estimated the accuracy measures of SDAI and CDAI using DAS28-ESR_4 as a gold standard, or estimated the accuracy measures of DAS28-ESR_4 using another medical tool as a gold standard (see section **5.4**). Such studies found out the sensitivity of SDAI and CDAI were poor (<0.4) using newly recruited RA patients and their specificities were very good (> 0.8). The decision to analyse the RA-MAP baseline dataset using these assumptions was made based on the findings from previous studies and is described in section **5.4**. Analysing the RA-MAP clinical dataset under these assumptions showed that CDAI and SDAI have poor sensitivity (< 40%), whilst DAS28-ESR_4 has a better sensitivity (0.7), and the DAS28-ESR_4, SDAI and CDAI have excellent ability to rule in the classification of remission in RA patients as their specificities are above 0.9. Therefore, making all three scores preferred tests for classifying newly diagnosed RA patients into remission and non-remission in practice.

6.3. Strength of the research study

The strengths of this research study are discussed in line with the evaluative criteria by Lincoln and Guba[382] but applying this idea to quantitative research. This includes the credibility, applicability, consistency and neutrality of the research study.

Firstly, the systematic review was carried out following the PRISMA[58] guidance. All databases related to medicine (both veterinary and human studies) were searched. Searching through all databases related to medicine enabled a wider number of relevant articles to be included in the review. Appropriate keywords such as imperfect, no or absence of gold standard, diagnostic accuracy, sensitivity, and specificity amongst others (see section **2.2.2**) were used to identify as many potentially relevant articles for the review as possible. Articles obtained from the databases were screened by three reviewers independently (including myself) to ensure that the articles included in the review matched the inclusion criteria and relevant articles were not missed. Undertaking the systematic review through this process ensured that the results and conclusions obtained are original and unbiased. Moreover, this approach ensures that the results obtained are not subjective, but a critical evaluation of different approaches proposed.

Secondly, the simulation studies were performed using the suggested guidelines by Morris et al[285]. These guidelines provide step by step recommendations to ensure that a simulation study is well-thought out, well-planned and executed accurately so that the results are reproducible, and the aim is achieved. In following these guidelines, the aim of my simulation studies was defined a priori, and the theories and underlying assumptions associated to each method and model were taken into consideration when simulating the datasets and analysing the simulated datasets. For example, in chapter three where two correction methods were compared: firstly, the aim I desired to achieve from the simulation studies was well-defined. For this, some of the datasets were simulated under the assumption of conditional independence, which is the key assumption underlying the development of Brenner and Staquet et al correction methods. The processes followed for the simulations were well-detailed as were the results obtained to ensure reproducibility. The process of generating the datasets as well as analysing the generated datasets were performed using the RStudio[341] statistical software and the OpenBugs[342] software, which are recognised statistical software employed to carry out this kind of statistical analysis. The codes used to simulate and analyse the datasets were peer reviewed in detail by one of my

supervisors (KW) to ensure there were no errors. The data-generating and data-analysis processes were repeated more than three times on two occasions to ensure that the results obtained are consistent and can be reproduced using the methods outlined in this study. Furthermore, the codes for the simulation and analysis of the data are reported in Appendix B.1, C.1, C.2 and D.5 so that others can critique in detail my methods and findings.

I have also used real world clinical datasets to support my findings from the simulation studies in chapter three and to provide a clinical application of the LCMs explored in chapter four. Using the real-world datasets is quite different from simulated datasets as they reflect the potential challenges or errors that can occur in the process of collecting the dataset, such as missing data. Although, this was not discussed in this thesis, it is a research to be carried out in the future. In addition, using real-world datasets can either support the findings from a simulation study or provide an avenue or scenarios where the findings from the simulation study can be flawed in real-time. The clinical dataset employed in this research was the RA-MAP clinical dataset. This clinical dataset was analysed using latent class models, which was deemed the most appropriate method. The LCMs considered the correlation that exists among the scores being evaluated. Various assumptions about the conditional dependence of the scores were considered, while the DIC from the different LCMs was explored. The parameters priors were obtained from peer-reviewed studies, whose populations matched the population associated with the clinical dataset. The SHELF[328, 330] elicitation web-app was employed to elicit the priors for the parameter of interest and this was cross-checked manually using the information from the peer reviewed articles. Using the SHELF provided a verification of the priors I obtained using the manual or analytic approach. In addition, it is a well-used tool for the elicitation of prior distributions for Bayesian analysis. Again, the Openbugs[324, 342] and RStudio[341] statistical software were employed to analyse the dataset. As with the analyses reported in chapter 4, the dataset was analysed three times to ensure reproducibility of the results obtained. Finally, critical thinking, findings from this study, and information from previous research studies were employed to provide recommendations and suggestions.

6.4. Limitations of the research study

The limitations of this study include: firstly, the focus of the study is on estimating the sensitivity and specificity of index test in the absence of a gold standard. There are other clinically important sensitivity and specificity employed to evaluate an index test, such as PPV and NPV (see section **1.1.5**) which are not discussed in this research study. Secondly, in the systematic review, only peer-reviewed published articles which proposed methods and diagnostic accuracy studies which estimate the sensitivity and specificity of index tests in the absence of a gold standard were identified. Other methods related to diagnostic accuracy such as estimating differences in sensitivity and specificity were not explored. Thirdly, in the comparison of methods identified from the systematic review, only the correction methods were compared based on their statistical properties. There were other methods identified in the review such as methods employed when a subsample of the participants does not undergo the gold standard. However, all these methods could not be explored due to the limited timeframe for this research. In addition, there could be other possible combinations of prevalence, sensitivity and specificity of RS and index test not explored in this study, the R-Code in Appendix **B.1** will aid readers to explore more. Fourthly, in the simulation study undertaken to explore various LCMs (chapter four), the generated datasets are limited to three tests. In addition, there are other possible combinations where the values for the sensitivity and specificity of the threes tests and the covariance terms are different from what is explored in chapter four. Thus, the comparison and inferences made in chapter four are by necessity limited.

Furthermore, a limitation with the analysis of the RA-MAP clinical dataset is that the three scores evaluated are correlated and the number of participants classified in remission by all the three scores was very small (one). This implied that, the estimated values of some of the parameters relied strongly on the informative priors used. This was shown in the sensitivity analysis carried out on the RA-MAP dataset. The estimated sensitivities of the scores were highly sensitive to changes in their prior distributions. However, the specificities of the scores were robust and insensitive to the changes in their prior distributions. In addition, the pattern of missingness of the RA-MAP clinical data was not taken into consideration when analysing the clinical dataset. This is a focus for future research.

Finally, this research study was impacted by COVID-19 which prevented the proposed group elicitation workshop from happening. The workshop was designed to work with

clinical rheumatologists in Newcastle and the North-East of England to elicit the parameters of interest. An aim of the workshop was to elicit prior distributions for the sensitivities and specificities of the three scores (SDAI, CDAI and DAS28-ESR_4) in discriminating between newly recruited RA patients in HDA and non-HDA at the standard cut-off. This information could not be obtained from published articles as some of the published article used cut-offs that were different from the standard cut-offs.

6.5. Further research

This work is only one step in explorations in this area. The following are notable next steps to follow on from this work:

- To investigate if statistical conditional independence implies biological conditional independence. From chapter three, where two correction methods were compared via simulation studies and clinical datasets, the assumption of conditional dependence or independence on the simulated datasets is done via mathematical model. However, in practice, the assumption of conditional independence or dependence is made on the biological component of the tests evaluated in the clinical datasets which may or may not follow the mathematical model.

- To explore the reason for obtaining illogical results via Staquet et al approach and exploring its impact in predicting if the index test and the reference standard are conditionally dependent given the true disease status. Being able to propose a test to check for conditional dependence between two tests will help to verify such an assumption and provide a guide on which statistical methods to employ in a diagnostic accuracy study. Furthermore, as observed from the analysis of the clinical datasets (Table 14 and Table 15), an illogical value was obtained for the estimated prevalence, which could have impacted the estimated specificities of the index tests. Thus, the Staquet et al approach could be further explored to ascertain other conditions that can make the Staquet et al approach produce illogical estimates, as well as possible implications where multiple conditions are satisfied simultaneously.

- To reanalyse the RA-MAP clinical dataset to adjust for missing data and explore if the missing data could affect the estimates obtained. Reanalysing the RA-MAP dataset taking into consideration the missingness of the missing data could change the posterior inference made about the scores.

- To develop a web-based application including most or all of the identified methods in the systematic review to encourage test evaluators and researchers to explore the various methods developed.

6.6. Conclusion

A variety of methods have been proposed to evaluate an index test in the absence of a gold standard. Some of these methods have been applied in diagnostic accuracy studies like the composite reference standard and Bayesian LCMs, amongst others, and some have not been applied like some of the Bias-correction approaches[62, 63] developed to estimate the accuracy measures of the index test when only a subsample of the participants undergo the gold standard. Some are computationally complex like the Bayesian LCMs and some require only basic mathematical knowledge like the Staquet et al correction method. However, all of these methods are built on underlying assumptions, which, if violated in a clinical dataset, could result in inaccurately estimated sensitivity and specificity, which could have dire consequences. This research has not only provided a guidance flowchart to help researchers decide on what choice of method to employ in their study. It has also compared some of these methods in diagnostic accuracy studies to explore how each method performs in different clinical scenarios and under different assumptions. This comparison and application to a clinical dataset will help researchers and test evaluators to select the best method to employ amongst other options.

Knowing and using the right method will provide appropriate estimates of the diagnostic accuracy of an index test, which can be used alongside other analysis related to the index test such as an economic evaluation to decide if the test could be adopted into clinical routine practice.

References

1. Department of Health AG. Clinical Evidence Guidlines (Medical Devices). In: Administration DoHTG, (ed.). Version 1.0 ed. Australian: Australian Government 2017, p. 165.
2. Graziadio S, Winter A, Lendrem BC, et al. How to ease the pain of taking a diagnostic point of care test to the market: a framework for evidence development. *Micromachines* 2020; 11: 291.
3. Bossuyt PM, Irwig L, Craig J, et al. Comparative accuracy: Assessing new tests against existing diagnostic pathways. *British Medical Journal* 2006; 332: 1089-1092. Note.
4. Deeks J, Takwoingi Y, Leeflang M, et al. Use of medical tests. Lesson 1.1: Cochrane Collaboration DTA Online Learning Materials. *The Cochrane Collaboration,September 2014Videocast (31 slides, 29 minutes, sound, colour) Available* < http://trainingcochraneorg> 2014.
5. Grimes DA and Schulz KF. Uses and abuses of screening tests. *Lancet* 2002; 359: 881-884. Review. DOI: 10.1016/S0140-6736(02)07948-5.
6. Deeks JJ. Systematic reviews of evaluations of diagnostic and screening tests. *BMJ* 2001; 323: 157. Article. DOI: 10.1136/bmj.323.7305.157.
7. Partridge EE, Abu-Rustum N, Giuliano A, et al. Cervical cancer screening. *Journal of the National Comprehensive Cancer Network* 2014; 12: 333-341.
8. Campbell JM, Klugar M, Ding S, et al. Diagnostic test accuracy: Methods for systematic review and meta-analysis. *International Journal of Evidence-Based Healthcare* 2015; 13: 154-162. Article. DOI: 10.1097/XEB.0000000000000061.
9. Locantore P, Ianni F and Pontecorvi A. Fine needle aspiration biopsy. *Minimally Invasive Therapies for Endocrine Neck Diseases*. Springer, 2016, pp.15-24.
10. Rector TS, Taylor BC and Wilt TJ. Chapter 12: Systematic review of prognostic tests. *Journal of General Internal Medicine* 2012; 27: S94-S101. Review. DOI: 10.1007/s11606-011-1899-y.
11. van de Vijver MJ. Molecular tests as prognostic factors in breast cancer. *Virchows Archiv* 2014; 464: 283-291.
12. Nicholson AG, Chansky K, Crowley J, et al. The International Association for the Study of Lung Cancer Lung Cancer Staging Project: proposals for the revision of the clinical and pathologic staging of small cell lung cancer in the forthcoming eighth edition of the TNM classification for lung cancer. *Journal of Thoracic Oncology* 2016; 11: 300-311.

13. Aletaha D and Smolen JS. The Simplified Disease Activity Index (SDAI) and Clinical Disease Activity Index (CDAI) to monitor patients in standard clinical care. *Best Practice and Research: Clinical Rheumatology* 2007; 21: 663-675. Review. DOI: 10.1016/j.berh.2007.02.004.

14. Brundage JF and Rubertone MV. Medical Surveillance Monthly Report: The first 20 years. *MSMR* 2015; 22: 2.

15. Reynolds D and Angevine E. Hand-arm vibration, part II: Vibration transmission characteristics of the hand and arm. *Journal of sound and vibration* 1977; 51: 255-265.

16. Glasziou P, Irwig L and Mant D. Monitoring in chronic disease: a rational approach. *Bmj* 2005; 330: 644-648.

17. McKenna RJ. Clinical aspects of cancer in the elderly. Treatment decisions, treatment choices, and follow-up. *Cancer* 1994; 74: 2107-2117.

18. Rosselli Del Turco M, Palli D, Cariddi A, et al. Intensive diagnostic follow-up after treatment of primary breast cancer: a randomized trial. *JAMA-Journal of the American Medical Association-US Edition* 1994; 271: 1593-1597.

19. Goldberg RM, Fleming TR, Tangen CM, et al. Surgery for recurrent colon cancer: strategies for identifying resectable recurrence and success rates after resection. *Annals of internal medicine* 1998; 129: 27-35.

20. Qiao Y-Y, Lin K-X, Zhang Z, et al. Monitoring disease progression and treatment efficacy with circulating tumor cells in esophageal squamous cell carcinoma: A case report. *World Journal of Gastroenterology: WJG* 2015; 21: 7921.

21. Aletaha D and Smolen JS. Diagnosis and management of rheumatoid arthritis: a review. *Jama* 2018; 320: 1360-1372.

22. Sunjaya AF and Sunjaya AP. Pooled Testing for Expanding COVID-19 Mass Surveillance. *Disaster Medicine and Public Health Preparedness* 2020: 1-2.

23. Knottnerus JA, Van Weel C and Muris JWM. Evaluation of diagnostic procedures. *BMJ* 2002; 324: 477-480. Article. DOI: 10.1136/bmj.324.7335.477.

24. Eusebi P. Diagnostic Accuracy Measures. *Cerebrovascular Diseases* 2013; 36: 267-272. DOI: 10.1159/000353863.

25. Mallett S, Halligan S, Matthew Thompson GP, et al. Interpreting diagnostic accuracy studies for patient care. *BMJ (Online)* 2012; 345. Review. DOI: 10.1136/bmj.e3999.

26. Hawkins DM, Garrett JA and Stephenson B. Some issues in resolution of diagnostic tests using an imperfect gold standard. *Statistics in Medicine* 2001; 20: 1987-2001. Article. DOI: 10.1002/sim.819.

27. Naaktgeboren CA, De Groot JAH, van Smeden M, et al. Evaluating Diagnostic Accuracy in the Face of Multiple Reference Standards. *Annals of Internal Medicine* 2013; 159: 195-+. DOI: 10.7326/0003-4819-159-3-201308060-00009.

28. Whiting PF, Rutjes AWS, Westwood ME, et al. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. *Journal of Clinical Epidemiology* 2013; 66: 1093-1104. DOI: 10.1016/j.jclinepi.2013.05.014.

29. Cardoso JR, Pereira LM, Iversen MD, et al. What is gold standard and what is ground truth? *Dental Press Journal of Orthodontics* 2014; 19: 27-30. DOI: 10.1590/2176-9451.19.5.027-030.ebo.

30. Bachmann LM, Jüni P, Reichenbach S, et al. Consequences of different diagnostic 'gold standards' in test accuracy research: Carpal Tunnel Syndrome as an example. *International journal of epidemiology* 2005; 34: 953-955.

31. Agin MA. Gold Standard. *Wiley Encyclopedia of Clinical Trials*. John Wiley & Sons, Inc., 2007.

32. Albert PS and Dodd LE. A Cautionary Note on the Robustness of Latent Class Models for Estimating Diagnostic Error without a Gold Standard. *Biometrics* 2004; 60: 427-435. DOI: 10.1111/j.0006-341X.2004.00187.x.

33. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Bmj-British Medical Journal* 2015; 351. DOI: 10.1136/bmj.h5527.

34. Wong HB and Lim GH. Measures of diagnostic accuracy: Sensitivity, specificity, PPV and NPV. *Proceedings of Singapore Healthcare* 2011; 20: 316-318. Note. DOI: 10.1177/201010581102000411.

35. Asif N, Ijaz A, Rafi T, et al. Diagnostic Accuracy of Serum Iron and Total Iron Binding Capacity (TIBC) in Iron Deficiency State. *Journal of the College of Physicians and Surgeons Pakistan* 2016; 26: 958-961. Article.

36. Sayão LB, Britto MCAD, Burity E, et al. Exhaled nitric oxide as a diagnostic tool for wheezing in preschool children: A diagnostic accuracy study. *Respiratory Medicine* 2016; 113: 15-21. Article. DOI: 10.1016/j.rmed.2016.02.008.

37. Mordiffi SZ, Goh ML, Phua J, et al. Confirming nasogastric tube placement: Is the colorimeter as sensitive and specific as X-ray? A diagnostic accuracy study. *International Journal of Nursing Studies* 2016; 61: 248-257. Article. DOI: 10.1016/j.ijnurstu.2016.06.011.

38. Perez-Warnisher MT, Gomez-Garcia T, Giraldo-Cadavid LF, et al. Diagnostic accuracy of nasal cannula versus microphone for detection of snoring. *Laryngoscope* 2017; 127: 2886-2890.

39. Glueck DH, Lamb MM, O'Donnell CI, et al. Bias in trials comparing paired continuous tests can cause researchers to choose the wrong screening modality. *Bmc Medical Research Methodology* 2009; 9. DOI: 10.1186/1471-2288-9-4.

40. De Groot JAHB, P. M. M.; Reitsma, J. B.; Rutjes, A. W. S.; Dendukuri, N.; Janssen, K. J. M.; Moons, K. G. M. Verification problems in diagnostic accuracy studies: Consequences and solutions. *BMJ (Online)* 2011; 343. Article. DOI: 10.1136/bmj.d4770.

41. Rutjes AW, Reitsma JB, Coomarasamy A, et al. Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health technology assessment (Winchester, England)* 2007; 11: iii, ix-51. Review.

42. Harel O and Zhou XH. Multiple imputation for correcting verification bias (vol 25, pg 3769, 2006). *Statistics in Medicine* 2008; 27: 4614-4615. DOI: 10.1002/sim.3322.

43. Albert PS. Imputation Approaches for Estimating Diagnostic Accuracy for Multiple Tests from Partially Verified Designs. *Biometrics* 2007; 63: 947-957. DOI: 10.1111/j.1541-0420.2006.00734.x.

44. Karch A, Koch A, Zapf A, et al. Partial verification bias and incorporation bias affected accuracy estimates of diagnostic studies for biomarkers that were part of an existing composite gold standard. *Journal of Clinical Epidemiology* 2016; 78: 73-82. Article. DOI: 10.1016/j.jclinepi.2016.03.022.

45. De Groot JAHD, N.; Janssen, K. J. M.; Reitsma, J. B.; Bossuyt, P. M. M.; Moons, K. G. M. Adjusting for differential-verification bias in diagnostic-accuracy studies: A Bayesian approach. *Epidemiology* 2011; 22: 234-241.

46. Thompson M and Van den Bruel A. Sources of Bias in Diagnostic Studies. *Diagnostic Tests Toolkit*. Wiley-Blackwell, 2011, pp.26-33.

47. De Groot JAH, Dendukuri N, Janssen KJM, et al. Adjusting for differential verification bias in diagnostic accuracy studies: A bayesian approach. *American Journal of Epidemiology* 2010; 11): S140. Conference Abstract.

48. Alonzo TA, Brinton JT, Ringham BM, et al. Bias in estimating accuracy of a binary screening test with differential disease verification. *Statistics in Medicine* 2011; 30: 1852-1864. DOI: 10.1002/sim.4232.

49. Kivlin D, Lim C, Ross C, et al. The Diagnostic and Treatment Patterns of Urologists in the United States for Interstitial Cystitis/Painful Bladder Syndrome. *Urology Practice* 2016; 3: 309-314. Article. DOI: 10.1016/j.urpr.2015.08.005.

50. Bogart LM, Suttorp MJ, Elliott MN, et al. Validation of a quality-of-life scale for women with bladder pain syndrome/interstitial cystitis. *Quality of Life Research* 2012; 21: 1665-1670. Article. DOI: 10.1007/s11136-011-0085-3.

51. Tirlapur SA, Priest L, Wojdyla D, et al. Bladder pain syndrome: Validation of simple tests for diagnosis in women with chronic pelvic pain: BRaVADO study protocol. *Reproductive Health* 2013; 10. Article. DOI: 10.1186/1742-4755-10-61.

52. Adams HJA, Kwee TC and Nievelstein RAJ. Influence of imperfect reference standard bias on the diagnostic performance of MRI in the detection of lymphomatous bone marrow involvement. *Clinical Radiology* 2013; 68: 750-751. DOI: 10.1016/j.crad.2013.01.022.

53. Lu Y, Dendukuri N, Schiller I, et al. A Bayesian approach to simultaneously adjusting for verification and reference standard bias in diagnostic test studies. *Statistics in Medicine* 2010; 29: 2532-2543. Article. DOI: 10.1002/sim.4018.

54. Vacek PM. The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics* 1985; 41: 959-968.

55. Garner P, Hopewell S, Chandler J, et al. When and how to update systematic reviews: consensus and checklist. *Bmj-British Medical Journal* 2016; 354. DOI: 10.1136/bmj.i3507.

56. Moher D, Tsertsvadze A, Tricco AC, et al. When and how to update systematic reviews. *Cochrane Database of Systematic Reviews* 2008. DOI: 10.1002/14651858.MR000023.pub3.

57. Chikere CMU, Wilson K, Graziadio S, et al. Diagnostic test evaluation methodology: A systematic review of methods employed to evaluate diagnostic tests in the absence of gold standard–An update. *PloS one* 2019; 14.

58. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ (Clinical research ed)* 2009; 339. Article. DOI: 10.1136/bmj.b2700.

59. Sayers A. Tips and tricks in performing a systematic review. *Br J Gen Pract* 2008; 58: 136-136.

60. ResearchSoft TI. EndNote (Version 8.0). *Berkeley, CA: Author* 2004.

61. Alonzo TA and Pepe MS. Assessing accuracy of a continuous screening test in the presence of verification bias. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2005; 54: 173-190. DOI: 10.1111/j.1467-9876.2005.00477.x.

62. Duc KT, Chiogna M and Adimari G. Bias–corrected methods for estimating the receiver operating characteristic surface of continuous diagnostic tests. *Electronic Journal of Statistics* 2016; 10: 3063-3113. Article. DOI: 10.1214/16-EJS1202.

63. Zhang Y, Alonzo TA and for the Alzheimer's Disease Neuroimaging I. Inverse probability weighting estimation of the volume under the ROC surface in the presence of verification bias. *Biometrical Journal* 2016; 58: 1338-1356. Article. DOI: 10.1002/bimj.201500225.

64. Begg CB and Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* 1983; 39: 207-215. Article.

65. Harel OZ, X. H. Multiple imputation for correcting verification bias. *Statistics in Medicine* 2006; 25: 3769-3786. DOI: 10.1002/sim.2494.

66. He H and McDermott MP. A robust method using propensity score stratification for correcting verification bias for binary tests. *Biostatistics* 2012; 13: 32-47. Article. DOI: 10.1093/biostatistics/kxr020.

67. Zhou XH. Maximum likelihood estimators of sensitivity and specificity corrected for verification bias. *Communications in Statistics - Theory and Methods* 1993; 22: 3177-3198. Article. DOI: 10.1080/03610929308831209.

68. Kosinski AS and Barnhart HX. Accounting for nonignorable verification bias in assessment of diagnostic tests. *Biometrics* 2003; 59: 163-171. Article. DOI: 10.1111/1541-0420.00019.

69. Kosinski AS and Barnhart HX. A global sensitivity analysis of performance of a medical diagnostic test when verification bias is present. *Statistics in Medicine* 2003; 22: 2711-2721. Article. DOI: 10.1002/sim.1517.

70. Martinez EZAA, J.; Louzada-Neto, F. Estimators of sensitivity and specificity in the presence of verification bias: A Bayesian approach. *Computational Statistics and Data Analysis* 2006; 51: 601-611. Article. DOI: 10.1016/j.csda.2005.12.021.

71. Buzoianu M and Kadane JB. Adjusting for verification bias in diagnostic test evaluation: A Bayesian approach. *Statistics in Medicine* 2008; 27: 2453-2473. Article. DOI: 10.1002/sim.3099.

72. Hajivandi A, Shirazi HRG, Saadat SH, et al. A Bayesian analysis with informative prior on disease prevalence for predicting missing values due to verification

bias. *Open Access Macedonian Journal of Medical Sciences* 2018; 6: 1225-1230. Article. DOI: 10.3889/oamjms.2018.296.

73. Zhou XH. Comparing accuracies of two screening tests in a two-phase study for dementia. *Journal of the Royal Statistical Society Series C: Applied Statistics* 1998; 47: 135-147. Article.

74. Lloyd CJ and Frommer DJ. An application of multinomial logistic regression to estimating performance of a multiple-screening test with incomplete verification. *Journal of the Royal Statistical Society Series C-Applied Statistics* 2008; 57: 89-102. DOI: 10.1111/j.1467-9876.2007.00602.x.

75. Albert PS and Dodd LE. On estimating diagnostic accuracy from studies with multiple raters and partial gold standard evaluation. *Journal of the American Statistical Association* 2008; 103: 61-73. DOI: 10.1198/016214507000000329.

76. Martinez EZ, Achcar JA and Louzada-Neto F. Bayesian estimation of diagnostic tests accuracy for semi-latent data with covariates. *Journal of Biopharmaceutical Statistics* 2005; 15: 809-821.

77. Xue X, Kim MY, Castle PE, et al. A new method to address verification bias in studies of clinical screening tests: Cervical cancer screening assays as an example. *Journal of Clinical Epidemiology* 2014; 67: 343-353. Article. DOI: 10.1016/j.jclinepi.2013.09.013.

78. Walter SD. Estimation of test sensitivity and specificity when disease confirmation is limited to positive results. *Epidemiology* 1999: 67-72.

79. Böhning D and Patilea V. A capture–recapture approach for screening using two diagnostic tests with availability of disease status for the test positives only. *Journal of the American Statistical Association* 2008; 103: 212-221.

80. Chu HZ, Yijie; Cole, Stephen R.; Ibrahim, Joseph G. On the estimation of disease prevalence by latent class models for screening studies using two screening tests with categorical disease status verified in test positives only. *Statistics in Medicine* 2010; 29: 1206-1218. DOI: 10.1002/sim.3862.

81. Baker SG. Evaluating multiple diagnostic tests with partial verification. *Biometrics* 1995; 51: 330-337. Article. DOI: 10.2307/2533339.

82. Van Geloven NB, K. A.; Opmeer, B. C.; Mol, B. W.; Zwinderman, A. H. How to deal with double partial verification when evaluating two index tests in relation to a reference test? *Statistics in Medicine* 2012; 31: 1265-1276.

83. Van Geloven N, Broeze KA, Opmeer BC, et al. Correction: How to deal with double partial verification when evaluating two index tests in relation to a reference test? *Statistics in Medicine* 2012; 31: 3787-3788. Erratum.

84. Aragon DC, Martinez EZ and Alberto Achcar J. Bayesian estimation for performance measures of two diagnostic tests in the presence of verification bias. *Journal of Biopharmaceutical Statistics* 2010; 20: 821-834. Article. DOI: 10.1080/10543401003618868.

85. Gray R, Begg CB and Greenes RA. Construction of receiver operating characteristic curves when disease verification is subject to selection bias. *Medical Decision Making* 1984; 4: 151-164.

86. Zhou XH. A nonparametric maximum likelihood estimator for the receiver operating characteristic curve area in the presence of verification bias. *Biometrics* 1996; 52: 299-305. Article. DOI: 10.2307/2533165.

87. Rodenberg C and Zhou XH. ROC curve estimation when covariates affect the verification process. *Biometrics* 2000; 56: 1256-1262. Review. DOI: 10.1111/j.0006-341X.2000.01256.x.

88. Zhou XH and Rodenberg CA. Estimating an ROC curve in the presence of non-ignorable verification bias. *Communications in Statistics - Theory and Methods* 1998; 27: 635-657. Article. DOI: 10.1080/03610929808832118.

89. Hunink MG, Richardson DK, Doubilet PM, et al. Testing for fetal pulmonary maturity: ROC analysis involving covariates, verification bias, and combination testing. *Medical Decision Making* 1990; 10: 201-211.

90. He HL, Jeffrey M.; McDermott, Michael P. Direct estimation of the area under the receiver operating characteristic curve in the presence of verification bias. *Statistics in Medicine* 2009; 28: 361-376. DOI: 10.1002/sim.3388.

91. Adimari G and Chiogna M. Nearest-neighbor estimation for ROC analysis under verification bias. *International Journal of Biostatistics* 2015; 11: 109-124. Article. DOI: 10.1515/ijb-2014-0014.

92. Adimari G and Chiogna M. Nonparametric verification bias-corrected inference for the area under the ROC curve of a continuous-scale diagnostic test. *Statistics and its Interface* 2017; 10: 629-641. Article. DOI: 10.4310/SII.2017.v10.n4.a8.

93. Gu J, Ghosal S and Kleiner DE. Bayesian ROC curve estimation under verification bias. *Statistics in Medicine* 2014; 33: 5081-5096. Article. DOI: 10.1002/sim.6297.

94. Fluss RR, Benjamin; Faraggi, David; Rotnitzky, Andrea. Estimation of the ROC Curve under Verification Bias. *Biometrical Journal* 2009; 51: 475-490. DOI: 10.1002/bimj.200800128.

95. Rotnitzky A, Faraggi D and Schisterman E. Doubly robust estimation of the area under the receiver-operating characteristic curve in the presence of verification bias. *Journal of the American Statistical Association* 2006; 101: 1276-1288.

96. Fluss R, Reiser B and Faraggi D. Adjusting ROC curves for covariates in the presence of verification bias. *Journal of Statistical Planning and Inference* 2012; 142: 1-11.

97. Liu DZ, Xiao-Hua. A Model for Adjusting for Nonignorable Verification Bias in Estimation of the ROC Curve and Its Area with Likelihood-Based Approach. *Biometrics* 2010; 66: 1119-1128. DOI: 10.1111/j.1541-0420.2010.01397.x.

98. Yu W, Kim JK and Park T. Estimation of area under the ROC Curve under nonignorable verification bias. *Statistica Sinica* 2018; 28: 2149-2166. Article. DOI: 10.5705/ss.202016.0315.

99. Page JH and Rotnitzky A. Estimation of the disease-specific diagnostic marker distribution under verification bias. *Computational Statistics and Data Analysis* 2009; 53: 707-717. Article. DOI: 10.1016/j.csda.2008.06.021.

100. Liu DZ, Xiao-Hua. Covariate Adjustment in Estimating the Area Under ROC Curve with Partially Missing Gold Standard. *Biometrics* 2013; 69: 91-100. DOI: 10.1111/biom.12001.

101. Liu D and Zhou XH. Semiparametric Estimation of the Covariate-Specific ROC Curve in Presence of Ignorable Verification Bias. *Biometrics* 2011; 67: 906-916. Article. DOI: 10.1111/j.1541-0420.2011.01562.x.

102. Yu BZ, Chuan. Assessing the accuracy of a multiphase diagnosis procedure for dementia. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2012; 61: 67-81. DOI: 10.1111/j.1467-9876.2011.00771.x.

103. Chi Y-YZ, Xiao-Hua. Receiver operating characteristic surfaces in the presence of verification bias. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2008; 57: 1-23. DOI: 10.1111/j.1467-9876.2007.00597.x.

104. Duc KT, Chiogna M and Adimari G. Nonparametric Estimation of ROC Surfaces Under Verification Bias. 2016.

105. To Duc K. bcROCsurface: An R package for correcting verification bias in estimation of the ROC surface and its volume for continuous diagnostic tests. *BMC Bioinformatics* 2017; 18. Article. DOI: 10.1186/s12859-017-1914-3.

106. Zhang Y, Alonzo TA and for the Alzheimer's Disease Neuroimaging I. Estimation of the volume under the receiver-operating characteristic surface adjusting for non-ignorable verification bias. *Statistical Methods in Medical Research* 2018; 27: 715-739. Article. DOI: 10.1177/0962280217742541.

107. Zhu R and Ghosal S. Bayesian Semiparametric ROC surface estimation under verification bias. *Computational Statistics and Data Analysis* 2019; 133: 40-52. Article. DOI: 10.1016/j.csda.2018.09.003.

108. To Duc K, Chiogna M, Adimari G, et al. Estimation of the volume under the ROC surface in presence of nonignorable verification bias. *Statistical Methods and Applications* 2019. Article in Press. DOI: 10.1007/s10260-019-00451-3.

109. Lu YD, Nandini; Schiller, Ian; Joseph, Lawrence. A Bayesian approach to simultaneously adjusting for verification and reference standard bias in diagnostic test studies. *Statistics in Medicine* 2010; 29: 2532-2543. DOI: 10.1002/sim.4018.

110. Capelli GN, A.; Nardelli, S.; di Regalbono, A. F.; Pietrobelli, M. Validation of a commercially available cELISA test for canine neosporosis against an indirect fluorescent antibody test (IFAT). *Preventive Veterinary Medicine* 2006; 73: 315-320. DOI: 10.1016/j.prevetmed.2005.10.001.

111. Ferreccio C, Barriga MI, Lagos M, et al. Screening trial of human papillomavirus for early detection of cervical cancer in Santiago, Chile. *International Journal of Cancer* 2012; 132: 916-923. Article. DOI: 10.1002/ijc.27662.

112. Iglesias-Garriz I, Rodríguez MA, García-Porrero E, et al. Emergency Nontraumatic Chest Pain: Use of Stress Echocardiography to Detect Significant Coronary Artery Stenosis. *Journal of the American Society of Echocardiography* 2005; 18: 1181-1186. DOI: https://doi.org/10.1016/j.echo.2005.07.020.

113. Cronin AM and Vickers AJ. Statistical methods to correct for verification bias in diagnostic studies are inadequate when there are few false negatives: A simulation study. *BMC Medical Research Methodology* 2008; 8. Article. DOI: 10.1186/1471-2288-8-75.

114. de Groot JAH, Janssen KJM, Zwinderman AH, et al. Correcting for Partial Verification Bias: A Comparison of Methods. *Annals of Epidemiology* 2011; 21: 139-148. DOI: 10.1016/j.annepidem.2010.10.004.

115. Heida A, Van De Vijver E, Van Ravenzwaaij D, et al. Predicting inflammatory bowel disease in children with abdominal pain and diarrhoea: Calgranulin-C versus calprotectin stool tests. *Archives of Disease in Childhood* 2018; 103: 565-571. Article. DOI: 10.1136/archdischild-2017-314081.

116. Viola A, Fontana A, Belvedere A, et al. Diagnostic accuracy of faecal calprotectin in a symptom-based algorithm for early diagnosis of inflammatory bowel disease adjusting for differential verification bias using a Bayesian approach. *Scandinavian Journal of Gastroenterology* 2020; 55: 1176-1184. Article. DOI: 10.1080/00365521.2020.1807599.

117. Brenner H. Correcting for exposure misclassification using an alloyed gold standard. *Epidemiology* 1996; 7: 406-410. Article.

118. Gart JJ and Buck AA. COMPARISON OF A SCREENING TEST AND A REFERENCE TEST IN EPIDEMIOLOGIC STUDIES .2. A PROBABILISTIC MODEL FOR COMPARISON OF DIAGNOSTIC TESTS. *American Journal of Epidemiology* 1966; 83: 593-&. DOI: 10.1093/oxfordjournals.aje.a120610.

119. Staquet M, Rozencweig M, Lee YJ, et al. Methodology for the assessment of new dichotomous diagnostic tests. *Journal of Chronic Diseases* 1981; 34: 599-610. Article. DOI: 10.1016/0021-9681(81)90059-X.

120. Albert PS. Estimating diagnostic accuracy of multiple binary tests with an imperfect reference standard. *Statistics in Medicine* 2009; 28: 780-797. DOI: 10.1002/sim.3514.

121. Emerson SC, Waikar SS, Fuentes C, et al. Biomarker validation with an imperfect reference: Issues and bounds. *Statistical Methods in Medical Research* 2018; 27: 2933-2945. Article. DOI: 10.1177/0962280216689806.

122. Thibodeau L. Evaluating diagnostic tests. *Biometrics* 1981: 801-804.

123. Hahn AL, Marc; Landt, Olfert; Schwarz, Norbert Georg; Frickmann, Hagen. Comparison of one commercial and two in-house TaqMan multiplex real-time PCR assays for detection of enteropathogenic, enterotoxigenic and enteroaggregative Escherichia coli. *Tropical Medicine & International Health* 2017; 22: 1371-1376. DOI: 10.1111/tmi.12976.

124. Matos RN, T. F.; Braga, M. M.; Siqueira, W. L.; Duarte, D. A.; Mendes, F. M. Clinical performance of two fluorescence-based methods in detecting occlusal caries lesions in primary teeth. *Caries Research* 2011; 45: 294-302. Article. DOI: 10.1159/000328673.

125. Mathews WC, Cachay ER, Caperna J, et al. Estimating the accuracy of anal cytology in the presence of an imperfect reference standard. *PLoS ONE* 2010; 5. Article. DOI: 10.1371/journal.pone.0012284.

126. Hadgu A, Dendukuri N and Hilden J. Evaluation of nucleic acid amplification tests in the absence of a perfect gold-standard test: a review of the statistical and epidemiologic issues. *Epidemiology* 2005: 604-612.

127. Hawkins DMG, J. A.; Stephenson, B. Some issues in resolution of diagnostic tests using an imperfect gold standard. *Statistics in Medicine* 2001; 20: 1987-2001. Article. DOI: 10.1002/sim.819.

128. Hui SL and Zhou XH. Evaluation of diagnostic tests without gold standards. *Statistical Methods in Medical Research* 1998; 7: 354-370. Review.

129. Hagenaars JA. Latent structure models with direct effects between indicators: local dependence models. *Sociological Methods & Research* 1988; 16: 379-405.

130. Uebersax JS. Probit latent class analysis with dichotomous or ordered category measures: Conditional independence/dependence models. *Applied Psychological Measurement* 1999; 23: 283-297.

131. Yang I and Becker MP. Latent variable modeling of diagnostic accuracy. *Biometrics* 1997: 948-958.

132. Qu Y, Tan M and Kutner MH. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics* 1996; 52: 797-810. DOI: 10.2307/2533043.

133. Albert PS, McShane LM, Shih JH, et al. Latent class modeling approaches for assessing diagnostic error without a gold standard: with applications to p53 immunohistochemical assays in bladder tumors. *Biometrics* 2001; 57: 610-619.

134. Zhang BC, Z.; Albert, P. S. Estimating Diagnostic Accuracy of Raters Without a Gold Standard by Exploiting a Group of Experts. *Biometrics* 2012; 68: 1294-1302.

135. Xu HB, Michael A.; Craig, Bruce A. Evaluating accuracy of diagnostic tests with intermediate results in the absence of a gold standard. *Statistics in Medicine* 2013; 32: 2571-2584. DOI: 10.1002/sim.5695.

136. Wang Z, Zhou X-H and Wang M. Evaluation of diagnostic accuracy in detecting ordered symptom statuses without a gold standard. *Biostatistics* 2011; 12: 567-581. DOI: 10.1093/biostatistics/kxq075.

137. Wang ZZ, Xiao-Hua. Random effects models for assessing diagnostic accuracy of traditional Chinese doctors in absence of a gold standard. *Statistics in Medicine* 2012; 31: 661-671. DOI: 10.1002/sim.4275.

138. Liu WZ, B.; Zhang, Z. W.; Chen, B. J.; Zhou, X. H. A pseudo-likelihood approach for estimating diagnostic accuracy of multiple binary medical tests. *Computational Statistics & Data Analysis* 2015; 84: 85-98. DOI: 10.1016/j.csda.2014.11.006.

139. Xue X, Oktay M, Goswami S, et al. A method to compare the performance of two molecular diagnostic tools in the absence of a gold standard. *Statistical Methods in Medical Research* 2019; 28: 419-431. Article. DOI: 10.1177/0962280217726804.

140. Branscum AJ, Gardner IA and Johnson WO. Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling. *Preventive veterinary medicine* 2005; 68: 145-163.

141. Nérette P, Stryhn H, Dohoo I, et al. Using pseudogold standards and latent-class analysis in combination to evaluate the accuracy of three diagnostic tests. *Preventive veterinary medicine* 2008; 85: 207-225.

142. Dendukuri N, Hadgu A and Wang L. Modeling conditional dependence between diagnostic tests: a multiple latent variable model. *Statistics in medicine* 2009; 28: 441-461.

143. Johnson WO, Gastwirth JL and Pearson LM. Screening without a "gold standard": The Hui-Walter paradigm revisited. *American Journal of Epidemiology* 2001; 153: 921-924. Article. DOI: 10.1093/aje/153.9.921.

144. Martinez EZL-N, F.; Derchain, S. F. M.; Achcar, J. A.; Gontijo, R. C.; Sarian, L. O. Z.; Syrjänen, K. J. Bayesian estimation of performance measures of cervical cancer screening tests in the presence of covariates and absence of a gold standard. *Cancer Informatics* 2008; 6: 33-46. Article.

145. Zhang J, Cole SR, Richardson DB, et al. A Bayesian approach to strengthen inference for case-control studies with multiple error-prone exposure assessments. *Statistics in medicine* 2013; 32: 4426-4437.

146. Spiegelhalter DJ, Best NG, Carlin BP, et al. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2002; 64: 583-639.

147. Pereira da Silva HD, Ascaso C, Gonçalves AQ, et al. A Bayesian approach to model the conditional correlation between several diagnostic tests and various replicated subjects measurements. *Statistics in Medicine* 2017; 36: 3154-3170. DOI: 10.1002/sim.7339.

148. Zhou X-HC, Pete; Zhou, Chuan. Nonparametric Estimation of ROC Curves in the Absence of a Gold Standard. *Biometrics* 2005; 61: 600-609. DOI: 10.1111/j.1541-0420.2005.00324.x.

149. Henkelman RM, Kay I and Bronskill MJ. Receiver operator characteristic (ROC) analysis without truth. *Medical Decision Making* 1990; 10: 24-29.

150. Beiden SV, Campbell G, Meier KL, et al. The problem of ROC analysis without truth: The EM algorithm and the information matrix. In: *Medical Imaging 2000: Image Perception and Performance* 2000, pp.126-135. International Society for Optics and Photonics.

151. Choi YK, Johnson WO, Collins MT, et al. Bayesian inferences for receiver operating characteristic curves in the absence of a gold standard. *Journal of Agricultural, Biological, and Environmental Statistics* 2006; 11: 210-229. Article. DOI: 10.1198/108571106X110883.

152. Wang C, Turnbull BW, Gröhn YT, et al. Nonparametric estimation of ROC curves based on Bayesian models when the true disease state is unknown. *Journal of Agricultural, Biological, and Environmental Statistics* 2007; 12: 128-146. Article. DOI: 10.1198/108571107X178095.

153. Branscum AJJ, Wesley O.; Hanson, Timothy E.; Gardner, Ian A. Bayesian semiparametric ROC curve estimation and disease diagnosis. *Statistics in Medicine* 2008; 27: 2474-2496. DOI: 10.1002/sim.3250.

154. Erkanli AS, Minje; Jane Costello, E.; Angold, Adrian. Bayesian semi-parametric ROC analysis. *Statistics in Medicine* 2006; 25: 3905-3928. DOI: 10.1002/sim.2496.

155. García Barrado L, Coart E and Burzykowski T. Development of a diagnostic test based on multiple continuous biomarkers with an imperfect reference test. *Statistics in Medicine* 2016; 35: 595-608. Article. DOI: 10.1002/sim.6733.

156. Coart E, Barrado LG, Duits FH, et al. Correcting for the Absence of a Gold Standard Improves Diagnostic Accuracy of Biomarkers in Alzheimer's Disease. *Journal of Alzheimer's Disease* 2015; 46: 889-899. Article. DOI: 10.3233/JAD-142886.

157. Jafarzadeh SR, Johnson WO and Gardner IA. Bayesian modeling and inference for diagnostic accuracy and probability of disease based on multiple diagnostic biomarkers with and without a perfect reference standard. *Statistics in Medicine* 2016; 35: 859-876. Article. DOI: 10.1002/sim.6745.

158. Hwang BS and Chen Z. An Integrated Bayesian Nonparametric Approach for Stochastic and Variability Orders in ROC Curve Estimation: An Application to Endometriosis Diagnosis. *Journal of the American Statistical Association* 2015; 110: 923-934. Article. DOI: 10.1080/01621459.2015.1023806.

159. Alonzo TA and Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Statistics in Medicine* 1999; 18: 2987-3003. Article. DOI: 10.1002/(SICI)1097-0258(19991130)18:22<2987::AID-SIM205>3.0.CO;2-B.

160. Schiller IvS, M.; Hadgu, A.; Libman, M.; Reitsma, J. B.; Dendukuri, N. Bias due to composite reference standards in diagnostic accuracy studies. *Statistics in Medicine* 2016; 35: 1454-1470.

161. Naaktgeboren CA, Bertens LC, van Smeden M, et al. Value of composite reference standards in diagnostic research. *Bmj* 2013; 347: f5605.

162. Tang S, Hemyari P, Canchola JA, et al. Dual composite reference standards (dCRS) in molecular diagnostic research: A new approach to reduce bias in the presence of Imperfect reference. *Journal of Biopharmaceutical Statistics* 2018; 28: 951-965. Article. DOI: 10.1080/10543406.2018.1428613.

163. Bertens LC, Broekhuizen BD, Naaktgeboren CA, et al. Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. *PLoS medicine* 2013; 10: e1001531.

164. Juhl DV, A.; Luhm, J.; Ziemann, M.; Hennig, H.; Görg, S. Comparison of the two fully automated anti-HCMV IgG assays: Abbott Architect CMV IgG assay and Biotest anti-HCMV recombinant IgG ELISA. *Transfusion Medicine* 2013; 23: 187-194. DOI: 10.1111/tme.12036.

165. Nateghi Rostami MHR, B.;, Aghsaghloo F and Nazari R. Comparison of clinical performance of antigen based-enzyme immunoassay (EIA) and major outer membrane protein (MOMP)-PCR for detection of genital Chlamydia trachomatis infection. *International Journal of Reproductive Biomedicine* 2016; 14: 411-420.

166. Spada EP, Daniela; Baggiani, Luciana; Bagnagatti De Giorgi, Giada; Perego, Roberta; Ferro, Elisabetta. Evaluation of an immunochromatographic test for feline AB system blood typing. *Journal of Veterinary Emergency and Critical Care* 2016; 26: 137-141. DOI: 10.1111/vec.12360.

167. Brocchi E, Bergmann IE, Dekker A, et al. Comparative evaluation of six ELISAs for the detection of antibodies to the non-structural proteins of foot-and-mouth disease virus. *Vaccine* 2006; 24: 6966-6979. DOI: https://doi.org/10.1016/j.vaccine.2006.04.050.

168. Williams GJM, Petra; Kerr, Marianne; Fitzgerald, Dominic A.; Isaacs, David; Codarini, Miriam; McCaskill, Mary; Prelog, Kristina; Craig, Jonathan C. Variability and accuracy in interpretation of consolidation on chest radiography for diagnosing pneumonia in children under 5 years of age. *Pediatric Pulmonology* 2013; 48: 1195-1200. DOI: 10.1002/ppul.22806.

169. Asselineau J, Paye A, Bessède E, et al. Different latent class models were used and evaluated for assessing the accuracy of campylobacter diagnostic tests:

Overcoming imperfect reference standards? *Epidemiology and Infection* 2018; 146: 1556-1564. Article. DOI: 10.1017/S0950268818001723.

170. Sobotzki CR, M.; Kennerknecht, N.; Hulsse, C.; Littmann, M.; White, A.; Von Kries, R.; Von Kotnig, C. H. W. Latent class analysis of diagnostic tests for adenovirus, Bordetella pertussis and influenza virus infections in German adults with longer lasting coughs. *Epidemiology and Infection* 2016; 144: 840-846. DOI: 10.1017/s0950268815002149.

171. Poynard TDL, V.; Zarski, J. P.; Stanciu, C.; Munteanu, M.; Vergniol, J.; France, J.; Trifan, A.; Le Naour, G.; Vaillant, J. C.; Ratziu, V.; Charlotte, F. Relative performances of FibroTest, Fibroscan, and biopsy for the assessment of the stage of liver fibrosis in patients with chronic hepatitis C: A step toward the truth in the absence of a gold standard. *Journal of Hepatology* 2012; 56: 541-548. Article. DOI: 10.1016/j.jhep.2011.08.007.

172. De La Rosa GDV, M. L.; Arango, C. M.; Gomez, C. I.; Garcia, A.; Ospina, S.; Osorno, S.; Henao, A.; Jaimes, F. A. Toward an operative diagnosis in sepsis: A latent class approach. *BMC Infectious Diseases* 2008; 8 (no pagination).

173. Xie YC, Zhen; Albert, Paul S. A crossed random effects modeling approach for estimating diagnostic accuracy from ordinal ratings without a gold standard. *Statistics in Medicine* 2013; 32: 3472-3485. DOI: 10.1002/sim.5784.

174. See CWA, W.; Melese, M.; Zhou, Z.; Porco, T. C.; Shiboski, S.; Gaynor, B. D.; Eng, J.; Keenan, J. D.; Lietman, T. M. How reliable are tests for trachoma? - A latent class approach. *Investigative Ophthalmology and Visual Science* 2011; 52: 6133-6137.

175. Nérette P, Dohoo I and Hammell L. Estimation of specificity and sensitivity of three diagnostic tests for infectious salmon anaemia virus in the absence of a gold standard. *Journal of Fish Diseases* 2005; 28: 89-99. Article. DOI: 10.1111/j.1365-2761.2005.00612.x.

176. Pak SIK, D. Evaluation of diagnostic performance of a polymerase chain reaction for detection of canine Dirofilaria immitis. *Journal of Veterinary Clinics* 2007; 24: 77-81. Article.

177. Jokinen J, Snellman M, Palmu AA, et al. Testing Pneumonia Vaccines in the Elderly: Determining a Case Definition for Pneumococcal Pneumonia in the Absence of a Gold Standard. *American Journal of Epidemiology* 2018; 187: 1295-1302. Article. DOI: 10.1093/aje/kwx373.

178. Santos FLN, Campos ACP, Amorim LDAF, et al. Highly accurate chimeric proteins for the serological diagnosis of chronic chagas disease: A latent class

analysis. *American Journal of Tropical Medicine and Hygiene* 2018; 99: 1174-1179. Article. DOI: 10.4269/ajtmh.17-0727.

179. Mamtani M, Jawahirani A, Das K, et al. Bias-corrected diagnostic performance of the naked eye single tube red cell osmotic fragility test (NESTROFT): An effective screening tool for β-thalassemia. *Hematology* 2006; 11: 277-286. Article. DOI: 10.1080/10245330600915875.

180. Karaman BF, Açıkalın A, Ünal İ, et al. Diagnostic values of KOH examination, histological examination, and culture for onychomycosis: a latent class analysis. *International Journal of Dermatology* 2019; 58: 319-324. Article. DOI: 10.1111/ijd.14255.

181. Yan Q, Karau MJ, Greenwood-Quaintance KE, et al. Comparison of diagnostic accuracy of periprosthetic tissue culture in blood culture bottles to that of prosthesis sonication fluid culture for diagnosis of prosthetic joint infection (PJI) by use of Bayesian latent class modeling and IDSA PJI criteria for classification. *Journal of Clinical Microbiology* 2018; 56. Article. DOI: 10.1128/JCM.00319-18.

182. Lurier T, Delignette-Muller ML, Rannou B, et al. Diagnosis of bovine dictyocaulosis by bronchoalveolar lavage technique: A comparative study using a Bayesian approach. *Preventive Veterinary Medicine* 2018; 154: 124-131. Article. DOI: 10.1016/j.prevetmed.2018.03.017.

183. Falley BN, Stamey JD and Beaujean AA. Bayesian estimation of logistic regression with misclassified covariates and response. *Journal of Applied Statistics* 2018; 45: 1756-1769. Article. DOI: 10.1080/02664763.2017.1391182.

184. Dufour SD, J.; Dubuc, J.; Dendukuri, N.; Hassan, S.; Buczinski, S. Bayesian estimation of sensitivity and specificity of a milk pregnancy-associated glycoprotein-based ELISA and of transrectal ultrasonographic exam for diagnosis of pregnancy at 28–45 days following breeding in dairy cows. *Preventive Veterinary Medicine* 2017; 140: 122-133. Article. DOI: 10.1016/j.prevetmed.2017.03.008.

185. Bermingham MLH, I. G.; Glass, E. J.; Woolliams, J. A.; Bronsvoort, B. M. D. C.; McBride, S. H.; Skuce, R. A.; Allen, A. R.; McDowell, S. W. J.; Bishop, S. C. Hui and Walter's latent-class model extended to estimate diagnostic test properties from surveillance data: A latent model for latent data. *Scientific Reports* 2015; 5. Article. DOI: 10.1038/srep11861.

186. Busch EL, Don PK, Chu H, et al. Diagnostic accuracy and prediction increment of markers of epithelial-mesenchymal transition to assess cancer cell detachment from primary tumors. *BMC Cancer* 2018; 18. Article. DOI: 10.1186/s12885-017-3964-3.

187. de Araujo Pereira GL, F.; de Fatima Barbosa, V.; Ferreira-Silva, M. M.; Moraes-Souza, H. A general latent class model for performance evaluation of diagnostic tests in the absence of a gold standard: an application to Chagas disease. *Computational and mathematical methods in medicine* 2012; 2012: 487502.

188. Hubbard RA, Huang J, Harton J, et al. A Bayesian latent class approach for EHR-based phenotyping. *Statistics in Medicine* 2019; 38: 74-87. Article. DOI: 10.1002/sim.7953.

189. Caraguel C, Stryhn H, Gagné N, et al. Use of a third class in latent class modelling for the diagnostic evaluation of five infectious salmon anaemia virus detection tests. *Preventive Veterinary Medicine* 2012; 104: 165-173. DOI: https://doi.org/10.1016/j.prevetmed.2011.10.006.

190. De Waele V, Berzano M, Berkvens D, et al. Age-Stratified Bayesian Analysis To Estimate Sensitivity and Specificity of Four Diagnostic Tests for Detection of Cryptosporidium Oocysts in Neonatal Calves. *Journal of Clinical Microbiology* 2011; 49: 76-84. DOI: 10.1128/jcm.01424-10.

191. Dendukuri N, Wang LL and Hadgu A. Evaluating Diagnostic Tests for Chlamydia trachomatis in the Absence of a Gold Standard: A Comparison of Three Statistical Methods. *Statistics in Biopharmaceutical Research* 2011; 3: 385-397. DOI: 10.1198/sbr.2011.10005.

192. Habib IS, I.; Uyttendaele, M.; De Zutter, L.; Berkvens, D. A Bayesian modelling framework to estimate Campylobacter prevalence and culture methods sensitivity: application to a chicken meat survey in Belgium. *Journal of Applied Microbiology* 2008; 105: 2002-2008. DOI: 10.1111/j.1365-2672.2008.03902.x.

193. Vidal EM, A.; Bertolini, E.; Cambra, M. Estimation of the accuracy of two diagnostic methods for the detection of Plum pox virus in nursery blocks by latent class models. *Plant Pathology* 2012; 61: 413-422. DOI: 10.1111/j.1365-3059.2011.02505.x.

194. Aly SSA, R. J.; Whitlock, R. H.; Adaska, J. M. Sensitivity and Specificity of Two Enzyme-linked Immunosorbent Assays and a Quantitative Real-time Polymerase Chain Reaction for Bovine Paratuberculosis Testing of a Large Dairy Herd. *International Journal of Applied Research in Veterinary Medicine* 2014; 12: 1-7.

195. Rahman AKMA, Saegerman C, Berkvens D, et al. Bayesian estimation of true prevalence, sensitivity and specificity of indirect ELISA, Rose Bengal Test and Slow Agglutination Test for the diagnosis of brucellosis in sheep and goats in Bangladesh. *Preventive Veterinary Medicine* 2013; 110: 242-252.

196. Praet NV, Jaco J.; Mwape, Kabemba E.; Phiri, Isaac K.; Muma, John B.; Zulu, Gideon; van Lieshout, Lisette; Rodriguez-Hidalgo, Richar; Benitez-Ortiz, Washington; Dorny, Pierre; Gabriël, Sarah. Bayesian modelling to estimate the test characteristics of coprology, coproantigen ELISA and a novel real-time PCR for the diagnosis of taeniasis. *Tropical Medicine & International Health* 2013; 18: 608-614. DOI: 10.1111/tmi.12089.

197. Espejo LA, Zagmutt FJ, Groenendaal H, et al. Evaluation of performance of bacterial culture of feces and serum ELISA across stages of Johne's disease in cattle using a Bayesian latent class model. *Journal of dairy science* 2015; 98: 8227-8239.

198. Haley C, Wagner B, Puvanendiran S, et al. Diagnostic performance measures of ELISA and quantitative PCR tests for porcine circovirus type 2 exposure using Bayesian latent class analysis. *Preventive veterinary medicine* 2011; 101: 79-88.

199. Menten JB, Marleen; Lesaffre, Emmanuel. Bayesian latent class models with conditionally dependent diagnostic tests: A case study. *Statistics in Medicine* 2008; 27: 4469-4488. DOI: 10.1002/sim.3317.

200. Tasony-Wagener EA. *Evaluation of Antigen Detection Assays for the Avian Influenza Virus*. Ph.D., University of Prince Edward Island (Canada), Ann Arbor, 2012.

201. Weichenthal S, Joseph L, Bélisle P, et al. Bayesian Estimation of the Probability of Asbestos Exposure from Lung Fiber Counts. *Biometrics* 2010; 66: 603-612. DOI: 10.1111/j.1541-0420.2009.01279.x.

202. Jafarzadeh SR, Warren DK, Nickel KB, et al. Bayesian estimation of the accuracy of ICD-9-CM- and CPT-4-based algorithms to identify cholecystectomy procedures in administrative data without a reference standard. *Pharmacoepidemiology and Drug Safety* 2016; 25: 263-268. DOI: 10.1002/pds.3870.

203. Diab RG, Tolba MM, Ghazala RA, et al. Intestinal schistosomiasis: Can a urine sample decide the infection? *Parasitology International* 2021; 80. Article. DOI: 10.1016/j.parint.2020.102201.

204. Dannemiller NG, Kechejian S, Kraberger S, et al. Diagnostic Uncertainty and the Epidemiology of Feline Foamy Virus in Pumas (Puma concolor). *Scientific Reports* 2020; 10. Article. DOI: 10.1038/s41598-020-58350-7.

205. Guerriero M, Bisoffi Z, Poli A, et al. Prevalence of asymptomatic SARS-CoV-2-positive individuals in the general population of northern Italy and evaluation of a diagnostic serological ELISA test: A cross-sectional study protocol. *BMJ Open* 2020; 10. Article. DOI: 10.1136/bmjopen-2020-040036.

206. Bonelli P, Loi F, Cancedda MG, et al. Bayesian analysis of three methods for diagnosis of cystic echinococcosis in sheep. *Pathogens* 2020; 9: 1-9. Article. DOI: 10.3390/pathogens9100796.

207. Bisoffi Z, Pomari E, Deiana M, et al. Sensitivity, specificity and predictive values of molecular and serological tests for COVID-19: A longitudinal study in emergency room. *Diagnostics* 2020; 10. Article. DOI: 10.3390/diagnostics10090669.

208. Aminu OR, Lembo T, Zadoks RN, et al. Practical and effective diagnosis of animal anthrax in endemic low-resource settings. *PLoS Neglected Tropical Diseases* 2020; 14: 1-17. Article. DOI: 10.1371/journal.pntd.0008655.

209. Amini M, Kazemnejad A, Zayeri F, et al. Diagnostic accuracy of maternal serum multiple marker screening for early detection of gestational diabetes mellitus in the absence of a gold standard test. *BMC Pregnancy and Childbirth* 2020; 20. Article. DOI: 10.1186/s12884-020-03068-7.

210. Amini M, Kazemnejad A, Zayeri F, et al. Application of bayesian latent variable model for early detection of gestational diabetes mellitus without a perfect reference standard test by β-human chorionic gonadotropin. *Iranian Journal of Epidemiology* 2020; 16: 71-80. Article.

211. Pfukenyi DM, Meletis E, Modise B, et al. Evaluation of the sensitivity and specificity of the lateral flow assay, Rose Bengal test and the complement fixation test for the diagnosis of brucellosis in cattle using Bayesian latent class analysis. *Preventive Veterinary Medicine* 2020; 181. Article. DOI: 10.1016/j.prevetmed.2020.105075.

212. Rijckaert J, Raes E, Buczinski S, et al. Accuracy of transcranial magnetic stimulation and a Bayesian latent class model for diagnosis of spinal cord dysfunction in horses. *Journal of Veterinary Internal Medicine* 2020; 34: 964-971. Article. DOI: 10.1111/jvim.15699.

213. Jekarl DW, Choi H, Kim JY, et al. Evaluating diagnostic tests for helicobacter pylori infection without a reference standard: Use of latent class analysis. *Annals of Laboratory Medicine* 2020; 40: 68-71. Article. DOI: 10.3343/alm.2020.40.1.68.

214. Islam MA, Rony SA, Kitazawa H, et al. Bayesian latent class evaluation of three tests for the screening of subclinical caprine mastitis in Bangladesh. *Tropical Animal Health and Production* 2020. Article. DOI: 10.1007/s11250-020-02263-0.

215. Saxena P, Choudhary H, Muthu V, et al. Which Are the Optimal Criteria for the Diagnosis of Allergic Bronchopulmonary Aspergillosis? A Latent Class Analysis.

Journal of Allergy and Clinical Immunology: In Practice 2020. Article. DOI: 10.1016/j.jaip.2020.08.043.

216. Hamard A, Greffier J, Bastide S, et al. Ultra-low-dose CT versus radiographs for minor spine and pelvis trauma: a Bayesian analysis of accuracy. *European Radiology* 2020. Article. DOI: 10.1007/s00330-020-07304-8.

217. Jansen MD, Guarracino M, Carson M, et al. Field Evaluation of Diagnostic Test Sensitivity and Specificity for Salmonid Alphavirus (SAV) Infection and Pancreas Disease (PD) in Farmed Atlantic salmon (Salmo salar L.) in Norway Using Bayesian Latent Class Analysis. *Frontiers in Veterinary Science* 2019; 6. Article. DOI: 10.3389/fvets.2019.00419.

218. Hahn A, Schwarz NG and Frickmann H. Comparison of screening tests without a gold standard—A pragmatic approach with virtual reference testing. *Acta Tropica* 2019; 199. Article. DOI: 10.1016/j.actatropica.2019.105118.

219. Toft N, Halasa T, Nielsen SS, et al. Composite or aseptic quarter milk samples: Sensitivity and specificity of PCR and bacterial culture of Staphylococcus aureus based on Bayesian latent class evaluation. *Preventive Veterinary Medicine* 2019; 171. Article. DOI: 10.1016/j.prevetmed.2019.05.002.

220. Soares Filho PM, Ramalho AK, de Moura Silva A, et al. Evaluation of post-mortem diagnostic tests' sensitivity and specificity for bovine tuberculosis using Bayesian latent class analysis. *Research in Veterinary Science* 2019; 125: 14-23. Article. DOI: 10.1016/j.rvsc.2019.04.014.

221. O'Hagan MJH, Ni H, Menzies FD, et al. Test characteristics of the tuberculin skin test and post-mortem examination for bovine tuberculosis diagnosis in cattle in Northern Ireland estimated by Bayesian latent class analysis with adjustments for covariates. *Epidemiology and Infection* 2019; 147: 1-8. Article. DOI: 10.1017/S0950268819000888.

222. García Barrado L, Coart E and Burzykowski T. Estimation of diagnostic accuracy of a combination of continuous biomarkers allowing for conditional dependence between the biomarkers and the imperfect reference-test. *Biometrics* 2017; 73: 646-655. Article. DOI: 10.1111/biom.12583.

223. Jafarzadeh SR, Johnson WO, Utts JM, et al. Bayesian estimation of the receiver operating characteristic curve for a diagnostic test with a limit of detection in the absence of a gold standard. *Statistics in Medicine* 2010; 29: 2092-2106.

224. Saugar JM, Merino FJ, Martin-Rabadan P, et al. Application of real-time PCR for the detection of Strongyloides spp. in clinical samples in a reference center in Spain. *Acta tropica* 2015; 142: 20-25.

225. Peterson LRY, S. A.; Davis, T. E.; Wang, Z. X.; Duncan, J.; Noutsios, C.; Liesenfeld, O.; Osiecki, J. C.; Lewinski, M. A. Evaluation of the cobas cdiff test for detection of toxigenic clostridium difficile in stool samples. *Journal of Clinical Microbiology* 2017; 55: 3426-3436. Article. DOI: 10.1128/JCM.01135-17.

226. Fiebrich HBB, A. H.; Kerstens, M. N.; Pijl, M. E. J.; Kema, I. P.; De Jong, J. R.; Jager, P. L.; Elsinga, P. H.; Dierckx, R. A. J. O.; Van Der Wal, J. E.; Sluiter, W. J.; De Vries, E. G. E.; Links, T. P. 6-[F-18]fluoro-L-dihydroxyphenylalanine positron emission tomography is superior to conventional imaging with123I-metaiodobenzylguanidine scintigraphy, computer tomography, and magnetic resonance imaging in localizing tumors causing catecholamine excess. *Journal of Clinical Endocrinology and Metabolism* 2009; 94: 3922-3930. Article. DOI: 10.1210/jc.2009-1054.

227. Karch AK, A.; Zapf, A.; Zerr, I.; Karch, A. Partial verification bias and incorporation bias affected accuracy estimates of diagnostic studies for biomarkers that were part of an existing composite gold standard. *Journal of Clinical Epidemiology* 2016; 78: 73-82. DOI: 10.1016/j.jclinepi.2016.03.022.

228. Wu HM, Cordeiro SM, Harcourt BH, et al. Accuracy of real-time PCR, Gram stain and culture for Streptococcus pneumoniae, Neisseria meningitidis and Haemophilus influenzae meningitis diagnosis. *BMC Infectious Diseases* 2013; 13 (1) (no pagination).

229. Dendukuri N, Schiller I, De Groot J, et al. Concerns about composite reference standards in diagnostic research. *BMJ (Online)* 2018; 360. Article. DOI: 10.1136/bmj.j5779.

230. Driesen M, Kondo Y, de Jong BC, et al. Evaluation of a novel line probe assay to detect resistance to pyrazinamide, a key drug used for tuberculosis treatment. *Clinical Microbiology and Infection* 2018; 24: 60-64. Article. DOI: 10.1016/j.cmi.2017.05.026.

231. Bessède E, Asselineau J, Perez P, et al. Evaluation of the diagnostic accuracy of two immunochromatographic tests detecting campylobacter in stools and their role in campylobacter infection diagnosis. *Journal of Clinical Microbiology* 2018; 56. Article. DOI: 10.1128/JCM.01567-17.

232. Alcántara R, Fuentes P, Antiparra R, et al. MODS-Wayne, a colorimetric adaptation of the Microscopic-Observation Drug Susceptibility (MODS) assay for

detection of mycobacterium tuberculosis pyrazinamide resistance from sputum samples. *Journal of Clinical Microbiology* 2019; 57. Article. DOI: 10.1128/JCM.01162-18.

233. Stratigaki E, Jost FN, Kühnisch J, et al. Clinical validation of near-infrared light transillumination for early proximal caries detection using a composite reference standard. *Journal of Dentistry: X* 2020; 4. Article. DOI: 10.1016/j.jjodo.2020.100025.

234. Ziswiler HR, Reichenbach S, Vögelin E, et al. Diagnostic value of sonography in patients with suspected carpal tunnel syndrome: A prospective study. *Arthritis and Rheumatism* 2005; 52: 304-311. Article. DOI: 10.1002/art.20723.

235. Taylor SA, Mallett S, Bhatnagar G, et al. Diagnostic accuracy of magnetic resonance enterography and small bowel ultrasound for the extent and activity of newly diagnosed and relapsed Crohn's disease (METRIC): a multicentre trial. *The Lancet Gastroenterology and Hepatology* 2018; 3: 548-558. Article. DOI: 10.1016/S2468-1253(18)30161-4.

236. Eddyani M, Sopoh GE, Ayelo G, et al. Diagnostic accuracy of clinical and microbiological signs in patients with skin lesions resembling buruli ulcer in an endemic region. *Clinical Infectious Diseases* 2018; 67: 827-834. Article. DOI: 10.1093/cid/ciy197.

237. Lerner EB, McKee CH, Cady CE, et al. A consensus-based gold standard for the evaluation of mass casualty triage systems. *Prehospital Emergency Care* 2015; 19: 267-271. Conference Paper. DOI: 10.3109/10903127.2014.959222.

238. van Houten CB, de Groot JAH, Klein A, et al. A host-protein based assay to differentiate between bacterial and viral infections in preschool children (OPPORTUNITY): a double-blind, multicentre, validation study. *The Lancet Infectious Diseases* 2017; 17: 431-440. Article. DOI: 10.1016/S1473-3099(16)30519-9.

239. Van Dyck E, Buvé A, Weiss HA, et al. Performance of commercially available enzyme immunoassays for detection of antibodies against herpes simplex virus type 2 in African populations. *Journal of Clinical Microbiology* 2004; 42: 2961-2965. Article. DOI: 10.1128/JCM.42.7.2961-2965.2004.

240. Elliott DG, Applegate LJ, Murray AL, et al. Bench-top validation testing of selected immunological and molecular Renibacterium salmoninarum diagnostic assays by comparison with quantitative bacteriological culture. *Journal of Fish Diseases* 2013; 36: 779-809. DOI: 10.1111/jfd.12079.

241. Bland JM and Altman DG. Validating scales and indexes. *Bmj* 2002; 324: 606-607.

242. Zaki R, Bulgiba A, Ismail R, et al. Statistical methods used to test for agreement of medical instruments measuring continuous variables in method comparison studies: a systematic review. *PloS one* 2012; 7: e37908.

243. Hsia ECS, Neil; Cush, John J.; Chaisson, Richard E.; Matteson, Eric L.; Xu, Stephen; Beutler, Anna; Doyle, Mittie K.; Hsu, Benjamin; Rahman, Mahboob U. Interferon-γ release assay versus tuberculin skin test prior to treatment with golimumab, a human anti-tumor necrosis factor antibody, in patients with rheumatoid arthritis, psoriatic arthritis, or ankylosing spondylitis. *Arthritis & Rheumatism* 2012; 64: 2068-2077. DOI: 10.1002/art.34382.

244. Itza F, Zarza D, Salinas J, et al. Turn-amplitude analysis as a diagnostic test for myofascial syndrome in patients with chronic pelvic pain. *Pain Research and Management* 2015; 20: 96-100.

245. Booi ANM, Jerome; Norton, H. James; Anderson, William E.; Ellis, Amy C. Validation of a Screening Tool to Identify Undernutrition in Ambulatory Patients With Liver Cirrhosis. *Nutrition in Clinical Practice* 2015; 30: 683-689. DOI: 10.1177/0884533615587537.

246. von Heymann W, Moll H and Rauch G. Study on sacroiliac joint diagnostics: Reliability of functional and pain provocation tests. *Manuelle Medizin* 2018; 56: 239-248. Article. DOI: 10.1007/s00337-018-0405-6.

247. Schliep KC, Stanford JB, Chen Z, et al. Interrater and intrarater reliability in the diagnosis and staging of endometriosis. *Obstetrics and Gynecology* 2012; 120: 104-112. Article. DOI: 10.1097/AOG.0b013e31825bc6cf.

248. Pérez-Warnisher MTG-G, Teresa; Giraldo-Cadavid, Luis Fernando; Troncoso Acevedo, Maria Fernanda; Rodríguez Rodríguez, Paula; Carballosa de Miguel, Pilar; González Mangado, Nicolás. Diagnostic accuracy of nasal cannula versus microphone for detection of snoring. *The Laryngoscope* 2017; 127: 2886-2890. DOI: 10.1002/lary.26710.

249. Soltan MA, Tsai YL, Lee PYA, et al. Comparison of electron microscopy, ELISA, real time RT-PCR and insulated isothermal RT-PCR for the detection of Rotavirus group A (RVA) in feces of different animal species. *Journal of Virological Methods* 2016; 235: 99-104. Article. DOI: 10.1016/j.jviromet.2016.05.006.

250. Palit ST, N.; Knowles, C. H.; Lunniss, P. J.; Bharucha, A. E.; Scott, S. M. Diagnostic disagreement between tests of evacuatory function: a prospective study of 100 constipated patients. *Neurogastroenterology & Motility* 2016; 28: 1589-1598. DOI: 10.1111/nmo.12859.

251. Alonzo TA. Verification bias-impact and methods for correction when assessing accuracy of diagnostic tests. *Revstat Statistical Journal* 2014; 12: 67-83. Article.

252. Bhaskaran K and Smeeth L. What is the difference between missing completely at random and missing at random? *International journal of epidemiology* 2014; 43: 1336-1339.

253. Little RJ and Rubin DB. Bayes and multiple imputation. *Statistical analysis with missing data* 2002: 200-220.

254. Naaktgeboren CAdG, J. A.; van Smeden, M.; Moons, K. G.; Reitsma, J. B. Evaluating diagnostic accuracy in the face of multiple reference standards. *Annals of Internal Medicine* 2013; 159: 195-202. Evaluation Studies; Research Support, Non-U.S. Gov't.

255. Glueck DHL, M. M.; O'Donnell, C. I.; Ringham, B. M.; Brinton, J. T.; Muller, K. E.; Lewin, J. M.; Alonzo, T. A.; Pisano, E. D. Bias in trials comparing paired continuous tests can cause researchers to choose the wrong screening modality. *BMC medical research methodology* 2009; 9: 4.

256. Dendukuri N, Wang L and Hadgu A. Evaluating diagnostic tests for Chlamydia trachomatis in the absence of a gold standard: A comparison of three statistical methods. *Statistics in Biopharmaceutical Research* 2011; 3: 385-397. Article. DOI: 10.1198/sbr.2011.10005.

257. Pepe MS and Janes H. Insights into latent class analysis of diagnostic test performance. *Biostatistics* 2006; 8: 474-484.

258. Nortunen T, Puustinen J, Luostarinen L, et al. Validation of the finnish version of the montreal cognitive assessment test. *Acta Neuropsychologica* 2018; 16: 353-360. Article. DOI: 10.5604/01.3001.0012.7964.

259. Cheng MF, Guo YL, Yen RF, et al. Clinical Utility of FDG PET/CT in Patients with Autoimmune Pancreatitis: A Case-Control Study. *Scientific Reports* 2018; 8. Article. DOI: 10.1038/s41598-018-21996-5.

260. Gorman SLR, S.; Melnick, M. E.; Abrams, G. M.; Byl, N. N. Development and validation of the function in sitting test in adults with acute stroke. *Journal of Neurologic Physical Therapy* 2010; 34: 150-160. DOI: 10.1097/NPT.0b013e3181f0065f.

261. Young GP, Senore C, Mandel JS, et al. Recommendations for a step-wise comparative approach to the evaluation of new screening tests for colorectal cancer. *Cancer* 2016; 122: 826-839. Review. DOI: 10.1002/cncr.29865.

262. Reitsma JBR, A. W. S.; Khan, K. S.; Coomarasamy, A.; Bossuyt, P. M. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. *Journal of Clinical Epidemiology* 2009; 62: 797-806. Review.

263. Nofuentes JAR and Del Castillo JDL. Comparing the likelihood ratios of two binary diagnostic tests in the presence of partial verification. *Biometrical Journal* 2005; 47: 442-457. Article. DOI: 10.1002/bimj.200410134.

264. Nofuentes JAR and Del Castillo JdDL. Comparison of the likelihood ratios of two binary diagnostic tests in paired designs. *Statistics in Medicine* 2007; 26: 4179-4201. DOI: 10.1002/sim.2850.

265. Nofuentes JAR and Del Castillo JDL. EM algorithm for comparing two binary diagnostic tests when not all the patients are verified. *Journal of Statistical Computation and Simulation* 2008; 78: 19-35. Article. DOI: 10.1080/10629360600938102.

266. Nofuentes JARDC, J. D. L.; Marzo, P. F. Computational methods for comparing two binary diagnostic tests in the presence of partial verification of the disease. *Computational Statistics* 2009; 24: 695-718. DOI: 10.1007/s00180-009-0155-y.

267. Nofuentes JARDC, J. D. L.; Jimenez, A. E. M. Comparison of the accuracy of multiple binary tests in the presence of partial disease verification. *Journal of Statistical Planning and Inference* 2010; 140: 2504-2519. DOI: 10.1016/j.jspi.2010.02.026.

268. Marin-Jimenez AE and Roldan-Nofuentes JA. Global hypothesis test to compare the likelihood ratios of multiple binary diagnostic tests with ignorable missing data. *Sort-Statistics and Operations Research Transactions* 2014; 38: 305-323.

269. Harel O and Zhou XH. Multiple imputation for the comparison of two screening tests in two-phase Alzheimer studies. *Statistics in Medicine* 2007; 26: 2370-2388. Article. DOI: 10.1002/sim.2715.

270. Zhou XH and Castelluccio P. Nonparametric analysis for the ROC areas of two diagnostic tests in the presence of nonignorable verification bias. *Journal of Statistical Planning and Inference* 2003; 115: 193-213. Article. DOI: 10.1016/S0378-3758(02)00146-5.

271. Wang C, Turnbull BW, Nielsen SS, et al. Bayesian analysis of longitudinal Johne's disease diagnostic data without a gold standard test. *Journal of Dairy Science* 2011; 94: 2320-2328. DOI: 10.3168/jds.2010-3675.

272. Masaebi F, Zayeri F, Nasiri M, et al. Contrastive analysis of diagnostic tests evaluation without gold standard: Review article. *Tehran University Medical Journal* 2019; 76: 708-714. Review.

273. Beeley C. *Web application development with R using Shiny*. Packt Publishing Ltd, 2013.
274. Lim C, Wannapinij P, White L, et al. Using a web-based application to define the accuracy of diagnostic tests when the gold standard is imperfect. *PloS one* 2013; 8: e79489.
275. Kinner S, Pickhardt PJ, Riedesel EL, et al. Diagnostic accuracy of MRI versus CT for the evaluation of acute appendicitis in children and young adults. *American Journal of Roentgenology* 2017; 209: 911-919.
276. Ganiyusufoglu A, Onat L, Karatoprak O, et al. Diagnostic accuracy of magnetic resonance imaging versus computed tomography in stress fractures of the lumbar spine. *Clinical radiology* 2010; 65: 902-907.
277. Eze J, Ohagwu C, Ugwuanyi D, et al. Diagnostic accuracy of ultrasound scans for the diagnosis of pelvic inflammatory disease keeping laboratory high vaginal swab/urine microscopy culture as gold standard in Anambra State, Nigeria. *International Journal of Medicine and Medical Sciences* 2018; 10: 94-99.
278. Chidambaram JD, Prajna NV, Larke NL, et al. Prospective study of the diagnostic accuracy of the in vivo laser scanning confocal microscope for severe microbial keratitis. *Ophthalmology* 2016; 123: 2285-2293.
279. Nielsen LR, Toft N and Ersbøll AK. Evaluation of an indirect serum ELISA and a bacteriological faecal culture test for diagnosis of Salmonella serotype Dublin in cattle using latent class models. *Journal of Applied Microbiology* 2004; 96: 311-319. Article. DOI: 10.1046/j.1365-2672.2004.02151.x.
280. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in medicine* 1998; 17: 857-872.
281. Wilson DS. Confidence intervals for motion and deformation of the Juan de Fuca plate. *Journal of Geophysical Research: Solid Earth* 1993; 98: 16053-16071.
282. Agresti A and Coull BA. Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician* 1998; 52: 119-126.
283. Efron B. Nonparametric standard errors and confidence intervals. *canadian Journal of Statistics* 1981; 9: 139-158.
284. Efron B and Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science* 1986: 54-75.
285. Morris TP, White IR and Crowther MJ. Using simulation studies to evaluate statistical methods. *Statistics in medicine* 2019.

286. Wang Z, Dendukuri N, Zar HJ, et al. Modeling conditional dependence among multiple diagnostic tests. *Statistics in medicine* 2017; 36: 4843-4859.

287. Voinov VGe and Nikulin MS. *Unbiased Estimators and Their Applications: Volume 1: Univariate Case*. Springer Science & Business Media, 2012.

288. Lehmann EL and Casella G. *Theory of point estimation*. Springer Science & Business Media, 2006.

289. Newey W. Chapter 36: Large sample estimation and hypothesis testing.(RF Engle and DL McFadden, eds.). Handbook of Econometrics 4 2111–2245. Elsevier, Edition, 1994.

290. Team RC. R: A language and environment for statistical computing. 2013.

291. Dendukuri N. *Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests*. Ph.D., McGill University (Canada), Ann Arbor, 1999.

292. Gardner IA, Stryhn H, Lind P, et al. Conditional dependence between tests affects the diagnosis and surveillance of animal diseases. *Preventive veterinary medicine* 2000; 45: 107-122.

293. Byrom J, Douce G, Jones P, et al. Should punch biopsies be used when high-grade disease is suspected at initial colposcopic assessment? A prospective study. *International Journal of Gynecologic Cancer* 2006; 16.

294. Jablonski-Momeni A, Stachniss V, Ricketts D, et al. Reproducibility and accuracy of the ICDAS-II for detection of occlusal caries in vitro. *Caries research* 2008; 42: 79-87.

295. Braga M, Mendes F, Martignon S, et al. In vitro comparison of Nyvad's system and ICDAS-II with lesion activity assessment for evaluation of severity and activity of occlusal caries lesions in primary teeth. *Caries research* 2009; 43: 405-412.

296. Rodrigues J, Hug I, Diniz M, et al. Performance of fluorescence methods, radiographic examination and ICDAS II on occlusal surfaces in vitro. *Caries research* 2008; 42: 297.

297. Diniz MB, Rodrigues JA, Hug I, et al. Reproducibility and accuracy of the ICDAS-II for occlusal caries detection. *Community dentistry and oral epidemiology* 2009; 37: 399-404.

298. Bader JD and Shugars DA. A systematic review of the performance of a laser fluorescence device for detecting caries. *Journal of the American Dental Association* 2004; 135: 1413-1426. Review. DOI: 10.14219/jada.archive.2004.0051.

299. Stein CM. On the coverage probability of confidence sets based on a prior distribution. *Banach Center Publications* 1985; 16: 485-514.

300. Collins J and Huynh M. Estimation of diagnostic test accuracy without full verification: a review of latent class methods. *Statistics in Medicine* 2014; 33: 4141-4169. DOI: 10.1002/sim.6218.

301. de Araujo Pereira G, Louzada F, de Fátima Barbosa V, et al. A general latent class model for performance evaluation of diagnostic tests in the absence of a gold standard: an application to Chagas disease. *Computational and mathematical methods in medicine* 2012; 2012: 487502. Article.

302. Mori JK, Yutaka; Yoshizaki, Masahiro; Fukinbara, Satoru. Latent class models for medical diagnostic tests in multicenter trials. *Statistics in Medicine* 2013; 32: 5091-5105. DOI: 10.1002/sim.5962.

303. Garrett ES and Zeger SL. Latent class model diagnosis. *Biometrics* 2000; 56: 1055-1067.

304. Greene WH and Hensher DA. A latent class model for discrete choice analysis: contrasts with mixed logit. *Transportation Research Part B: Methodological* 2003; 37: 681-698.

305. Hui SL and Walter SD. Estimating the error rates of diagnostic tests. *Biometrics* 1980; 36: 167-171. Article.

306. Erosheva EA and Joutard C. Estimating Diagnostic Error without a Gold Standard: A Mixed Membership Approach. *Handbook of Mixed Membership Models and Their Applications*. Chapman and Hall/CRC, 2014, pp.175-192.

307. Godfrey K and DiStefano III J. Identifiability of model parameter. *IFAC Proceedings Volumes* 1985; 18: 89-114.

308. Dempster AP. Covariance selection. *Biometrics* 1972: 157-175.

309. Limmathurotsakul D, Jamsen K, Arayawichanont A, et al. Defining the true sensitivity of culture for the diagnosis of melioidosis using Bayesian latent class models. *PloS one* 2010; 5: e12485.

310. Berkvens D, Speybroeck N, Praet N, et al. Estimating disease prevalence in a Bayesian framework using probabilistic constraints. *Epidemiology* 2006; 17: 145-153.

311. Hadgu A and Qu Y. A biomedical application of latent class models with random effects. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 1998; 47: 603-616.

312. Qu Y and Hadgu A. A model for evaluating sensitivity and specificity for correlated diagnostic tests in efficacy studies with an imperfect reference test. *Journal of the American Statistical Association* 1998; 93: 920-928.

313. Goetghebeur E, Liinev J, Boelaert M, et al. Diagnostic test analyses in search of their gold standard: Latent class analyses with random effects. *Statistical Methods in Medical Research* 2000; 9: 231-248. Review. DOI: 10.1177/096228020000900304.
314. Albert PS. Misclassification Models. *Encyclopedia of Biostatistics*. John Wiley & Sons, Ltd, 2005.
315. Albert PS. Misclassification Models. *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, Ltd, 2014.
316. Berger JO. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.
317. Gelman A, Stern HS, Carlin JB, et al. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.
318. Broemeling LD. *Advanced Bayesian methods for medical test accuracy*. 2016, p.1-460.
319. O'Hagan A. Expert knowledge elicitation: subjective but scientific. *The American Statistician* 2019; 73: 69-81.
320. Azzalini A. *Statistical inference based on the likelihood*. Routledge, 2017.
321. Held L and Sabanés Bové D. Applied statistical inference. *Springer, Berlin Heidelberg, doi* 2014; 10: 16.
322. Carlo CM. Markov chain monte carlo and gibbs sampling. *Lecture notes for EEB* 2004; 581.
323. Lunn D, Spiegelhalter D, Thomas A, et al. The BUGS project: Evolution, critique and future directions. *Statistics in medicine* 2009; 28: 3049-3067.
324. Spiegelhalter D, Thomas A, Best N, et al. OpenBUGS user manual, version 3.0. 2. *MRC Biostatistics Unit, Cambridge* 2007.
325. Carlin BP and Louis TA. *Bayesian methods for data analysis*. CRC Press, 2008.
326. Johnson SR, Tomlinson GA, Hawker GA, et al. Methods to elicit beliefs for Bayesian priors: a systematic review. *Journal of clinical epidemiology* 2010; 63: 355-369.
327. Alhussain ZA and Oakley JE. Eliciting judgements about uncertain population means and variances. *arXiv preprint arXiv:170200978* 2017.
328. Oakley JE and O'Hagan A. SHELF: the Sheffield Elicitation Framework (version 4). . *School of Mathematics and Statistics, University of Sheffield, UK* 2019.
329. Raiffa H and Schlaifer R. Applied statistical decision theory. 1961.
330. Gosling JP. SHELF: the Sheffield elicitation framework. *Elicitation*. Springer, 2018, pp.61-93.

331. Morris DE, Oakley JE and Crowe JA. A web-based tool for eliciting probability distributions from experts. *Environmental Modelling & Software* 2014; 52: 1-4.
332. Authority EFS. Guidance on expert knowledge elicitation in food and feed safety risk assessment. *EFSA Journal* 2014; 12: 3734.
333. Jaynes ET. *Probability theory: The logic of science*. Cambridge university press, 2003.
334. Jaynes ET. Prior probabilities. *IEEE Transactions on systems science and cybernetics* 1968; 4: 227-241.
335. Zhu M and Lu AY. The counter-intuitive non-informative prior for the Bernoulli family. *Journal of Statistics Education* 2004; 12.
336. Spiegelhalter DJ, Best NG, Carlin BP, et al. *Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models*. 1998. Research Report, 98-009.
337. Brooks SP and Gelman A. General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics* 1998; 7: 434-455.
338. Mengersen KL, Robert CP and Guihenneuc-Jouyaux C. MCMC convergence diagnostics: a reviewww. *Bayesian statistics* 1999; 6: 415-440.
339. Ashby D. Bayesian statistics in medicine: a 25 year review. *Statistics in Medicine* 2006; 25: 3589-3631. DOI: 10.1002/sim.2672.
340. Allaire J. RStudio: integrated development environment for R. *Boston, MA* 2012; 770.
341. Team R. RStudio: integrated development for R. *RStudio, Inc, Boston, MA URL http://www rstudio com* 2015; 42: 14.
342. Surhone LM, Tennoe MT and Henssonow SF. OpenBUGS. 2010.
343. Sturtz S, Ligges U and Gelman A. R2OpenBUGS: a package for running OpenBUGS from R. *URL http://cran rproject org/web/packages/R2OpenBUGS/vignettes/R2OpenBUGS pdf* 2010.
344. OpenBUGS S and Thomas MN. Package 'R2OpenBUGS'. 2013.
345. Ligges U, Kerman J, OpenBUGS S, et al. Package 'R2OpenBUGS'. 2017.
346. Team SD. RStan: the R interface to Stan. *R package version* 2016; 2.
347. Hoffman MD and Gelman A. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J Mach Learn Res* 2014; 15: 1593-1623.

348. Dendukuri N and Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics* 2001; 57: 158-167. Article. DOI: 10.1111/j.0006-341X.2001.00158.x.

349. Thisted RA. *Elements of statistical computing: Numerical computation*. Routledge, 2017.

350. Thisted RA. *Elements of statistical computing: Numerical computation*. CRC Press, 1988.

351. Tom B and Consortium R-M. Characterization of disease course and remission in early seropositive rheumatoid arthritis. *medRxiv* 2020.

352. Vander Cruyssen B, Van Looy S, Wyns B, et al. DAS28 best reflects the physician's clinical judgment of response to infliximab therapy in rheumatoid arthritis patients: validation of the DAS28 score in patients under infliximab treatment. *Arthritis research & therapy* 2005; 7: R1063.

353. Prevoo M, Van'T Hof MA, Kuper H, et al. Modified disease activity scores that include twenty-eight-joint counts development and validation in a prospective longitudinal study of patients with rheumatoid arthritis. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology* 1995; 38: 44-48.

354. Aletaha D, Nell VP, Stamm T, et al. Acute phase reactants add little to composite disease activity indices for rheumatoid arthritis: validation of a clinical activity score. *Arthritis research & therapy* 2005; 7: R796-806. Article.

355. Fransen J, Welsing P, De Keijzer R, et al. Disease activity scores using C-reactive protein: CRP may replace ESR in the assessment of RA disease activity. *Ann Rheum Dis* 2004; 62: 151.

356. Slama IB, Allali F, Lakhdar T, et al. Reliability and validity of CDAI and SDAI indices in comparison to DAS-28 index in Moroccan patients with rheumatoid arthritis. *BMC Musculoskeletal Disorders* 2015; 16. Article. DOI: 10.1186/s12891-015-0718-8.

357. Arya V, Malaviya A and Raja R. CDAI (clinical disease activity index) in rheumatoid arthritis: cut-off values for classification into different grades of disease activity. *Indian Journal of Rheumatology* 2007; 2: 91-94.

358. Gaujoux-Viala C, Mouterde G, Baillet A, et al. Evaluating disease activity in rheumatoid arthritis: which composite index is best? A systematic literature analysis of studies comparing the psychometric properties of the DAS, DAS28, SDAI and CDAI. *Joint Bone Spine* 2012; 79: 149-155.

359. Aletaha D and Smolen J. The Simplified Disease Activity Index (SDAI) and the Clinical Disease Activity Index (CDAI): a review of their usefulness and validity in rheumatoid arthritis. *Clinical and experimental rheumatology* 2005; 23: S100.

360. Medeiros M and Quixadá R. Correlation of rheumatoid arthritis activity indexes (Disease Activity Score 28 measured with ESR and CRP, Simplified Disease Activity Index and Clinical Disease Activity Index) and agreement of disease activity states with various cut-off points in a Northeastern Brazilian population. *Revista brasileira de reumatologia* 2015; 55: 477-484.

361. Kumar BS, Suneetha P, Mohan A, et al. Comparison of Disease Activity Score in 28 joints with ESR (DAS28), Clinical Disease Activity Index (CDAI), Health Assessment Questionnaire Disability Index (HAQ-DI) & Routine Assessment of Patient Index Data with 3 measures (RAPID3) for assessing disease activity in patients with rheumatoid arthritis at initial presentation. *The Indian journal of medical research* 2017; 146: S57.

362. Park SY, Lee H, Cho SK, et al. Evaluation of disease activity indices in Korean patients with rheumatoid arthritis. *Rheumatology International* 2012; 32: 545-549. Article. DOI: 10.1007/s00296-011-1798-x.

363. Malibiradar S, Singh AK, Kumar A, et al. Comparative validation of clinical disease activity index (CDAI) and simplified disease activity index (SDAI) in rheumatoid arthritis in India. *Indian Journal of Rheumatology* 2013; 8: 102-106. DOI: http://dx.doi.org/10.1016/j.injr.2013.05.002.

364. Martins FM, Da Silva JAP, Santos MJ, et al. DAS28, CDAI and SDAI cut-offs do not translate the same information: results from the Rheumatic Diseases Portuguese Register Reuma. pt. *Rheumatology* 2014; 54: 286-291.

365. Sharma R, Thakare M, Thomas J, et al. Simplified Disease Activity Index as an index of disease activity in patients with rheumatoid arthritis: a comparison with DAS28. *Indian Journal of Rheumatology* 2009; 4: 11-14.

366. Gorial FI. Validity and reliability of CDAI in comparison to DAS28 in Iraqi patients with active rheumatoid arthritis. *Journal of the Faculty of Medicine* 2012; 54: 230-232.

367. Soubrier M. Should we revisit the definition of higher disease activity state in rheumatoid arthritis (RA)(abstract)? *Arthritis Rheum* 2004; 50: S386.

368. Legrand J, Kirchgesner T, Sokolova T, et al. Early clinical response and long-term radiographic progression in recent-onset rheumatoid arthritis: Clinical remission within six months remains the treatment target. *Joint Bone Spine* 2019; 86: 594-599.

369. Ben Abdelghani K, Miladi S, Souabni L, et al. AB0285 Which Score is Better to Assess Remission in Rheumatoid Arthritis? *Annals of the Rheumatic Diseases* 2014; 73: 898-898. DOI: 10.1136/annrheumdis-2014-eular.2958.

370. Rintelen B, Sautner J, Haindl P, et al. Remission in rheumatoid arthritis: a comparison of the 2 newly proposed ACR/EULAR remission criteria with the rheumatoid arthritis disease activity index-5, a patient self-report disease activity index. *The Journal of rheumatology* 2013; 40: 394-400.

371. Altman DG and Bland JM. Diagnostic tests 1: Sensitivity and specificity. *British Medical Journal* 1994; 308: 1552. Note.

372. Raiffa H. Decision analysis: Introductory lectures on choices under uncertainty. 1968.

373. Peizer DB and Pratt JW. A normal approximation for binomial, F, beta, and other common, related tail probabilities, I. *Journal of the American Statistical Association* 1968; 63: 1416-1456.

374. Feller W. On the normal approximation to the binomial distribution. *Selected Papers I*. Springer, 2015, pp.655-665.

375. Mood AM. Introduction to the Theory of Statistics. 1950.

376. Wan X, Wang W, Liu J, et al. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC medical research methodology* 2014; 14: 135.

377. Smolen JS, Breedveld FC, Schiff MH, et al. A simplified disease activity index for rheumatoid arthritis for use in clinical practice. *Rheumatology* 2003; 42: 244-257. Article. DOI: 10.1093/rheumatology/keg072.

378. Singh H, Kumar H, Handa R, et al. Use of clinical disease activity index score for assessment of disease activity in rheumatoid arthritis patients: an Indian experience. *Arthritis* 2011; 2011.

379. Cowles MK and Carlin BP. Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association* 1996; 91: 883-904.

380. Berger J. The robust Bayesian viewpoint (with discussion). Robustness of Bayesian Analysis,(J. Kadane, Ed.). Amsterdam: North Holland, 1984.

381. Rujes AWS, Reitsma JB, Coomarasamy A, et al. Evaluation of diagnostic test when there is no gold standard. A review of methods. *Health Technology Assessment* 2007; 11: 1-+.

382. Cohen DJ and Crabtree BF. Evaluative criteria for qualitative research in health care: controversies and recommendations. *The Annals of Family Medicine* 2008; 6: 331-339.

383. Hsia EC, Schluger N, Cush JJ, et al. Interferon-γ release assay versus tuberculin skin test prior to treatment with golimumab, a human anti–tumor necrosis factor antibody, in patients with rheumatoid arthritis, psoriatic arthritis, or ankylosing spondylitis. *Arthritis & Rheumatism* 2012; 64: 2068-2077. DOI: 10.1002/art.34382.

384. Bielak LF, Rumberger JA, Sheedy Ii PF, et al. Probabilistic model for prediction of angiographically defined obstructive coronary artery disease using electron beam computed tomography calcium score strata. *Circulation* 2000; 102: 380-385. Article. DOI: 10.1161/01.CIR.102.4.380.

385. Punglia RS, D'Amico AV, Catalona WJ, et al. Effect of verification bias on screening for prostate cancer by measurement of prostate-specific antigen. *New England Journal of Medicine* 2003; 349: 335-342.

386. van Geloven N, Broeze KA, Opmeer BC, et al. How to deal with double partial verification when evaluating two index tests in relation to a reference test? *Statistics in medicine* 2012; 31: 1265-1276.

387. Nishikawa H, Imanaka Y, Sekimoto M, et al. Influence of verification bias on the assessment of MRI in the diagnosis of meniscal tear. *American Journal of Roentgenology* 2009; 193: 1596-1602.

388. Nishikawa H, Imanaka Y, Sekimoto M, et al. Verification bias in assessment of the utility of MRI in the diagnosis of cruciate ligament tears. *American Journal of Roentgenology* 2010; 195: W357-W364.

389. Ahmadi F, Rashidy Z, Haghighi H, et al. Uterine cavity assessment in infertile women: Sensitivity and specificity of three-dimensional Hysterosonography versus Hysteroscopy. *Iranian journal of reproductive medicine* 2013; 11: 977.

390. Little RJA and Rubin DB. The analysis of social science data with missing values. *Sociological Methods & Research* 1989; 18: 292-326.

391. Rubin DB. Multiple imputation after 18+ years. *Journal of the American statistical Association* 1996; 91: 473-489.

392. Harel O and Zhou XH. Multiple imputation for correcting verification bias. *Statistics in medicine* 2006; 25: 3769-3786.

393. Lloyd CJF, Donald J. An application of multinomial logistic regression to estimating performance of a multiple-screening test with incomplete verification.

Journal of the Royal Statistical Society: Series C (Applied Statistics) 2008; 57: 89-102. DOI: 10.1111/j.1467-9876.2007.00602.x.

394. Albert PSD, L. E. On estimating diagnostic accuracy from studies with multiple raters and partial gold standard evaluation. *Journal of the American Statistical Association* 2008; 103: 61-73. DOI: 10.1198/016214507000000329.

395. Barnhart HX and Kosinki AS. Evaluating medical diagnostic tests at the subunit level in the presence of verification bias. *Statistics in Medicine* 2003; 22: 2161-2176. Article. DOI: 10.1002/sim.1436.

396. Lin CY, Barnhart HX and Kosinski AS. The weighted generalized estimating equations approach for the evaluation of medical diagnostic test at subunit level. *Biometrical Journal* 2006; 48: 758-771. Article. DOI: 10.1002/bimj.200510199.

397. Dorfman DD and Alf Jr E. Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals—rating-method data. *Journal of mathematical psychology* 1969; 6: 487-496.

398. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of mathematical psychology* 1975; 12: 387-415.

399. Rodenberg CZ, X. H. ROC curve estimation when covariates affect the verification process. *Biometrics* 2000; 56: 1256-1262. Review.

400. Gu J and Ghosal S. Bayesian ROC curve estimation under binormality using a rank likelihood. *Journal of Statistical Planning and Inference* 2009; 139: 2076-2083. DOI: https://doi.org/10.1016/j.jspi.2008.09.014.

401. Ning J and Cheng PE. A comparison study of nonparametric imputation methods. *Statistics and Computing* 2012; 22: 273-285. Article. DOI: 10.1007/s11222-010-9223-y.

402. Liu DZ, Xiao-Hua. Semiparametric Estimation of the Covariate-Specific ROC Curve in Presence of Ignorable Verification Bias. *Biometrics* 2011; 67: 906-916. DOI: 10.1111/j.1541-0420.2011.01562.x.

403. Hadgu A. The discrepancy in discrepant analysis. *Lancet* 1996; 348: 592-593. DOI: 10.1016/s0140-6736(96)05122-7.

404. Joseph LG, T. W.; Coupal, L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology* 1995; 141: 263-272. Article.

405. Georgiadis MP, Johnson WO, Gardner IA, et al. Correlation-adjusted estimation of sensitivity and specificity of two diagnostic tests. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2003; 52: 63-76.

406. Dendukuri NW, L.; Hadgu, A. Evaluating diagnostic tests for Chlamydia trachomatis in the absence of a gold standard: A comparison of three statistical methods. *Statistics in Biopharmaceutical Research* 2011; 3: 385-397.

407. Zhang JC, Kathryn; McLinden, James H.; Stapleton, Jack T. Bayesian analysis and classification of two enzyme-linked immunosorbent assay tests without a gold standard. *Statistics in Medicine* 2013; 32: 4102-4117. DOI: 10.1002/sim.5816.

408. Dufour SD, J.; Dubuc, J.; Dendukuri, N.; Hassan, S.; Buczinski, S. Bayesian estimation of sensitivity and specificity of a milk pregnancy-associated glycoprotein-based ELISA and of transrectal ultrasonographic exam for diagnosis of pregnancy at 28-45 days following breeding in dairy cows. *Preventive Veterinary Medicine* 2017; 140: 122-133.

409. Aly SS, Anderson RJ, Whitlock RH, et al. Sensitivity and specificity of two enzyme-linked immunosorbent assays and a quantitative real-time polymerase chain reaction for bovine paratuberculosis testing of a large dairy herd. *International Journal of Applied Research in Veterinary Medicine* 2014; 12: 1-7. Article.

410. Nerette P, Dohoo I and Hammell L. Estimation of specificity and sensitivity of three diagnostic tests for infectious salmon anaemia virus in the absence of a gold standard. *Journal of Fish Diseases* 2005; 28: 89-99. DOI: 10.1111/j.1365-2761.2005.00612.x.

411. Nérette P, Hammell L, Dohoo I, et al. Evaluation of testing strategies for infectious salmon anaemia and implications for surveillance and control programs. *Aquaculture* 2008; 280: 53-59.

412. Sidibe CAKG, V.; Thiaucourt, F.; Niang, M.; Lesnoff, M.; Roger, F. Performance evaluation of two serological tests for contagious bovine pleuropneumonia (CBPP) detection in an enzootic area using a Bayesian framework. *Tropical Animal Health and Production* 2012; 44: 1233-1238. DOI: 10.1007/s11250-011-0063-3.

413. Speybroeck NP, N.; Claes, F.; van Hong, N.; Torres, K.; Mao, S.; van den Eede, P.; Thinh, T. T.; Gamboa, D.; Sochantha, T.; Thang, N. D.; Coosemans, M.; Buscher, P.; D'Alessandro, U.; Berkvens, D.; Erhart, A. True versus apparent Malaria infection prevalence: The contribution of a Bayesian approach. *PLoS ONE* 2011; 6 (2) (no pagination).

414. Jacobson MW, P.; Nordengrahn, A.; Merza, M.; Emanuelson, U. Evaluation of a blocking ELISA for the detection of antibodies against Lawsonia intracellularis in pig sera. *Acta veterinaria Scandinavica* 2011; 53: 23.

415. Busch ELD, P. K.; Chu, H. T.; Richardson, D. B.; Keku, T. O.; Eberhard, D. A.; Avery, C. L.; Sandler, R. S. Diagnostic accuracy and prediction increment of markers of epithelial-mesenchymal transition to assess cancer cell detachment from primary tumors. *Bmc Cancer* 2018; 18. DOI: 10.1186/s12885-017-3964-3.

416. McDonald JL and Hodgson DJ. Prior precision, prior accuracy, and the estimation of disease prevalence using imperfect diagnostic tests. *Frontiers in Veterinary Science* 2018; 5. Article. DOI: 10.3389/fvets.2018.00083.

417. Walter SD and Irwig LM. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *Journal of Clinical Epidemiology* 1988; 41: 923-937. Article. DOI: 10.1016/0895-4356(88)90110-2.

418. Kang L, Carter R, Darcy K, et al. A fast Monte Carlo EM algorithm for estimation in latent class model analysis with an application to assess diagnostic accuracy for cervical neoplasia in women with AGC. *Journal of applied statistics* 2013; 40: 2699.

419. Nielsen SS, Grønbæk C, Agger JF, et al. Maximum-likelihood estimation of sensitivity and specificity of ELISAs and faecal culture for diagnosis of paratuberculosis. *Preventive Veterinary Medicine* 2002; 53: 191-204. DOI: https://doi.org/10.1016/S0167-5877(01)00280-X.

420. Wang CT, B. W.; Grohn, Y. T.; Nielsen, S. S. Nonparametric estimation of ROC curves based on Bayesian models when the true disease state is unknown. *Journal of Agricultural Biological and Environmental Statistics* 2007; 12: 128-146. DOI: 10.1198/108571107x178095.

421. Jafarzadeh SRJ, Wesley O.; Utts, Jessica M.; Gardner, Ian A. Bayesian estimation of the receiver operating characteristic curve for a diagnostic test with a limit of detection in the absence of a gold standard. *Statistics in Medicine* 2010; 29: 2090-2106. DOI: 10.1002/sim.3975.

422. Hall P and Zhou X-H. Nonparametric estimation of component distributions in a multivariate mixture. *The annals of statistics* 2003; 31: 201-224.

423. Saugar JMM, F. J.; Martin-Rabadan, P.; Fernandez-Soto, P.; Ortega, S.; Garate, T.; Rodriguez, E. Application of real-time PCR for the detection of Strongyloides spp. in clinical samples in a reference center in Spain. *Acta Tropica* 2015; 142: 20-25.

424. Paine SK, Basu A, Choudhury RG, et al. Multiplex PCR from Menstrual Blood: A Non-Invasive Cost-Effective Approach to Reduce Diagnostic Dilemma for Genital Tuberculosis. *Molecular Diagnosis and Therapy* 2018; 22: 391-396. Article. DOI: 10.1007/s40291-018-0322-3.

425. Fleiss JL, Cohen J and Everitt BS. Large sample standard errors of kappa and weighted kappa. *Psychological bulletin* 1969; 72: 323.

426. Feuer EJ and Kessler LG. Test statistic and sample size for a two-sample McNemar test. *Biometrics* 1989: 629-636.

427. Keezer MRP, Amélie; Stechysin, Barbara; Veilleux, Martin; Jetté, Nathalie; Wolfson, Christina. The diagnostic test accuracy of a screening questionnaire and algorithm in the identification of adults with epilepsy. *Epilepsia* 2014; 55: 1763-1771. DOI: 10.1111/epi.12805.

Appendices

This section contains all the additional information supporting this research study.

A.1. PRISMA Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	20
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	NIL
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	20
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	NIL
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	21
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	21
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	21
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	22
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	22
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	23
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	A3 Data extraction form

Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	NIL
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	NIL
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	Narrative synthesis

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	NIL
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	NIL
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	24
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	NIL
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	NIL
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	NIL
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	NIL
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	NIL
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	NIL
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	25 - 34
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	35

Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	36
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	See Acknowledgement

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097 For more information, visit: **www.prisma-statement.org**. Page 2 of 2

A.2. Example of search on SCOPUS database

(TITLE-ABS-KEY ("no gold standard*" OR "without gold standard*" OR "missing gold standard*" OR "imperfect reference standard*" OR "no reference standard*" OR "missing reference standard*" OR "partial verification" OR "differential verification") AND TITLE-ABS-KEY (“diagnostic accuracy”) AND TITLE-ABS-KEY ("medical test*" OR "new test*" OR "index test*" OR "diagnostic test*" OR "screening test*" OR routine*)) AND (LIMIT-TO (PUBYEAR , 2019) OR LIMIT-TO (PUBYEAR , 2018) OR LIMIT-TO (PUBYEAR , 2017) OR LIMIT-TO (PUBYEAR , 2016) OR LIMIT-TO (PUBYEAR , 2015) OR LIMIT-TO (PUBYEAR , 2014) OR LIMIT-TO (PUBYEAR , 2013) OR LIMIT-TO (PUBYEAR , 2012) OR LIMIT-TO (PUBYEAR , 2011) OR LIMIT-TO (PUBYEAR , 2010) OR LIMIT-TO (PUBYEAR , 2009) OR LIMIT-TO (PUBYEAR , 2008) OR LIMIT-TO (PUBYEAR , 2007) OR LIMIT-TO (PUBYEAR , 2006) OR LIMIT-TO (PUBYEAR , 2005)).

Similar search terms were used in other databases.

A.3. Data extraction Sheet

Date collected	
Collector's ID	
Author(s)	
Year	
URL	
Title	
Source	
Database	
URL	
Aim/objective of study	
Type of study design	
"No gold standard" type	
Target Condition	
Index test	
Reference standard	
Data	
Time frame	
Method	
Notes	**Assumptions** **Was it met?** **Results**

A.4. Data Extraction Sheet (Example)

Date collected	15.06.2018
Collector's ID	CMU
Author (Year)	Hsia, Schluger [383]
URL	https://onlinelibrary.wiley.com/doi/epdf/10.1002/art.34382
Title	Interferon-γ release assay versus tuberculin skin test prior to treatment with golimumab, a human anti-tumour necrosis factor antibody, in patients with rheumatoid arthritis, psoriatic arthritis, or ankylosing spondylitis
Source	Arthritis & Rheumatism
Database	Wiley Online Library
Aim/objective of study	To evaluate the performance of an interferon-γ release assay (IGRA) versus the standard tuberculin skin test (TST) as a screening tool for latent TB (LTB).
Type of study design	Cohort Study
"No gold standard" type	No gold standard
Index test	IGRA and TST
Target Condition	Latent tuberculosis
Data	2282 patients
Method	Positivity rate – used to comparing two or more tests Argreement – Kappa statistic
Notes	• Patients were screened with three tests; TST, IGRA and Chest radiography. • Concordance of tests results was done using the kappa statistic / coefficient. • Rate of positivity is the proportion of participants that the test identified as positive (having the target condition) based on the test is defined as the cut-off. • Multivariate logistic regression analyses were performed to determine if there were any covariates or factor associated with the screening test.

A.5. Supplementary information

Diagnostic test evaluation methodology: A systematic review of methods employed to evaluate diagnostic tests in the absence of gold standard – an update

In this supplementary material, the identified methods from the systematic review are briefly discussed. Some strengths and weaknesses of each method are highlighted. In addition, clinical applications (which are published articles where the method was employed) and key references (which are articles that described the proposed method) of the proposed methods are included.

As noted in the discussion section of the review (section **2.5**), not all the methods have been applied outside the original publication of the developed method. Thus, those methods that have not been applied clinically outside the original publication of the developed method do not have clinical application.

The identified methods are presented in tables under three main indices (A.5.1, A.5.2 and A.5.3):

- A.5.1: include all methods proposed when the gold standard is missing for some of the participants in the study, and the diagnostic outcome is binary. This is from Table 45 to Table 50. The tables are stratified based on the number of index test being evaluated and how the test outcome is measured (binary, ordinal or continuous).
- A.5.2: includes all methods proposed to evaluate medical test with more than two diagnostic outcomes and the gold standard is missing for some of the participants in the study. Table 51 to Table 52 focuses on methods employed to estimate the on ROC surface and volume under the ROC surface of single index test with ordinal (Table 51) and continuous (Table 52) results; and Table 53 focus on estimating the sensitivities and specificities of multiple binary tests when the disease status is categorical. The methods listed under section A1 and A2 have the true disease status of some participants missing; hence, the methods are developed with the assumption that the missing disease status is either missing at random (MAR) or missing not at random (MNAR). And methods developed based on the assumption of missing not at random has the 'missing at random' as a special case.
- A.5.3: includes methods proposed to evaluate medical test(s) when the reference standard is imperfect or there is no-gold standard. This is from Table 54 to Table 57.
  For all methods in A.5.1, A.5.2 and A.5.3; when multiple tests are evaluated together, the tests could be conditional independent given the true disease status, or conditional

dependent given the true disease status. Thus, some methods are developed based on these assumptions.

Abbreviations

CI: conditional independence
CD: conditional dependence
MAR: missing at random
MNAR: missing not at random
AUROC: Area under the ROC curve
ROC: Receiver operating characteristic
IPW: inverse probability weight
FI: Full imputation
LCM: latent class model
WGEE: weighted generalised estimating equation
PG-BRL: partial gold Bayesian rank likelihood
DR: Doubly robust
VUS: volume under the surface
NI: non-ignorable verification
EM: Expectation maximization
MSI: Mean square imputation

Update of the systematic review till December 2020

Articles published from January 2019 to December 2020 that fulfilled the eligibility criteria as defined in section **2.2.1** were reviewed in December 2020. From the search, 22 articles were identified of which most of them (20) were clinical application using the Bayesian latent class model, one employed the CRS and another article employed the differential verification. The identified articles are cited under the clinical application column in appropriate tables below.

A.5.1: Tables of methods employed in evaluating medical test(s) with missing gold standard in a binary-class diagnostic outcome.

Table 45: Methods employed for single binary index test

Single binary index test						
Method (Year)	MAR / MNAR	CI / CD	Characteristics	Strength	Weaknesses	Key reference
Begg & Greenes (B&G) (1983)	MAR	CI	This method is based on the Bayes theorem and the missing disease status of the unverified participants are considered as a doubling sampling problem. The true disease status of unverified participants is viewed as double-sampling problem and are included in the analysis. **Clinical application:** [110-112, 384, 385]	• Easy to implement analytically and the results are easily interpretable. • Can adjust for binary or discrete covariates.	• Estimates of sensitivity and specificity could still be biased if there are few false negative in the study (Cronin and Vickers [113]). • In the presence of more than one covariate, parametric models could be considered. Hence, the estimates obtained could be prone to bias that arise from model-misspecification.	Begg and Greenes [64]
Likelihood-based (1993)	MNAR	CI	The conditional probability of verification given the true disease status, are assumed to be known and specified with two values. Hence, the boundary of this values can be estimated using the observed data, which is employed to obtain bounded values for the sensitivity and specificity of the index test under the MNAR assumption.	• There is no need to conduct sensitivity analysis because this method produces all possible value of the sensitivity and specificity of the index test under the assumption that the verification is non-ignorable	• It assumes that the verification is known which is not always true in practice. • The estimated value of the sensitivity and specificity of the test is not a single value (except if the verification is specified) but a range of value that is bounded. Thus, interpretation of estimates is not straightforward.	Zhou [67]

Table 44 cont.: Methods employed for single binary index test

Method (Year)	MAR / MNAR	CI / CD	Characteristics	Strength	Weaknesses	Key reference
Expectation maximization (EM) – based regression (2003)	MNAR	CI	Parametric models are used to model the joint distribution of the verification, disease and index test response. The conditional probability of disease, test, and verification are modelled using logistic regression. The estimates of the marginal sensitivity and specificity are obtained by maximising the log-likelihood of the observed data using the EM approach. The MAR assumption is a special case of this approach.	• This method produces a single value each as the estimated sensitivity and specificity of the index tests not bounded values. • Observed discrete covariates can be considered with this approach. • Standard error of the parameters can be estimated using the observed information matrix not via bootstrapping. The probability of verification is estimated from the observed data rather than assuming it is known.	• Due to the MNAR assumption, the model could be non-identifiable. Hence to ascertain if the estimates obtained are unique, the information matrix from the EM process need to be singular. • The choice of model-selection is limited because of the non-identifiability problem. This method is prone to bias if the model is misspecified.	Kosinski and Barnhart [68]
Global sensitivity analysis (2003)	MNAR	CI	This method provides all possible (but bounded) jointed values of the sensitivity and specificity of the index test under different assumptions of the missing mechanism (MCAR, MAR and MNAR) in a graphical form. The area produced by the bounded possible values is called the true ignorance region (TIR). It is a sensitivity analysis that help researchers see the amount of bias that can be made as a result of the MNAR assumption. **Clinical application**: [386-389]	• This method allows researcher to understand how different missing mechanism assumption can impact their inference on the diagnostic accuracy of the index test. • The graphical presentation of all the possible pair values of the sensitivity and specificity makes the interpretation easy. With this approach, there is no need to change the values of parameters in the fitted model to obtain all the possible pairs of sensitivity and specificity under the MNAR assumption.	This approach does not take into consideration observed covariates such as age, race, etc. that could affect the verification of the true disease status of the participants.	Kosinski and Barnhart [69]

Table 44 cont.: Methods employed for single binary index test

Method (Year)	MAR / MNAR	CI / CD	Characteristics	Strength	Weaknesses	Key reference
Multiple imputation (2006)	MAR	CI	The true disease status of the unverified participants is considered as a missing data problem. So, the missing disease status are imputed with M (>1) plausible simulated values based on the assumption of MAR. M complete datasets are obtained. Each dataset is analysed to obtain the estimate of sensitivity and specificity of the test. The estimates are pooled together to obtain a single estimate of the sensitivity, specificity and its variances using the Rubin's rule (Little and Rubin [390]; Rubin [391]).	• This approach is an alternative to Beggs & Greenes method. However, it is shown to be more flexible than B&G in incorporating more than one covariates [114]. Estimates obtained are easily interpretable.	If the MAR assumption do not hold, the MNAR assumption need to be modelled within the data augmentation process which will require specifying the right model and could be computationally demanding.	Harel and Zhou [392]
Propensity score (2012)	MAR	CI	This method defines the propensity score as "*the probability of verification given the test response and observed covariates*". It is assumed to be unknown and estimates using the observed data. Participants are stratified based on their propensity score. Estimate of sensitivity and specificity of the index test are obtained by pooling the estimates obtained from each stratum.	• Although, this approach is model based, the estimates obtained are less sensitive to model misspecification because the propensity score (verification model) is used to stratify the participants.	• The number of strata must be decided. This method requires the sample size to be sufficiently large because of the stratification.	He and McDermott [66]

Table 44 cont.: Methods employed for single binary test

Method (Year)	MAR / MNAR	CI / CD	Characteristics	Strength	Weaknesses	Key reference
Bayesian approach (2006, 2008, 2018)	MAR & MNAR	CI	Generally, Bayesian method combines prior information about the parameters of interest such as sensitivity and specificity of the index test in a parametric distribution together with the likelihood of the observed data to get the predictive or posterior distribution which is used to make inference about the parameters of interest. The three articles discussed here estimates the test's performance of a single test with binary response under MNAR and MAR assumption. The method by Martinez [70] is the Bayesian approach of the likelihood – based method by Zhou [67]. However, the verification quantities are taken as unknown and estimated alongside the sensitivity, specificity and prevalence of disease within the Bayesian process. This method does not employ observed covariates. Buzoianu and Kadane [71] used the data augmentation approach (which is a Bayesian imputation approach) to impute the missing disease status and simultaneously estimate the parameters of the models which is used to obtain the marginal sensitivity and specificity of the index test. This is the Bayesian approach of Kosinski and Barnhart [68]. The Bayesian approach by Hajivandi, Shirazi [72] is an extension of the approach by Martinez [70] and Buzoianu and Kadane [71]. This extended Bayesian method eliminates the verification variable. So that the probability of disease or test response do not depends on verification. Hence, a new model is developed to predict the true disease status for individuals who do not undergo the index test.	• Generally, Bayesian methods overcome the problem of model non-identifiability faced by the maximum likelihood approach. Because they use prior information about the parameters of interest which imposes some statistical restriction on the parameters. • The number of parameters to be estimated using Bayesian approach is not limited like the maximum likelihood approach. • Prior distribution of the sensitivity and specificity can be elicited from the estimates obtained from the B&G approach (Martinez [70]) or elicited from other sources like expert opinion or previous research.	• Sensitivity analysis is required to study the impact of the prior distribution on the estimates. • If the prior is non-informative then estimates tend towards likelihood. However, if informative priors are used, the obtained estimates tend to balance the idea. • Sufficient sample size is required to avoid the inference to coincide completely with the prior.	Martinez [70] Buzoianu and Kadane [71] Hajivandi, Shirazi [72]

Table 46: Methods employed for multiple binary index tests

Multiple binary index tests						
Method	**MAR / MNAR**	**CI / CD**	**Characteristics**	**Strength**	**Weaknesses**	**Key reference**
Baker et al (Maximum likelihood based) (1995)	MNAR	CI	The idea behind this approach is to use multiple tests to fit identifiable verification models to overcome the problem of non-identifiability that arises as a result of the MNAR assumption in the maximization of the log – likelihood function. The sensitivity and specificity of the combined multiple tests is obtained by choosing the disease model and verification model that fits the data. This method derives the ROC curve using the combined responses of the multiple index tests employed in the study to obtain the combination of the tests' response that maximizes TPRs at a given FPRs.	• The use of multiple index tests makes the likelihood of the observed data identifiable under the MNAR assumption. • The model can be extended to include covariates.	• This method does not estimate the sensitivity and specificity of the index tests individually but only collectively. • This method does not consider the conditional dependence of the multiple index tests. • The estimates obtained are sensitive to the MNAR assumption (via the verification model). Thus, they are prone to bias that could arise from model misspecification. Evidence of model misspecification can be checked by fitting various verification models.	Baker [81]
Maximum likelihood approach (1998)	MAR	CI	This method uses the maximum likelihood approach to estimate the sensitivities and specificities of two binary (screening) index tests under the assumption that both tests are conditional independent given the true disease status.	• Estimates are easy to compute or solve analytically. • The method can incorporate binary or categorical covariate to obtain the covariate specific diagnostic accuracy measures.	• Method cannot be employed in the case of more than two index tests. • Tests must be conditional independent.	Zhou [73]

Table 45 cont.: Methods employed for multiple binary index tests

Method	MAR / MNAR	CI / CD	Characteristics	Strength	Weaknesses	Key reference
Latent class model (LCM) (1999, 2008)	MAR	CI & CD	When two screening tests are used the screen negatives (those participants with negative responses in both tests) are considered to be negative or non-diseased. Hence, they are not referred for further testing with the gold standard because their true disease status is assumed to be negative. Methods (based on LCM) were developed by Walter [78], Böhning and Patilea [79] to estimate the sensitivities and specificities of the two binary index (screening) tests when all screen negatives do not get their disease status verified with the gold standard. The disease status of unverified participants are assumed to be latent or unobserved or missing but not negative. Because the two screening tests are imperfect. Walter [78] assumes both tests have dichotomised responses and they are conditional independent given the true disease status. Böhning and Patilea [79] is an extension of Walter [78] which relaxes the assumption of conditional independence and assumes that both tests are conditionally dependent given true disease status and that the conditional dependence is homogenous.	• The assumption that all participants with negative response in both tests are non-diseased is relaxed because both index (screening) tests are imperfect. • Sensitivities and specificities are calculated separately for the two tests evaluated. Böhning and Patilea [79] do not require the two tests' responses to be independent.	• The homogenous conditional dependence of the tests across all participants may not be true in practice. The estimates obtained can be biased if the model is misspecified.	Walter [78] Böhning and Patilea [79]

Table 45 cont.: Methods employed for multiple binary index tests

Method	MAR / MNAR	CI / CD	Characteristics	Strength	Weaknesses	Key reference
Bayesian approach (2005)	MAR	CI	This Bayesian approach estimates the sensitivity and specificity of two index tests which are applied to all participants while adjusting for observed covariates.	• Using informative priors circumvent the problem of non-identifiability that arises from likelihood estimation procedure. Correlation between tests is considered.	Misspecification of the prior model or distribution can bias estimates.	Martinez, Achcar [76]
Lloyd et al (2008)	MAR	NIL	This method uses multinomial logistic regression to estimate the joint sensitivity and specificity of all the index tests. The aim of this method is to evaluate the diagnostic accuracy of a combination of multiple dichotomised or binary tests to decide verification process.	The conditional dependence or independence of the tests is of no importance in this approach.	It does not provide the diagnostic accuracy of the individual index test applied in the study.	Lloyd [393]
Semi-latent class: Gaussian random effect (GRE) Finite mixture (FM) (2008)	MAR	CI & CD	This approach is a modification of the GRE by [132] and Finite mixture by Albert and Dodd [32] which was original developed to evaluate medical tests when the is no gold standard or the reference test is imperfect. This method is semi-latent because the disease status of those that were verified with the gold standard are taken to be known and participants not verified are assumed to be latent. This method is an alternative to the imputation and reweighting approach by Albert [43].	• This is more robust than the imputation approach by Albert [43] if the models are correctly specified. • It is a data-driven method.	Prone to bias that could arise from model misspecification.	Albert [394]

Table 45 cont.: Methods employed for multiple binary index tests

Method	MAR / MNAR	CI / CD	Characteristics	Strength	Weaknesses	Key reference
Imputation and reweighting: MSI, IPW, and SPE (2007)	MAR	CI & CD	This approach applies the idea of the estimators by Alonzo and Pepe [61] (MSI, IPW and SPE) to estimate the sensitivity and specificity of multiple binary tests. The joint and marginal sensitivities and specificities of the tests being evaluated are estimated under the assumption that the verification is known or fixed by design. However, if it isn't, then it can be estimated using the observed data. This approach is an alternative to the semi-latent methods by Albert [394]	• Comparing this approach to the semi-latent methods (GRE and FM) by Albert [394], it is simple and less prone to errors due to model misspecification. • It is less computational expensive compared to the semi-latent methods. Additional discussion about the advantages and disadvantages of the two methods are in this article.	• The assumption that the verification is being fixed by designed or known is very restrictive and may not always be true in practice. In extreme biased sampling (those with all negative index test response do not get their disease status verified), the imputation approach requires statistical adjustment like extrapolation.	Albert [43]
Bayesian approach (2010)	MNAR	CI	This is a modification of the Martinez [70] approach to include two index binary tests. It uses Markov – chain Monte Carlo (MCMC) methods to simulate samples for the joint posterior distribution. The probability of verification is modelled using eight parameters in the form described by Zhou [67].	• Employing two good sources of information rather than only the observed data can improve the accuracy of the estimates.	• Relatively large sample size is needed to reduce the weight of the prior on the posterior.	Aragon, Martinez [84]

Table 45 cont.: Methods employed for multiple binary index tests

Method	MAR / MNAR	CI / CD	Characteristics	Strength	Weaknesses	Key reference
Likelihood – based approach (addresses double partial verification) (2012)	MNAR	CI	This approach deals with double partial verification where not all the participants undergo all the index tests, and some do not undergo the gold standard. It is likelihood-based following the pattern of Baker [81], and Kosinski and Barnhart [68].	Rather than using only participants that undertook both index tests (probable discarding data from participants who did not undergo both index tests); this approach utilises every data from all participants whether they undertook either index tests.	• This method is model based, so it is prone to bias due to model misspecification. This method is computational expensive, especially when the number of index tests with conditional dependence increases and the number of missing participants within each test varies.	Van Geloven [82] Van Geloven, Broeze [83]
Weighted Generalise Estimating Equation (WGEE) based method (2014, 2003, 2006)	MAR	CD	WGEE is employed to deal with correlated data with missing data problem. Because not all participants undergo the gold standard the number of verified participants is used as the weight. The single model proposed by Xue, Kim [77] estimates the diagnostic accuracy of multiple index tests and compare them. The model can also combine data from different study to obtain the sensitivity and specificity of the index tests. The weighted least square (WLS)[395] method and WGEE method by Lin, Barnhart [396] are proposed to estimate the diagnostic accuracy of the index test when applied at subunit levels.	• This approach can estimate the diagnostic accuracy of more than two binary index tests that are correlated. Data from two or more study can be combined using this approach to obtain the parameters of interest.	• This approach requires large sample size because of the normal approximation assumption. • It is model – based; hence it is prone to error due to misspecification of models. With this approach, verification depends only on the test response not on the observed covariates. Thus, it cannot adjust for observed covariates.	Xue, Kim [77] Barnhart and Kosinki [395] Lin, Barnhart [396]

Table 47: Methods employed for single ordinal index test

Single ordinal index test						
Method	**MAR / MNAR**	**CI / CD**	**Characteristics**	**Strength**	**Weaknesses**	**Key reference**
Parametric AUROC (1984)	MAR	CI	Constructing a ROC curve using the pair of sensitivities and specificities estimated using any correction method for binary test (like the B&G method) at every cut-off of continuous or ordinal tests often produce a step function ROC (not smooth). Hence, this method employs the procedure of Dorfman and Alf Jr [397] to derive a smooth ROC curve. The method assumes that the index test result has an underlying continuous scale; the disease and non-diseased group of the underlying continuous scale of the test response are normally distributed (bi-normality) after monotonic transformation, and there are latent cut-offs. With the ROC curve derived, the AUROC is calculated graphically or using the trapezoidal rule by Bamber [398].	• This method produces a smooth ROC and the AUROC is estimated graphically or via the trapezoidal rule. • Estimate is easy to interpret.	• The underlying bi-normality assumption of the disease and non-diseases group may not always be true in practice. • This approach estimates only the AUROC. • Pairs of sensitivity and specificity on the smooth ROC cannot be obtained at any cut-off because they are latent and are employed for model purpose. • With the derived ROC curve, the test's response at the extreme of the upper ROC curve are indicative of non-diseased than diseased, which is not in practice. • The estimated AUROC depends on the number of verified participants; thus, a relative high sample size is advised. • This method can be prone to bias that arises from model-misspecification. • This method cannot incorporate any observed covariates as the verification of participants depends only on the test response.	Gray, Begg [85]

Table 46 cont.: Methods employed for single ordinal index tests

Method	MAR / MNAR	CI / CD	Characteristics	Strength	Weaknesses	Key reference
Non-parametric likelihood based (1996)	MAR	CI	This method estimates the AUROC of an ordinal test via likelihood approach.	• This approach is non-parametric so not prone to model-misspecification. • No need to make any assumption about the distribution of the disease and non-disease group. This approach can adjust for discrete covariates.	This approach estimates only the AUROC, not the ROC curve (the pairs of sensitivities and specificities at different cut-offs).	Zhou [86]
Parametric maximum likelihood based (1998)	MNAR	CI	This method employs the basic assumption of Dorfman and Alf Jr [397] procedure to derived a smooth ROC curve. The EM algorithm is employed to obtain the maximum likelihood estimator of the ROC curve. The method incorporates sensitivity analysis to evaluate the impact of the MNAR assumption.	• This approach produces range of possible values for the pair of sensitivity and specificity under the non-ignorable verification assumption. Hence, a separate sensitivity analysis is not needed. If the verification is known and accurate, then the estimates obtained are robust.	• Assumption of bi-normality of disease and non-disease group is made which may not be true in practice. • The estimates of sensitivity and specificity is bounded by a range of values and not a single value. The verification is assumed to be known, which may not be true in practice.	Zhou and Rodenberg [88]
Parametric maximum likelihood based (2000)	MAR	CI	This approach is an extension of Gray, Begg [85] method to include observed categorical covariates that can affect the verification of the disease status of the participants. This approach calculate the AUROC and the ROC curve applying the procedure and basic assumption of Dorfman and Alf Jr [397].	This method produces covariate-specific ROC and AUROC.	• The underlying assumption of bi-normality about the diseased and non-diseased group may not be true in practice. • This method is prone to bias that can arise from model – misspecification. Pairs of sensitivity and specificity on the smooth ROC cannot be obtained at any cut-off because they are latent and employed for model development.	Rodenberg [399]

Table 48: Methods employed for single continuous index test

Single continuous index test						
Method	**MAR / MNAR**	**CI / CD**	**Characteristics**	**Strength**	**Weaknesses**	**Key reference**
Hunink et al (I990)	MAR	CI	This approach evaluates the ROC curve of a continuous test based on the assumption that the selection for verification depend on observed covariates and not the result of the index test. Logistic regression analysis was employed to evaluate the probability of verification given the observed covariates to correct for verification bias.	• A single ROC curve that adjust for all the covariates employed in the study can be produced as well as covariate specific ROC curves.	• In practice, using only observed covariates such as signs and symptoms, age or sex may not be a strong evidence to send participants for disease verification.	Hunink, Richardson [89]
Doubly robust (DR): AUROC only and Empirical ROC (2006, 2009)	MNAR	CI	The doubly robust estimators discussed here produces marginal AUROC and ROC not covariate specific ROCs and AUROCs. The disease status of all the participants (verified or unverified) are replaced with the estimated disease which is a function of the probability of disease and verification probability. Both probabilities are specified parametrically. It is doubly robust because the correct specification of either the disease or verification model makes the estimator consistent. However, misspecification of both models makes the estimator inconsistent. Sensitivity analyses are undertaken to evaluate the impact of the MNAR assumption on the estimates.	• The DR approaches are consistent and asymptotic normal provided the either the verification or diseased model is correctly specified. • This approach can adjust for continuous covariate without having to dichotomised or discretize them. • The number of covariates that can be included in the model is not limited.	• The estimator is model based so misspecification of the models makes the estimator inconsistent. • There is efficiency cost with this approach; an estimator with correctly specified disease or verification model yields same consistent estimates as the DR estimator. However, the variance of the DR estimator is larger because of the additional model. • The non-ignorable parameter is assumed to be known which is not always true in practice. • The empirical ROC curve derived by Fluss [94] has non-monotonic property because the value of the ROC can lie outside the range of 0 and 1. Hence, the isotonic regression procedure is suggested to correct for the non-monotonicity of the ROC (estimated sensitivities and specificities).	Rotnitzky, Faraggi [95] (AUROC only) Fluss [94] (Empirical ROC)

Table 47 cont.: Methods employed for single continuous index test

Method	MAR / MNAR	CI / CD	Characteristics	Strength	Weaknesses	Key reference
Imputation and Reweighting: Full imputation (FI) Mean score imputation (MSI) Inverse probability weight (IPW) Semi-parametric efficient (SPE) (2005)	MAR	CI	The four estimators (FI, MSI, IPW and SPE) derive the ROC curve and AUROC from empirical data. The FI estimator imputes the probability of disease (disease status) for all participants in the study regardless if they were verified with the gold standard or not. The MSI imputes the probability of disease only for participants with missing disease status. The IPW uses participants whose true disease status were verified. It weights each observation from the verified with the inverse of its probability of verification. The SPE incorporate the probability of verification and probability of disease to obtain the estimates of sensitivity and specificity. That is why it is referred to as doubly robust. The AUROC is estimated using trapezoidal rule. A modification of the FI, MSI, IPW and SPE estimators that adjust for non-ignorable verification is constructed by Liu [97].	• No assumption is made about the distribution of the test response of the diseased and non-diseased group. • These methods are easy to implement, because the estimators only need regression model is fitted to the binary response – disease and or verification. • The four estimators are consistent, provided the disease model and/or verification is specified correctly. The SPE is doubly robust; the correct specification of either the verification or disease model make it a consistent estimator.	• The estimators are model-based, so prone to bias due to model misspecification. The estimators are inconsistent if the model is misspecified. The effect of model misspecification on each of the estimator is discussed in the article. • Estimates of sensitivity and specificity could still be biased if there are few false negative in the study (Cronin and Vickers [113]). • The SPE estimator does not produce a monotonic (increasing) ROC. Although, this can be corrected using isotonic regression (Fluss [94]). • The SPE is inconsistent if both verification and disease model is misspecified. • If there are observed covariate(s) that affected verification, this approach does not adjust for it. The IPW employs information from verified participants, hence there is loss of information from those unverified.	Alonzo and Pepe [61]
Propensity – score adjustment method (2018)	MNAR	CI	It's a parametric approach. The probability of verification for the diseased participants not the whole sample is model under some parametric assumption. The AUC of the index test is estimated.	Since this method uses parametric assumption to model the probability of verification for only verified participants than the whole sample; the approach seems to be quite simpler or straightforward.	The parametric assumption used to model the probability of verification for the verified participants could be misspecified or non-ideal in practice; thus, resulting in wrong estimated of the AUC.	Yu, Kim [98]

Table 47 cont.: Methods employed for single continuous index test

Method	MAR / MNAR	CI / CD	Characteristics	Strength	Weaknesses	Key reference
U-statistics estimator (2009)	MAR	CI	The AUROC is estimated based on U-statistics and inverse probability weighting (IPW) technique. The IPW technique is used to correct for verification bias and the probability of verification is assumed to be known. The test response of the diseased and the non-diseased group are assumed to follow the F distribution and the U-statistics estimator is constructed under this assumption.	• The estimate obtained with this method is equivalent to the estimate obtained using the IPW estimator by Alonzo and Pepe [61]. However, this estimator has a closed form variance. This method does not require the probability of disease as it depends mainly on the verification probability. And if this is known, then it is less prone to bias that could arise from model misspecification.	• The assumption that the test's response follows an F-distribution may not be true in practice. • This approach only estimates the AUROC not the ROC curve. The verification probability may not be known in practice.	He [90]
Likelihood based (imputation and reweighted): FI, MSI , IPW and Pseudo – DR (PDR) (2010)	MNAR	CI	This approach is an extension of Alonzo and Pepe [61] to account for non-ignorable verification. The non-ignorable parameter was estimated from the observed data (rather than taken to be known / specified) by using the whole participants (not only those verified) to model the disease model. The log-likelihood is a function of both the disease and verification model. The log-likelihood was solved using scoring equations. Estimates of the probability of disease and verification probability is obtained which is employed to estimate the ROC curve and AUROC. The Pseudo DR is not doubly robust as the SPE estimator [61] or the DR estimators [94, 95]; because the verification probability and disease probability are estimated from same likelihood function and correct specification of both model is required to make the PDR estimator consistent.	• The estimators do not specify the NI parameter or assumes it to be known like the DR approach; rather it is estimated from the observed data. • The estimators derive both ROC curve and AUROC. • The disease's estimator for the MSI and PDR methods are statistically unbiased. The estimators are consistent provided the underlying assumptions surrounding their development are fulfilled.	• The misspecification of the models will cause the estimator to be inconsistent. • The PDR estimator produces non-monotonic ROC curve. However, this can be corrected using isotonic regression technique (Fluss [94]) • None of the estimators including the PDR estimator have the doubly robust property. A reasonably large sample size is required because of the non-ignorable parameter that is estimated from the observed data.	Liu [97]

Table 47 cont.: Methods employed for single continuous index tests

Method	MAR / MNAR	CI / CD	Characteristics	Strength	Weaknesses	Key reference
Partial gold Bayesian rank likelihood (PG-BRL) (2014)	MAR	CI	This approach is a modification of Gu and Ghosal [400] originally developed to derive the ROC and AUROC of diagnostic tests with continuous response in full or complete verification of participants with the gold standard. The PG-BRL is constructed to correct for verification bias. The method uses Bayesian technique to estimate the posterior distribution of the bi-normal parameters (mean and variance) of the disease group only (because the non-disease group is assumed to follow the standard normal distribution) and the prevalence of the true disease status. The observed data are placed in rank and label. Labels are used to describe the true disease status of the verified participants (non-disease = 0 and disease – 1) and missingness of unverified participants (unverified = 2). The ranks are invariant. Due to the missing labels of some participants, the data argumentation technique is applied via Gibbs sampling to impute the missing labels. Inference about the ROC and AUROC are derived using the posterior distributions of the parameters estimated (mean and variance).	• This estimator derives both the ROC and AUROC under the bi-normality assumptions of the diseased and non-diseased group. The PG-BRL estimator is consider to perform equivalently in terms of accuracy when compared to some bias – correction estimators like the FI, MSI, IPW and SPE [61].	• The assumption of bi-normality of the disease and non-disease group after some transformation is a restrictive assumption in practice. • Adjusting for observed covariates with this approach is computational complex as various transformation of the data are required. To use this approach for MNAR the verification needs to be explicit known and must be reflected in the verification function.	Gu, Ghosal [93]

Table 47 cont.: Methods employed for single continuous index tests

Method	MAR / MNAR	CI / CD	Characteristics	Strength	Weaknesses	Key reference
K – nearest neighbour (2015, 2017)	MAR	CI	These non-parametric estimators employs the K nearest neighbour imputation approach (Ning and Cheng [401]) to impute the missing disease status of participants where the diseases status is not verified.	• It is a fully non-parametric approach, so not prone to model misspecification. • It is a non-parametric version of the MSI approach of Alonzo and Pepe [61]. The estimator is consistent and asymptotically normal under MAR assumption.	• The choice of K and distance measure could be quite challenging in obtaining an unbiased estimate of the TPR and FPR. This approach requires a reasonably high sample size.	Adimari and Chiogna [91] (ROC) Adimari and Chiogna [92] (AUROC)

Table 49: Methods employed for single continuous index test with focus on covariate-specific ROC

Single index test with continuous response but focus on covariate specific ROC						
Method	**MAR / MNAR**	**CI / CD**	**Characteristics**	**Strength**	**Weaknesses**	**Key reference**
Fully parametric (2009)	MNAR	CI	This method is a variation of the DR approach by Rotnitzky, Faraggi [95]. However, the joint distribution of the test, disease, observed covariates and verification are specified parametrically rather than using the likelihood of the joint distribution. The ROC curve and AUROC derived are covariate specific. The observed covariate can be in continuous form.	• The estimator is consistent provided the underlying assumptions involving their development is fulfilled. • The number of covariates are unlimited and the continuous form of the covariates are retained.	• Requires large sample size. • Model – misspecification makes the estimator inconsistent.	Page and Rotnitzky [99]
Semi-parametric: FI, IPW and PDR (2011)	MAR	CI	The three estimators proposed here are modifications of the imputation and reweighting approach of Alonzo and Pepe [61] and doubly robust approach by Rotnitzky, Faraggi [95] to produce ROC curves that are covariate specific.	• Compared to the fully parametric approach [99], the semi-parametric estimators is less impacted by model-misspecification.	• This approach only produces ROC curves but not AUROC because it does not have an explicit expression. • The continuous covariates have to be change to discrete form.	Liu [402]

Table 48 cont.: Methods employed for single continuous index test with focus on covariate-specific ROC

Method	MAR / MNAR	CI / CD	Characteristics	Strength	Weaknesses	Key reference
Doubly Robust (2012)	MNAR	CI	This method is an extension of the DR estimator by Fluss [94] to make the estimated pair of sensitivity and specificities (ROC) covariate specific. To achieve this, semi-parametric location scale model was used to model the effect of the observed covariates on the ROC curve. The location scale model models the test's response as function of the disease and covariates.	• The estimator is doubly robust • It is consistent and asymptotically normal. The covariate is in discrete form.	• Estimator is inconsistent if the both models (verification and disease) are misspecified. • The derive ROC is non – monotonic so the isotonic regression corrected is needed The non-ignorable parameter is specified so sensitivity analysis is required to study the impact of the MNAR assumption.	Fluss, Reiser [96]
Imputation & reweighting (2013)	MAR & MNAR	CI	This method is a modification of the semi-parametric method by Liu [402] to develop an estimators that can also estimates the covariate – specific AUROCs	• The method also estimates the AUROCs unlike the semiparametric approach. Adjust for covariates without having to change to discrete form.	It is model based, hence the estimator is inconsistent in the presence of model misspecification.	Liu [100]

Table 50: Methods employed for multiple ordinal or continuous index tests

Multiple index tests with ordinal or continuous responses						
Method	**MAR / MNAR**	**CI / CD**	**Characteristics**	**Strength**	**Weaknesses**	**Key reference**
Profile and EM method (2003)	MNAR	CI	This is a ML – based approach which employs the profile method combined with EM to obtain a global maximum likelihood estimator for the sensitivity and specificity of two index tests with ordinal results. Although the article goes further to compare the two tests with assumption that they are correlated. However, this review focused on the approach employed to estimate the AUC of each index test.	• This approach is non-parametric, so not prone to model misspecification.	• In estimating the AUC of each index test, the two tests are assumed to be independent given the true disease status. • The two index tests must have the same number of ordinal class.	Zhou and Castelluccio [270]
Likelihood based Imputation (MCEM) approach (2012)	MAR	CD	This estimator is developed to estimate the ROC curve and AUROC of diagnostic tests in a multi-phase trial with partial verification. That is a trial with more than two – phase. The gold standard is applied to participants in the last phase of the trial to confirm true disease status. In multi-phase trial with multiple screening tests, decision rule is made on how to define positive and negative to move participants unto the last testing stage. Often times, the tests could be correlated or repeated. This method derived two estimators of the ROC curve when the believe-the-positive (BP) decision rule is used or believe-the-negative (BN) rule is employed. It is a likelihood-based approach and the Monte Carlo Expectation Maximization (MCEM) is used to maximise the log-likelihood.	• This method estimates the diagnostic accuracy of screening (index) tests that could be correlated and are employed in multi-phase trials which adjusting for partial verification within each phase.	• This method is model based hence model misspecification makes the estimator inconsistent. • The method is computational complex as the MCEM is employed because direct maximization approach is difficult. • As the number of sequential tests employed in the study increases, the more complex the computational procedure of estimating the diagnostic accuracy especially if the tests are conditional dependent given true diseases status.	Yu [102]

A.5.2: Tables of methods employed to evaluate medical test when there is missing gold standard and the diagnostic outcomes is classified into three. Hence, focusing on ROC surface and volume of surface (VUS).

Table 51: Methods employed for single ordinal index test with ROC surface and VUS

Single ordinal index test						
Method	**MAR / MNAR**	**CI / CD**	**Characteristics**	**Strength**	**Weaknesses**	**Key reference**
Non-parametric likelihood-based approach (2008)	MAR	CI	This estimator estimates the empirical ROC surface and volume under the ROC surface (VUS) using likelihood – based approach.	• This estimator is not model based so bias due to model misspecification is eliminated. • Covariates can be adjusted; however, it must be in the discrete form to get covariate specific ROC surface.	• Sparse data affects the estimate derived.	Chi [103]

Table 52: Methods employed for single continuous index test with ROC surface and VUS

Single continuous index test						
Method	**MAR / MNAR**	**CI / CD**	**Characteristics**	**Strength**	**Weaknesses**	**Key reference**
Imputation and reweighting: MSI, FI, IPW and SPE (2016)	MAR	CI	This estimator is an extension of the imputation and reweighting approach (FI, MSI, IPW and SPE) by Alonzo and Pepe [61] to incorporate three disease class status.	• Same as the imputation and reweighting estimators of Alonzo and Pepe [61].	• Same as the imputation and reweighting estimators of Alonzo and Pepe [61].	Duc, Chiogna [62]
Non-parametric approach: KNN approach (2016, 2017)	MAR		This estimator is an extension of the KNN approach by Adimari and Chiogna [91].	Same as the KNN approach above.	Same as the KNN approach above.	Duc, Chiogna [104] To Duc [105]
IPW (2016)	MAR	CI	This estimator is an extension of the IPW approach by Alonzo and Pepe [61] which both derive the ROC surface and estimate the VUS. Only information from verified participants is employed in the analysis and each observation from verified participant is weighted with the inverse of the verification probability.	• The estimator is consistent given the verification of probability is accurately known or estimated.	• There is loss of information especially from participants whose disease status were not verified with the gold standard. • Same as the IPW estimator above.	Zhang, Alonzo [63]
IPW, DR and PDR estimators (2018)	MNAR	CI	Three estimators are developed to estimate the volume under the ROC surface (VUS) only. These estimators are extension of the IPW estimator under the framework of the doubly robust technique, the DR and PDR approach by Rotnitzky, Faraggi [95] and Liu [97] respectively.	• The IPW and DR correct for verification bias in considerable samples; however, the PDR approach requires large sample. • The DR estimator has the doubly robust property. • The PDR has same strength and weaknesses as above.	• Sensitivity analysis is required to study the effect of specifying the non-ignorable parameter in the DR and IPW approach.	Zhang, Alonzo [106]

Table 52 cont. Methods employed for single continuous index test with ROC surface and VUS

Method	MAR / MNAR	CI / CD	Characteristics	Strength	Weaknesses	Key reference
Bayesian Semi-parametric ROC (2018)	MAR	CI	The method is based on tri-normality assumption and it is an extension of the rank-based likelihood approach by Gu and Ghosal [400], Gu, Ghosal [93]	• This method uses the tri--normality assumption to model the three diagnostic outcomes, resulting in a smoother ROC surface.	• Adjusting for observed covariates is computational complex. It is a parametric approach, so deviation from the parametric assumptions like the tri-normality of the diagnostic outcomes leads to inconsistent estimators.	Zhu and Ghosal [107]
Parametric – based approach FI, MSI, PDR (2019)	MAR / MNAR	CI	Parametric regression model is used to model the probability of disease and verification using the whole sample and not only participants whose disease status were verified with the gold standard. It is an extension of the approach by Liu [97].	• The NI parameter is estimated from the observed data. The estimators are consistent provided the underlying assumptions surrounding their development are fulfilled.	• The misspecification of the models will cause the estimator to be inconsistent. The PDR estimator has the doubly robust properties if either the disease or verification model is specified correct under the MAR assumption. However, under the MNAR assumption the PDR estimator needs both the model to be correctly specified to be consistent.	To Duc, Chiogna [108]

Table 53: Methods employed for multiple binary index tests and categorical disease status.

Multiple binary tests and categorical disease status						
Method	**MAR / MNAR**	**CI / CD**	**Characteristics**	**Strength**	**Weaknesses**	**Key reference**
Latent class model (LCM) (2010)	MAR	CI & CD	Chu [80] proposed a LCM method (frequentist and Bayesian) to estimate the sensitivities and specificities of the two index (screening) tests when all participants with negative responses in both tests do not get their disease status verified with the gold standard. The disease status of unverified participants is assumed to be latent or unobserved. The tests are binary, but the disease status is categorical. This approach assumes that only participants with negative results in the two binary tests applied do not get their disease status verified.	• The Bayesian approach overcomes the non-identifiability problem encountered using the frequentist approach.	• The conditional dependence model is not robust in that different conditional dependent homogenous model produces different estimates of the sensitivity and specificity.	Chu [80]

A.5.3: Tables of methods employed in evaluating medical test(s) with an imperfect reference standard or no gold standard.

Table 54: Methods employed to evaluate index test(s) when the sensitivity and specificity of the imperfect reference standard is known precisely.

Methods	Characteristics	Strengths	Weakness	Key reference
Algebraic correction functions (1966, 1981, 1996) Bounded algebraic correction function (1981, 2018)	The estimators considered here are mathematical functions. The bias-corrected sensitivity and specificity of the new test is a function of the known sensitivity and specificity of the imperfect reference test. These methods are applied to evaluate medical test with binary response. The approach by Emerson, Waikar [121] aim to estimate the possible minimum and maximum values of the sensitivity and specificity of the index test when evaluated using an imperfect reference standard and the diagnostic accuracy of the reference standard is known. **Clinical Application**: Hahn [123], Matos [124], Mathews, Cachay [125]	• It is easy to implement analytically. • Incorporating information from a reliable source in addition to the available data can improve the accuracy of the estimated parameters. • With the bounded correction by Emerson, Waikar [121], there is no need to perform sensitivity analysis as this approach provides the lower and upper bound value of the sensitivity and specificity of the test being evaluated.	• If the information about the diagnostic accuracy of the imperfect test is not accurate then more bias is produced. • The approach by Emerson, Waikar [121] estimates the upper and lower bounds of the sensitivity and specificity of the test being evaluate but not an exact single value.	Gart and Buck [118] Brenner [117] Staquet, Rozencweig [119] Emerson, Waikar [121]

Table 54 cont.: Methods employed to evaluate index test(s) when the sensitivity and specificity of the imperfect reference standard is known precisely

Methods	Characteristics	Strengths	Weakness	Key reference
Gaussian Random Effect (GRE) Finite Mixture (FM), and Beta Binomial (2009)	These methods are model based. They are applied when there are multiple binary index tests to evaluate. These methods estimate the joint sensitivity and specificity of the index tests as well as their individual sensitivity and specificity while taking into consideration the conditional dependence structure across the index tests. The methods are modification of the GRE approach by Qu, Tan [132] and FM by Albert and Dodd [32].	• Can be applied to multiple binary tests with conditional dependence structure and conditional independence of the tests is a special case of this approach. • This approach is an alternative to the LCA method. The estimates obtained from this approach can be very robust provided the diagnostic accuracy of the RS is high and the dependence structure among the index tests given the RS is also high.	• Misspecification of the model can bias the estimate obtained. Inaccurate error rate of the imperfect tests can bias the estimates obtained. Thus, it is encouraged to obtain the accuracy measures of the imperfect reference standard in comparison with the gold standard (or participants with known disease status).	Albert [120]

Table 55: Methods employed to evaluate index test(s) when the sensitivity and specificity of the imperfect reference standard is unknown.

Methods	Characteristics	Strengths	Weaknesses	Key reference
Discrepancy analysis	All participants undergo both the index test and a reference standard which is imperfect. Then participants with discordant responses undergo another test called the resolver test, which is not a gold standard but assumed to have a high accuracy. This approach has been modified recently in that some samples with concordant responses are also verified with the resolver test or all positive responses to the index test and discordant responses are retested with the resolver test [164]. Another modification is by Hawkins [127] where participants that undergo the resolver test are sampled from the four groups (TP, FP, TN, FN) and the proportion of the verified and unverified samples are taken into consideration when estimating the sensitivity and specificity. **Clinical application** Van Dyck, Buvé [239]; Juhl [164]; Nateghi Rostami, Aghsaghloo [165] Spada [166]; Brocchi, Bergmann [167]	• Easy to implement analytically. • It is less expensive compared to where all participants have to undergo the three tests.	• This method has been shown to be biased; overestimating the sensitivity of the index test.	Hadgu [403] Hadgu, Dendukuri [126] Schiller [160] Hawkins [127]

Table 55 cont.: Methods employed to evaluate index test(s) when the sensitivity and specificity of the imperfect reference standard is unknown.

Methods	Characteristics	Strengths	Weaknesses	Key reference
Latent class model (LCM)	This approach assumes the disease status of all the participants are latent (unobserved) no test is a reference standard as. All tests are evaluated using statistical model. It is often assumed that the tests' responses are dichotomous. However, methods have been developed to derive the ROC and AUROC of medical tests when there is no gold standard. LCM can be broadly divided into frequentist and Bayesian approaches.	• LCM can be applied to study whose diagnostic outcome (disease status) is dichotomous or more. Unlike the discrepancy approach, correction methods which is applied to studies with only binary diagnostic outcomes. It evaluates all the tests employed in the study simultaneously, because none of the test employed in the study is a gold standard.	• Under the assumption that the tests are conditional dependent given the true disease status; the conditional dependence structure among the tests lacks robustness in that different dependence structure models fits the data but yield different estimates. • LCM are probability models that rely on parametric assumptions.	
Bayesian LCM (binary or dichotomized tests)	Bayesian LCM combines the likelihood of the observed data with prior information about the parameters to be estimated (sensitivity and specificity). The Bayesian approach was developed to make the LCMs identifiable while accurately estimating the diagnostic accuracy measures. **Examples of Bayesian methods** Joseph [404], Dendukuri and Joseph [348], Johnson, Gastwirth [143]; Georgiadis, Johnson [405]; Nérette, Stryhn [141]; Dendukuri, Hadgu [142]; Dendukuri [406] (multi-latent variable model); The MLVM uses both CI and CD assumptions. Martinez [144]; Zhang [407]; Dufour [408]; García Barrado, Coart [222]; Lu, Dendukuri [53] **Clinical Application of different Bayesian methods:** [115, 181-186, 190, 192, 193, 195, 196, 409-415]	• Bayesian LCM overcomes the problem of non-identifiability faced by the frequentist approach because they use probability constraints (called prior distribution) on the parameters to be estimated. Using an informative prior is recommended to make the model identifiable. It is rational to assume that combining information from two reliable source tend to produce accurate estimate and inference.	• LCMs with different conditional dependence structures can produce different estimates and still fit same data. • To reduce the impact of the prior information on the estimates being derive, it is rational to have sufficient data size. It is recommended that accurate and precise priors be used to yield accurate posterior inference. However, imprecise priors can be used provided they are initialize with accuracy (McDonald and Hodgson [416]).	Branscum, Gardner [140] Berkvens, Speybroeck [310]

Table 55 cont.: Methods employed to evaluate index test(s) when the sensitivity and specificity of the imperfect reference standard is unknown.

Methods	Characteristics	Strengths	Weaknesses	Key reference
Frequentist LCM	Frequentist LCMs use information from the observed data only to compute the estimate of the sensitivity and specificity such as the sensitivity, and specificity of the index test. **Examples of frequentist LCMs** ***Standard / Traditional two class latent class model (TLCM)*** by Hui and Zhou [128],[417] assumes the tests are conditional independent (CI) and their sensitivities and specificities are constant across all population. **Clinical application of the TLCM** 123, 168, 169, 174-178 • ***MECM approach***: Kang, Carter [418] applied binary, independent assumption • ***Log-linear latent class model (LLCM)*** (Hagenaars [129]) relaxes the CI assumption of the traditional LCM and models conditional dependence using log-linear models. • ***Probit latent class model (PLCM)*** (Uebersax [130]) relaxes the CI assumption and assumes that the K latent classes follow some multi-variate normal distribution with a mean vector and covariance matrix. • **Xu [135] *extended LLCM and PLCM*** to take into consideration intermediate responses of the tests if they are available. Hence, the diagnostic outcome is not dichotomized but three-classed.	• This approach is based on the observed data only. • Some of the frequentist approach are straight forward like the TLCM, LLCM; however, some are computational complex because of the conditional dependence structure across the tests like the PLCM, two-random effect LCM, • Some Latent class models have some advantages over the others as well as some disadvantages; thus, it is worth reading through the original articles where the models were proposed to select which models is appropriate for the analysis.	• Choice of model are restricted because of non-identifiability problem. Hence, strict assumptions are made to make the models identifiable. The estimates obtained from the LLCM approach is not interpretable as the diagnostic accuracy of the tests evaluated.	

Table 55 cont.: Methods employed to evaluate index test(s) when the sensitivity and specificity of the imperfect reference standard is unknown.

Methods	Characteristics	Strengths	Weaknesses	Key reference
Frequentist LCM	• ***Gaussian random effect*** (GRE) ***LCM*** by Qu, Tan [132] estimates the diagnostic accuracies of multiple tests (or raters) and uses the Gaussian distribution to model the conditional dependence structure across the tests. **Clinical application of the RE LCM**: [170-172, 179] • ***Finite mixture (FM) model*** by Albert, McShane [133] estimates the accuracies of multiple tests but use finite mixture to model the conditional dependence across the tests. GRE and FM are subject specific, and the test sensitivity and specificity are fixed. • ***Two-cross random effect LCM*** by Zhang [134] estimates the accuracies of multiple tests with conditional dependence by using the Monte Carlo Expectation Maximization (**MCEM**) algorithm to maximize the full likelihood of the data. **Clinical application of two-cross random effect LCM** Xie [173] • Alternative to the MCEM two-cross RE is the ***maximum pseudo-likelihood estimation via Newton Raphson (NR) algorithm*** by Liu [138]. This approach aims to reduce the computational time of two-cross RE especially if there are many participants and tests by constructing some class of pseudo-like function (pairwise likelihood, triple wise likelihood, hybrid likelihood, and dimensional-wise likelihood) which can be maximized using the expectation maximization (EM) or Newton Raphson algorithm. The latent class model developed by Xue, Oktay [139] estimates the accuracy of two tests (with conditional dependence) specifically designed to diagnosis tumor mutations (molecular testing). It is a modification of the Random effect LCM by Qu, Tan [132]	• This approach is based on the observed data only. • Some of the frequentist approach are straight forward like the TLCM, LLCM; however, some are computational complex because of the conditional dependence structure across the tests like the PLCM, two-random effect LCM, • Some Latent class models have some advantages over the others as well as some disadvantages; thus, it is worth reading through the original articles where the models were proposed to select which models is appropriate for the analysis.	• Choice of model are restricted because of non-identifiability problem. Hence, strict assumptions are made to make the models identifiable. The estimates obtained from the LLCM approach is not interpretable as the diagnostic accuracy of the tests evaluated.	

Table 55 cont.: Methods employed to evaluate index test(s) when the sensitivity and specificity of the imperfect reference standard is unknown.

Methods	Characteristics	Strengths	Weaknesses	Key reference
ROC curve approaches	Estimating the ROC curve using the standard latent class approach [128] for every cutoff and then plotting the curve using estimates obtained from this approach produces a ROC curve that is not monotonic [419]. However, there are methods developed to produce ROC curves in the case of imperfect reference test. They are: • **ROC with ordinal tests** Henkelman, Kay [149]; Beiden, Campbell [150]; Zhou [148]; Wang, Zhou [136] is an extension of [148] to incorporate multiple tests; and Wang [137] is an extension of Wang, Zhou [136] to incorporate conditional dependence structure across the multiple tests being evaluated. • **ROC with continuous tests** Choi, Johnson [151]; Wang [420]; Branscum [153]; Jafarzadeh [421]; Hall and Zhou [422]; Erkanli [154]	This approach estimates the overall diagnostic accuracy of the test(s) being evaluated and does not assume that the test(s) being evaluated is dichotomized.	• Some of the methods are model based so adequate specification of the model is important to derive unbiased estimates. The methods are computational tasking especially when there are multiple tests to evaluate and the assumption of conditional dependence is imposed on the tests being evaluated.	

Table 56: Construction of reference standard.

Methods	Characteristics	Strength	Weakness	Key reference
Composite reference standard (CRS)	CRS uses a predetermined rule to construct a reference standard test using multiple imperfect tests. The participants undergo all the tests (index test and the imperfect reference tests). The reference standard is formed using the responses from the imperfect tests excluding the index test. The index test is evaluated in comparison to the established reference test. The dual CRS (dCRS) proposed by Tang, Hemyari [162], employs the "any positive" rule and "all positive" rule to estimate the sensitivity and specificity of the index test. **Clinical application:** [225, 226, 234, 423]	• Combining different imperfect tests using a predetermined rule to rule in or rule out diagnosis.	• Can have incorporation bias if the index test is also part of the test employed to construct the reference standard. Paine, Basu [424] • Challenges can arise in deciding the number of tests to combine to make the constructed reference standard adequate to discriminate patients or participants with the target condition. • It could be burdensome, especially if the number of tests to combine is many. • The performance of this approach depends on the performance of each test employed as reference standard and the conditional dependence between the tests.	Schiller [160] Naaktgeboren, Bertens [161] Tang, Hemyari [162]
Expert or panel or consensus opinion	This approach employs the decision(s) of expert(s) of a health condition (disease) as the reference standard to evaluate the new or index test. Often, observed covariates like signs and symptoms of the participants or test response of another test (not the index test) can be used together with expert(s) decisions to ascertain the disease status of the participants. **Clinical application**:[234]	• The accuracy of the experts can be close to gold standard because they have good knowledge of the target condition.	• There could be discrepancy across the experts' decisions in confirm the diagnosis of the participants. • This approach could be time-consuming especially if there are large number of participants.	Bertens, Broekhuizen [163]

Table 57: Table of other methods employed to evaluate medical test(s)

Method	Characteristics	Strengths	Weaknesses	Key reference
Study of agreement	The study of agreement looks at how two or more test responses agree or disagree. Often, this approach is used alongside other types of methods like the latent class analysis. Commonly used agreement measures in a diagnostic accuracy study is the Kappa statistic [425] (Cohen kappa and Fleiss or Scott Kappa) and McNemar test [426]. **Clinical application**:[193, 243, 246]	• It is a way to explore the observed data to understand the relationship between the tests' responses.	• Using the approach as a measure of accuracy is not encouraged; because the disagreement or agreement between two or more tests does not imply that one test is better or more efficient than the other.	Zaki, Bulgiba [242]
Validation	With this approach the disease status of the participants in the study are known (case and control). The estimates obtained are often referred to as analytical sensitivity and specificity [240]. This approach assesses the test based on what it is supposed or designed to do. **Clinical application**: [167, 244, 245, 260, 427]	• This approach provides a basic knowledge on the performance of the index test. The estimates can be employed as prior information in Bayesian analysis when the test is applied to participants with unknown disease status.	• In practice, the diagnostic test is designed to be applied to participants who may have or may have not shown signs or symptoms of the disease; thus, the estimated diagnostic accuracy using this approach may not reflect the true diagnostic accuracy of the index test.	Elliott, Applegate [240]
Test positivity rate	This approach estimates the proportion of participants who have positive result in a test. It is often taken that a test with the highest positivity rate compared to other tests have better accuracy[225]. However, this may not be true because these tests are prone to misclassification error (that is they are not gold standards) and the number of participants with positive result could depend on the prevalence of the target condition in the sample (sub-population) that is being studied. **Clinical application**: [225, 239, 243]	• This approach gives possible estimates of the sensitivities of the tests employed in the study. • It has been used to assess whether to include or exclude some tests in diagnostic accuracy studies (Van Dyck, Buvé [239]).	• This approach should not be used as a standalone analysis to decide the accuracy of index test; because having a positive result to an index test does not imply presence of target condition especially if the test is imperfect.	

B.1. R Code Chapter three – comparison of correction methods

This section explains the R-Code employed to simulate the different datasets explored in this paper and to analyse the clinical datasets.

1. Calculate the cell probabilities for multinomial distribution using the fixed effect modelling approach[54, 286]

```{r cell probabilities}
proba<- function(pd,sRS,spRS, sIT, spIT, cova1, cova2){
#sRS and spRS are sensitivity and specificity of RS respectively
#sIT and spIt are sensitivity and specificity of IT respectively
#cova1 is the covariance term among the diseased group
#cova2 is the covariance term among the non-diseased group
#pd is the prevalence of the target condition
a<- pd*(sRS*sIT + cova1)+((1 -pd)*(1-spRS)*(1-spIT) + cova2)
  c<- pd*(sRS*(1 - sIT) - cova1) + ((1 -pd)*( 1- spRS)*spIT + cova2)
  b<- pd*((1 - sRS)*sIT - cova1) + ((1 -pd)*spRS*(1-spIT) + cova2)
  d<- pd*((1 - sRS)*(1 - sIT) + cova1) + ((1 -pd)*spRS*spIT + cova2)
prom<- c(a, c, b, d)
return(prom)
}
```

2. Code employed to estimate the unadjusted and corrected sensitivity and specificity of the index test.

```{r fun1}
cal<- function(dtab, sRS, spRS){
#dtab is the 2 by 2 matrix simulated using the multinomial distribution and the cell probability function (#1)
Np<- sum(dtab[1,1],dtab[1,2],dtab[2,1],dtab[2,2]) # total number of participants
  e<- sum(dtab[1,1], dtab[2,1]) # a+c total RS positive
  f<- sum(dtab[1,2], dtab[2,2]) # b+d total RS negative
  g<- sum(dtab[1,1], dtab[1,2]) #a+b total IT positive
  h<- sum(dtab[2,1], dtab[2,2]) # c+d # total IT negative
  prev<- e/Np # sample prevalence
  senIT <- dtab[1,1]/ e  # unadjusted sensitivity of index test
  specIT<- dtab[2,2]/f # unadjusted specificity of index test
  senbre<- (prev*sRS*senIT + (1 - prev)*(1 - spRS)*(1 - specIT))/(prev*sRS + (1 - prev)*(1 - spRS)) # Brenner corrected sensitivity
  specbre<- (prev*(1 - sRS)*(1-senIT) + (1 - prev)*(spRS)*(specIT))/(prev*(1-sRS) + (1 - prev)*spRS) # Brenner corrected specificity
  senstaq<- (g*spRS - dtab[1,2])/ (Np*(spRS - 1) + e) # Staquet et al corrected sensitivity
```

```
  specstaq<- (h*sRS - dtab[2,1])/(Np*sRS - e) # Staquet et al corrected specificity
  estpre<- (prev + spRS - 1)/(sRS + spRS - 1) # estimated prevalence
  result<- c(senIT, specIT, senbre, specbre, senstaq, specstaq, estpre, prev)
}
```

3. Code employed to estimate the covariance inequalities (boundary to decide the choice of covariance terms given that the index test and reference standard are conditionally dependent)

```{r covabound}
covabound<- function(sRS, spRS, sIT, spIT){
  lcovsen<- (-sRS * sIT) + max(0, sRS + sIT -1) #lower value for covariance among diseased group
  ucovsen<- min(sRS,sIT) - (sRS * sIT) #upper value of covariance among the diseased group
  lcovspec<- -spRS*spIT + max(0, spRS +spIT - 1) #lower value for covariance among non – diseased group
    ucovaspec<- min(spRS,spIT) - (spRS * spIT) #upper value of covariance among the diseased group
return (cbind(c("lower sen", "upper sen", "lower spec", "upper spec"),c(lcovsen, ucovsen, lcovspec, ucovaspec)))
}
```

4. Code to generate random samples of 2 by 2 tables under the assumption of conditional independence and conditional depedence using the possible covariance terms using the cell probabilties function.

```{r multinomial}
sim<- function(numb,n,pd,sRS,spRS, sIT, spIT, cova1, cova2){
  # n is the sample size (number of participants)
  # numb is the number of samples simulated with size n
prom<- proba(pd,sRS,spRS, sIT, spIT, cova1, cova2)
exp<- rmultinom(numb,n,prom)
tab<- list()
  for(i in 1:numb){
    tab[[i]]<- matrix(exp[,i],2, 2)
  }
return(tab)
}
```

Example

```
set.seed(1235679)
exsim<- sim(1,100,0.9,1,1,0.8,0.7,0.05, 0.05)
```

5. Estimate the unadjusted and corrected sensitivty and specficty from numb samples simulated

```{r solve}
sol<- function(numb,n,pd,sRS,spRS, sIT, spIT, cova1, cova2){
  tabu<- sim(numb, n, pd, sRS, spRS, sIT, spIT, cova1, cova2)
  mat<- matrix(NA, numb, 8)
  for (i in 1:numb){
   mat[i,] <-  cal(tabu[[i]], sRS,spRS)
  }
  colnames(mat) <- c("Unadjsen","Unadjspec", "Sen.Brenner", "Spec.Brenner",
"Sen.Staquet", "Spec.Staquet", "EstPre", "Sam.Prev")
  tabmat<- list(tabu, mat)
  return(tabmat)
}
```

Example
set.seed(1235679)
sol(10, 50, 0.3, 0.9, 0.9, 0.8, 0.7, 0,0)[[2]][,1]

6. Obtain the mean values of estimates, standard deviation, mean square error (MSE) and bias.

#Call up the required packages in R
```{r call}
library(gstat)
library(e1071)
library(hydroGOF)
```

Function to estimate the Mean, MSE, SD and bias of the unadjusted and corrected sensitivity and specificity of index test
```{r descriptive}
desol<- function(numb,n,pd,sRS,spRS, sIT, spIT, cova1, cova2){
  msol<- sol(numb,n,pd,sRS,spRS, sIT, spIT, cova1, cova2)[[2]]
msol1<- msol[!rowSums(!is.finite(msol)),]# remove rows with inf values or non- finite values.
msol2<- msol1[!rowSums(msol1 > 2),] # remove rows with any value above 2.
#Values above 1 or below 0 are obtained via the Staquet et al approach.
mval<- apply(msol1, 2, mean)
sval<- apply(msol1, 2, sd)
new<- msol1
numb1<- length(msol1[,1])
```

```
para<- cbind(rep(sIT, numb1),rep(spIT, numb1),rep(sIT, numb1), rep(spIT, numb1),rep(sIT, numb1), rep(spIT, numb1),rep(pd, numb1),rep(pd, numb1))
msqerror<- mse(new, para)
realval<- c(sIT, spIT, sIT, spIT, sIT, spIT, pd, pd)
BiasEP<- abs(mval - realval)
MCerr<- sval/sqrt(numb)
tog<- cbind(mval,sval, msqerror, BiasEP, MCerr)
### returns only the mean value, standard deviation and MSE
return(tog)
}
```

Simulated examples

Imperfect test RS better than IT, IT and RS are conditionally independent

```{r example1}
set.seed(1235679)
example0<- desol(200,50,0.3, 0.9, 0.9, 0.8, 0.7, 0.00, 0.00)
example1<- desol(200,80,0.3, 0.9, 0.9, 0.8, 0.7, 0.00, 0.00)
example2<- desol(200,100,0.3,0.9, 0.9, 0.8, 0.7, 0.00, 0.00)
example3<- desol(200,120,0.3, 0.9, 0.9, 0.8, 0.7, 0.00, 0.00)
example4<- desol(200,150,0.3, 0.9, 0.9, 0.8, 0.7, 0.00, 0.00)
example5<- desol(200,180,0.3, 0.9, 0.9, 0.8, 0.7, 0.00, 0.00)
example6<- desol(200,200,0.3, 0.9, 0.9, 0.8, 0.7, 0.00, 0.00)
example7<- desol(200,250,0.3, 0.9, 0.9, 0.8, 0.7, 0.00, 0.00)
example8<- desol(200,300,0.3, 0.9, 0.9, 0.8, 0.7, 0.00, 0.00)
example9<- desol(200,350,0.3, 0.9, 0.9, 0.8, 0.7, 0.00, 0.00)
example10<- desol(200,400,0.3, 0.9, 0.9, 0.8, 0.7, 0.00, 0.00)
example11<- desol(200,500,0.3, 0.9, 0.9, 0.8, 0.7, 0.00, 0.00)
example12<- desol(200,600,0.3, 0.9, 0.9, 0.8, 0.7, 0.00, 0.00)
example13<- desol(200,700,0.3, 0.9, 0.9, 0.8, 0.7, 0.00, 0.00)
example14<- desol(200,800,0.3, 0.9, 0.9, 0.8, 0.7, 0.00, 0.00)
example15<- desol(200,900,0.3, 0.9, 0.9, 0.8, 0.7, 0.00, 0.00)
example16<- desol(200,1000,0.3, 0.9, 0.9, 0.8, 0.7, 0.00, 0.00)
```

```{r example2}
### Imperfect test RS worse than IT, IT and RS are conditionally independent
set.seed(1235679)
example0<- desol(200,50,0.3, 0.8, 0.7,0.9, 0.9,  0.00, 0.00)
example1<- desol(200,80,0.3, 0.8, 0.7,0.9, 0.9, 0.00, 0.00)
example2<- desol(200,100,0.3,0.8, 0.7,0.9, 0.9, 0.00, 0.00)
example3<- desol(200,120,0.3, 0.8, 0.7, 0.9, 0.9, 0.00, 0.00)
example4<- desol(200,150,0.3, 0.8, 0.7, 0.9, 0.9, 0.00, 0.00)
example5<- desol(200,180,0.3,0.8, 0.7, 0.9, 0.9,  0.00, 0.00)
```

```
example6<- desol(200,200,0.3,0.8, 0.7, 0.9, 0.9,  0.00, 0.00)
example7<- desol(200,250,0.3, 0.8, 0.7, 0.9, 0.9,  0.00, 0.00)
example8<- desol(200,300,0.3, 0.8, 0.7, 0.9, 0.9,  0.00, 0.00)
example9<- desol(200,350,0.3, 0.8, 0.7, 0.9, 0.9,  0.00, 0.00)
example10<- desol(200,400,0.3, 0.8, 0.7, 0.9, 0.9,  0.00, 0.00)
example11<- desol(200,500,0.3, 0.8, 0.7, 0.9, 0.9,  0.00, 0.00)
example12<- desol(200,600,0.3,0.8, 0.7,  0.9, 0.9,  0.00, 0.00)
example13<- desol(200,700,0.3,0.8, 0.7, 0.9, 0.9,  0.00, 0.00)
example14<- desol(200,800,0.3, 0.8, 0.7,0.9, 0.9,  0.00, 0.00)
example15<- desol(200,900,0.3, 0.8, 0.7,0.9, 0.9,  0.00, 0.00)
example16<- desol(200,1000,0.3,0.8, 0.7, 0.9, 0.9,  0.00, 0.00)
```

Put performance measures in a single data frame

```{r data1}
toptab<- cbind(example0[,2],example1[,2], example2[,2], example3[,2], example4[,2], example5[,2], example6[,2], example7[,2], example8[,2], example9[,2], example10[,2], example11[,2], example12[,2], example13[,2], example14[,2], example15[,2], example16[,2])
samsize<- c(50, 80,100, 120, 150, 180, 200, 250, 300, 350, 400, 500, 600, 700, 800, 900, 1000)
ttab1<- t(toptab)
ttab<- ttab1[,1:6]
colnames(ttab)<- c("sdUnadjsen","sdUnadjspec", "sdSen.Brenner", "sdSpec.Brenner", "sdSen.Staquet", "sdSpec.Staquet")

dimtab<- cbind(example0[,1],example1[,1], example2[,1], example3[,1], example4[,1], example5[,1], example6[,1], example7[,1], example8[,1], example9[,1], example10[,1], example11[,1], example12[,1], example13[,1], example14[,1], example15[,1], example16[,1])
tdimtab1<- t(dimtab)
tdimtab<- tdimtab1[,1:6]
colnames(tdimtab)<- c("meanUnadjsen","meanUnadjspec", "meanSen.Brenner", "meanSpec.Brenner", "meanSen.Staquet", "meanSpec.Staquet")

msqtab<- cbind(example0[,3],example1[,3], example2[,3], example3[,3], example4[,3], example5[,3], example6[,3], example7[,3], example8[,3], example9[,3], example10[,3], example11[,3], example12[,3], example13[,3], example14[,3], example15[,3], example16[,3])
tmsqtab1<- t(msqtab)
tmsqtab<- tmsqtab1[,1:6]
colnames(tmsqtab)<- c("msqUnadjsen","msqUnadjspec", "msqSen.Brenner", "msqSpec.Brenner", "msqSen.Staquet", "msqSpec.Staquet")
```

biastab<- cbind(example0[,4],example1[,4], example2[,4], example3[,4], example4[,4], example5[,4], example6[,4], example7[,4], example8[,4], example9[,4], example10[,4], example11[,4], example12[,4], example13[,4], example14[,4], example15[,4], example16[,4])

tbtab1<- t(biastab)

tbtab<- tbtab1[,1:6]

colnames(tbtab)<- c("biasUnadjsen","biasUnadjspec", "biaSen.Brenner", "biaSpec.Brenner", "biaSen.Staquet", "biaSpec.Staquet")

MCerrtab<- cbind(example0[,5],example1[,5], example2[,5], example3[,5], example4[,5], example5[,5], example6[,5], example7[,5], example8[,5], example9[,5], example10[,5], example11[,5], example12[,5], example13[,5], example14[,5], example15[,5], example16[,5])

MCtab1<- t(MCerrtab)

MCtab<- MCtab1[,1:6]

colnames(MCtab)<- c("MCerrUnadjsen","MCerrUnadjspec", "MCerrSen.Brenner", "MCerrSpec.Brenner", "MCerrSen.Staquet", "MCerrSpec.Staquet")

samtab<- round(data.frame(samsize, ttab,tdimtab, tmsqtab, tbtab, MCtab),4)

```

Other possible variations or conditions can be explored by changing the values of the sensitivities and specificities of IT and or RS and prevalence.

7. Plot the unadjusted and corrected mean sensitivty and specificty of IT alongside the SD, Bias and MSE

## call up required R packages

```{r library}
library(dplyr)
library(tidyr)
library(ggplot2)
library(reshape2)
library(gridExtra)
```

plot performance measures against sample size

```{r plot2}
My_Theme = theme(axis.title.x = element_text(size = 16),axis.text.x = element_text(size = 14),axis.title.y = element_text(size = 16), axis.text.y = element_text(size = 14), legend.title=element_text(size=12),legend.text=element_text(size=12))

df <- melt(samtab[,c("samsize","sdUnadjsen", "sdSen.Brenner", "sdSen.Staquet")], id="samsize")

```
pg <- df    # Copy data into new data frame
# Rename the column and the values in the factor
levels(pg$variable)[levels(pg$variable)=="sdUnadjsen"] <- "Unadjusted"
levels(pg$variable)[levels(pg$variable)=="sdSen.Brenner"] <- "Brenner"
levels(pg$variable)[levels(pg$variable)=="sdSen.Staquet"] <- "Staquet"
names(pg)[names(pg)=="variable"]  <- "Standard.Error"

p<- ggplot(pg, aes(x=samsize, y=value, col= Standard.Error)) + geom_line() + labs(x ="Sample size", y = "SE sensitivity")   + geom_line(size = 1) +
coord_cartesian(ylim=c(0.0,0.3))+ My_Theme
#+ geom_hline(yintercept=0.9, linetype="dashed", color = "yellow", size = 2)

##### specificity
df1 <- melt(samtab[,c("samsize","sdUnadjspec", "sdSpec.Brenner", "sdSpec.Staquet")], id="samsize")
pg1 <- df1   # Copy data into new data frame
# Rename the column and the values in the factor
levels(pg1$variable)[levels(pg1$variable)=="sdUnadjspec"] <- "Unadjusted"
levels(pg1$variable)[levels(pg1$variable)=="sdSpec.Brenner"] <- "Brenner"
levels(pg1$variable)[levels(pg1$variable)=="sdSpec.Staquet"] <- "Staquet"
names(pg1)[names(pg1)=="variable"]  <- "Standard.Error"

p1<- ggplot(pg1, aes(x=samsize, y=value, col= Standard.Error)) + geom_line() + labs(x ="Sample size", y = "SE specificity")  + geom_line(size = 1) +
coord_cartesian(ylim=c(0.0,0.07))+ My_Theme
#+ geom_hline(yintercept=0.9, linetype="dashed", color = "yellow", size = 2)

### put both plot as one
grid.arrange(p, p1, nrow=2)
```

8. Estimate the mean sensitivity and specificity of IT at varying prevalences

```
## Estimate only the mean sensitivity and specificity of IT
```{r meansol}
meansol<- function(numb,n,pd,sRS,spRS, sIT, spIT, cova1, cova2){
  msol<- sol(numb,n,pd,sRS,spRS, sIT, spIT, cova1, cova2)[[2]]
  msol1<- msol[!rowSums(!is.finite(msol)),]# remove rows with inf values or non- finite values.
  msol2<- msol1[!rowSums(msol1 > 2),] # exclude rows greater than 2
  msol3<- msol2[!rowSums(msol2 < 0),] #exclude rows less than zero
meanval<- apply(msol3, 2, mean)
return(meanval)
```
```

```
}
```

Code that estimates the mean values of the estimator at different prevalence

```{r soldiff}
soldiff<- function (z,numb,n,sRS,spRS, sIT, spIT, cova1, cova2){
  pd<- seq(0.00, 1, length.out = z)
  top<- list()
  for(i in 1:z){
    top[[i]]<- meansol(numb,n,pd[i],sRS,spRS,sIT,spIT, cova1, cova2)
  }
  tim<- matrix(NA,z, 8)
  for(i in 1:z){
    tim[i,]<- top[[i]]
  }
  ss<- rep(n, z)
 colnames(tim) <- c("Unadjsen","Unadjspec", "Sen.Brenner", "Spec.Brenner",
"Sen.Staquet", "Spec.Staquet", "EstPre", "Sam.Prev")
 timpd<- cbind(pd, tim,ss)#data.frame
return(timpd)
}
```

Simulate 1000 participants, 200 multiple samples, 100 prevelances. RS is better than IT and RS is imperfect

```{r data}
set.seed(1235679)
preout<- soldiff(100,200,1000, 0.9, 0.9, 0.8, 0.8, 0.00, 0.00)
preout<- data.frame(preout)
```

plot the mean sensitivity and specificity

```{r plot2}
My_Theme = theme(axis.title.x = element_text(size = 16),axis.text.x =
element_text(size = 14),axis.title.y = element_text(size = 16), axis.text.y =
element_text(size = 14),
legend.title=element_text(size=12),legend.text=element_text(size=12))

df <- melt(preout[,c("pd","Unadjsen", "Sen.Brenner", "Sen.Staquet")], id="pd")
pg <- df    # Copy data into new data frame
# Rename the column and the values in the factor
levels(pg$variable)[levels(pg$variable)=="Unadjsen"] <- "Unadjusted"
levels(pg$variable)[levels(pg$variable)=="Sen.Brenner"] <- "Brenner"
```

```
levels(pg$variable)[levels(pg$variable)=="Sen.Staquet"] <- "Staquet"
names(pg)[names(pg)=="variable"]  <- "Mean"

p<- ggplot(pg, aes(x=pd, y=value, col= Mean)) + geom_line()+ geom_point() + labs(x ="Prevelance", y = "Mean sensitivity")   + geom_line(size = 1) + coord_cartesian(ylim=c(0.2,1))+ My_Theme + geom_hline(yintercept=0.8, linetype="dashed", color = "yellow", size = 2)

##### specificity
df1 <- melt(preout[,c("pd","Unadjspec", "Spec.Brenner", "Spec.Staquet")], id="pd")
pg1 <- df1   # Copy data into new data frame
# Rename the column and the values in the factor
levels(pg1$variable)[levels(pg1$variable)=="Unadjspec"] <- "Unadjusted"
levels(pg1$variable)[levels(pg1$variable)=="Spec.Brenner"] <- "Brenner"
levels(pg1$variable)[levels(pg1$variable)=="Spec.Staquet"] <- "Staquet"
names(pg1)[names(pg1)=="variable"]  <- "Mean"

p1<- ggplot(pg1, aes(x=pd, y=value, col= Mean)) + geom_line() + geom_point() + labs(x ="Prevelance", y = "Mean specificity")  + geom_line(size = 1) + coord_cartesian(ylim=c(0.4,1.2))+ My_Theme + geom_hline(yintercept=0.8, linetype="dashed", color = "yellow", size = 2)



### put both plot as one
grid.arrange(p, p1, nrow=2)
```

9. Estimate the mean corrected and unadjusted sensitivty and specificty of IT assuming sensitivity of RS (or specificty of RS) varies from 0 to 1. Unlike the function in 6 and 8 above the sensitivty of RS and IT, and the specificity of RS and IT are fixed. This allows more possible combinations to examine how the corrections method perform.

```
#### estimate mean sensitivity and specificity by fixing one of these parameters – sRS, spRS, sIT, spIT- and varying the others
```{r soldiff1}
soldiff1<- function (z,pd,numb,n,sRS, spRS, sIT, cova1, cova2){
##in this stated function the spIT is varied and the others are fixed
## to vary another parameter change it appropraitely
  spIT<- seq(0, 1, length.out = z)
  top<- list()
  for(i in 1:z){
    top[[i]]<- meansol(numb,n,pd,sRS,spRS,sIT,spIT[i], cova1, cova2)
  }
  tim<- matrix(NA,z, 8)
```
```

```
  for(i in 1:z){
    tim[i,]<- top[[i]]
  }
  ss<- rep(n, z)
 colnames(tim) <- c("Unadjsen","Unadjspec", "Sen.Brenner", "Spec.Brenner", "Sen.Staquet", "Spec.Staquet", "EstPre", "Sam.Prev")
 timpd<- cbind(spIT, tim,ss)#data.frame
return(timpd)
}
```

explore1000 partcipants, 200 multiple samples

```{r explore}
set.seed(1235679)
preout<- soldiff1(100,0.3,200,1000, 0.9, 0.9, 0.8, 0.00, 0.00)
preout<- data.frame(preout)
```

plot mean sensitivity and specificity

```{r plot2}
My_Theme = theme(axis.title.x = element_text(size = 16),axis.text.x = element_text(size = 14),axis.title.y = element_text(size = 16), axis.text.y = element_text(size = 14), legend.title=element_text(size=12),legend.text=element_text(size=12))

df <- melt(preout[,c("spIT","Unadjsen", "Sen.Brenner", "Sen.Staquet")], id="spIT")
pg <- df    # Copy data into new data frame
# Rename the column and the values in the factor
levels(pg$variable)[levels(pg$variable)=="Unadjsen"] <- "Unadjusted"
levels(pg$variable)[levels(pg$variable)=="Sen.Brenner"] <- "Brenner"
levels(pg$variable)[levels(pg$variable)=="Sen.Staquet"] <- "Staquet"
names(pg)[names(pg)=="variable"]  <- "Mean"

p<- ggplot(pg, aes(x=spIT, y=value, col= Mean)) + geom_line()+ geom_point() + labs(x ="Specificity IT", y = "Mean sensitivity") + geom_line(size = 1) + coord_cartesian(ylim=c(0.4,1))+ My_Theme + geom_hline(yintercept=0.8, linetype="dashed", color = "yellow", size = 2)

##### specificity
df1 <- melt(preout[,c("spIT","Unadjspec", "Spec.Brenner", "Spec.Staquet")], id="spIT")
pg1 <- df1   # Copy data into new data frame
# Rename the column and the values in the factor
levels(pg1$variable)[levels(pg1$variable)=="Unadjspec"] <- "Unadjusted"
levels(pg1$variable)[levels(pg1$variable)=="Spec.Brenner"] <- "Brenner"
levels(pg1$variable)[levels(pg1$variable)=="Spec.Staquet"] <- "Staquet"
```

```
names(pg1)[names(pg1)=="variable"]  <- "Mean"

p1<- ggplot(pg1, aes(x=spIT, y=value, col= Mean)) + geom_line() + geom_point() + labs(x ="Specificity IT", y = "Mean specificity")  + geom_line(size = 1) + coord_cartesian(ylim=c(0,1))+ My_Theme +geom_abline(intercept = 0,slope = 1, color="yellow", linetype="dashed", size=2)
#+ geom_hline(yintercept=0.8, linetype="dashed", color = "yellow", size = 2)

### put both plot as one
grid.arrange(p, p1, nrow=2)
```

10. Code to estimate the sensitivty and specificty of IT which include the Brenner second estimators for positively correlated IT and RS. This pair of estimators is explored in Appendix File 4.

```{r Brenner positive fun}
calpos<- function(dtab, sRS, spRS){
Np<- sum(dtab[1,1],dtab[1,2],dtab[2,1],dtab[2,2]) # total number of participants
  e<- sum(dtab[1,1], dtab[2,1]) # a+c total RS positive
  f<- sum(dtab[1,2], dtab[2,2]) # b+d total RS negative
  g<- sum(dtab[1,1], dtab[1,2]) #a+b total IT positive
  h<- sum(dtab[2,1], dtab[2,2]) # c+d # total IT negative
  prev<- e/Np # prevalence of the diseased in sample of study
  senIT <- dtab[1,1]/ e  # sensitivity of index test unadjusted
  specIT<- dtab[2,2]/f # specificity of index test unadjusted
  senpos<- (prev * senIT +(1 - prev)*(1 - spRS))/(prev*sRS + (1 - prev)*(1 - spRS)) # corrected sensitivity of IT using the positively correlated pair of estimator by Brenner
  specpos<- (prev * ( 1 - sRS) +(1 - prev)*specIT)/(prev*( 1 - sRS) + (1 - prev)*spRS) # corrected sensitivity of IT using the positively correlated pair of estimator by Brenner
  senbre<- (prev*sRS*senIT + (1 - prev)*(1 - spRS)*(1 - specIT))/(prev*sRS + (1 - prev)*(1 - spRS))
  specbre<- (prev*(1 - sRS)*(1-senIT) + (1 - prev)*(spRS)*(specIT))/(prev*(1-sRS) + (1 - prev)*spRS)
  senstaq<- (g*spRS - dtab[1,2])/ (Np*(spRS - 1) + e)
  specstaq<- (h*sRS - dtab[2,1])/(Np*sRS - e)
  estpre<- (prev + spRS - 1)/(sRS + spRS - 1)
  return( c(senpos, specpos, senIT, specIT, senbre, specbre, senstaq, specstaq))
}
```

Code to generate random samples of estimate the mean values

```{r Brenner positive}
```

```
solpos<- function(numb,n,pd,sRS,spRS, sIT, spIT, cova1, cova2){
  tabu<- sim(numb, n, pd, sRS, spRS, sIT, spIT, cova1, cova2)
  mat<- matrix(NA, numb, 8)
  for (i in 1:numb){
   mat[i,] <-  calpos(tabu[[i]], sRS,spRS)
  }
  colnames(mat) <- c("Sen.pos.Bre", "Spec.pos.Bre", "Unadjsen",
"UnadjSpec","Sen.Brenner","Spec.Brenner","Sen.Staquet", "Spec.Staquet")
  meanmat<- apply(mat, 2, mean)
  #tabmat<- list(tabu, mat)
  return(meanmat)
}
```

Code to estimate the mean values at different prevelances and result

```{r soldiff}
solposdiff<- function (z,numb,n,sRS,spRS, sIT, spIT, cova1, cova2){
  pd<- seq(0, 1, length.out = z)
  top<- list()
  for(i in 1:z){
    top[[i]]<- solpos(numb,n,pd[i],sRS,spRS,sIT,spIT, cova1, cova2)
  }
  ss<- rep(n, z)
  tim<- matrix(NA,z, 8)
  for(i in 1:z){
    tim[i,]<- top[[i]]
  }
 colnames(tim) <- c("Sen.pos.Bre", "Spec.pos.Bre", "Unadjsen",
"UnadjSpec","Sen.Brenner", "Spec.Brenner","Sen.Staquet", "Spec.Staquet")
 timpd<- data.frame(pd, tim, ss)
return(timpd)
}
```

simulate dataset with IT and RS conditionally dependent and covariance terms among the disease and non-diseased group are 0.05.

```{r solposdiff}
set.seed(1235679)
preout<- solposdiff(100,200,1000,0.9,0.9,0.8,0.8,0.05,0.05)
preout<- data.frame(preout)
```

plot the mean values against the prevalences

```{r plot2}
My_Theme = theme(axis.title.x = element_text(size = 16),axis.text.x = element_text(size = 14),axis.title.y = element_text(size = 16), axis.text.y = element_text(size = 14), legend.title=element_text(size=12),legend.text=element_text(size=12))

df <- melt(preout[,c("pd","Unadjsen", "Sen.Brenner", "Sen.Staquet", "Sen.pos.Bre")], id="pd")
pg <- df    # Copy data into new data frame
# Rename the column and the values in the factor
levels(pg$variable)[levels(pg$variable)=="Unadjsen"] <- "Unadjusted"
levels(pg$variable)[levels(pg$variable)=="Sen.Brenner"] <- "Brenner"
levels(pg$variable)[levels(pg$variable)=="Sen.Staquet"] <- "Staquet"
levels(pg$variable)[levels(pg$variable)=="Sen.pos.Bre"] <- "BrennerPos"
names(pg)[names(pg)=="variable"]  <- "Mean"

p<- ggplot(pg, aes(x=pd, y=value, col= Mean)) + geom_line()+ geom_point() + labs(x ="Prevelance", y = "Mean sensitivity")   + geom_line(size = 1) + coord_cartesian(ylim=c(0.0,1))+ My_Theme + geom_hline(yintercept=0.8, linetype="dashed", color = "yellow", size = 2)

##### specificity
df1 <- melt(preout[,c("pd","UnadjSpec", "Spec.Brenner", "Spec.Staquet", "Spec.pos.Bre")], id="pd")
pg1 <- df1   # Copy data into new data frame
# Rename the column and the values in the factor
levels(pg1$variable)[levels(pg1$variable)=="UnadjSpec"] <- "Unadjusted"
levels(pg1$variable)[levels(pg1$variable)=="Spec.Brenner"] <- "Brenner"
levels(pg1$variable)[levels(pg1$variable)=="Spec.Staquet"] <- "Staquet"
levels(pg1$variable)[levels(pg1$variable)=="Spec.pos.Bre"] <- "BrennerPos"
names(pg1)[names(pg1)=="variable"]  <- "Mean"

p1<- ggplot(pg1, aes(x=pd, y=value, col= Mean)) + geom_line() + geom_point() + labs(x ="Prevelance", y = "Mean specificity")  + geom_line(size = 1) + coord_cartesian(ylim=c(0.0,1))+ My_Theme + geom_hline(yintercept=0.8, linetype="dashed", color = "yellow", size = 2)

### put both plot as one
grid.arrange(p, p1, nrow=2)
```

11. Calculate the sensitivity and specificity of the clinical dataset

```{r clinical1}
### use the cal function and put the sRS and spRS appropriately.
## Matos et al NC dataset
tabLF1<- matrix(c(241, 110,6,26), 2, 2) # matrix for LFpen
tabFC1<- matrix(c(156, 195,2,30), 2, 2) # matrix for FC
tiLF1<- cal(tabLF1, 0.796,0.799) # estimates for LFpen
tiFC1<- cal(tabFC1, 0.796,0.799) # estimates for FC
tiLF1
tiFC1
```

D3 classification
```{r clinical2}
tabLF1<- matrix(c(20, 1,45,341), 2, 2) # matrix for LFpen
tabFC1<- matrix(c(21, 0,38,348), 2, 2) # matrix for FC
tiLF1<- cal(tabLF1, 0.786, 0.995) # estimates for LFpen
tiFC1<- cal(tabFC1, 0.786, 0.995) # estimates for FC
tiLF1
tiFC1
```

12. Calculate the 95% confidence interval of clinical dataset using the Wilson score interval

Code Wilson score interval
```{r wilson}
wilfun<- function(p, n){
  z<- qnorm(1-0.05/2)
  rt<- 1/(1 + (z^2 /n))
  rt1<- p + (z^2/(2*n))
  rtp<- rt*rt1
  vart<- ((p*(1 - p))/n) + ((z^2)/(4*(n^2)))
  zvar<- (z/(1 + ((z^2)/n))) * sqrt(vart)
  LL<- rtp-zvar
  UL<- rtp+zvar
  return(c(LL, UL))
}
```

calculate 95%CI Mathew dataset
```{r cal}
wilsu<- wilfun(0.65, 62) # sensitivity unadjusted
wilsb<- wilfun(0.5,62 ) # Brenner corrected sensitivity
```

```
wilss<- wilfun(0.89,62)# Staquet corrected sensitivity
wilpu<- wilfun(0.89, 199)# unadjusted specificity
wilpb<- wilfun(0.85, 199) # Brenner corrected specificity
wilps<- wilfun(0.96, 199) # Staquet et al corrected specificity
wilsu
wilsb
wilss
wilpu
wilpb
wilps
```

calculate 95% CI of the sensitivity of LFpen for NC detection

```{r cal}
wilu<- se(351, 32, 0.81, 0.69)
wilb<- se(351, 32,0.44, 0.68)
wils<- se(351, 32, 0.04, 0.70)
wilu
wilb
wils
```

calculate 95% CI of the sensitivity of FC for NC detection

```{r cal}
wilu<- se(351, 32, 0.91, 0.44)
wilb<- se(351, 32, 0.65, 0.44)
wils<- se(351, 32, 0.36, 0.45)
wilu
wilb
wils
```

calculate the 95% CI for unadjusted specificity FC- D3 dataset

```{r test}
install.packages("DescTools")
library("DescTools")
BinomCI(348, 386, 0.95, sides = "two.sided", method = "wilson")
```

Present result using barchart

D3 dataset

```{r barchart}

## sensitivity for Lfpen

dat<- data.frame(Examiners = factor(c("1","2", "1","2", "1","2")), levels = (c("1","2")),
Method = factor(c("Unadjusted", "Unadjusted", "Brenner", "Brenner","Staquet", "Staquet")),
Sensitivity = c(0.952,1, 0.865, 0.91, 1.037, 1.087))

### sensitivty for FC

bat<- data.frame(Examiners = factor(c("1","2", "1","2", "1","2")), levels = (c("1","2")),
Method = factor(c("Unadjusted", "Unadjusted", "Brenner", "Brenner","Staquet", "Staquet")),
Sensitivity = c(1, 0.905, 0.905, 0.822, 1.092, 0.985))

## specificty  for Lfpen

dat1<- data.frame(Examiners = factor(c("1","2", "1","2", "1","2")), levels = (c("1","2")),
Method = factor(c("Unadjusted", "Unadjusted", "Brenner", "Brenner","Staquet", "Staquet")),
Specificity = c(0.883, 0.860, 0.874, 0.850, 0.896, 0.873))

## specificity for FC

bat1<- data.frame(Examiners = factor(c("1","2", "1","2", "1","2")), levels = (c("1","2")),
Method = factor(c("Unadjusted", "Unadjusted", "Brenner", "Brenner","Staquet", "Staquet")),
Specificity = c(0.902, 0.881, 0.891, 0.872, 0.915, 0.893))

```

C.1. R Code Chapter Four – Simulation of datasets for investigation of LCMs

In chapter four, various latent class models were explored to understand how they perform when employed to estimate the sensitivity and specificity of three tests that are correlated among themselves. The function (R-Code) employed to simulate the dataset are presented below:

Using the fixed effect modelling approach to generate or simulate the data of concerns
Calculate covariance boundaries

```{r covabound}

covabounds<- function(s1,s2,s3){

#s1, s2, s3 are sensitivity however they can be interchanged with specificity in this function

  Lcs111<- -(s1*s2*s3)

  Ucs111<- min(s1, min(s2, s3))-s1*s2*s3

  Lcs011<- -((1 - s1)*s2*s3)

  Ucs011<- min(1 - s1, min(s2, s3))-((1 -s1)*s2*s3)

  Lcs001<- -((1 - s1)*(1 - s2) *s3)

  Ucs001<- min((1 - s1), min(1 - s2, s3))-((1 - s1)*(1 - s2)*s3)

  Lcs000<- -((1 - s1)*(1 - s2) * (1 - s3))

  Ucs000<- min((1 - s1), min(1 - s2, 1 - s3))-((1-s1)*(1 -s2)*(1 - s3))

  Lcs100<- -(s1*(1 - s2)*(1 -s3))

  Ucs100<- min(s1, min(1 - s2, 1 - s3))-(s1*(1 - s2)*(1 - s3))

  Lcs101<- -(s1*(1 - s2)*s3)

  Ucs101<- min(s1, min(1 - s2, s3))-(s1*(1 - s2)*s3)

  Lcs110<- -(s1*s2*(1 - s3))

  Ucs110<- min(s1, min(s2, 1 - s3))-(s1*s2*(1 -s3))

  Lcs010<- -(s1*(1 - s2)*s3)

  Ucs010<- min(s1, min(1 - s2, s3))-(s1*(1 - s2)*s3)

  cbs<- matrix(c(Lcs010, Lcs110, Lcs101, Lcs100, Lcs000, Lcs001, Lcs011, Lcs111, Ucs010, Ucs110, Ucs101, Ucs100, Ucs000, Ucs001, Ucs011, Ucs111), ncol=8, byrow= TRUE)

   rownames(cbs)<- c("Lower", "Upper")

   colnames(cbs)<- c("010", "110", "101","100", "000", "001", "011", "111")

  return(cbs)

}
```

```
### check out if the covariances sums up to zero as defined by Wang et al.
```{r probT}
# check out something
checkcov<- function(cs011,cs111,cs000,cs001,cc011, cc111,cc000,cc001){
  #covariance term for diseased group
cs100<- cs011+cs111-cs000
cs101<- -(cs001+cs011+cs111)
cs110<- cs000+cs001-cs111
cs010<- -(cs000+cs001+cs011)
# for non-diseased group
cc100<- cc011+cc111-cc000
cc101<- -(cc001+cc011+cc111)
cc110<- cc001+cc000-cc111
cc010<- -(cc000+cc001+cc011)
rim<- c(cs100, cs101, cs110, cs010, cc100,cc101, cc110, cc010)
rimn<- c("cs100", "cs101", "cs110", "cs010", "cc100", "cc101", "cc110", "cc010")
rrim<- rbind(rimn, rim)
return(rrim)
}
```
Function for cell probabilities
```{r proba}
proba<- function(pd,s1,s2,s3,c1,c2,c3,cs011,cs111,cs000,cs001,cc011,
cc111,cc000,cc001){
#s1, s2, s3 are sensitivities of the three tests
#c1, c2, c3 are specificities of the three tests
  #covarance term for diseased group
cs100<- cs011+cs111-cs000
cs101<- -(cs001+cs011+cs111)
cs110<- cs000+cs001-cs111
cs010<- -(cs000+cs001+cs011)
```
```

```
# for non-diseased group
cc100<- cc011+cc111-cc000
cc101<- -(cc001+cc011+cc111)
cc110<- cc001+cc000-cc111
cc010<- -(cc000+cc001+cc011)
  a<- pd*(s1*s2*s3 + cs111) + (1 -pd)*((1-c1)*(1-c2)*(1 -c3) + cc111)#111
  b<- pd*(s1*s2*(1 -s3) + cs110) + (1 -pd)*((1-c1)*(1-c2)*c3 + cc110)#110
  c<- pd*(s1*(1 - s2)*s3 + cs101) + (1 -pd)*((1-c1)*c2*(1 -c3) + cc101)#101
  d<- pd*(s1*(1 - s2)*(1 -s3) + cs100) + (1 -pd)*((1-c1)*c2*c3 + cc100)#100
  e<- pd*((1 -s1)*s2*s3 + cs011) + (1 -pd)*(c1*(1 - c2)*(1 - c3) + cc011)#011
  f<- pd*((1 -s1)*s2*(1 -s3) + cs010) + (1 -pd)*(c1*(1 - c2)*c3 + cc010)#010
  g<- pd*((1 -s1)*(1 - s2)*s3 + cs001) + (1 -pd)*(c1*c2*(1 - c3) + cc001)#001
  h<- pd*((1 -s1)*(1 - s2)*(1 - s3) + cs000) + (1 -pd)*(c1*c2*c3 + cc000)#000
prom<- c(a, b,c,d,e,f,g,h)
return(prom)
}
```

Find the pairwise covariance of the two tests

```{r pairwise}
covs<- function(cs011,cs111,cs000,cs001,cc011, cc111,cc000,cc001){
   #covarance term for diseased group
cs100<- cs011+cs111-cs000
cs101<- -(cs001+cs011+cs111)
cs110<- cs000+cs001-cs111
cs010<- -(cs000+cs001+cs011)
# for non-diseased group
cc100<- cc011+cc111-cc000
cc101<- -(cc001+cc011+cc111)
cc110<- cc001+cc000-cc111
cc010<- -(cc000+cc001+cc011)
#CD pairwise tests diseased group
```

```
cs12<- cs111+cs110
cs13<-cs111+ cs101
cs23<- cs011+cs111
# CD pairwise non-diseased group
cc12<- cc000+cc001
cc13<-cc000+ cc010
cc23<- cc000+ cc100
cvar<- c(cs12, cs13, cs23, cc12, cc13, cc23)
nam<- c("covs12", "covs13", "covs23", "covc12", "covc13", "covc23")
cvarn<- rbind(nam, cvar)
return(cvarn)
}
```

simulate many samples

```{r pract}
set.seed(1235679)
exps<- rmultinom(100,500,pros)# all tests are positively coralated diseased thrid order
expci<- rmultinom(100,500, proci) # all tests are CI
exp23d<- rmultinom(100,500,pro23d) #cs23 = 0.11 # positively correlated but only among test 2 amd 3
exp23cd<- rmultinom(100, 500, pro23cd) #cs23 = 0.11 # positively correlated but only among test 2 amd 3
```

take average of simulated values above

```{r averagesim}
expsm<- matrix(round(apply(exps,1, mean)), 8, 1)
expcim<- matrix(round(apply(expci,1, mean)), 8, 1)
exp23m<- matrix(round(apply(exp23d,1, mean)), 8, 1)
exp23cdm<- matrix(round(apply(exp23cd, 1, mean)),8,1)
```

create tests responses in tablular form

```{r tabular}
```

```
tab<- matrix(c(1,1,1,1,1,0,1,0,1,1,0,0,0,1,1,0,1,0,0,0,1,0,0,0), ncol=3, byrow=TRUE)
res1m<- cbind(tab, expsm)
res2m<- cbind(tab,expcim)
res3m<- cbind(tab, exp23m)
res4m<- cbind(tab,exp23cdm)
write.table(res1m, file = "res1m.txt", sep="\t", row.names = FALSE, col.names = FALSE)
write.table(res2m, file = "res2m.txt", sep="\t", row.names = FALSE, col.names = FALSE)
write.table(res3m, file = "res3m.txt", sep="\t", row.names = FALSE, col.names = FALSE)
write.table(res4m, file = "res4m.txt", sep="\t", row.names = FALSE, col.names = FALSE)
```

put resm1 data on a full dataset style to allow for REM

```{r data1}
#library(mefa)
ai<- data.frame("test 1" =1, "test 2" = 1, "test 3" =1) #111
aip<- rep(ai,77)
bi<- data.frame("test 1" =0, "test 2" = 0, "test 3" =0) #000
bip<- rep(bi,228)
ci<- data.frame("test 1" =1, "test 2" = 1, "test 3" =0) #110
cip<- rep(ci, 39)
di<- data.frame("test 1" =1, "test 2" = 0, "test 3" =1) #101
dip<- rep(di, 21)
ei<- data.frame("test 1" =1, "test 2" = 0, "test 3" =0) #100
eip<- rep(ei, 34)
fi<- data.frame("test 1" =0, "test 2" = 1, "test 3" =1) #011
fip<- rep(fi, 14)
gi<- data.frame("test 1" =0, "test 2" = 1, "test 3" =0) #010
gip<- rep(gi, 60)
hi<- data.frame("test 1" =0, "test 2" = 0, "test 3" =1) #001
hip<- rep(hi, 27)
dataci<- rbind.data.frame(aip,cip, dip, eip, fip, gip, hip,bip)
write.table(dataci, file = "dataci.txt", sep=",", row.names = FALSE, col.names = FALSE)
```

C.2. Openbugs code employed to analyse the simulated dataset

```
# Model 0 - Hui Walter CI assumption – Openbugs
model{
# Likelihood
freqobs[1:8]~dmulti(p[1:8],500)
for (i in 1:8){
    p[i]<-prev*(positive[i])+(1-prev)*(negative[i])
    positive[i]<-(s[1]*a[i]+(1-s[1])*(1-a[i])) * (s[2]*b[i]+(1-s[2])*(1-b[i]))
                * (s[3]*c[i]+(1-s[3])*(1-c[i]))
    negative[i]<-((1-sp[1])*a[i]+sp[1]*(1-a[i])) * ((1-sp[2])*b[i]+sp[2]*(1-b[i]))
                * ((1-sp[3])*c[i]+sp[3]*(1-c[i]))
    }

# Prior
prev~dbeta(1,1)
for (j in 1:3){
    s[j]~dbeta(1,1)
    sp[j]~dbeta(1,1)
    }
}

# Model 1 – FEM – Wang et al Style (assuming CI) – Openbugs
model{
### Likelihood of observed data
freqobs[1:8] ~ dmulti(p[1:8],500)
#=========================================
# prior distributions of prevalence, sensitivities and specificities
#=========================================
prev ~ dbeta(1,1)
for (i in 1:3) {
s[i] ~ dbeta(1,1)
c[i] ~ dbeta (1,1)}
```

```
#=======================================================
# probabilities of observing different cross-classifications
# of two dichotomous diagnostic tests
#=======================================================
p[1]<-prev*(s[1]*s[2]*s[3]+cs111)+(1-prev)*((1-c[1])*(1-c[2])*(1-c[3])+cc111) # 111
p[2]<-prev*(s[1]*s[2]*(1-s[3])+cs110)+(1-prev)*((1-c[1])*(1-c[2])*c[3]+cc110) #110
p[3]<-prev*(s[1]*(1-s[2])*s[3]+cs101)+(1-prev)*((1-c[1])*c[2]*(1-c[3])+cc101) #101
p[4]<-prev*(s[1]*(1-s[2])*(1-s[3])+cs100)+(1-prev)*((1-c[1])*c[2]*c[3]+cc100) #100
p[5]<-prev*((1-s[1])*s[2]*s[3]+cs011)+(1-prev)*(c[1]*(1-c[2])*(1-c[3])+cc011) # 011
p[6]<-prev*((1-s[1])*s[2]*(1-s[3])+cs010)+(1-prev)*(c[1]*(1-c[2])*c[3]+cc010) #010
p[7]<-prev*((1-s[1])*(1-s[2])*s[3]+cs001)+(1-prev)*(c[1]*c[2]*(1-c[3])+cc001) #001
p[8]<-prev*((1-s[1])*(1-s[2])*(1-s[3])+cs000)+(1-prev)*(c[1]*c[2]*c[3]+cc000) #000
#=========================================
# prior distributions covariance term
#=========================================
cs111<-0
cs110<-0
cs101<-0
cs100<-0
cs011<-0
cs010<-0
cs001<-0
cs000<-0
cc111<-0
cc110<-0
cc101<-0
cc100<-0
cc011<-0
cc010<-0
cc001<-0
cc000<-0
```

```
}
```

Model 2: REM model probit style - Openbugs

```
model{
# Likelihood
for (i in 1:500){
    status[i]~dbern(prev)
    z[i]~dnorm(0,1)
for(j in 1:3){
    y[i,j]~dbern(p[i,j])
    }
    }
### expand likelihood as three scores are correlated
for (i in 1:500){
  #p[i,1]<- phi(status[i]*alpha[1] + ((1-status[i])*beta[1]))
for (j in 1:3) {
   p[i,j]<- phi(status[i]*alpha[j] + ((1-status[i])*beta[j]) + (status[i]*b1*z[i]))
    }
    }

##### Prior
prev~dbeta(1,1)
b1~ dnorm(0,0.01)I(0,)
#b1<- 0 # use for CI
#b2<- 0 # Use for CI

### use below in CI
#for (j in 1:3){
# alpha[j] ~ dnorm(0,0.1)I(-1,)
#beta[j] ~ dnorm(0,0.1)I(,1)
# }

## Priors for sensitivity and specificity test (centred on simulated truth)
```

```
#s[1] ~ dbeta(6,0.667)
#sp[1] ~ dbeta(4.05, 0.45)
#s[2]~dbeta(4,1)
#sp[2]~dbeta(4,1)
#sp[2]~dbeta(1,1)
#s[3]~dbeta(3.5, 1.5)
#sp[3]~dbeta(1,1)
#sp[3]~dbeta(4.05, 0.45)

## priors for sensitivity and specificity test (not centred on simulated truth)
s[1] ~ dbeta(4,1)
sp[1] ~ dbeta(3.5, 1.5)
s[2] ~ dbeta(3.5, 1.5)
sp[2] ~ dbeta (4.05, 0.45)
s[3] ~dbeta(6,0.667)
sp[3] ~ dbeta (4,1)


### Posterior calculation of parameters
### used for CI assumption
#for(j in 1:3){
  #  s[j]<-phi(alpha[j]/sqrt(1+pow(b1,2)))
#sp[j]<- phi(-beta[j]/sqrt(1+pow(b2,2)))
#}
}
```

Model 3: REM model logit style – Openbugs

```
model {
##### Likelihood
for (i in 1:500){
    status[i]~dbern(prev)
    z[i]~dnorm(0,1)
for(j in 1:3){
    y[i,j]~dbern(p[i,j])
    }
    }

### expand likelihood as three scores are correlated
for (i in 1:500){
#logit(p[i,1])<-status[i]*alpha[1]+(1-status[i])*beta[1]
for(j in 1:3) {
    logit(p[i,j])<-status[i]*(alpha[j] + status[i]*b1*z[i]) + (1-status[i])*beta[j]
    }
}

######    Prior    distribution ################

prev~dbeta(1,1) ## flat or non informed prior
#b1<- 0 # use for CI
#b2<- 0 # use for CI
b1~dnorm(0,0.1)I(0,) # under conditional dependence among diseased group
#b2~dnorm(0,0.1)I(0,) # use if the assumption of non-diseased correlated

###### use under conditional independence assumption
#for(j in 1:3){
#alpha[j]~dnorm(0,0.1)I(-1,)
#beta[j]~dnorm(0,0.1)I(,1)
#}
```

```
####### Use if priors distribution centred on the simulated truth ######
#s[1] ~ dbeta(6,0.667)
#sp[1] ~ dbeta(4.05, 0.45)
#s[2]~dbeta(4, 1)
#sp[2]~dbeta(1,1)
#sp[2]~dbeta(4,1)
#s[3]~dbeta(3.5, 1.5)
#sp[3]~dbeta(4.05, 0.45)
#sp[3]~dbeta(1,1)

## priors for sensitivity and specificity test (not centred on simulated truth)
s[1] ~ dbeta(4,1)
sp[1] ~ dbeta(3.5, 1.5)
s[2] ~ dbeta(3.5, 1.5)
sp[2] ~ dbeta (4.05, 0.45)
s[3] ~dbeta(6,0.667)
sp[3] ~ dbeta (4,1)

for (j in 1:3){
alpha[j] <- logit(s[j])
beta[j]<--logit(sp[j])
}
}
```

Model 4: Finite Mixture model – Openbugs

```
model{
for (i in 1:500)
{
 d[i] ~ dbern(prev)
l[i,1] ~ dbern(eta1[i])
l[i,2] ~ dbern (eta0[i])
eta1[i]<- d[i]*tau[1] # diseased group
eta0[i]<- (1-d[i])*tau[2] # non diseased group
for (j in 1:3){
 y[i,j] ~ dbern(p[i,j])
 p[i,j]<- d[i]*(l[i,1]+(1 - l[i,1])*w[j,1])+(1-d[i])*(1 - (l[i,2]+(1 -l[i,2])*(1-w[j,2])))
}
}

## prior
prev ~ dbeta(1,1)
#for (k in 1:2){tau[k]<- 0} # assuming CI
#for (k in 1:2){tau[k] ~ dbeta(1,1)}
tau[1] ~ dbeta(0.5,0.5)
tau[2]<- 0
for (j in 1:3){
  w[j,1]~dbeta(0.5,0.5)
 w[j,2]~dbeta(0.5,0.5)
}

############################################################
########## Use if priors distribution centred on the truth ################
#s[1] ~ dbeta(6,0.667)
#sp[1] ~ dbeta(4.05, 0.45)
#s[2]~dbeta(4, 1)
#sp[2]~dbeta(4,1)
```

```
#s[3]~dbeta(3.5, 1.5)
#sp[3]~dbeta(4.05, 0.45)
#######################################################

## priors for sensitivity and specificity test (not centred on simulated truth)
s[1] ~ dbeta(4,1)
sp[1] ~ dbeta(3.5, 1.5)
s[2] ~ dbeta(3.5, 1.5)
sp[2] ~ dbeta (4.05, 0.45)
s[3] ~dbeta(6,0.667)
sp[3] ~ dbeta (4,1)

### Posterior calculation of parameters
## used under assumption of CI
#for (j in 1:3){
#s[j] <- tau[1]+(1-tau[1])*w[j,1]
#sp[j] <- tau[2] + (1-tau[2])*(1 - w[j,2])
#}

}
```

Model 5: Fixed effect model Wang et al style - RStan

Start using stan

```{r stan3}
library("StanHeaders")
library(ggplot2)
library("rstan") # observe startup messages
library("inline")
library("matrixStats")
```

enter data

Prevalence is 0.3

```{r stan5}
FEMCI<- list(T = 8,  freqobs= c(77,39,21,34,14,60,27,228), N= 500)
FEM23<- list(T = 8,  freqobs= c(92,23,5,50,15,61,28,226), N= 500)
FEM123<-list(T = 8,  freqobs= c(100,20,8,45,11,62,23,231), N= 500)
```

Code for analysis

```{r stancode}
FEMCD1<- "
data {
  int<lower=0>T; // number of possible combination of the tests response
        int<lower=0>freqobs[T];// the frequency of each possible combinations
}

// The parameters accepted by the model.
parameters {
 real<lower=0, upper=1> pi;          // prevalence
        real<lower=0, upper=1>s1;// sensitivity of T1
        real<lower=0, upper=1>s2;// sensitivity of T2
        real<lower=0, upper=1>s3;// sensitivity of T3
```

```
        real<lower=0, upper=1>c1;// specificity of T1
        real<lower=0, upper=1>c2;// specificity of T2
        real<lower=0, upper=1>c3;// specificity of T3



        // conditional dependence term for result (T1=1, T2=1, T3=1 | D=1)
        real<lower=-s1*s2*s3,upper=fmin(s1, fmin(s2, s3))-s1*s2*s3> covS111;
        real<lower=-(1-s1)*s2*s3,upper=fmin(1-s1, fmin(s2, s3))-(1-s1)*s2*s3> covS011;
        real<lower=-(1-s1)*(1-s2)*s3,upper=fmin(1-s1, fmin(1-s2, s3))-(1-s1)*(1-s2)*s3>
covS001;
        real<lower=-(1-s1)*(1-s2)*(1-s3),upper=fmin(1-s1, fmin(1-s2, 1-s3))-(1-s1)*(1-
s2)*(1-s3)> covS000;
}

transformed parameters {
real<lower=-s1*(1-s2)*(1-s3),upper=fmin(s1, fmin(1-s2, 1-s3))-s1*(1-s2)*(1-s3)> covS100;
real<lower=-s1*(1-s2)*s3,upper=fmin(s1, fmin(1-s2, s3))-s1*(1-s2)*s3> covS101;
real<lower=-s1*s2*(1-s3),upper=fmin(s1, fmin(s2, 1-s3))-s1*s2*(1-s3)> covS110;
real<lower=-(1-s1)*s2*(1-s3),upper=fmin(1-s1, fmin(s2, 1-s3))-(1-s1)*s2*(1-s3)> covS010;
vector<lower=0,upper=1>[T] pr;    // joint probability of each type of possible test result

// the pairwise conditional dependence between
        real covST12;         // test 1 and test 2
        real covST13;         // test 1 and test 3
        real covST23;         // test 2 and test 3

        // calculate the transformed conditional dependence terms
        covS100 = covS011 + covS111 - covS000;
        covS101 = -(covS001 + covS011 + covS111);
        covS110 = covS000 + covS001 - covS111;
        covS010 = -(covS000 + covS001 + covS011);

        // calculate the pairwise conditional dependence terms
```

```
        covST12 = covS111 + covS110;
        covST13 = covS111 + covS101;
        covST23 = covS011 + covS111;

        // probability of having combination of test response
        pr[1] = pi*(s1*s2*s3+covS111)+(1-pi)*((1-c1)*(1-c2)*(1-c3));    // 111
        pr[2] = pi*(s1*s2*(1-s3)+covS110)+(1-pi)*((1-c1)*(1-c2)*(c3));  // 110
        pr[3] = pi*(s1*(1-s2)*s3+covS101)+(1-pi)*((1-c1)*(c2)*(1-c3));  // 101
        pr[4] = pi*(s1*(1-s2)*(1-s3)+covS100)+(1-pi)*((1-c1)*c2*c3);    // 100
        pr[5] = pi*((1-s1)*s2*s3+covS011)+(1-pi)*((c1)*(1-c2)*(1-c3));  // 011
        pr[6] = pi*((1-s1)*s2*(1-s3)+covS010)+(1-pi)*(c1*(1-c2)*(c3));  // 010
        pr[7] = pi*((1-s1)*(1-s2)*s3+covS001)+(1-pi)*(c1*(c2)*(1-c3));  // 001
        pr[8] = pi*((1-s1)*(1-s2)*(1-s3)+covS000)+(1-pi)*((c1)*c2*c3);  // 000

}
// The model to be estimated. We model the output
// 'y' to be normally distributed with mean 'mu'
// and standard deviation 'sigma'.

model {
  // quartile method
        // priors:
        pi ~ beta(1,1);
//s1 ~ beta(1,1);
//c1 ~ beta(1,1);
//s2 ~ beta(1,1);
//c2 ~ beta (1,1);
//s3 ~ beta(1,1);
//c3 ~ beta (1,1);

// Use if priors distribution for 3 CD tests
s1 ~ beta(6,0.667);
```

```
c1 ~ beta(4.05,0.45);

s2 ~ beta(4,1);

c2 ~ beta (4,1);

//c2 ~ beta(4,1);

s3 ~ beta(3.5, 1.5);

c3 ~ beta (4.05, 0.45);

//c3 ~ beta(4.05,0.45);



        // likelihood function

        freqobs ~ multinomial(pr);

}"
```

Run analysis

```{r stanexam1}

fit<- stan(model_code = FEMCD1, data= FEM23, iter =100000, warmup = 2000, chains =3)

#fit<- stan(model_code = FEMCD1, data= FEMBR, iter = 100000, warmup = 2000, chains = 3, control = list(adapt_delta = 0.90)) #iter = 10000, warmup = 1000, chains = 2, verbose = TRUE,  max_treedepth = 15)
```

Print result

```{r diag}

print(fit, pars=c("pi","s1", "s2", "s3", "c1","c2", "c3", "covST12", "covST23", "covST13"), digits.summary = 5)#, , probs=c(.1,.5,.9)
```

plot necessary plots

```{r stanplot}

plot(fit, pars=c("pi","s1", "s2", "s3", "c1","c2", "c3"), ci_level = 0.95, outer_level = 0.999)
```

```
plot(fit, show_density = TRUE, pars=c("pi","s1", "s2", "s3", "c1","c2", "c3"))
#plot(fit, show_density = TRUE, pars="pi",ci_level = 0.95,outer_level = 0.999)
#plot(fit, show_density = TRUE, pars="s1", ci_level = 0.95,outer_level = 0.999)
#plot(fit, show_density = TRUE, pars= "s2", ci_level = 0.95,outer_level = 0.999)
#plot(fit, show_density = TRUE, pars= "s3", ci_level = 0.95,outer_level = 0.999)
#plot(fit, show_density = TRUE, pars= "c1", ci_level = 0.95,outer_level = 0.999)
#plot(fit, show_density = TRUE, pars="c2", ci_level = 0.95,outer_level = 0.999)
#plot(fit, show_density = TRUE, pars= "c3", ci_level = 0.95,outer_level = 0.999)
traceplot(fit, pars=c("pi","s1", "s2", "s3", "c1","c2", "c3"))
```

C.3. Diagnostic plots of 23CD dataset under CI assumption

The 23CD dataset is the simulated dataset where test 2 and test 3 are conditionally dependent given the true disease status and both tests are conditionally independent with test 1 given the true disease status.

Figure 30: Trace plots of the sensitivities and specificities of the three tests assuming that all tests are conditionally independent.

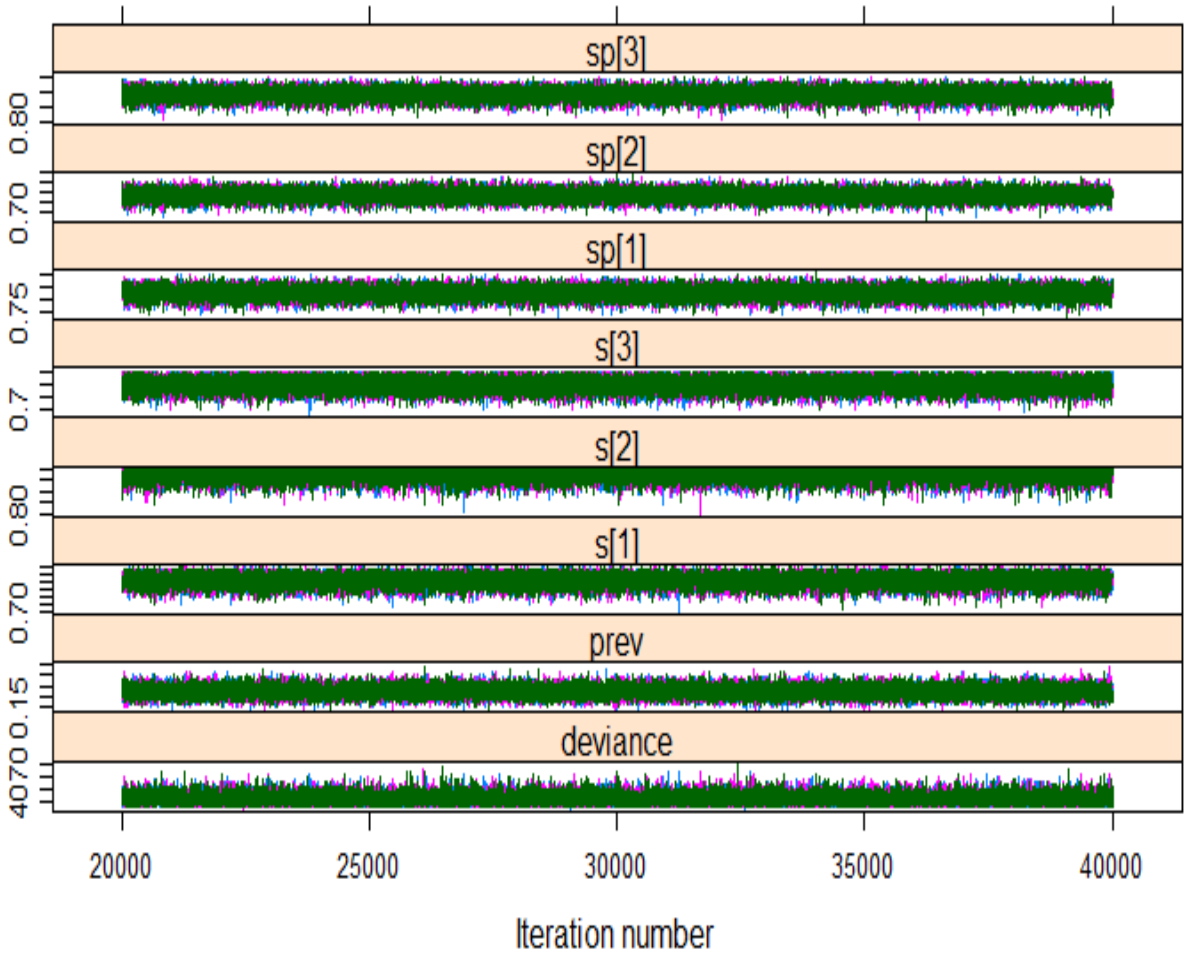


Figure 31: Density plots of the sensitivities and specificities of the three tests assuming that all tests are conditionally independent.

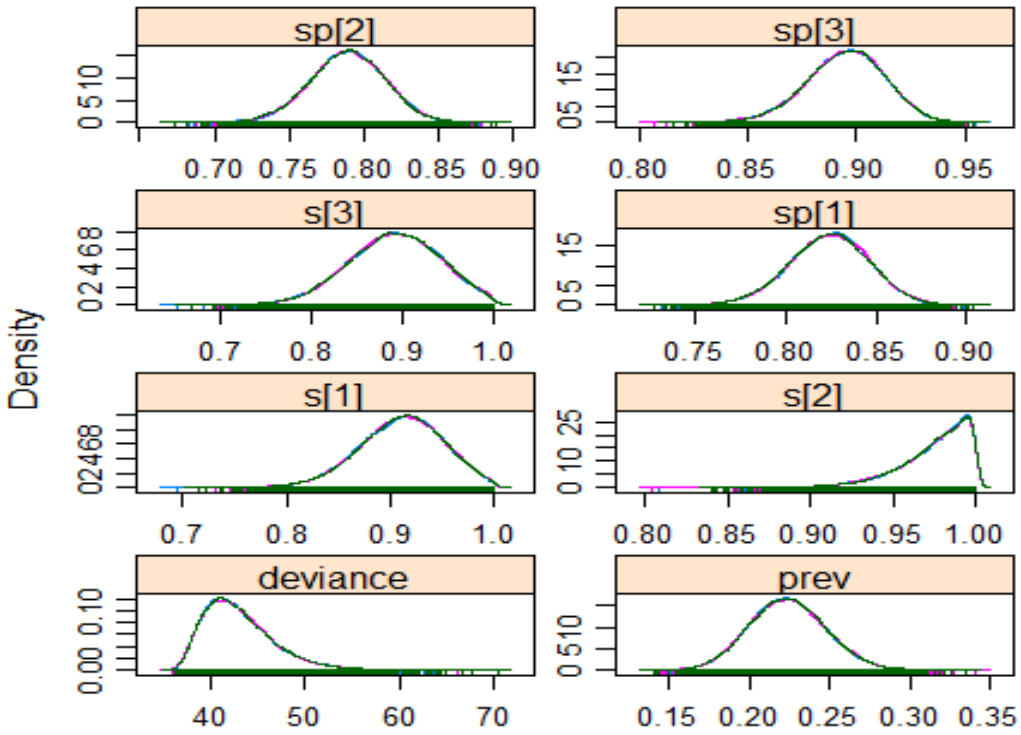


Figure 32: Autocorrelation plots of the sensitivities and specificities of the three tests assuming that all tests are conditionally independent.

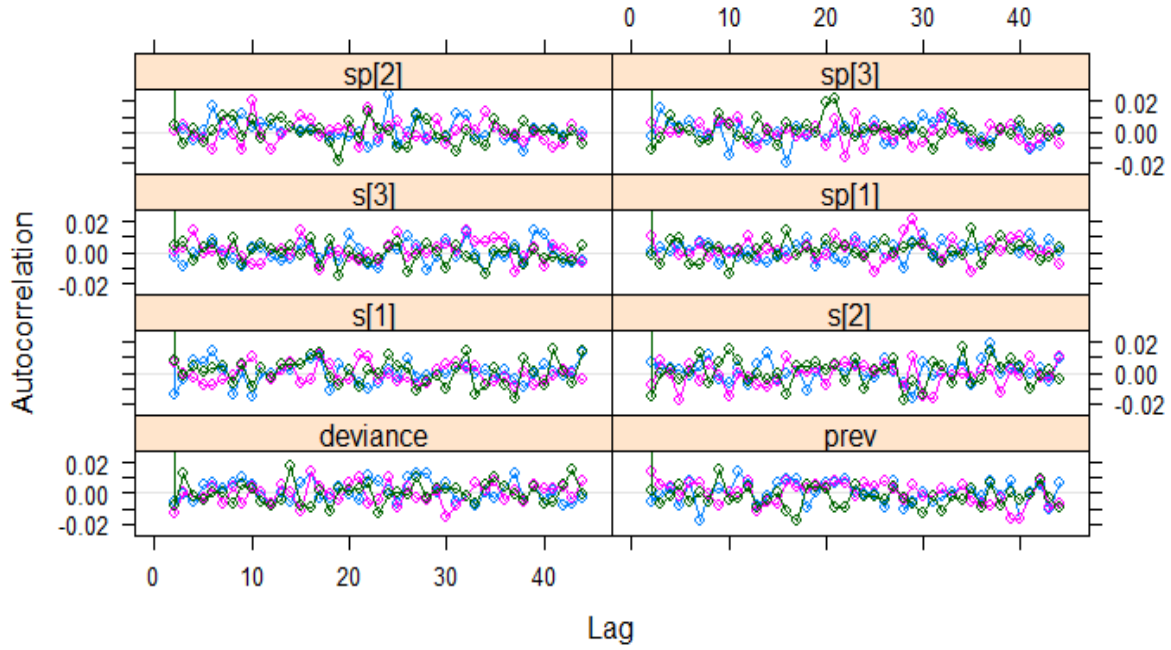
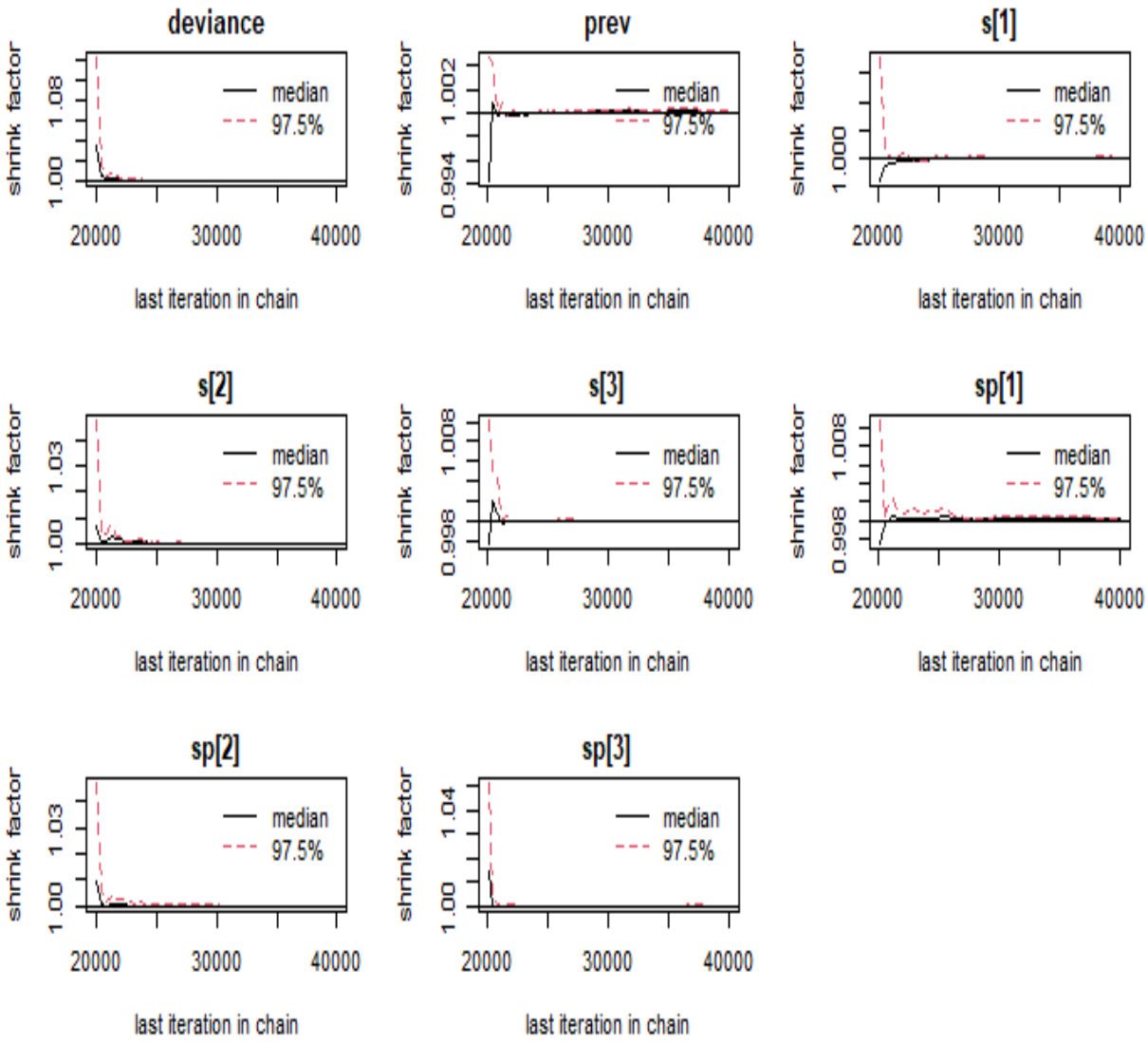


Figure 33: Gelman diagnostic plots of the sensitivities and specificities of the three tests assuming that all tests are conditionally independent.



C.4. Diagnostic plots of 23CD dataset under the assumption of CD (FEM)

Figure 34: Auto-correlation plots of the prevalence, sensitivities and specificities of the tests assuming that all tests are conditionally dependent.

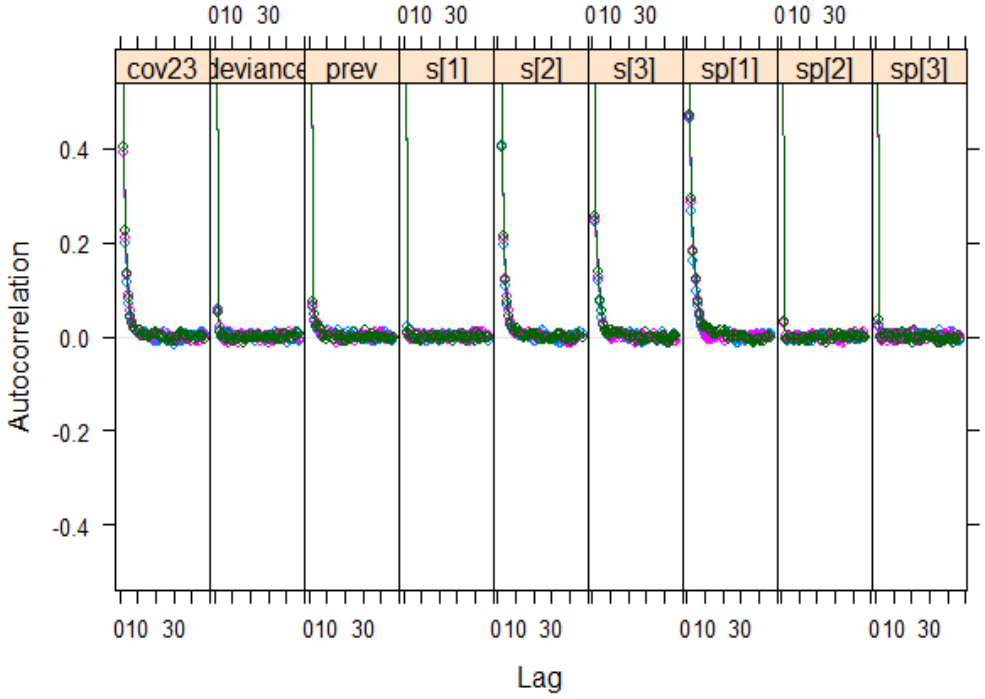


Figure 35: Trace plots for the sensitivities and specificities of test 1, test 2 and test 3, and prevalence assuming that all tests are conditionally dependent.

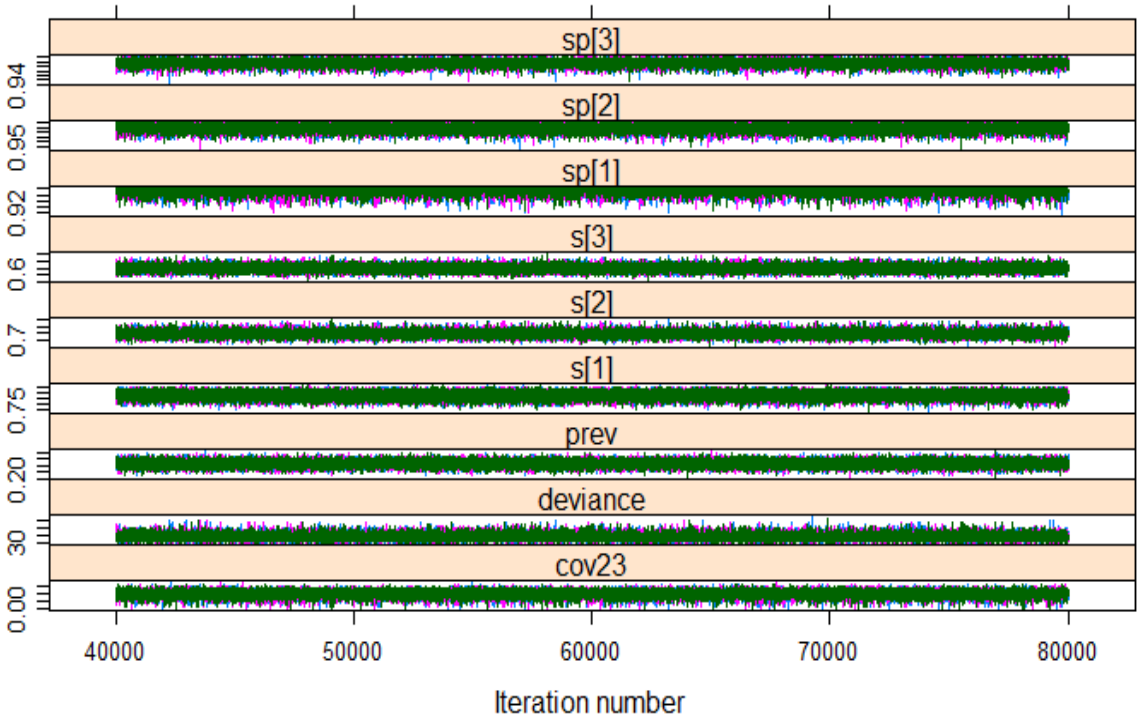


Figure 36: Density plots of the sensitivities and specificities of the three tests and the prevalence assuming that all tests are conditionally dependent.

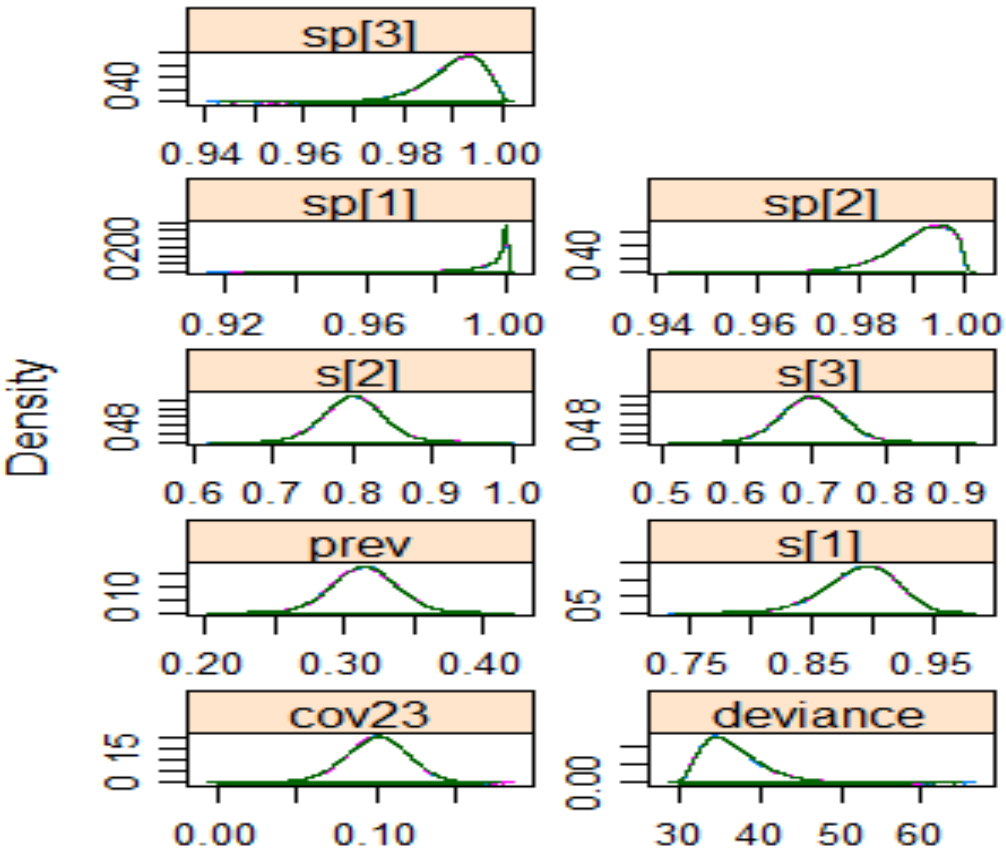
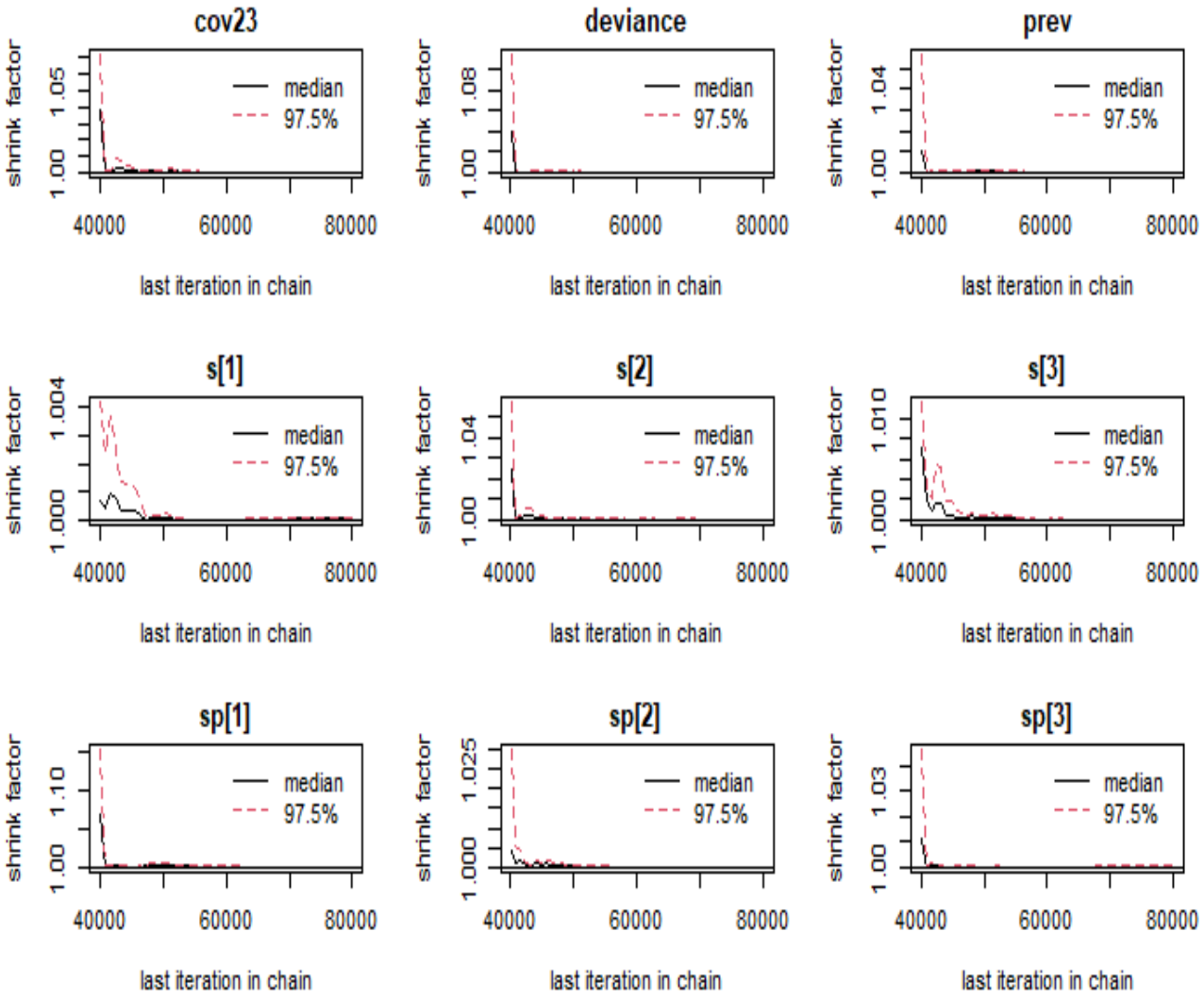
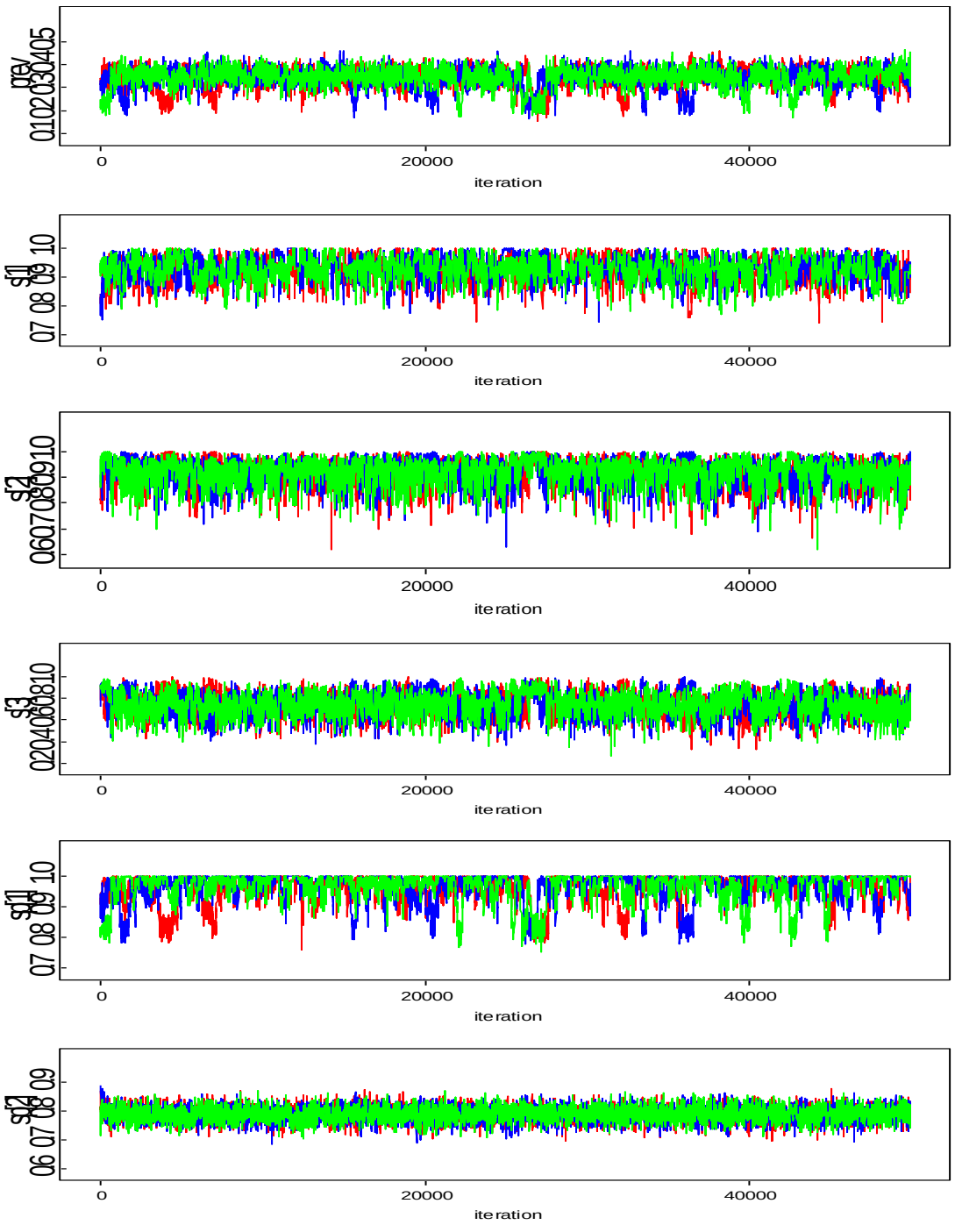


Figure 37: Gelman diagnostic plots of prevalence, sensitivities and specificities of the three tests



C.5. Diagnostic plots of 23CD dataset under the assumption of CD (REML)

Figure 38: Trace plots of sensitivities and specificities of the three tests, and the prevalence assuming that all tests are conditionally dependent.



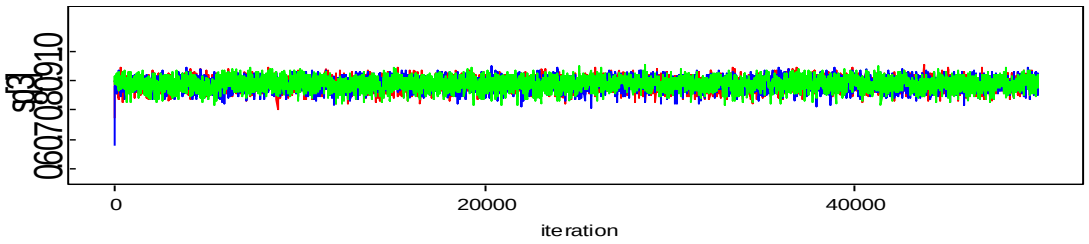
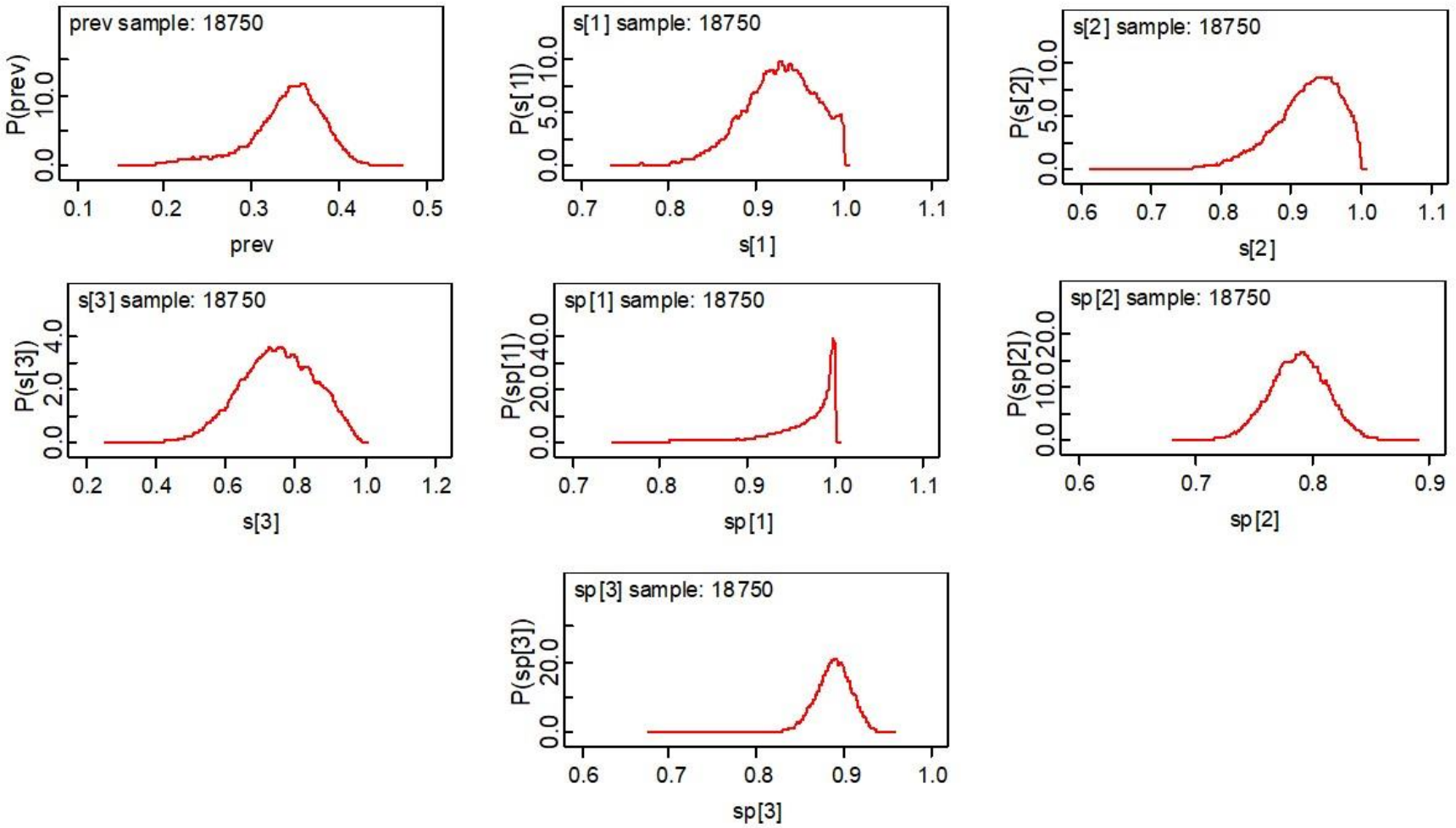

0
20000
40000
iteration

Figure 39: Density plots of sensitivities and specificities of the three tests, and the prevalence assuming that all tests are conditionally dependent.



C.6. Diagnostic plots of 23CD under the assumption of CD using informative priors not centred on the truth (FEM)

Figure 40: Trace plots of the sensitivities and specificities of the three tests

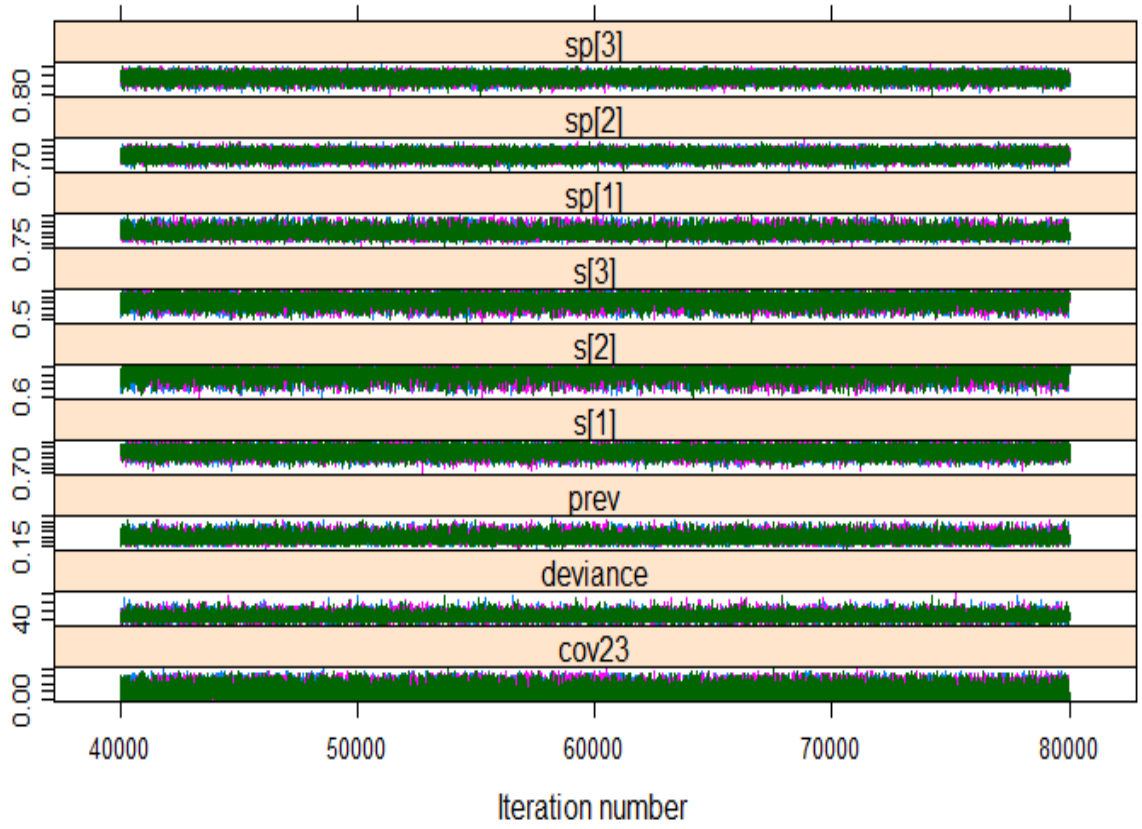


Figure 41: Autocorrelation of prevalence, sensitivities and specificities of the three tests

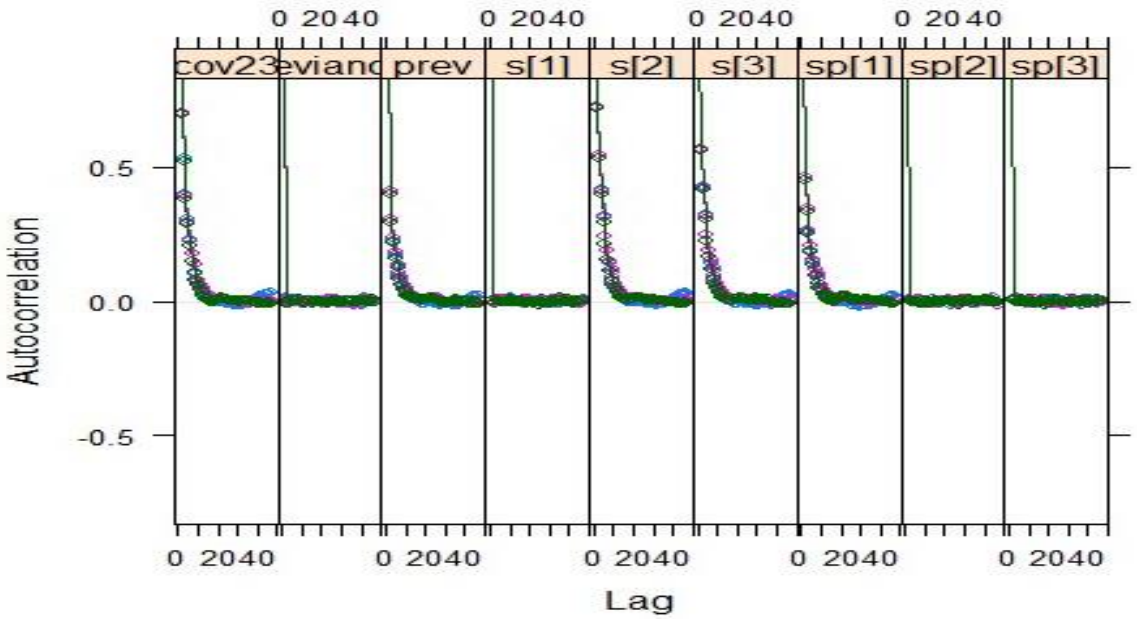


Figure 42: Density plots of the sensitivities and specificities of the three tests

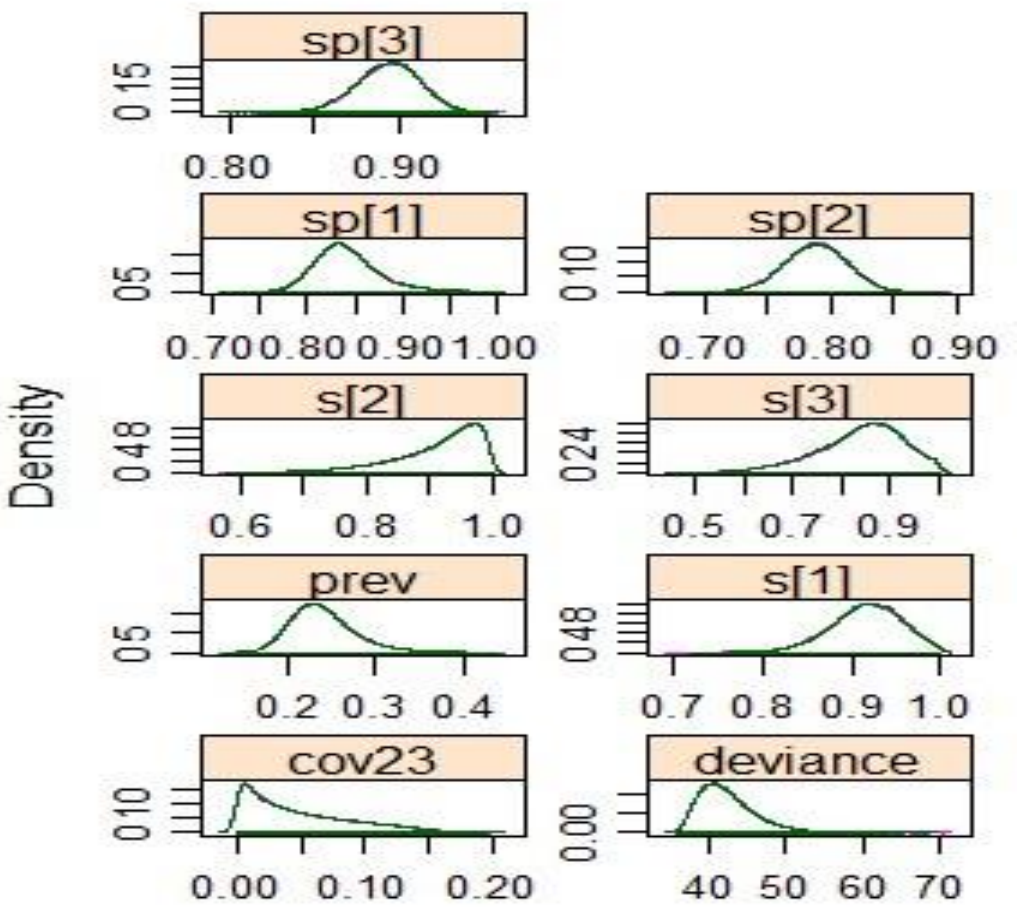
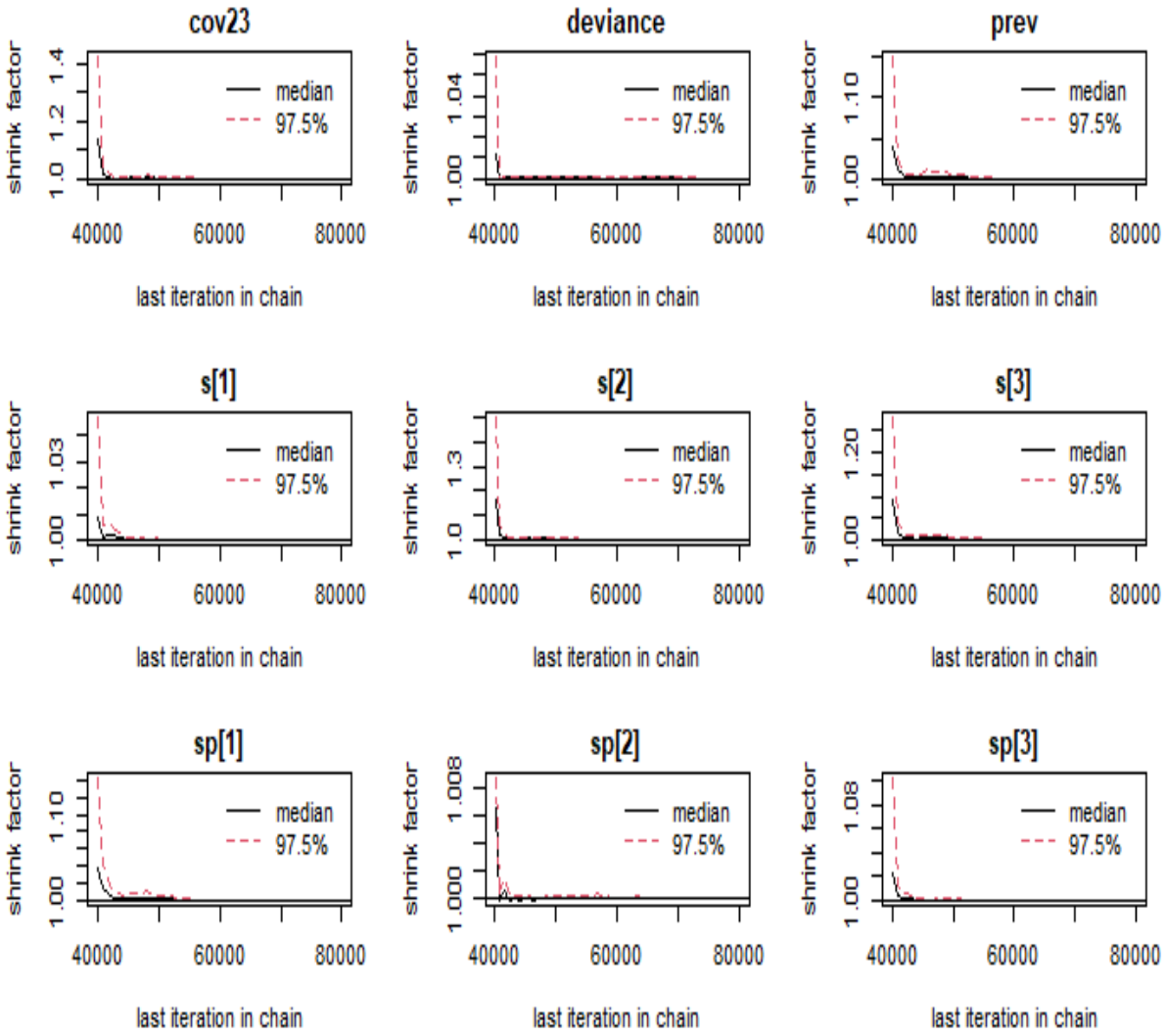


Figure 43: Gelman diagnostic plots for the sensitivities and specificities of three tests



C.7. Diagnostic plots of 123CD under the assumption of CI (FEM).

The 123CD dataset is the simulated dataset where test 1, test 2 and test 3 are conditionally dependent given the true disease status.

Figure 44: Trace plots of sensitivities and specificities of the three tests

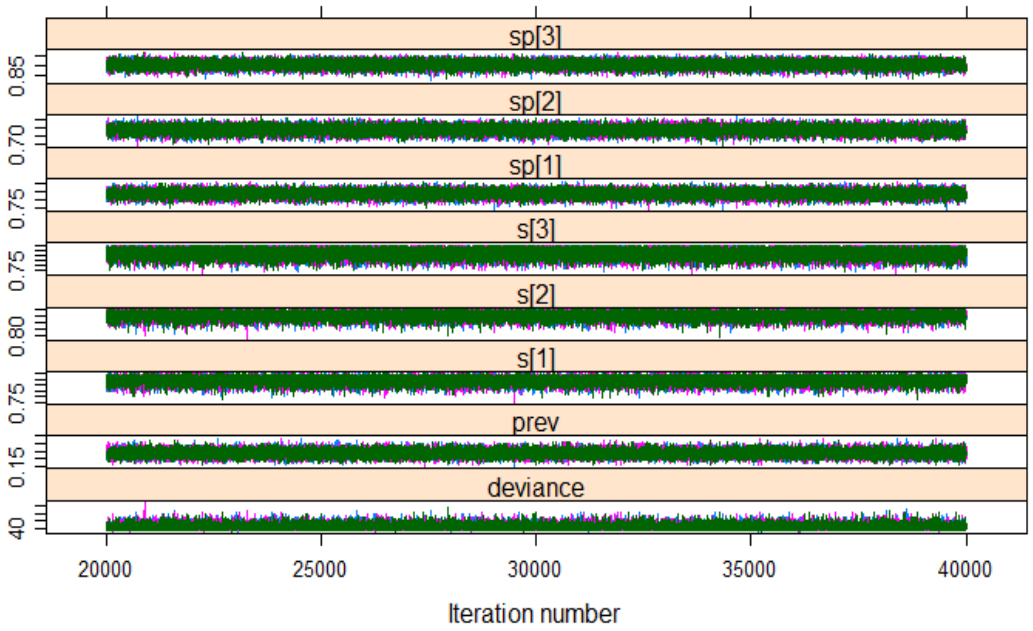


Figure 45: Density plots of sensitivities and specificities of the three tests

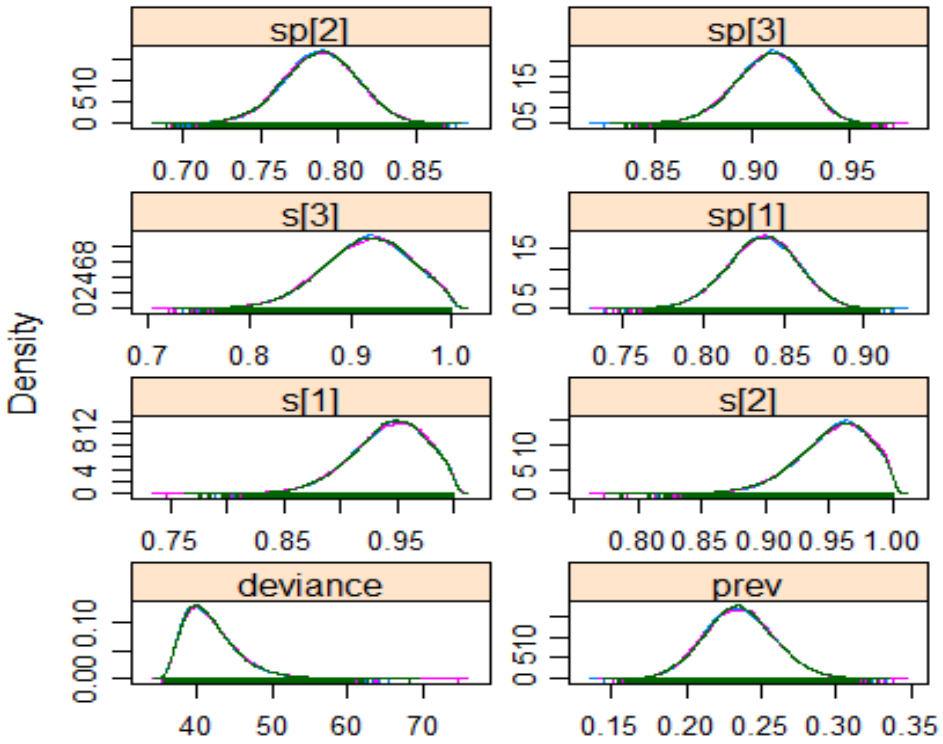


Figure 46: Auto-correlation plots of sensitivities and specificities of the three tests

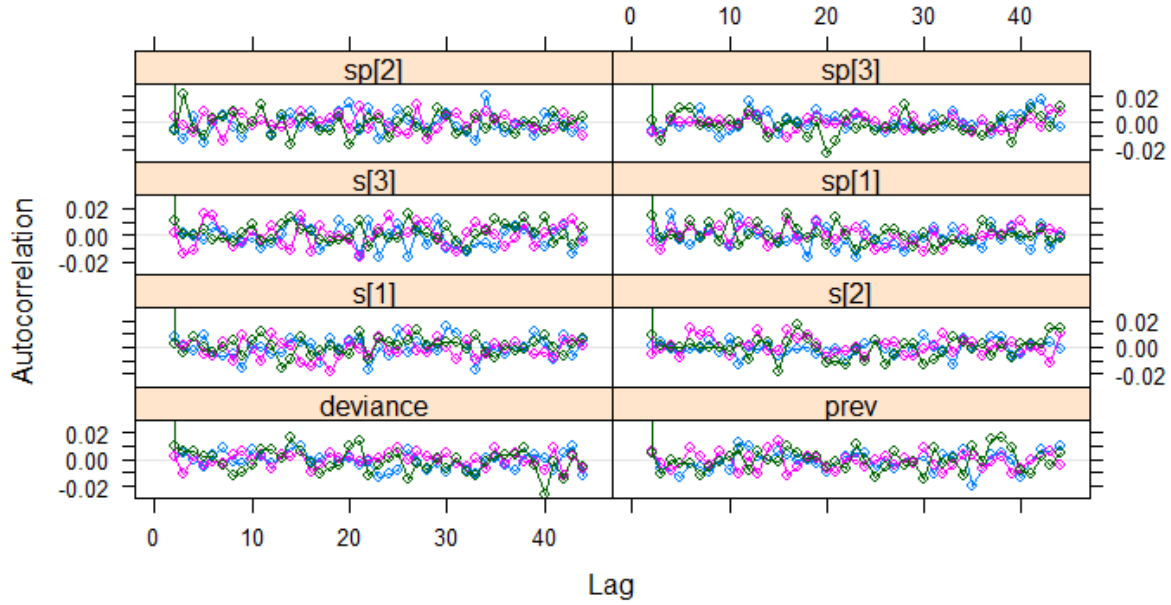
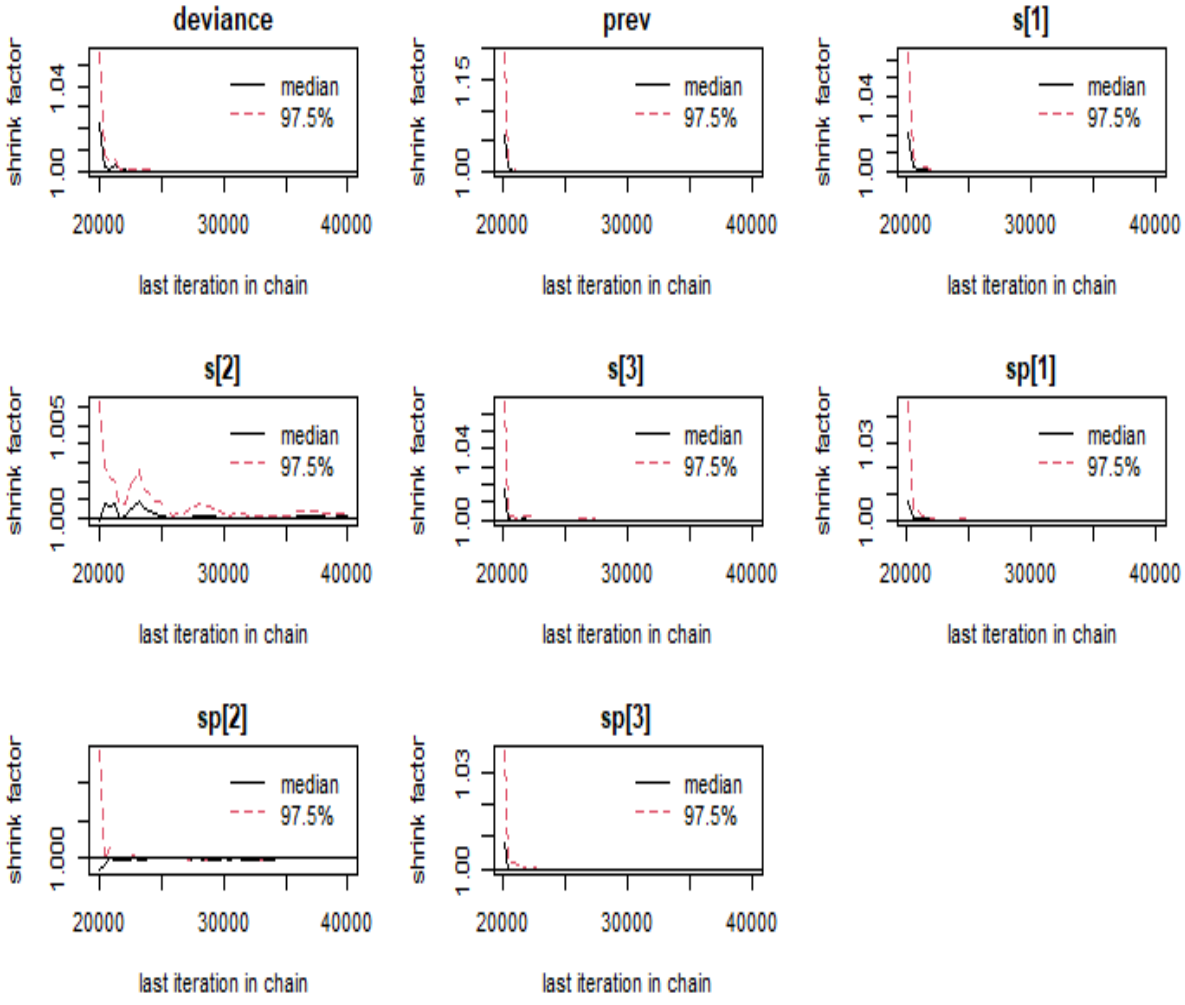


Figure 47: Gelman Diagnostic plot of sensitivity and specificity of the three tests



C.8. Diagnostic plots of 123CD under the assumption of CD (REML)

Figure 48: Trace plots of sensitivities and specificities of three tests

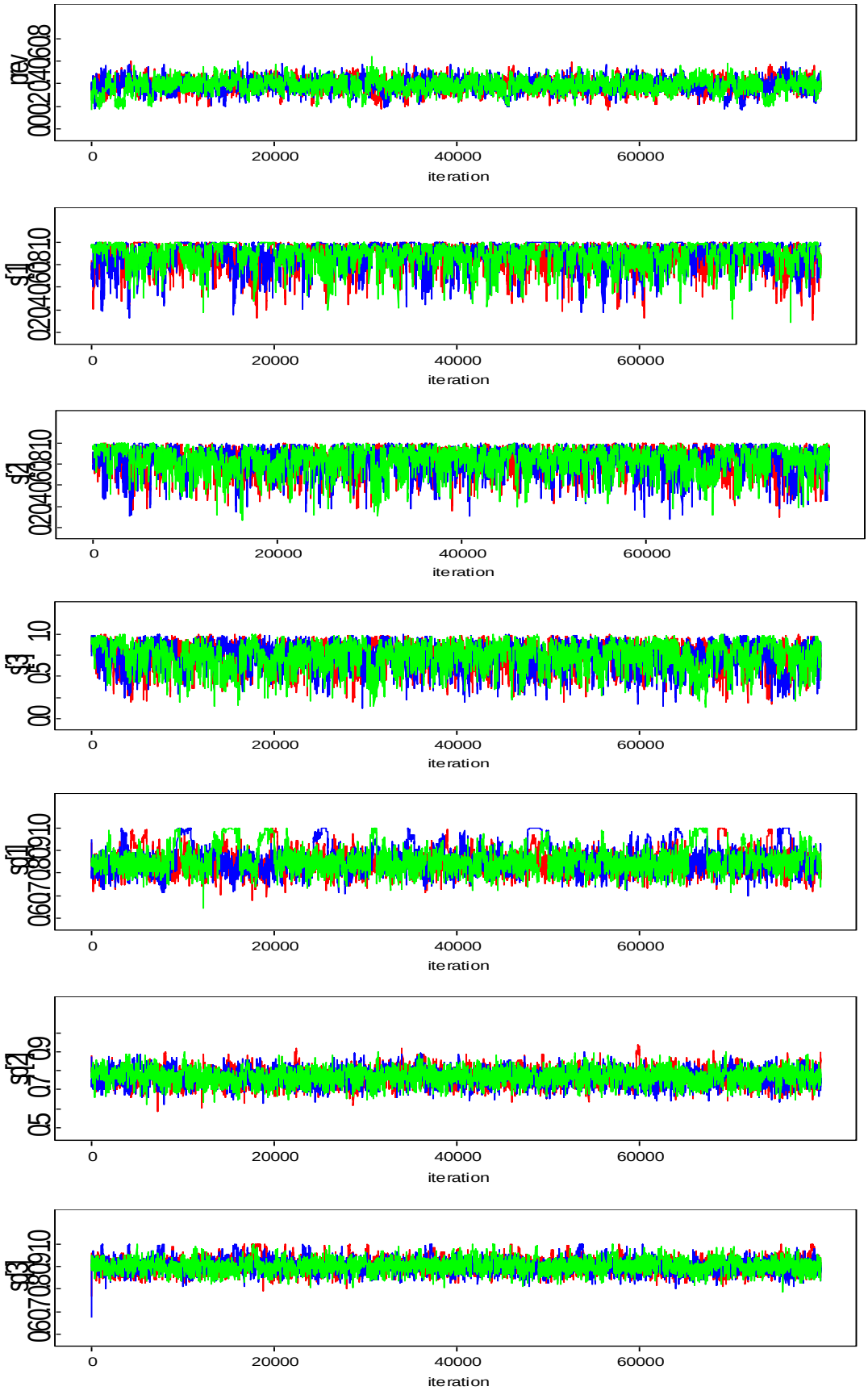
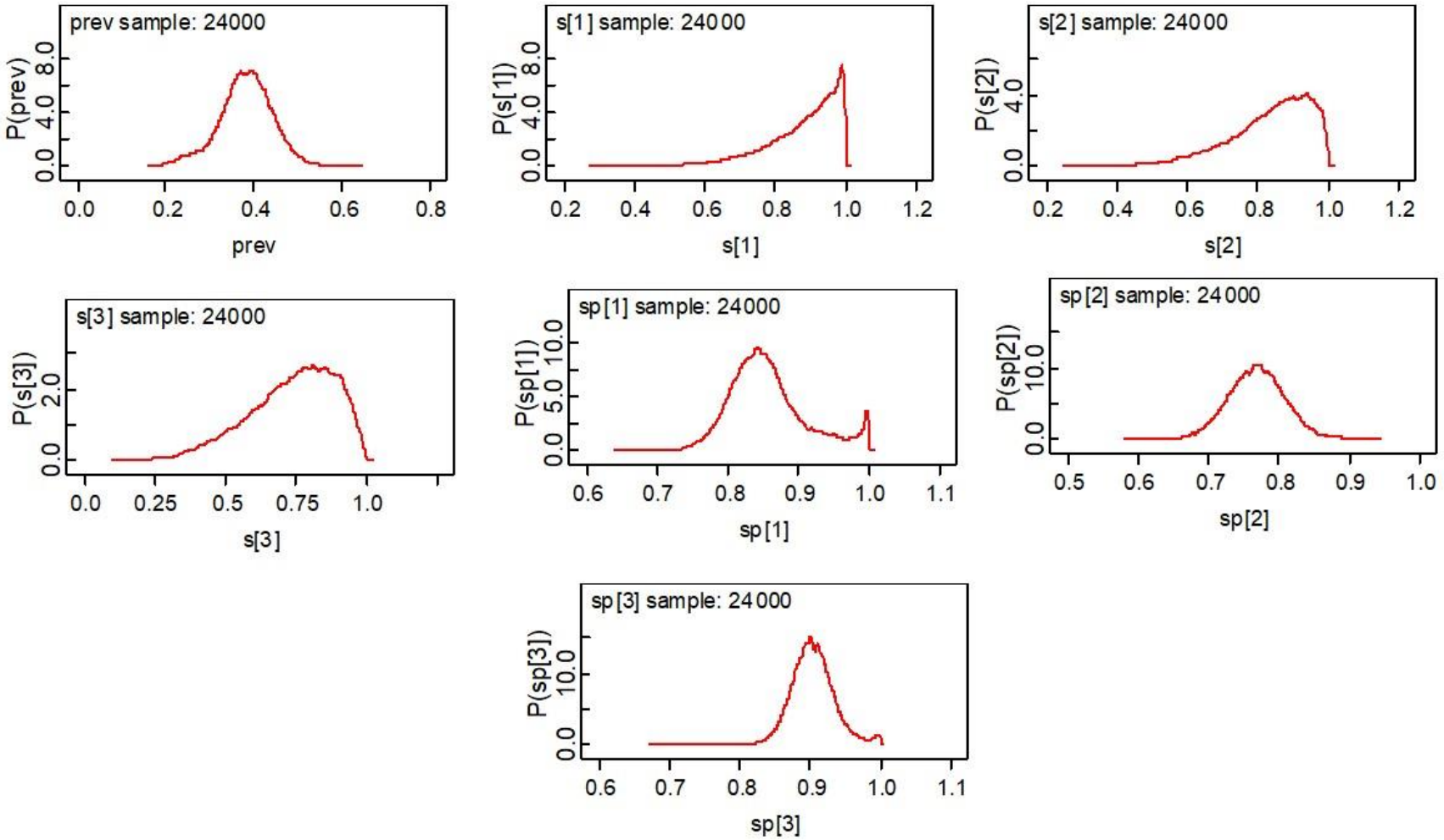


Figure 49: Density plots of the sensitivity and specificities of the three tests under the CD assumption (REML) using priors centred on the simulated truth.



C.9. Diagnostic plots of 123CD under the assumption of CD (FEM_W) using informative priors centred on the simulated truth

Figure 50: Trace plots of sensitivities and specificities of the three tests, and the prevalence via the FEM_W model

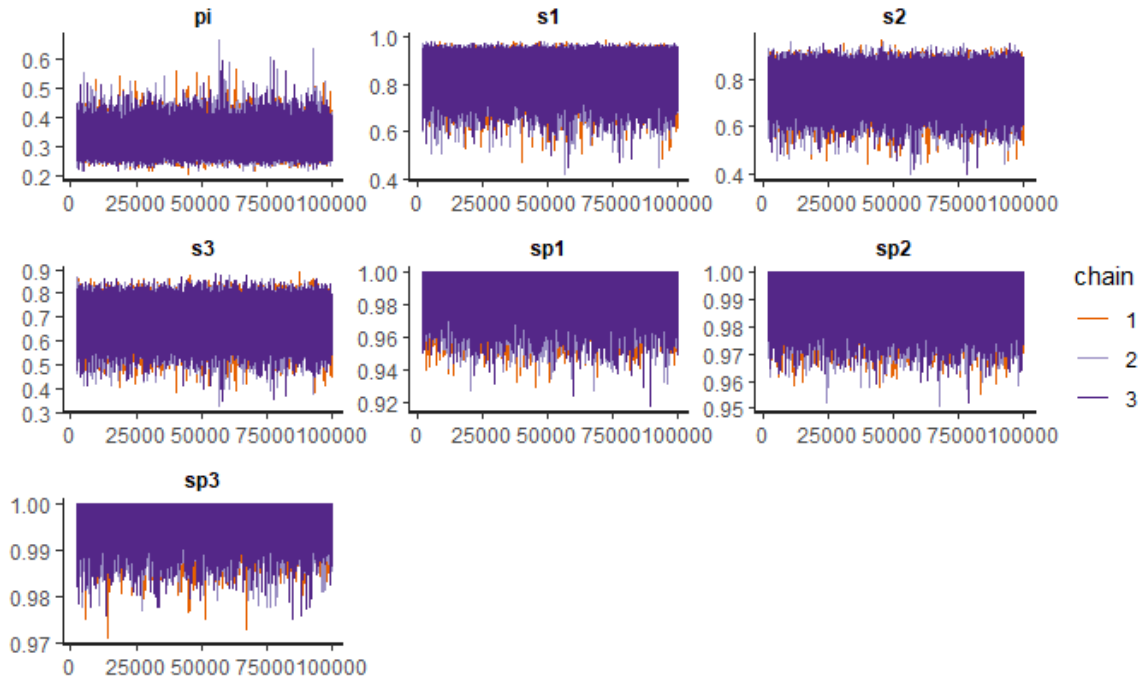
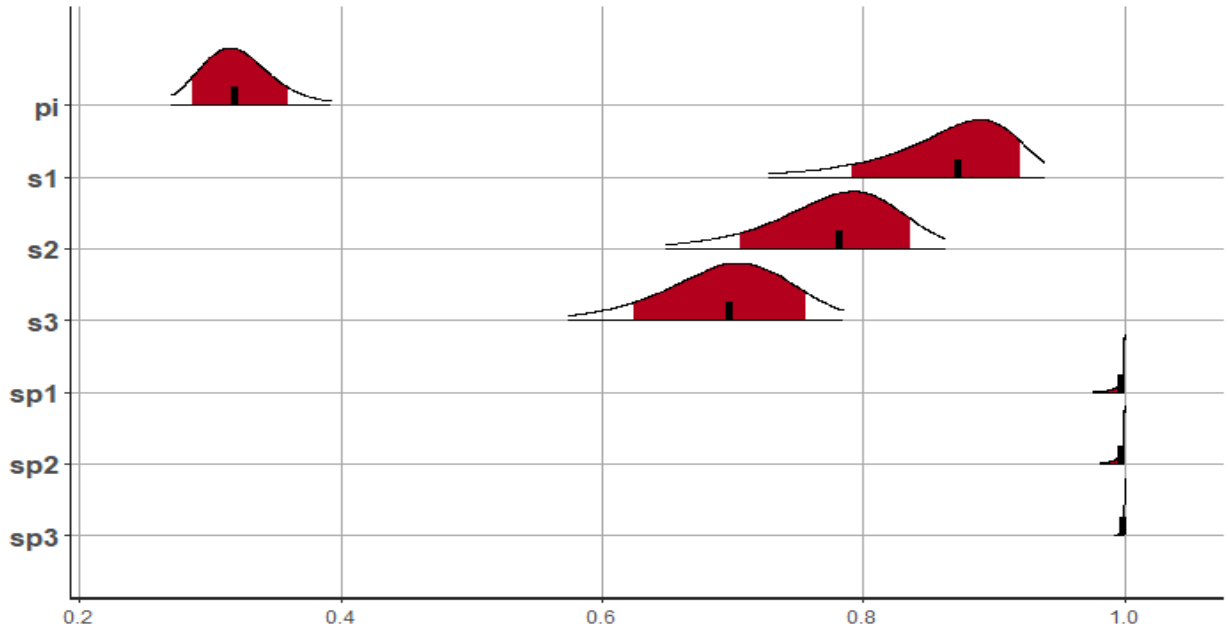


Figure 51: Density plots of the sensitivities and specificities of the three tests using the FEM_W model and assuming all tests are conditionally dependent



The red colour in under the curve is the 95% confidence interval

C.10. Diagnostic plots of 123CD under the assumption of CD (REML)

Figure 52: Trace plots of sensitivities and specificities of the three tests

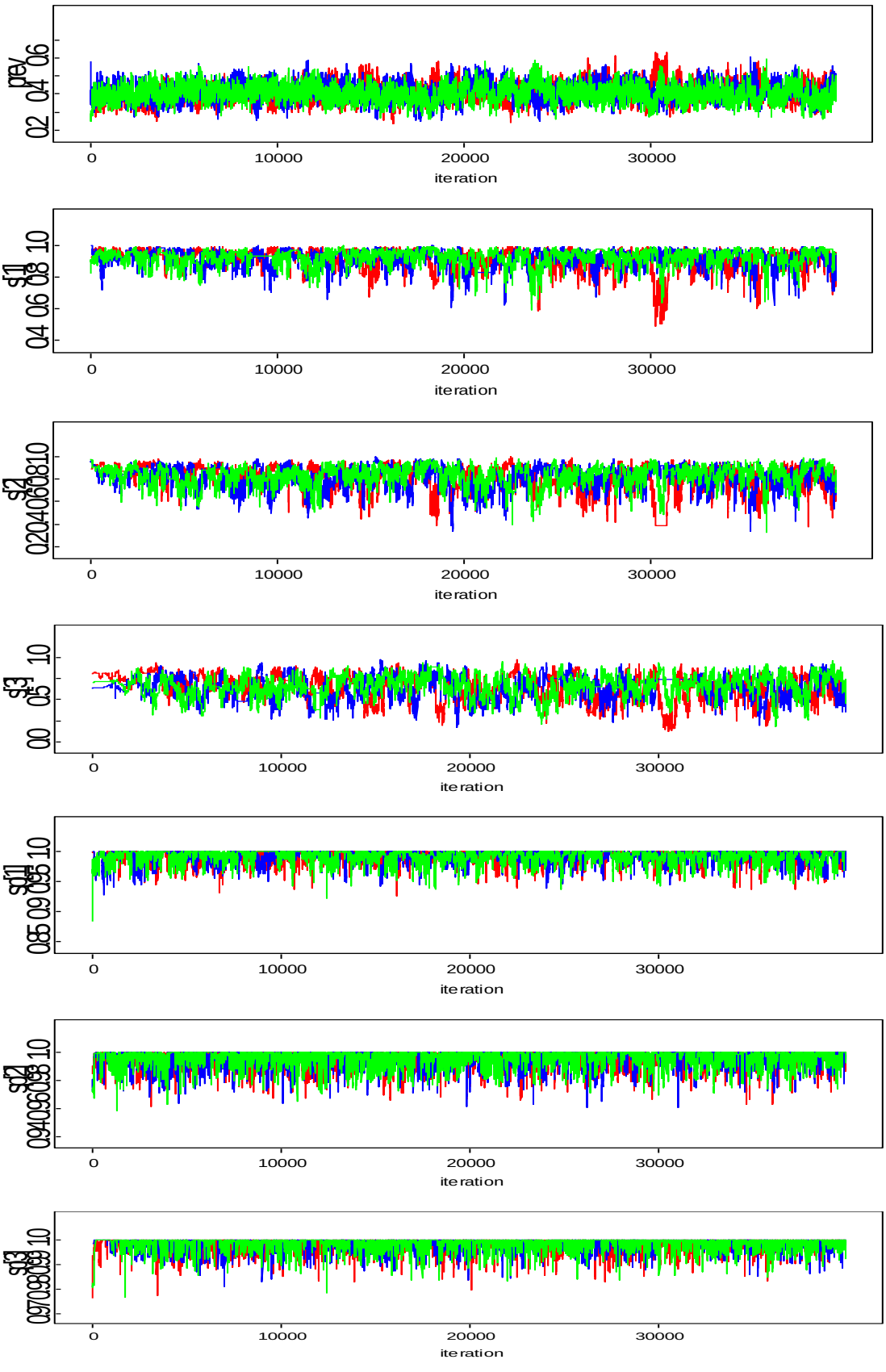
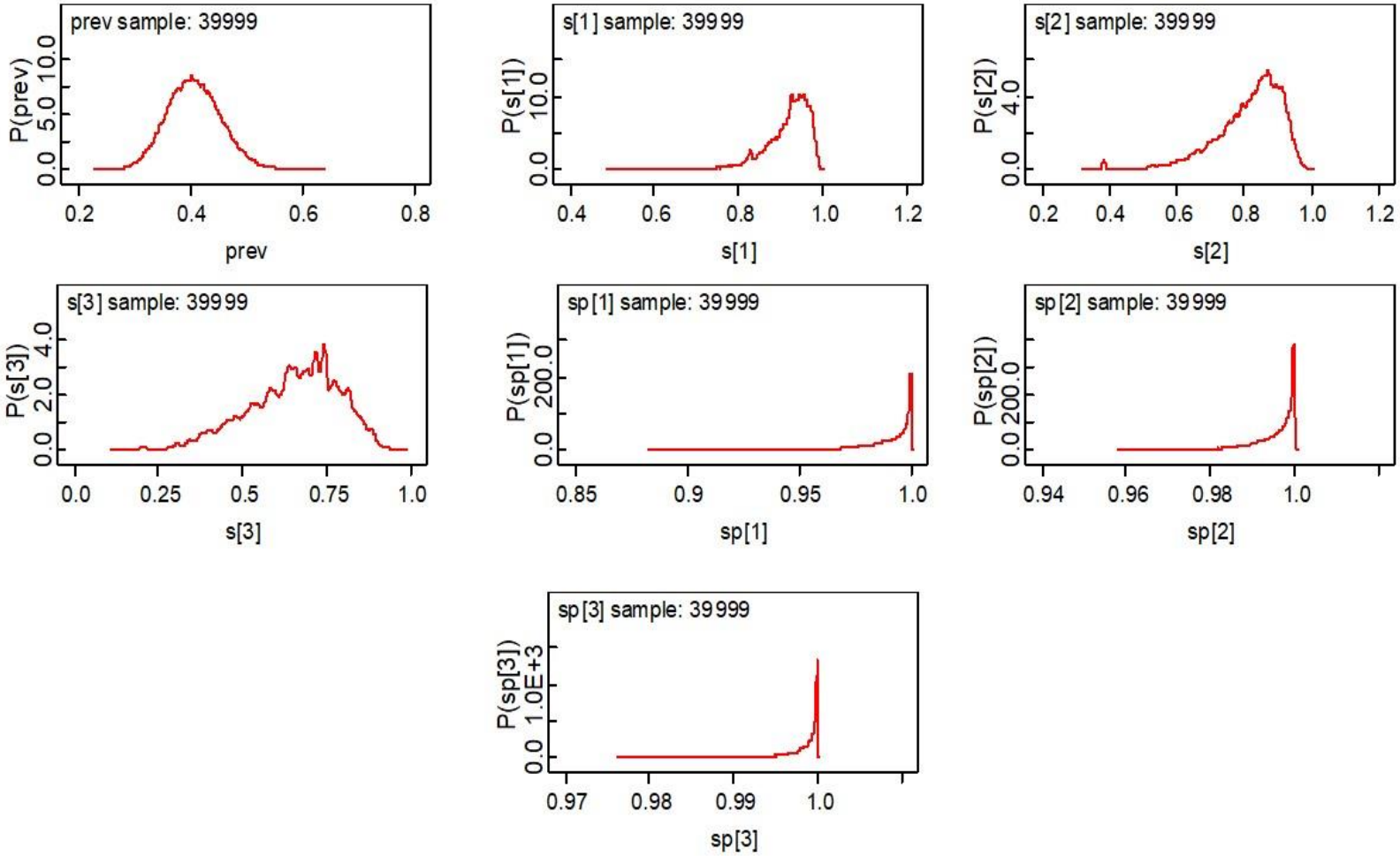
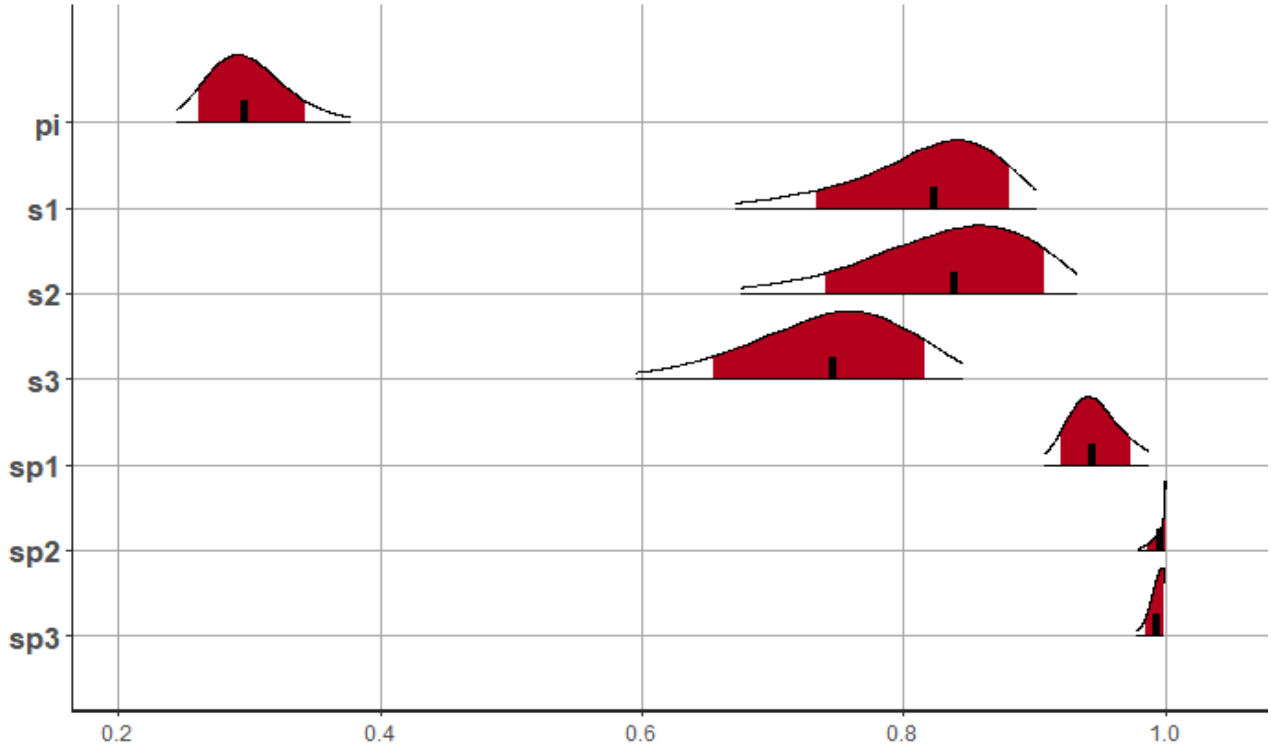


Figure 53: Density plots of sensitivities and specificities of the three tests, and the prevalence under the assumption of conditionally dependence (REML)



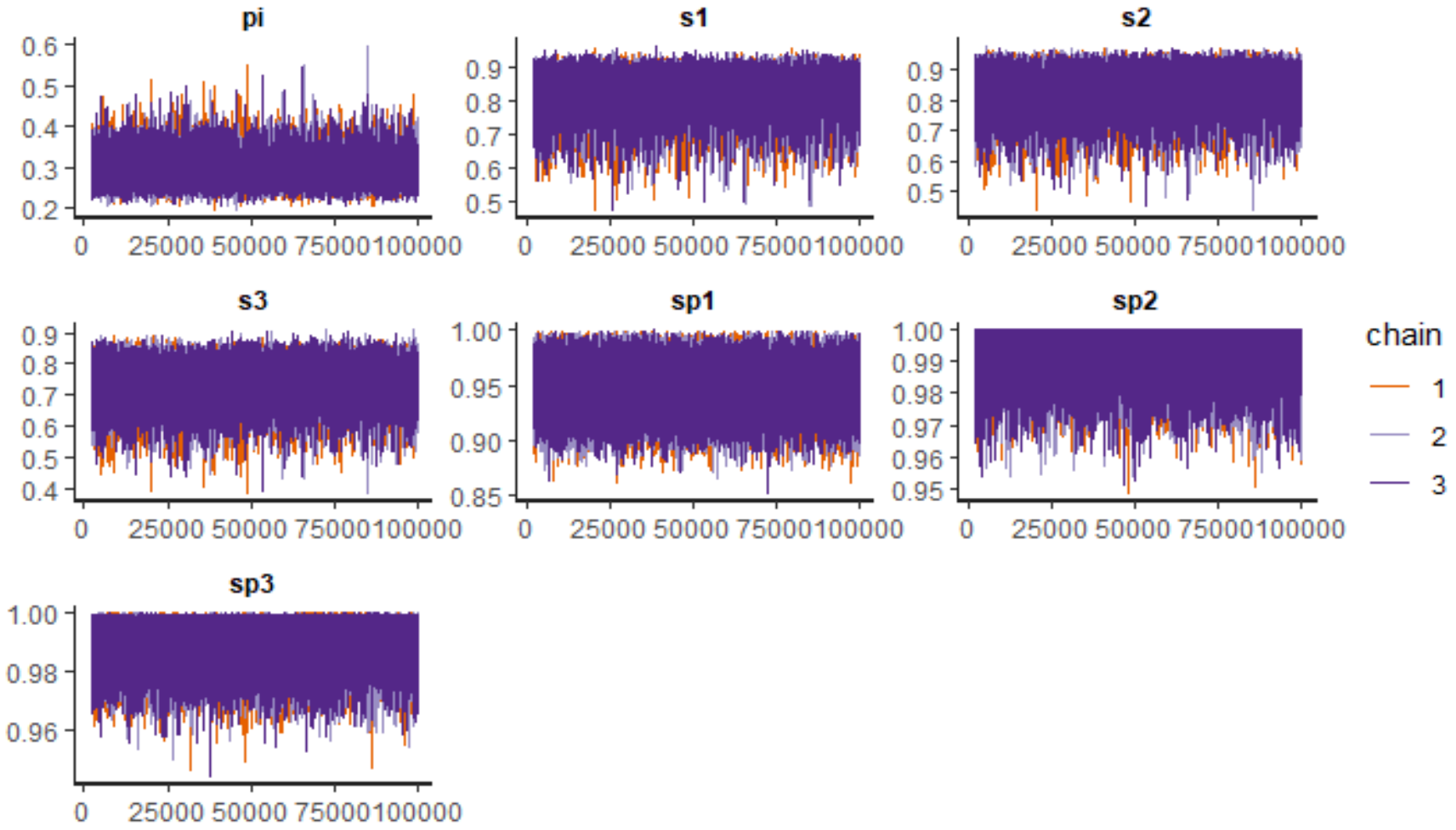
C.11. Diagnostic plots of 123CD under CD assumption and the priors are not centred on the simulated truth.

Figure 54: Density Diagnostic plots of 123CD under the assumption of conditional dependence using the FEMw.



The red colour under the curve is the 95% confidence interval

Figure 55: Trace Diagnostic plots of 123CD under the assumption of conditional dependence using the FEMw.



D.1. Diagnostic plots of the RABR dataset under CI assumption

When analysing the RABR dataset, DAS28-ESR$_4$ is denoted as the test 1, SDAI is demoted as test 2 and CDAI is denoted as test 3. This implies that s[1] represent the sensitivity of DAS28-ESR$_4$, s[2] represent the sensitivity of SDAI and s[3] represent the sensitivity of CDAI. Similarly the specificity of DAS28-ESR$_4$, SDAI and CDAI is denoted as sp[1], sp[2], and sp[3] respectively.

Figure 56: Trace plots of sensitivities and specificities of DAS28-ESR$_4$, SDAI and CDAI

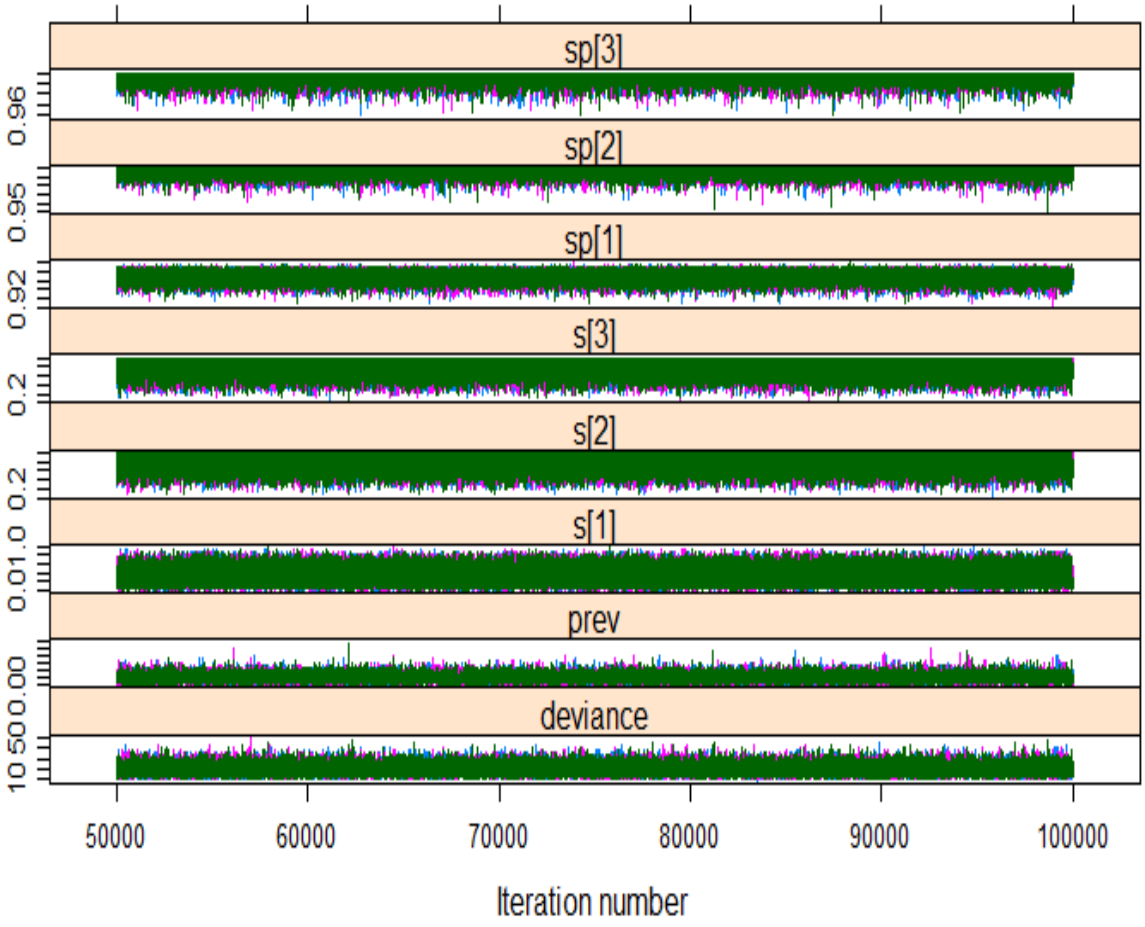


Figure 57: Auto-correlation plots of sensitivities and specificities of DAS28-ESR4, SDAI and CDAI

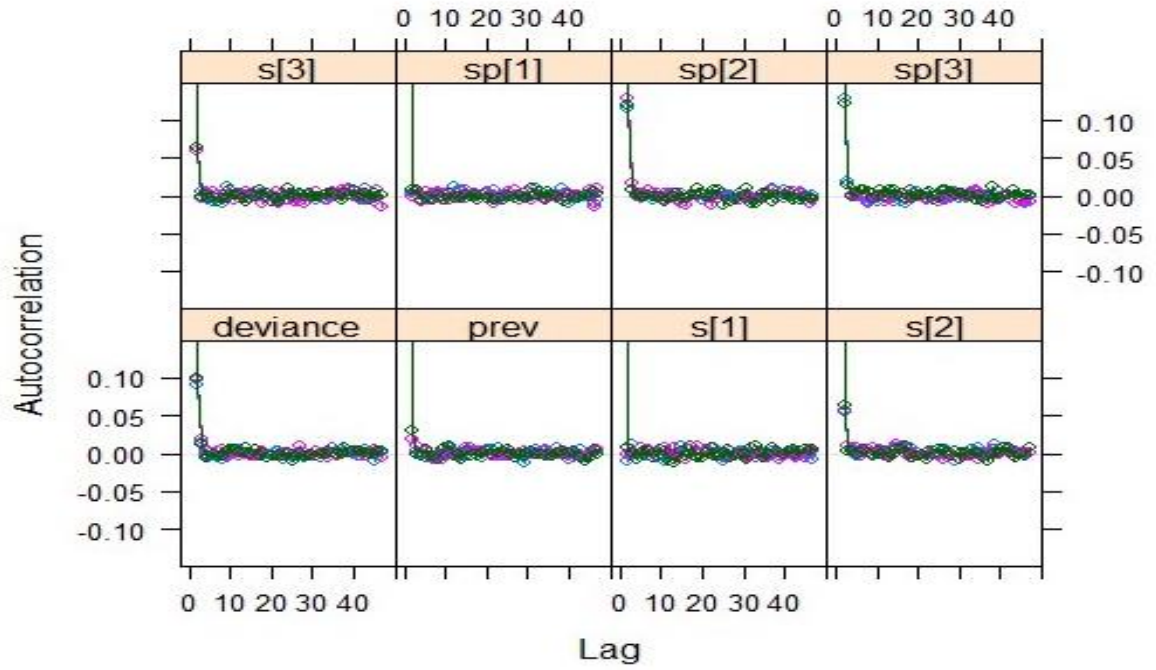


Figure 58: Density plots of sensitivities and specificities of DAS28-ESR4, SDAI and CDAI

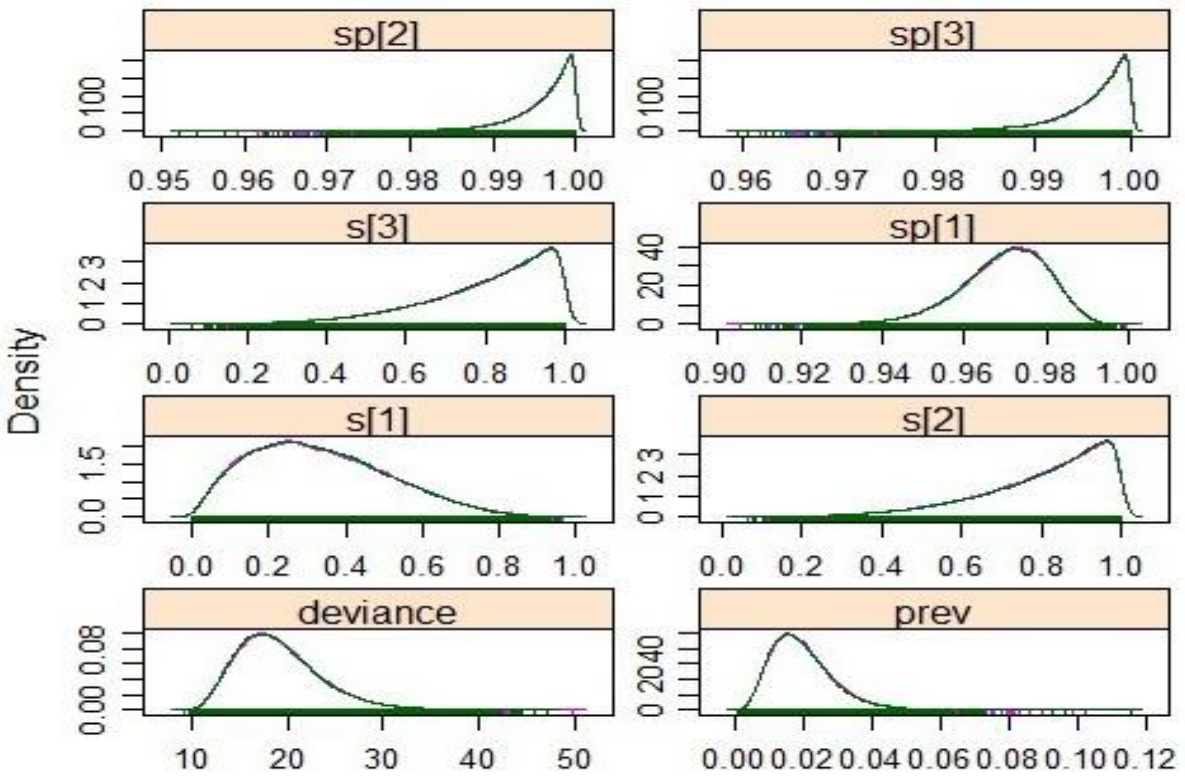
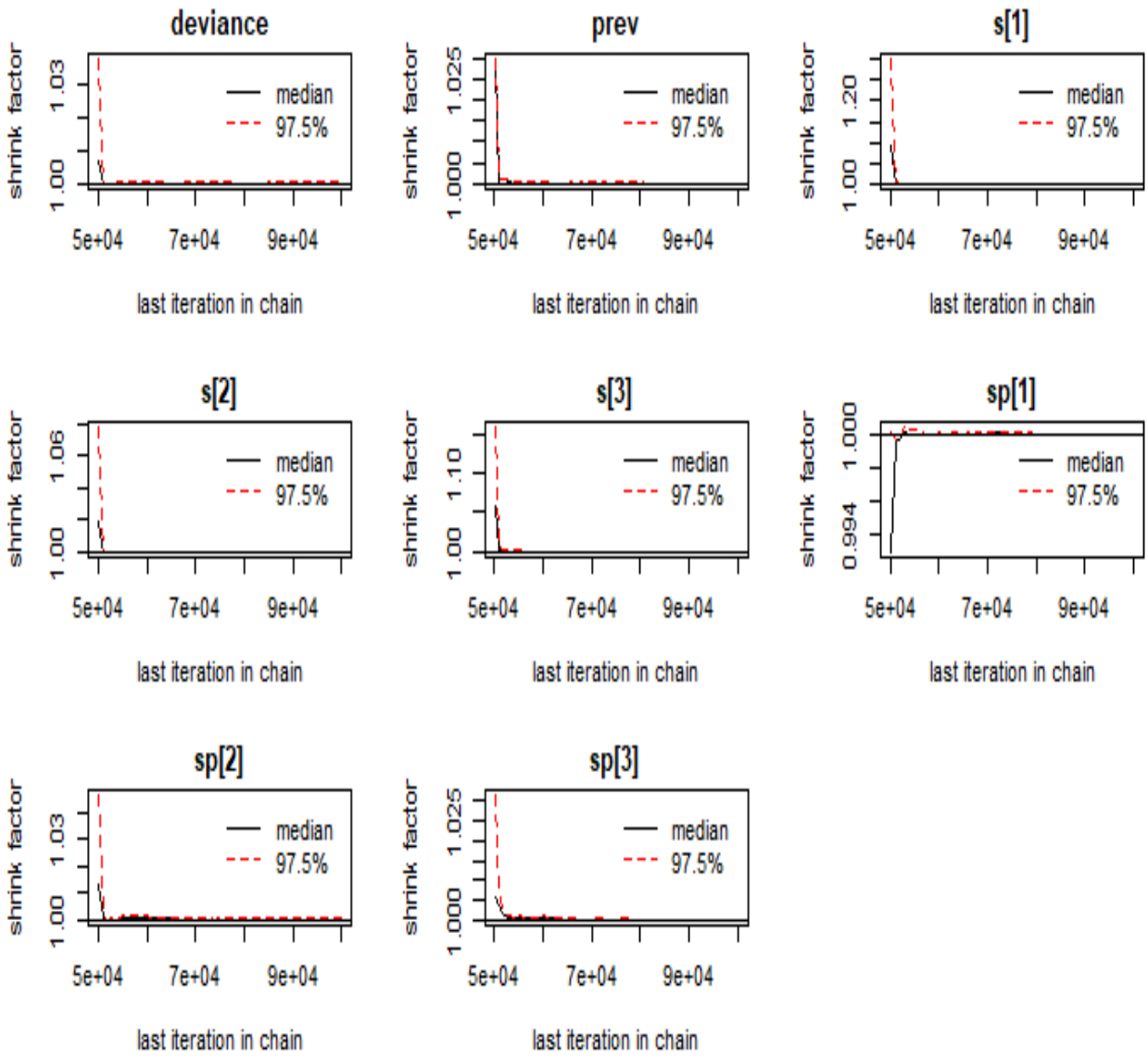


Figure 59: Gelman diagnostic plots of sensitivities and specificities of DAS28-ESR4, SDAI and CDAI



D.2. Density plots of the prior and posterior distribution of DAS28-ESR$_4$, SDAI and CDAI using the REML on the RABR dataset

Figure 60: Density plots of the prior and posterior distribution of the sensitivities of DAS28-ESR$_4$, SDAI and CDAI using the REML on the RABR dataset

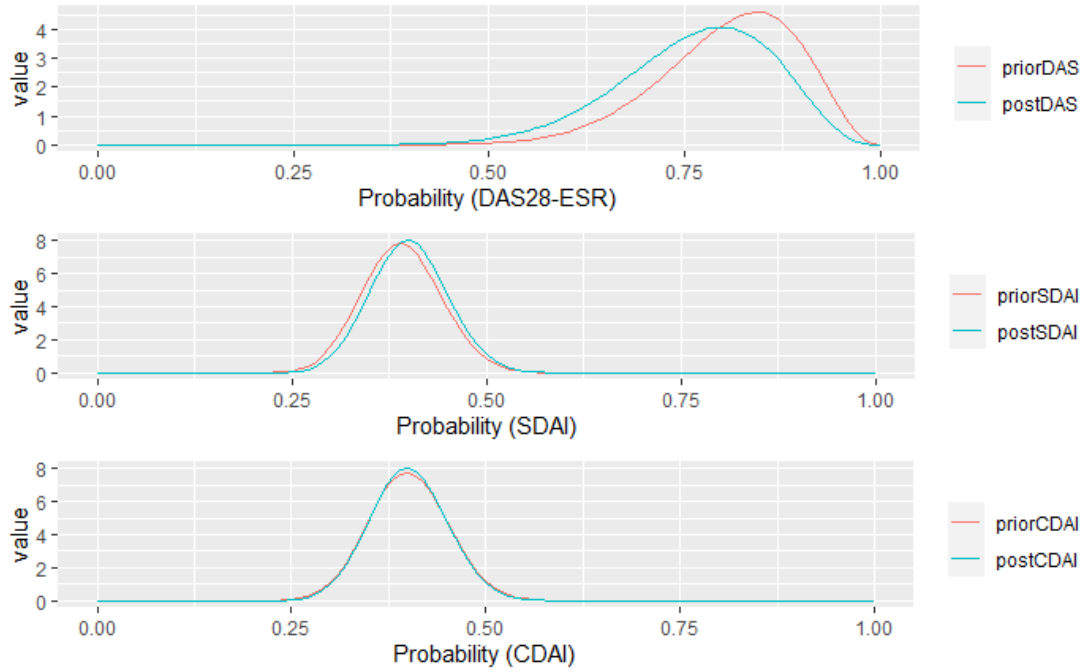
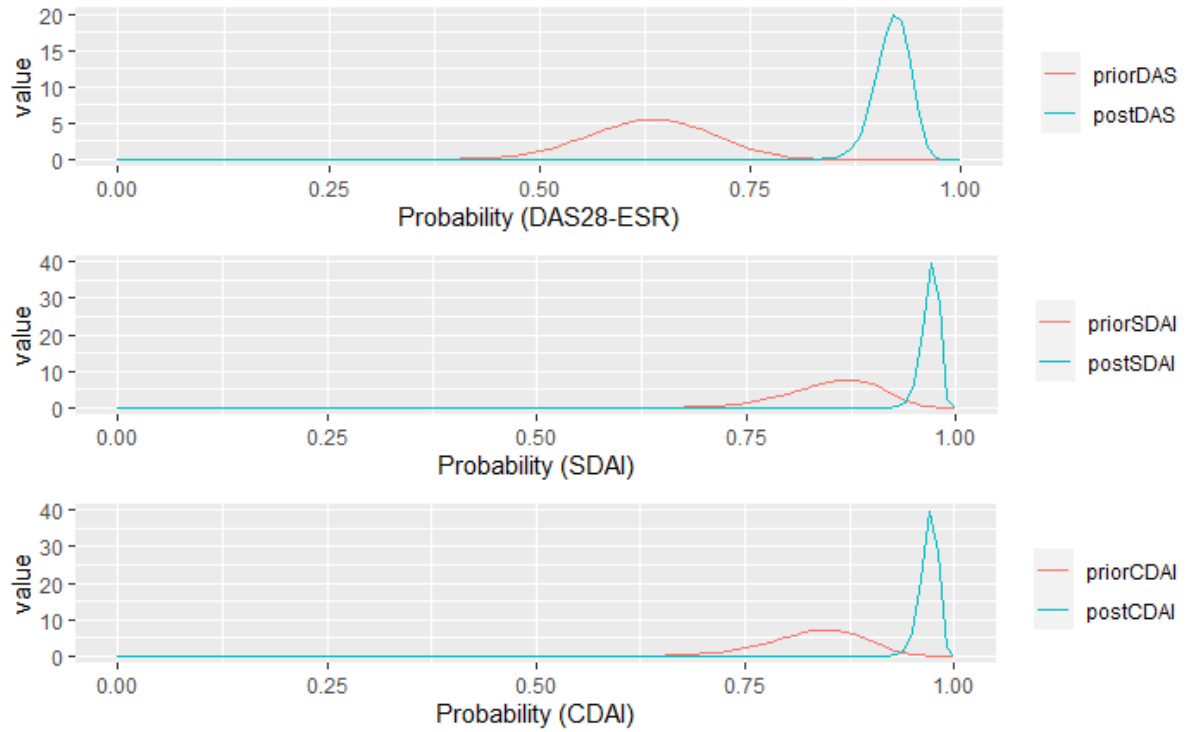


Figure 61: Density plots of the prior and posterior distribution of the specificities of DAS28-ESR$_4$, SDAI and CDAI using the REML on the RABR dataset



D.3. Diagnostic plots of the RABR dataset under CD assumption

Figure 62: Trace plots of sensitivities and specificities of DAS28-ESR4, SDAI and CDAI (REML)

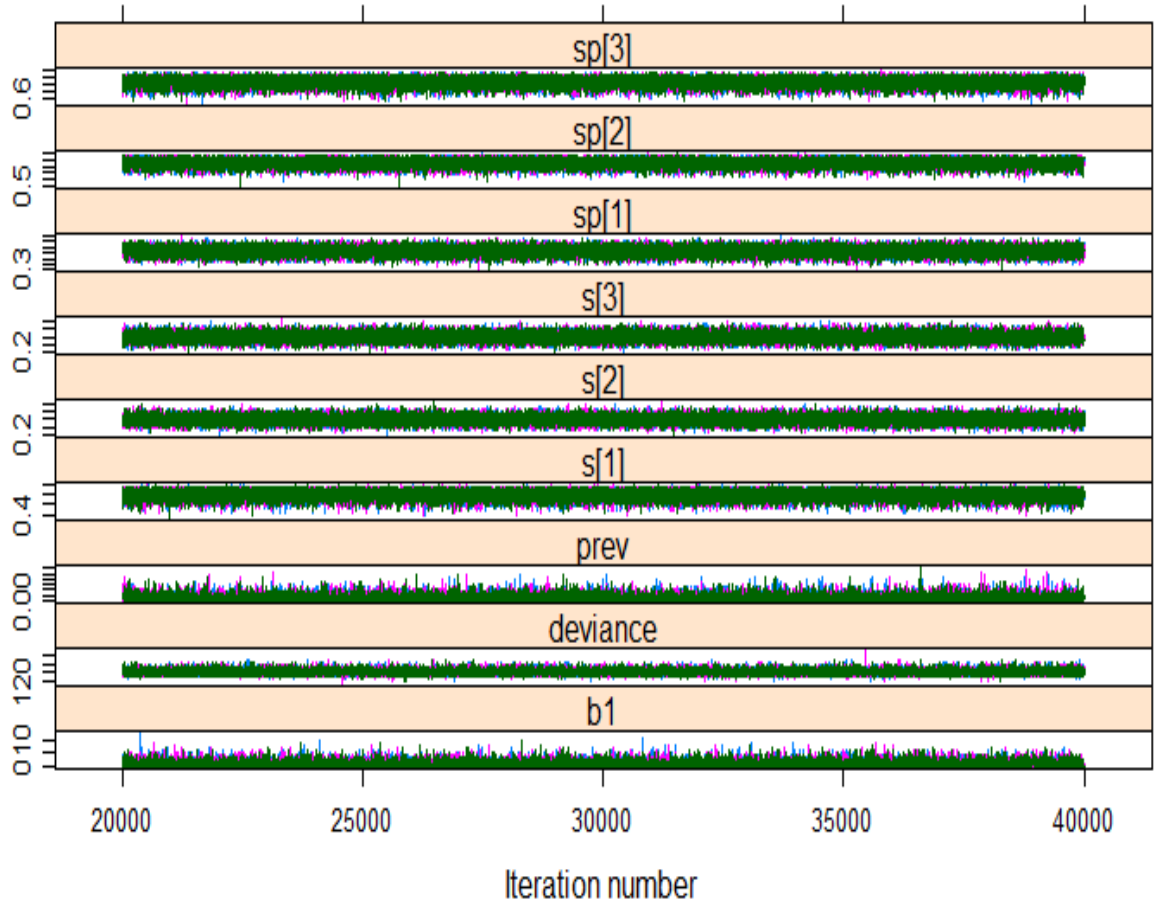


Figure 63: Density plots of sensitivities and specificities of DAS28-ESR4, SDAI and CDAI (REML)

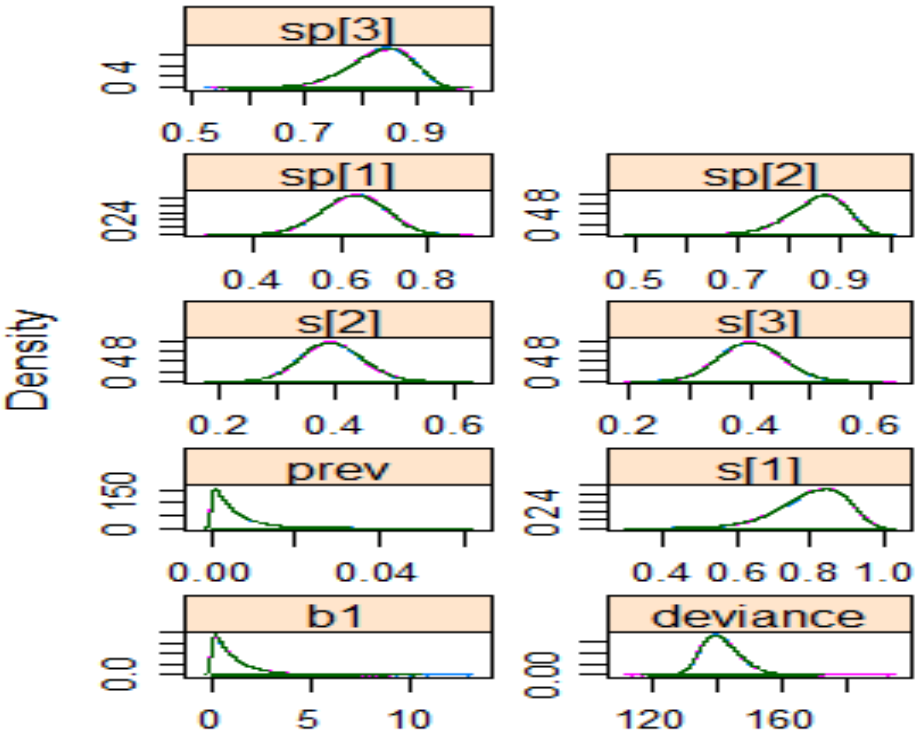


Figure 64: Auto-correlation plots of sensitivities and specificities of DAS28-ESR4, SDAI and CDAI (REML)

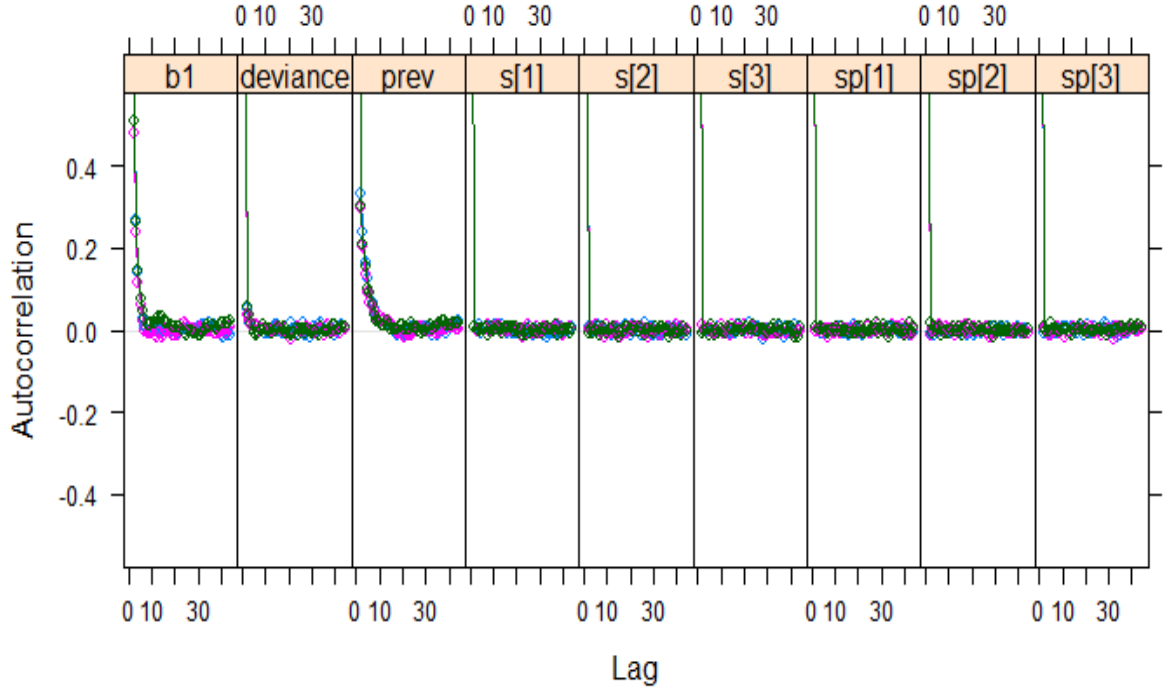


Figure 65: Gelman diagnostic plots of sensitivities and specificities of DAS28-ESR$_4$, SDAI and CDAI (REML)

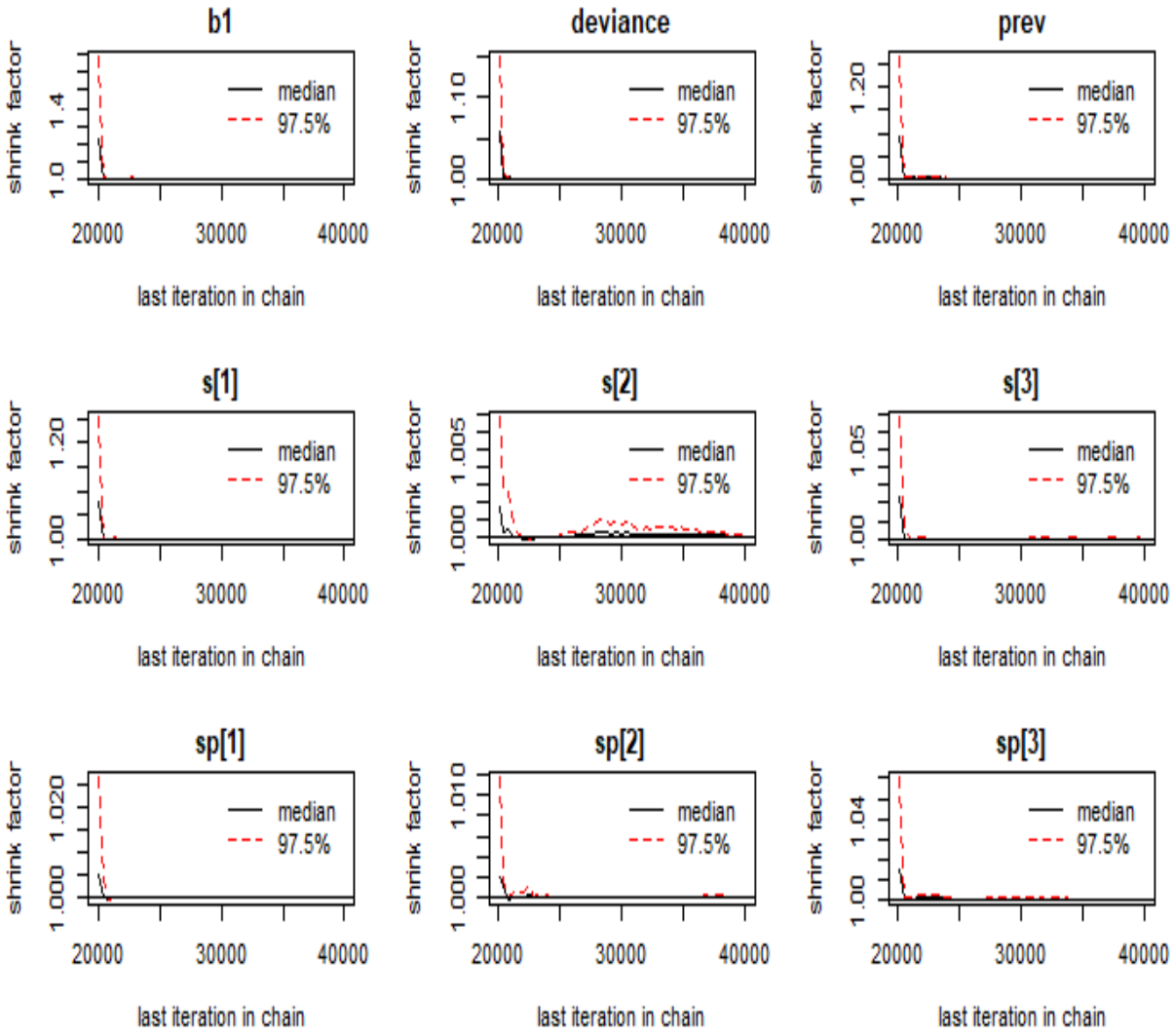


Figure 66: Trace plots of sensitivities and specificities of DAS28-ESR4, SDAI and CDAI FEMW

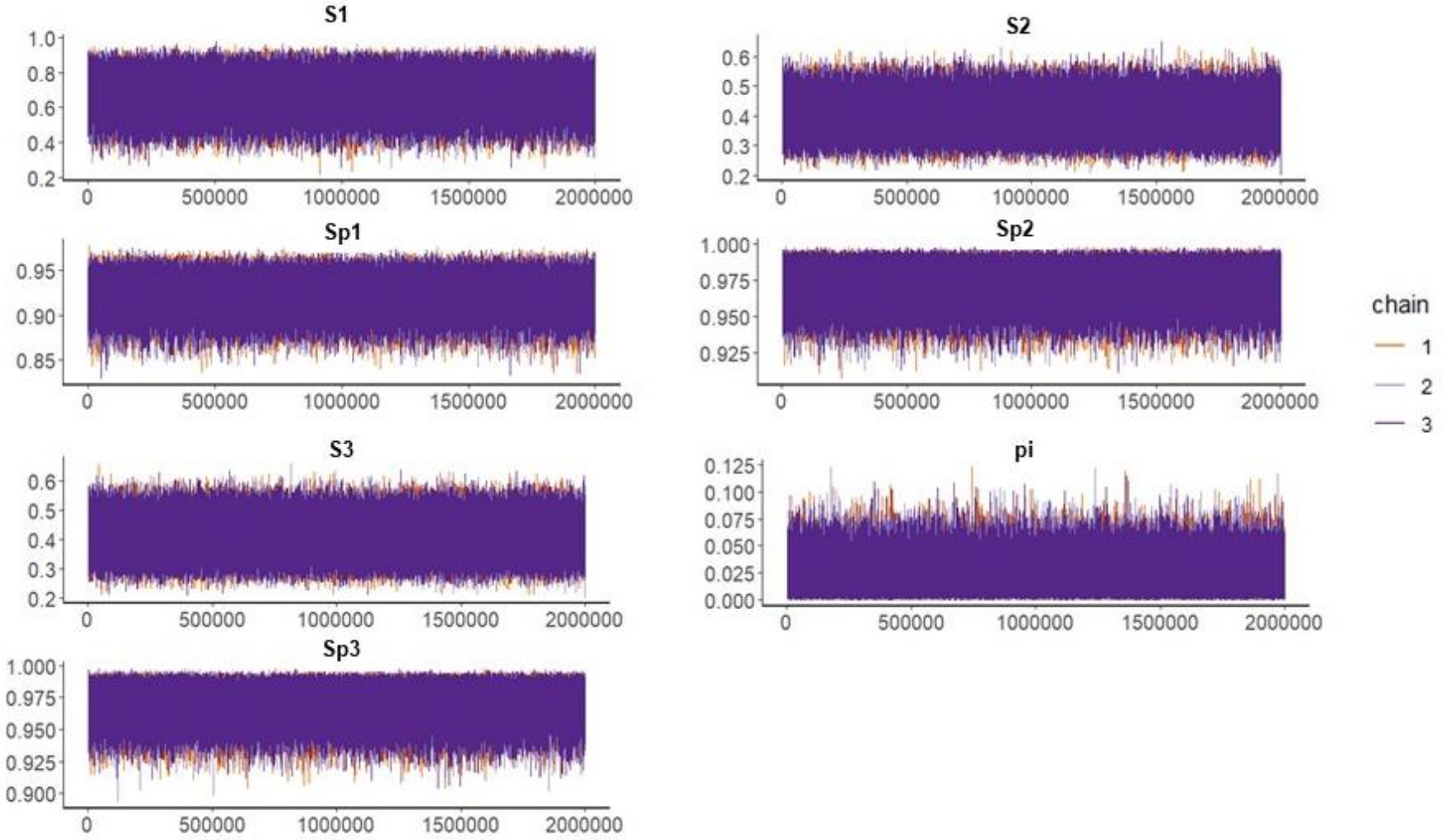
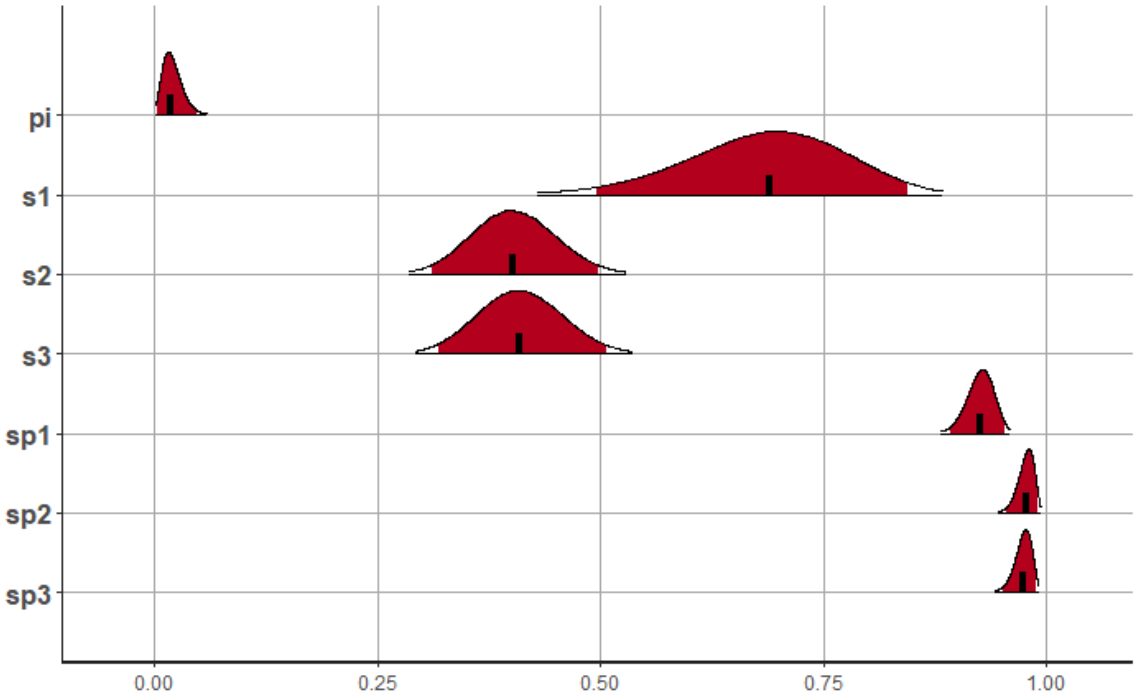


Figure 67: Density plots of sensitivities and specificities of DAS28-ESR4, SDAI and CDAI FEMw



The red colour under the curve is the 95% confidence interval

D.4. Diagnostic plots of the sensitivity analysis of RABR

Figure 68: Trace plots of the sensitivities and specificities of SDAI, CDAI and DAS28-ESR$_4$ (FEM$_W$).

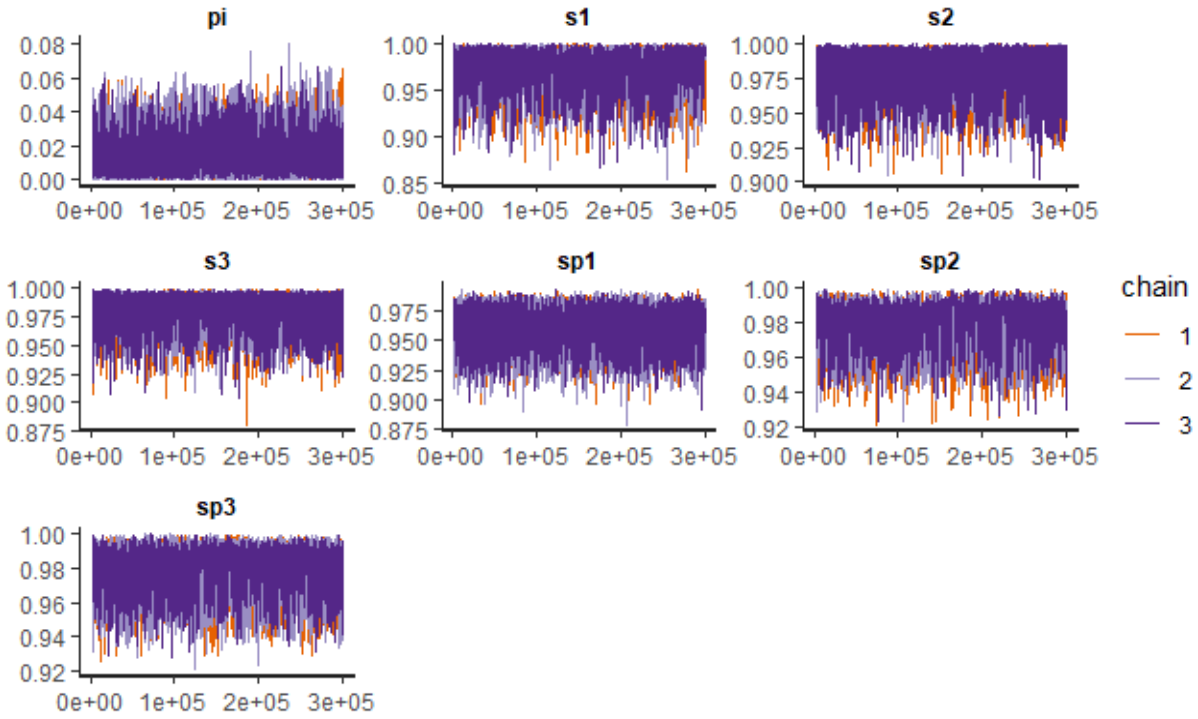
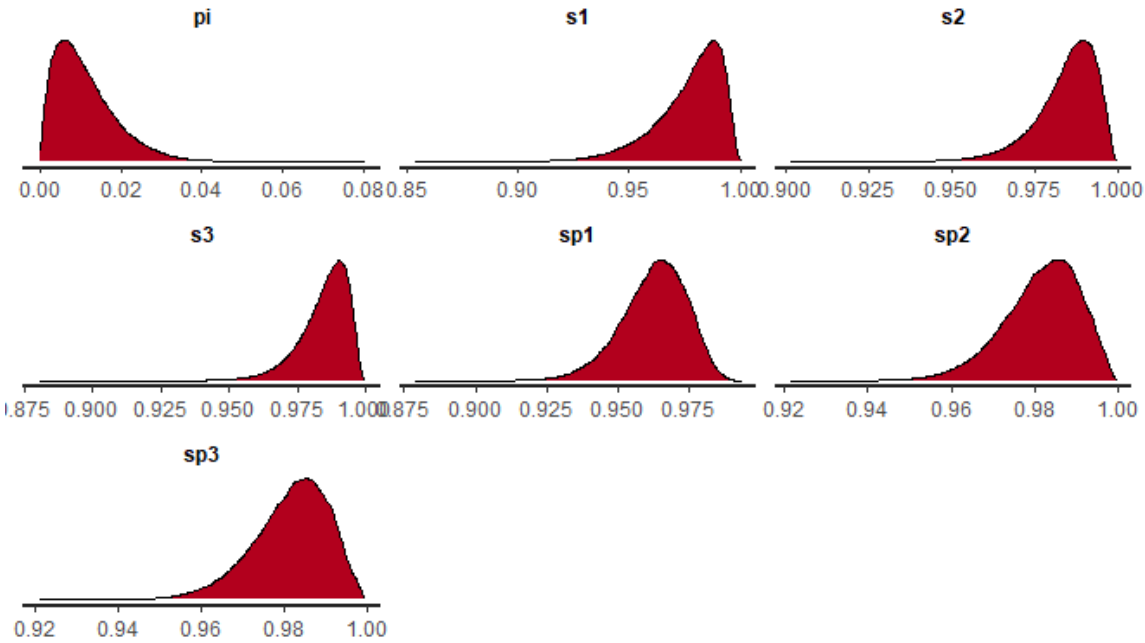


Figure 69: Density plots of the sensitivities and specificities of SDAI, CDAI and DAS28-ESR$_4$ (FEM$_W$).



D.5. R-Code for clinical dataset analysis

When analysing the DAS28-ESR$_4$ is denoted as the test 1, SDAI is demoted as test 2 and CDAI is denoted as test 3. This implies that s[1] represent the sensitivity of DAS28-ESR$_4$, s[2] represent the sensitivity of SDAI and s[3] represent the sensitivity of CDAI. Similarly the specificity of DAS28-ESR$_4$, SDAI and CDAI is denoted as sp[1], sp[2], and sp[3] respectively.

Start R code

Model 6: RStan clinical dataset RABR

enter data

```{r stan5}
FEMBR<- list(T = 8,  freqobs= c(1,0,0,7,3,0,0,253), N= 264)
## freqobs is the observed frequency of the combination of test responses
```

Code the model for analysis

```{r stancode}
FEMCD1<- "
data {
  int<lower=0>T; // number of possible combination of the tests response
        int<lower=0>freqobs[T];// the frequency of each possible combinations
        //int<lower=0>N; // total number of observation
}

// The parameters accepted by the model.
parameters {
 real<lower=0, upper=1> pi;         // prevalence
        real<lower=0, upper=1>s1;// sensitivity of T1
        real<lower=0, upper=1>s2;// sensitivity of T2
        real<lower=0, upper=1>s3;// sensitivity of T3
        real<lower=0, upper=1>c1;// specificity of T1
        real<lower=0, upper=1>c2;// specificity of T2
        real<lower=0, upper=1>c3;// specificity of T3

        // conditional dependence term for result (T1=1, T2=1, T3=1 | D=1)
```

```
        real<lower=-s1*s2*s3,upper=fmin(s1, fmin(s2, s3))-s1*s2*s3> covS111;
        real<lower=-(1-s1)*s2*s3,upper=fmin(1-s1, fmin(s2, s3))-(1-s1)*s2*s3> covS011;
        real<lower=-(1-s1)*(1-s2)*s3,upper=fmin(1-s1, fmin(1-s2, s3))-(1-s1)*(1-s2)*s3> covS001;
        real<lower=-(1-s1)*(1-s2)*(1-s3),upper=fmin(1-s1, fmin(1-s2, 1-s3))-(1-s1)*(1-s2)*(1-s3)> covS000;
}

transformed parameters {
real<lower=-s1*(1-s2)*(1-s3),upper=fmin(s1, fmin(1-s2, 1-s3))-s1*(1-s2)*(1-s3)> covS100;
real<lower=-s1*(1-s2)*s3,upper=fmin(s1, fmin(1-s2, s3))-s1*(1-s2)*s3> covS101;
real<lower=-s1*s2*(1-s3),upper=fmin(s1, fmin(s2, 1-s3))-s1*s2*(1-s3)> covS110;
real<lower=-(1-s1)*s2*(1-s3),upper=fmin(1-s1, fmin(s2, 1-s3))-(1-s1)*s2*(1-s3)> covS010;
vector<lower=0,upper=1>[T] pr;    // joint probability of each type of possible test result

// the pairwise conditional dependence between
        real covST12;         // test 1 and test 2
        real covST13;         // test 1 and test 3
        real covST23;         // test 2 and test 3

        // calculate the transformed conditional dependence terms
        covS100 = covS011 + covS111 - covS000;
        covS101 = -(covS001 + covS011 + covS111);
        covS110 = covS000 + covS001 - covS111;
        covS010 = -(covS000 + covS001 + covS011);

        // calculate the pairwise conditional dependence terms
        covST12 = covS111 + covS110;
        covST13 = covS111 + covS101;
        covST23 = covS011 + covS111;

        // probability of having combination of test response
        pr[1] = pi*(s1*s2*s3+covS111)+(1-pi)*((1-c1)*(1-c2)*(1-c3));    // 111
```

```
        pr[2] = pi*(s1*s2*(1-s3)+covS110)+(1-pi)*((1-c1)*(1-c2)*(c3));  // 110
        pr[3] = pi*(s1*(1-s2)*s3+covS101)+(1-pi)*((1-c1)*(c2)*(1-c3));  // 101
        pr[4] = pi*(s1*(1-s2)*(1-s3)+covS100)+(1-pi)*((1-c1)*c2*c3);    // 100
        pr[5] = pi*((1-s1)*s2*s3+covS011)+(1-pi)*((c1)*(1-c2)*(1-c3));  // 011
        pr[6] = pi*((1-s1)*s2*(1-s3)+covS010)+(1-pi)*(c1*(1-c2)*(c3));  // 010
        pr[7] = pi*((1-s1)*(1-s2)*s3+covS001)+(1-pi)*(c1*(c2)*(1-c3));  // 001
        pr[8] = pi*((1-s1)*(1-s2)*(1-s3)+covS000)+(1-pi)*((c1)*c2*c3);  // 000

}
// The model to be estimated. We model the output
// 'y' to be normally distributed with mean 'mu'
// and standard deviation 'sigma'.

model {
  // quartile method
        // priors:
        //s1 ~ beta (217,2.43); //s1~beta(15.2,3.68); // RABR
  //c1 ~ beta (62.5,177);
 // c1~beta(27.6,16.20); // RABR
  //s2 ~ beta (75.6,1.77);
 // s2 ~ beta(35.8,55.7);//RABR
  //c2 ~ beta (540,80.80);
  //c2 ~ beta(36.3,6.26); // RABR
  //s3 ~ beta (49.9,8.19);
 // s3 ~ beta(36.6,54.80); // RABR
  //c3 ~ beta (35.6,6.14);
  //c3 ~ beta(35.9,7.20); //RABR

        // likelihood function
freqobs ~ multinomial(pr);
}"
```

perform the analysis

```{r stanexam1}

#install.packages("Rtools")

fit<- stan(model_code = FEMCD1, data= FEMBR, iter = 100000, warmup = 2000, chains = 3, control = list(adapt_delta = 0.90))

#iter = 10000, warmup = 1000, chains = 2, verbose = TRUE,  max_treedepth = 15)
```

Print out results

```{r diag}

print(fit,  pars=c("pi","s1",  "s2",  "s3",  "c1","c2",  "c3",  "covST12",  "covST23",  "covST13", "log_lik"), digits.summary = 5)#, , probs=c(.1,.5,.9)
```

plot necessary plots

```{r stanplot}

plot(fit, pars=c("pi","s1", "s2", "s3", "c1","c2", "c3"), ci_level = 0.95, outer_level = 0.999)

plot(fit, show_density = TRUE, pars=c("pi","s1", "s2", "s3", "c1","c2", "c3"))

#plot(fit, show_density = TRUE, pars="pi",ci_level = 0.95,outer_level = 0.999)

#plot(fit, show_density = TRUE, pars="s1", ci_level = 0.95,outer_level = 0.999)

#plot(fit, show_density = TRUE, pars= "s2", ci_level = 0.95,outer_level = 0.999)

#plot(fit, show_density = TRUE, pars= "s3", ci_level = 0.95,outer_level = 0.999)

#plot(fit, show_density = TRUE, pars= "c1", ci_level = 0.95,outer_level = 0.999)

#plot(fit, show_density = TRUE, pars="c2", ci_level = 0.95,outer_level = 0.999)

#plot(fit, show_density = TRUE, pars= "c3", ci_level = 0.95,outer_level = 0.999)

traceplot(fit, pars=c("pi","s1", "s2", "s3", "c1","c2", "c3"))
```