Bayesian evidence synthesis methods for two diagnostic tests: application to Alzheimer's disease dementia

by

Athena McBride

Submitted to Swansea University in fulfilment of the requirements for the Degree of Doctor of Philosophy

Swansea University

2025

Abstract

This thesis considers a range of methodological challenges related to the synthesis of comparative diagnostic accuracy studies that evaluate two tests in the same patients, and aims to address them through the development of novel meta-analysis methodology in a Bayesian framework. The novel methods are applied to a real-world example in Alzheimer's disease dementia, for which test comparisons are important in optimising diagnostic pathways and improving detection. Firstly, the thesis introduces complex dependence structures present in meta-analyses of comparative diagnostic accuracy studies; in particular, within-study associations arising between sensitivities and specificities when patients undergo both tests of interest. This thesis assesses the impact of accounting for within-study dependencies on key test accuracy parameters by fitting a meta-analysis model that treats the two tests as independent to simulated data in which the associations are known. Ignoring within-study dependencies is shown to lead to underestimation of joint sensitivity and specificity, which measure the agreement between tests and enable modelling of diagnostic test combinations and pathways. This motivates the need for methodological development to jointly model the accuracy of two diagnostic tests and the associations between them. Novel Bayesian meta-analysis models for synthesising evidence on the accuracy of two diagnostic tests are developed, capturing within-study dependencies using bivariate copulas. Motivated by an example in Alzheimer's disease dementia, the bivariate copula framework is shown to lead to improved model fit compared to the approach that does not account for within-study associations. The bivariate copula models are extended to incorporate individual participant data on combined test performance, capturing within-study dependencies through trivariate copulas. These methods can be used to inform optimal combinations of diagnostic tests for health care policy and decision-making.

Declarations and statements

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed: Athena McBride

Date: 21st February 2025

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by explicit references. A bibliography is appended.

Signed: Athena McBride

Date: 21st February 2025

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed: Athena McBride

Date: 21st February 2025

Acknowledgements

I would like to express my gratitude to my supervisory team, including Professor Rhiannon Owen (Swansea University), Professor Rich Fry (Swansea University), Professor Terry Quinn (University of Glasgow) and Professor Sylwia Bujkiewicz (University of Leicester) for their guidance, support, and encouragement throughout my studies. Particular thanks are due to Professor Rhiannon Owen, for helping me navigate my MSc in Medical Statistics, my first research role and now my PhD. Thank you for always advocating for me and your endlessly positive attitude through the ups and downs of the last four years.

I would like to thank my funder, Health Data Research UK (HDR UK), for the opportunity to undertake this research. I would also like to thank the National Institute for Health Research (NIHR) for their support early in my academic career, without which I would not be in this position. Thank you to Professor Jon Deeks and Professor Yemisi Takwoingi for encouraging me to pursue a career in medical statistics and for sparking my interest in evidence synthesis with your passion and enthusiasm. I am also grateful to Professor Penny Whiting, Professor Hayley Jones and the team at NIHR Applied Research Collaboration West for their support when I embarked upon this PhD in 2019. Particular thanks to Dr Tasos Papanikos for generously sharing your insider knowledge of copulas, which saved me months of frustration.

Thank you to my Mum, Dad, and Emily-Jane for your love and confidence (and I forgive you for each time you asked, without malicious intent, when I would finish my PhD). Thank you to Philip, for your incessant support in the form of encouragement, reassurance, and, when all else failed, snack-based incentives.

Contents

List of Figures

List of Tables

Abbreviations

Abbreviation	Definition
10-CS	10-point cognitive screener
6CIT	6-item cognitive impairment test
AIC	Akaike information criterion
A$\boldsymbol{\beta}$	Amyloid-β
ANOVA	Analysis of variance
BRMA	Bivariate random effects meta-analysis
CDF	Cumulative distribution function
C-PIB	C-labelled Pittsburgh Compound B
CrI	Credible interval
CSF	Cerebrospinal fluid
CT	Computerised tomography
DIC	Deviance information criterion
DOR	Diagnostic odds ratio
DSM	Diagnostic and Statistical Manual of Mental Disorders
D-vine	Drawable vine
F-FDG	Flourine-fluorodeoxyglucose
FPR	False positive rate
FDA	Food and Drug Administration
HMC	Hamiltonian Monte Carlo
HPD	Highest posterior density
HSROC	Hierarchical summary receiver operating characteristic
HTA	Health technology assessment
ICD	International Classification of Disease
IPD	Individual participant data
IQR	Interquartile range
MCI	Mild cognitive impairment
MCMC	Markov Chain Monte Carlo

MMSE	Mini Mental State Examination
MoCA	Montreal Cognitive Assessment
MRI	Magnetic resonance imaging
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NINCDS-ADRDA	National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association
NUTS	No-U-Turn Sampler
RMSE	Root mean square error
PET	Positron emission tomography
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
P-tau	Phosphorylated tau
RCT	Randomised controlled trial
SROC	Summary receiver operating characteristic
SUVR	Standardised uptake value ratio
TPR	True positive rate
TSD	Technical Support Document
T-tau	Total tau
TYM	Test Your Memory
STARD	Standards for Reporting of Diagnostic Accuracy Studies
WAIC	Widely applicable information criterion
WHO	World Health Organization

Chapter 1

Introduction

1.1 Aims of the thesis

The increasing availability of diagnostic tests in contemporary healthcare has motivated the development of meta-analysis models to jointly synthesise the accuracy of multiple tests. The gold standard of evidence on the comparative or combined performance of two diagnostic tests is a comparative diagnostic accuracy study, which evaluates the accuracy of both tests in the same patients. Comparative diagnostic accuracy studies most commonly use a within-subject design, in which all patients undergo all tests of interest. Under this paired design, however, within-study associations arise between the sensitivities and specificities of the two tests.

Currently, the impact of accounting for these associations on key test accuracy parameters is unknown. While a number of meta-analysis models have been proposed to synthesise data on the accuracy of two diagnostic tests, many of these approaches do not account for correlations between multiple tests evaluated using a paired design. This is partially driven by limited reporting of fully cross-classified data, needed to account for within-study associations, in comparative diagnostic accuracy studies. The models that do account for the associations often make restrictive distributional assumptions and their application is restricted by availability of cross-classified data. This thesis considers a range of methodological challenges related to the meta-analysis of paired diagnostic accuracy studies. The limitations of existing methods are addressed through the development of novel meta-analysis models that account for within-study associations through a flexible Bayesian framework, making use of the full range of possible reporting formats for comparative diagnostic accuracy

studies.

The aims of the thesis are as follows:

1. Evaluate the impact of ignoring within-study associations between sensitivities and specificities in a meta-analysis of paired diagnostic accuracy studies.

2. Develop Bayesian meta-analysis models to evaluate the comparative diagnostic accuracy of two tests assessed against a common reference standard through a paired design, using a combination of study- and individual-level data to capture within-study associations.

3. Extend the models to evaluate the combined accuracy of two tests compared to a common reference standard under a paired design, using data at the individual-level to account for within-study associations.

The Bayesian evidence synthesis methods developed in this thesis will be applied to a case study in Alzheimer's disease dementia. Current research in Alzheimer's disease indicates that changes in certain biomarkers may precede clinically recognisable disease by decades.[1, 2] These biomarkers are beginning to be utilised in the diagnostic pathway for Alzheimer's disease dementia, allowing for earlier diagnosis and the possibility of preventative treatment. Through the application of methodology developed in this thesis to biomarker data, it is possible to evaluate their comparative and combined diagnostic accuracy, enabling the assessment of different testing strategies and diagnostic pathways. This could help to provide urgent answers on the utility of biomarkers as diagnostic tools and improve the diagnosis of patients with, or at risk of, Alzheimer's disease dementia. Methodological developments undertaken in this thesis will be generalisable to a wide range of disease areas.

In the remainder of this chapter, background information is provided on concepts that are key to understanding this thesis. Section 1.2 describes important features of meta-analysis of diagnostic accuracy studies, Section 1.3 introduces Bayesian statistical methodology, and Section 1.4 highlights the challenges faced in the diagnosis and treatment of Alzheimer's disease dementia and how biomarkers can help to address them. Section 1.5 describes the structure of the thesis and how each chapter relates to these aims.

1.2 Meta-analysis of diagnostic test accuracy studies

Meta-analysis is the statistical combination of results from two or more studies that answer the same research question, used to derive conclusions on the body of research and, in the context of healthcare decision-making, to inform evidence-based clinical practice.[3] Data from primary studies, identified through a systematic search of the available literature, are synthesised in a single analysis with the aim to produce a summary estimate of a parameter of interest, for example the effectiveness of an intervention for treating a condition or the accuracy of a diagnostic test for confirming or excluding it.

As well as allowing the integration of findings from multiple studies in a quantitative and coherent review, meta-analysis may produce a more precise estimate of the effect or accuracy measure compared to single studies that contribute to the analysis.[4] Meta-analysis can aid in understanding often-variable study results by identifying differences attributable to chance, those explained by observed differences in study or patient characteristics, and those likely to be real.[4] The degree of conflict in the literature is assessed by estimating the between-study standard deviation, and the uncertainty in the summary estimate is summarised using a confidence interval (or credible interval in a Bayesian framework, see Section 1.3).

Meta-analysis is not without its limitations if bias is introduced to the study pool or if statistical models are misused, however. If primary studies are biased due to their design, or if bias arises at the review stage through inappropriate research methodology, then the summary estimate produced by the resulting meta-analysis may not reflect the actual effect and lead to erroneous conclusions. Therefore, it is important to assess to what extent each primary study is at risk of bias using a validated tool, and to ensure that systematic reviews are conducted in line with guidelines such as the Cochrane Handbook for Systematic Reviews of Interventions or the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy.[5, 4]

Diagnostic test accuracy studies evaluate the ability of a test to correctly classify participants as having the target condition or not. The results (positive or negative) of an index test are compared to a reference standard test, which is assumed to perfectly classify participants' disease status. Diagnostic accuracy studies typically produce a pair of test accuracy measures that evaluate a test's ability to confirm or exclude a target condition, with sensitivity (the proportion of participants with

the target condition that are correctly identified by the test) and specificity (the proportion of participants without the target condition that are correctly identified by the test) being the most common metrics used.[6]

Heterogeneity often arises between sensitivity and specificity when diagnostic accuracy studies are combined in a meta-analysis due to differences in patient or test characteristics across studies. This between-study heterogeneity induces a dependence between sensitivity and specificity. A common source of between-study heterogeneity is test threshold, a criteria (such as a numerical cut-off) applied to a test to define a positive or negative result. As threshold for test positivity increases, specificity increases while sensitivity decreases. This trade-off results in a negative between-studies correlation between sensitivity and specificity, that may lead to their underestimation unless the association is adequately captured.[7] Sensitivity and specificity in each study are estimated using distinct patient populations: those with the target condition and those without. Therefore, in a meta-analysis of diagnostic accuracy studies evaluating a single test against a reference standard, the measures are considered independent at the within-study level.

1.2.1 Methodological challenges in meta-analyses of multiple diagnostic tests

In comparison to evidence synthesis methods for single diagnostic tests, methodological development to jointly evaluate the accuracy of two tests against a common reference standard is relatively novel and as such guidance and best practice are ever-evolving. Most clinically relevant questions are comparative in nature; whether a novel diagnostic test is sufficiently accurate to use in practice depends on the accuracy of other, existing tests. Joint synthesis of multiple tests also enables assessment of the accuracy of different testing strategies and diagnostic pathways. Diagnostic accuracy studies, however, often focus on evaluating single tests, making direct comparison challenging and potentially introducing bias where non-comparative studies are combined to answer comparative questions on test accuracy.[8]

Unlike intervention studies, in which patients are frequently randomised to independent treatment groups, participants of comparative diagnostic accuracy studies of two tests often undergo both tests of interest plus a reference standard. Under this paired design, within-study dependencies arise between sensitivities and specificities of the two tests. Cross-classified data for each study - containing all possible com-

binations of test results compared to true disease status, equivalent to individual participant data (IPD) - are typically required to account for the associations between multiple tests, but are rarely reported in practice.[9, 10, 11, 4] Assessing the importance of accounting for these dependencies in an evidence synthesis framework - whether they impact summary accuracy estimates, and under which conditions - represents a novel area of methodological development. The need to investigate whether more complex models for synthesising comparative diagnostic accuracy studies, including methods that incorporate within-study associations between multiple tests, outperform simpler models, which assume independence between tests, has been highlighted.[12]

As well as additional, complex dependence structures present in paired studies, computational difficulties may also arise as a consequence of synthesising data on multiple tests in a single analysis. As the number of tests increases, the number of estimated parameters increases exponentially. This can lead to computationally demanding models with long run times, imprecise estimates of key parameters, and issues with model convergence, particularly if the number of participants or studies is low. Choice of model parametrisation, Bayesian sampler (see Section 1.3) and software are important considerations when jointly modelling multiple tests and the dependence structures between them.

1.2.2 Health technology assessment of diagnostic tests

Health technology assessment (HTA) is a systematic process for evaluating the properties of a health technology used in the prevention, diagnosis or treatment of a condition. HTA assesses the clinical and cost-effectiveness of an intervention, as well as its impact, both direct and indirect, on the broader health system. It is a multidisciplinary process that aims to inform healthcare decision-makers, and acts as a bridge between scientific research and policy-makers.[13] The National Institute for Health and Care Excellence (NICE) determines whether novel health technologies are reimbursed by the National Health Service (NHS) in England and Wales, balancing high quality health and social care with value for money by producing evidence-based guidance to improve outcomes for both individual patients and society as a whole.

The accuracy of a novel diagnostic technology may be evaluated through one of two NICE programmes. The first is the Medical Technologies Evaluation Programme, which covers tests that offer similar health outcomes at less cost to the NHS or

improved outcomes as the same cost at current NHS practice. The second is the Diagnostics Assessment Programme,[14] which covers tests that have the potential to improve health outcomes but at an increased cost to the NHS, meaning their evaluation is more complex and may require clinical and cost-effectiveness analysis or assessment in comparison to other technologies. The latter can require the use of complex evidence synthesis methods, particularly those that allow the joint meta-analysis of diagnostic accuracy data on two or more tests.

A report by the NICE Decision Support Unit published in 2020, commissioned to support NICE HTA programmes, found there was no clear guidance on the synthesis of comparative diagnostic accuracy studies that use a paired design.[15] The report noted that cross-classified data (discussed in detail in Section 2.3.4) are required to account for within-study associations between tests. The authors also highlighted that research questions prioritised by the NICE Diagnostics Assessment Programme often relate to the joint accuracy of multiple diagnostic tests used in combination, estimation of which requires modelling techniques that account for within-study dependencies. Existing models that allow explicit estimation of joint accuracy parameters were reported to often run into computational difficulties, even for seemingly simplistic scenarios.

1.3 Bayesian methodology

Bayesian methodology possesses several defining features that make it a natural approach for healthcare decision-making. There are two distinct schools of thought in statistical inference, which differ primarily in their handling of uncertainty: the frequentist approach and the Bayesian approach.[16] In a frequentist framework, parameters are assumed to have fixed, but unknown, true values. Inference is made about the probability of the observed data at specific parameter values, rather than the parameter itself. In a Bayesian framework, each parameter in the model is assumed to be unknown and is estimated from a probability distribution; both the model parameters and the data are treated as random elements, meaning uncertainty around each is captured.[17] Bayesian methods are a natural fit for the clinical and cost-effectiveness analyses that form a vital component of the HTA process, as healthcare decision-making is often relatively conservative and required under a level of uncertainty.[3] Using Bayesian modelling, direct probability statements about posterior parameters can be made, allowing conclusions to be presented in a more

intuitive form that aids decision-making.

Bayesian models can be updated in light of new knowledge regarding a parameter or set of parameters, forming posterior distributions that summarise both the observed data via specification of a likelihood and external evidence via a prior distribution. This feature of Bayesian methodology offers a flexible approach to modelling complex clinical scenarios, while making efficient use of the available literature base. The development of accessible and user-friendly software to implement different Bayesian samplers, including Markov Chain Monte Carlo and Hamiltonian Monte Carlo simulation, make it relatively straightforward to fit increasingly complex Bayesian models.[16]

1.4 Alzheimer's disease dementia

Dementia describes a progressive, clinical syndrome in which there is impairment in cognitive function beyond what might be expected during the usual process of ageing. It encompasses a range of cognitive, psychological and behavioural symptoms, including difficulties with memory loss, problems with language and communication, and change in personality.[18] Dementia is characterised by the progressive loss of ability to perform activities of daily living;[19] people with dementia will eventually be unable to look after themselves, and may lose the ability to communicate. Dementia is thought to progress through a series of stages, from biologically active but clinically silent, to impairments of memory and thinking that are not sufficient to interfere with daily activity (mild cognitive impairment), to overt dementia.

According to recent WHO statistics, dementia is currently the seventh leading cause of death among all diseases worldwide and is recognised as a global health priority, subject to the WHO 'Global action plan on the public health response to dementia 2017–2025'.[20, 21] A report published by the Alzheimer's Society in 2019 estimated that there are approximately 885,000 older people living with dementia in the UK, increasing by 80% to an estimated 1.6 million by 2040.[22] The same report estimated that dementia cost the UK £34.7 billion in 2019, with social care (£15.7 billion), unpaid care (£13.9 billion) and healthcare (£4.9 billion) accounting for the majority of the expense. By 2040, this figure is expected to rise to £94.1 billion - an increase of 172%.[22]

As well as the challenge to health and social care provision, dementia has consid-

erable impact on the physical, mental and social well-being of people living with the condition, along with their carers and families. Caring is often a stressful and mentally demanding role, reflected in the increased rates of depression and anxiety found in dementia carers.[23, 24] The psychological effect of caring is exacerbated by increased social isolation and lack of support. The effects of caring for a family member with dementia on physical health are well documented, with carers found to be at increased risk of cardiovascular problems, lowered immunity, and chronic conditions such as diabetes and arthritis.[25] Carers are less able to partake in preventative lifestyle activities such as exercise and more likely to smoke, drink alcohol and have poor sleep patterns.[25] Considerable financial effects are felt by those with dementia and their carers, including both direct costs, such as health and social care, and indirect costs, such as loss of earnings, unpaid care, and the mortality burden. In England, an estimated 60.6% of the cost of social care is met by service users and their families.[22]

Alzheimer's disease is the most common cause of dementia, estimated to be attributable in 62% of dementia cases in people aged 65+ years in the UK.[26] Alzheimer's disease is thought to be caused by the build up of abnormal protein structures in the brain, such as amyloid plaques or tau tangles, which disrupt the connection between nerve cells. Alzheimer's disease dementia is characterised by impairment in the ability to learn and recall recent information.[27] Other subtypes include vascular dementia (17%), dementia with Lewy bodies (4%), frontotemporal dementia (2%) and Parkinson's dementia (2%).[26] The boundaries between different types of dementia can be indistinct. 'Mixed' forms of dementia, which occur when a person simultaneously has two or more types, account for around 10% of dementias in the UK ,[26] and the majority of dementia found in patients aged 80 years or older.[28, 29]

1.4.1 Diagnosis of Alzheimer's disease dementia

Dementia is underdiagnosed and typically detected at a relatively late stage in the disease process. Alzheimer's Disease International estimates that there are 41 million undiagnosed cases of dementia globally, representing 75% of people with dementia worldwide.[30] This rate is as high as 90% in some lower- and middle-income countries. COVID-19 has exacerbated the issue further, causing, for the first time, a sustained drop in dementia diagnoses. Alzheimer's Society report that, from January 2020 to January 2022, dementia diagnosis rates in England dropped from 67.6%

to 61.6%, falling short of the UK government objective to maintain a minimum of two-thirds diagnosis rate.[31] Due to the backlog caused by the COVID-19 pandemic, it is estimated that there are over 30,000 people living without the dementia diagnosis they would otherwise have received,[32] restricting their access to essential support services. International healthcare systems are looking to redesign their diagnostic pathways for dementia, and new technologies such as biomarkers are expanding the diagnostic options available.[33]

Early diagnosis of dementia enhances the potential for good quality of life for people with dementia, carers, and their families. From the perspective of the person living with dementia, diagnosis increases access to vital care resources, aids symptom management through the support of a multidisciplinary care team, and provides an explanation for potentially distressing cognitive and behavioural changes. Prompt detection enables people with dementia and their carers to plan and prepare for the future, and ensure appropriate levels of care and legal powers are in place when they are needed so that quality of life does not suffer. Early diagnosis empowers the person living with dementia to be involved in decisions around their care when they are most able to do so. From a therapeutic perspective, a confirmatory diagnosis of Alzheimer's disease dementia is required for any pharmacological intervention to be prescribed. Improved diagnostic systems can also help to identify participants with early cognitive impairment for trials of new treatments.[2] With the emergence of new therapeutic medications for the first time in decades,[34, 35] early diagnosis may provide vital access to drugs that delay symptom progression and cognitive decline in the near future.

Under current NICE guidelines in the UK,[36] patients with suspected dementia undergo cognitive testing using a validated brief structured cognitive instrument, such as the Mini-Cog or Memory Impairment Screen. Dementia subtype is determined by further cognitive testing in a specialist setting; most patients will receive an imaging test and, where uncertainty around the diagnosis persists, may undergo additional biomarker testing. Clinical diagnosis can be particularly challenging in the earlier stages of Alzheimer's disease dementia when symptoms are mild. Where there are several similar tests for the same condition, such as cognitive questionnaires for diagnosing Alzheimer's disease dementia, models that estimate the comparative diagnostic accuracy of the tests are key to determining which should be used in clinical practice. To compare the accuracy of different diagnostic pathways or testing strategies comprising of multiple components, statistical methods that estimate

combined test accuracy are required. For example, the combination of two or more biomarker tests for Alzheimer's disease dementia may increase their sensitivity or specificity compared to each test applied in isolation. However, this can only be assessed through the joint synthesis of diagnostic accuracy data on all tests of interest, using a model that accounts for dependencies between multiple tests.

More work is needed to ensure that an accurate and timely diagnosis is available to each person living with dementia. An Alzheimer's Society survey of people with dementia or close to someone with dementia found that 91% saw at least one benefit to getting a diagnosis, such as access to medication or practical support from health and social care services, with many wishing they had received their diagnosis earlier.[37] Identifying the condition early in the disease process improves quality of life for people with dementia and those around them, and, as new treatments become available, offers an opportunity to manage symptom progression. If and when a preventative treatment is available through the NHS, early diagnosis will be key to identify patients who would benefit from the intervention.

1.4.2 Biomarkers

Biomarkers are characteristics that can be objectively measured and used to evaluate a biological process. Current research in Alzheimer's disease indicates that changes in certain biomarkers may precede clinically recognisable disease by decades.[1, 2] Figure 1.1 gives a theoretical example of how biomarkers could move from normal to abnormal over time as an individual moves from a cognitively normal state to a state of overt dementia.

Biomarkers can be used as early diagnostic indicators of changes in Alzheimer's disease pathology. A review by Hampel et al [38] summarised the existing literature on cerebrospinal fluid (CSF) biomarkers as diagnostic tools for Alzheimer's disease dementia. There is evidence that amyloid-beta (Aβ) peptides, measured in the CSF, decline as amyloid plaques in the brain accumulate and cognition becomes increasingly impaired, indicating their potential as diagnostic tool for Alzheimer's disease dementia. At the time of the review around 20 studies of Aβ had been conducted in approximately 2,000 Alzheimer's disease dementia patients and controls, with the estimated sensitivity and specificity of the marker ranging between 80-90%. The same review found that sensitivity and specificity of total tau (t-tau) protein, another characteristic pathology of Alzheimer's disease dementia measured in the CSF,

ranged between 80-90%, respectively, based on around 50 studies in approximately 5,000 patients and controls.

In a more recent review, Ritchie et al [39] found that sensitivity and specificity of CSF Aβ biomarkers for Alzheimer's disease dementia ranged between 36-100% and 29-91%, respectively, in their analysis of 14 studies. In another review, Ritchie et al [40] found that sensitivity of t-tau biomarkers in the CSF ranged between 51-90% and specificity ranged between 48-88%, based on their analysis of 6 studies. Considerable uncertainty around the accuracy of combinations of biomarkers for diagnosing Alzheimer's disease dementia remains.

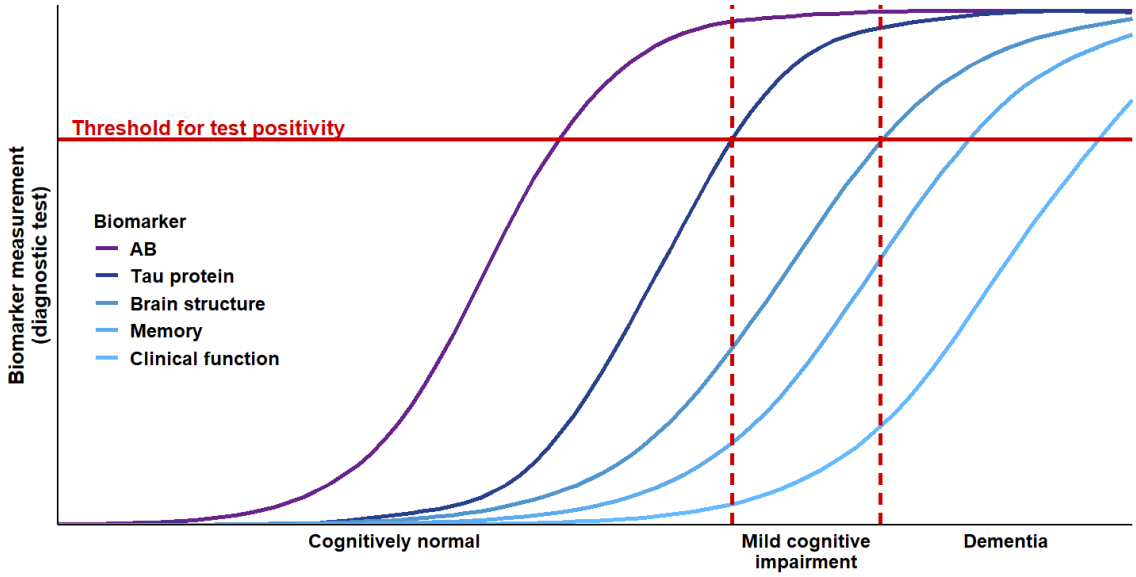


Figure 1.1: Diagram showing the theoretical progression of different biomarkers over time, adapted from a schematic diagram by Jack et al (2010).[1]

Biomarkers for Alzheimer's disease dementia have other potential uses beyond their utility as diagnostic tests. They may also be used to develop a mechanistic understanding of the underlying biology of the disease, conduct population-level screening to detect Alzheimer's disease risk, measure treatment effects on Alzheimer's disease pathology over time, and act as a surrogate endpoint for treatment effect on cognitive impairment in clinical trials of pharmacological interventions for Alzheimer's disease dementia. Application of biomarkers to predict treatment effects is still at an early stage compared to their use as diagnostic tests. The FDA approved the use of biomarkers as surrogate endpoints in clinical trials of treatments for early Alzheimer's disease dementia in 2018,[41] and the first treatments have recently been approved in the US using this accelerated pathway.[34, 35] Appropriate evaluation of novel

biomarkers as reliable predictors of clinical benefit through a post-approval study is required, and is indeed a condition of the accelerated approval pathway.

1.5 Structure of the thesis

The thesis is organised into eight chapters. The first provides an introduction to current methodological challenges in the field of diagnostic test evaluation that this thesis aims to address. Important clinical definitions are established and the urgent need to assess the comparative and combined diagnostic accuracy of biomarkers in Alzheimer's disease dementia is highlighted.

Chapter 2 outlines the statistical methods relevant to the thesis, covering the diagnostic test evaluation framework, meta-analysis, and Bayesian statistics.

Chapter 3 describes the methodological review undertaken to identify and describe existing statistical models to jointly analyse the accuracy of two or more diagnostic tests. Key strengths and limitations of current methods are highlighted and discussed. These methods will be formally evaluated and extended, representing novel methodological development, in Chapters 5, 6 and 7.

Chapter 4 describes the literature review undertaken to identify comparative diagnostic accuracy studies of cognitive, imaging and biomarker tests for Alzheimer's disease dementia. These data are used in Chapters 5, 6 and 7 to enable development of novel methodology and to evaluate the accuracy of different tests and testing combinations for diagnosing Alzheimer's disease dementia.

Chapter 5 presents the simulation study undertaken to evaluate the impact of accounting for within-study dependencies between sensitivities and specificities in a meta-analysis of comparative diagnostic accuracy studies, addressing the first aim of this thesis. The currently recommended meta-regression approach for synthesising diagnostic accuracy data on two tests assessed using a paired design, which treats tests as independent, is fit to simulated data where within-study dependencies are known. Data are simulated based on a motivating example in Alzheimer's disease dementia, extracted as part of the literature review in Chapter 4. The findings of this chapter, alongside the methodological review in Chapter 3, motivate the need for methodological development to jointly model two diagnostic tests and the within-study dependencies between them, using a flexible and computationally efficient meta-analysis approach.

Chapter 6 describes model development undertaken in a Bayesian meta-analysis framework to jointly evaluate the accuracy of two diagnostic tests. Aggregate, study-level data are fed into the models, which use bivariate copulas to flexibly account for within-study dependencies between the sensitivities and specificities of two tests assessed using a paired design. The models can utilise external evidence, as well as IPD from a single study or a subset of studies, to estimate copula dependence parameters that capture association between multiple tests, improving upon limitations of existing methods highlighted in Chapter 3. The models are fit to data on the comparative accuracy of two diagnostic tests for Alzheimer's disease dementia, demonstrating their contribution to evaluation of novel biomarker tests and applicability to this disease area. The performance of the bivariate copula models are compared to a meta-regression model that assumes independence between the two tests and therefore does not account for within-study dependencies.

Chapter 7 extends the models developed in Chapter 6 to fully account for within-study dependencies between two diagnostic tests evaluated under a paired design. Trivariate copulas are used to directly estimate both the marginal and joint sensitivities and specificities of the two tests. The models incorporate IPD from all studies, making full use of the available evidence base. The extension of the bivariate copula methodology to a trivariate copula approach addresses questions on the combined accuracy of the tests, allowing the assessment of realistic diagnostic pathways and testing strategies. Motivated by an example in Alzheimer's disease dementia, the models are fit to individual-level data on the diagnostic accuracy of two tests simulated in Chapter 5.

Chapter 8 concludes the thesis with a discussion. The findings and conclusions from Chapters 5, 6 and 7 are summarised. The applicability of the proposed methodology to improve the diagnosis of Alzheimer's disease dementia and its generalisability to other disease areas are considered. Opportunities for further research are discussed.

Chapter 2

Methodological background

2.1 Chapter overview

Chapter 2 introduces key statistical theories and methodologies that are used and further developed upon in the remaining chapters of this thesis. Section 2.2 describes the assumptions and advantages of Bayesian methodology, as well as discussing Bayesian statistical software for its implementation. Section 2.3 outlines the conduct of diagnostic test accuracy studies and Section 2.4 describes meta-analysis models for such studies.

2.2 Bayesian methodology

There are several features of Bayesian methodology that make it a natural approach for clinical decision-making (see Section 1.2.2); for this reason, the remainder of this thesis will focus on Bayesian evidence synthesis methods.

Bayesian methods offer a flexible approach to modelling complex clinical situations. A key feature of Bayesian methodology is the ability to make direct probability statements about parameters of interest, while capturing uncertainty around the model parameters and the data by treating both as random elements.[17] Bayesian models can be updated with new evidence via the prior distribution, making best use of all the available information and allowing for efficient evaluation of an intervention or test.[3]

In frequentist statistics, an unknown parameter θ is assumed to have a fixed true

value, e.g. θ_0, so all probability statements relate to observed random variables, e.g. data (y). Inference is therefore focussed on estimating the probability of data at specific parameter values, e.g. $p(y \mid \theta)$, known as the likelihood function. A 95% confidence interval can be estimated to quantify the uncertainty around an estimate and is interpreted as follows: in a series of confidence intervals produced by replicating the study a large number of times, 95% of these intervals will contain the true value.[42]

In Bayesian methodology, the reverse is true. Bayesian statistics are derived from Bayes' theorem, which states that for some unknown quantity of interest, θ:

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)} \tag{2.1}$$

where $p(\theta)$ is the prior distribution of θ and the conditional probability $p(\theta \mid y)$ is the posterior distribution of θ. As in frequentist likelihood-based models, the conditional probability $p(y \mid \theta)$ is the likelihood function. $p(y)$ is the normalising constant which ensures that the posterior distribution integrates to 1. As it is not usually necessary to calculate $p(y)$ in order to infer θ, Bayes theorem is often reduced to the following [3]:

$$\begin{aligned} &p(\theta \mid y) \propto p(y \mid \theta)p(\theta) \\ &e.g. Posterior \propto Likelihood \times Prior \end{aligned} \tag{2.2}$$

Bayesian methods allow the formulation of direct probability statements about posterior parameter estimates, conditional on both the observed data and prior knowledge or beliefs. Parameter estimates derived in a Bayesian framework are typically presented with a 95% credible interval (CrI), which quantifies the uncertainty around the posterior estimate. A 95% CrI indicates that there is a 95% probability that the true, unobserved parameter value lies within the interval given the observed data and prior evidence. The lower and upper limits of the interval can be estimated using the 2.5% and 97.5% percentiles of the posterior distribution,[16] known as an equal-tailed interval. Highest posterior density (HPD) intervals, estimated using the shortest interval with 95% probability of containing the true parameter value, are an alternative method of estimating the CrI. HDI intervals may be more meaningful when a posterior distribution is multimodal.[43]

Prior distributions can be based on external evidence from findings in previous stud-

ies or meta-analyses, or on expert opinion.[17] Minimally informative priors, where a prior distribution is chosen to represent a feasible range of values a parameter is expected to take while incorporating a large amount of uncertainty around where that value lies, can be chosen to allow data-driven estimation of posterior distributions.[44] It is assumed that the likelihood of observed data informs the posterior estimates of parameters, while the prior distributions have minimal impact. This assumption should be tested, however, as it has been demonstrated that choice of prior can have a large impact on posterior estimates, particularly variance parameters, despite intending to be non-informative.[44]

When the prior and likelihood distributions combine to result in a posterior distribution that belongs to the same family as the prior, it is possible to express the posterior distribution (Equation 2.1) in closed form.[16] These pairs of distributions, known as conjugates, allow the straightforward application of Bayes' theorem. When this is not the case, more complex computational methods are required to estimate the posterior.[16] These methods are described in Section 2.2.1 and 2.2.2 below.

2.2.1 Markov Chain Monte Carlo and Gibbs sampling

Markov Chain Monte Carlo (MCMC) is an efficient, iterative method for drawing random samples from a probability distribution, based on repeatedly drawing values of θ and then improving on those draws to better approximate the posterior distribution, $p(\theta \mid y)$.[45, 46, 47] The sampler starts at an initial guess for each of the parameters of interest. Sampling that follows is sequential, with subsequent values dependent on the previous value drawn, forming what is known as a Markov chain. Markov chains satisfy the property that the current value of the chain depends on the previous value only.[16] While MCMC methods were first proposed in the 1950s, they were not widely adopted until the development of MCMC simulation methods in the 1980s.

The Gibbs sampler is the simplest Markov chain simulation method.[48] At each iteration of the Gibbs sampler, a sample is drawn from the posterior distribution of each parameter in the model, conditional on the current values of the other parameters. These distributions, known as full conditional distributions, are typically easier to sample from than the joint distribution due to their univariate qualities. Initially, these samples may not be representative of the target posterior distribution; however, it can be shown that with sufficient iterations a chain of simulated

values will converge to the posterior distribution.[16] Simulations prior to convergence, known as 'burn-in' samples, are discarded. After convergence is reached, a large number of samples are taken from the joint posterior distribution, from which summary measures are calculated. There are several pieces of software for performing MCMC simulations using Gibbs sampling, including WinBUGS (Windows operating systems),[49] OpenBUGS (Linux and Windows operating systems),[50] and JAGS.[51]

While this 'random walk' behaviour can be efficient under certain circumstances, the Gibbs sampler can require long burn-in and sampling phases. Conditional sampling algorithms, such as Gibbs sampling, can be slow when parameters are highly correlated in the posterior distribution.[47] This is often true when jointly synthesising diagnostic accuracy data on multiple tests, which features complex dependence structures both between and within studies (see Section 2.4.2 and 2.4.3). Alternative sampling methods can offer improved performance through faster convergence and shorter sampling times.

2.2.2 Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC) is an alternative MCMC method to Gibbs sampling. HMC suppresses the random walk behaviour of the MCMC algorithm, allowing the sampler to move more rapidly through the posterior distribution. A momentum parameter is updated simultaneously with θ at each iteration, enabling rapid movement through the space of θ that preserves the 'energy' of the trajectory.[47] For more complex models, particularly those with highly correlated parameters, HMC has been found to be more efficient and stable than Gibbs sampling.[47] The No-U-Turn Sampler (NUTS) [52] was developed to implement HMC, and packaged into the Stan software.[53]

2.2.3 Stan

The Stan computer program [53] was developed to apply HMC given a Bayesian model, and the package *rstan* created as an R interface to Stan.[54] Due to the presence of highly correlated parameters within diagnostic accuracy data on two tests, all models within this thesis will be fit in a Bayesian framework using HMC simulation, implemented in Stan. Stan, unlike WinBUGS and OpenBUGS, also allows

user-defined functions and distributions, which are crucial to model development in Chapters 6 and 7.

2.2.3.1 Model specification

Stan models are organised into an ordered sequence of named blocks, in a way that differs from other Bayesian samplers whose layout is less stringent. The first section is the function block, which contains user-defined functions that will be called upon in other parts of the model. The data block defines the variables that will be read into the model as data, while the transformed data block allows their transformation prior to model fitting. Variables defined in the parameters block are the parameters that will be sampled by Stan; further variable derivation based on these parameters is specified in the transformed parameters block. Parameters can be constrained within a specified upper and/or lower bound (e.g. standard deviation > 0). Statements within the model block specify the priors and likelihood of the model. The generated quantities block derives additional quantities post-sampling based on the parameters or data, transformed or otherwise.

2.2.3.2 Model initialization

Stan, like other Bayesian samplers, requires the specification of a set of initial values, which act as a starting point for the NUTS sampler. Initial values can be specified for all unobserved parameters, and choosing initial values that are likely under the posterior distribution can speed up convergence. Where no initial values are provided, Stan will initialise unconstrained parameters with values drawn from a uniform distribution bounded between (-2, 2).

2.2.3.3 Divergent transitions

Divergent transitions are a common issue encountered when applying HMC. Divergent transitions arise when a region of the posterior distribution is difficult for the sampler to explore, often caused by a highly degree of curvature.[47] When a large number divergent transitions are present, or even a small number that are concentrated in a particular area of the distribution, sample estimates are unreliable and valid conclusions cannot be drawn from them.[53]

The number of divergent transitions may be reduced by decreasing the step size of

the sampler, causing the sampler to explore the posterior distribution more carefully (and by extension more slowly). In some cases, such as fitting a hierarchical model when data are sparse or parameters from different levels of the model are highly correlated in the posterior, this will not be sufficient and reparameterisation of the model will be required.

2.2.3.4 Non-centred parameterisation

Reparameterisation, a process in which a model is re-expressed in a mathematically equivalent form, can reduce computation times for hierarchical (e.g. meta-analysis) models, and, under certain scenarios, may improve convergence and chain mixing. Choice of parameterisation is an important consideration for any MCMC or HMC sampler, but is crucial when specifying a model in Stan due to its impact on the sampler performance.[53]

The non-centred parameterisation is a flexible method for re-expressing a hierarchical model in order to separate the dependence between its layers.[55] The more sparse the data, as is common in meta-analytic data sets where the number of studies may be small, or more complex the hierarchical model, the greater the need for reparameterisation. Where Stan indicates a large number of divergent transitions are present within the sample, the non-centred parameterisation can make the posterior curvature more manageable, reducing their number or removing them entirely.

Consider the random effects meta-analysis model for synthesising study-level data on a single outcome, e.g. treatment effect, described in Section 2.4 and used here to illustrate the centred and non-centred parameterisation. The observed treatment effects in study $i = 1, ..., I$, y_i, are assumed to be normally distributed around mean study-specific true treatment effects, μ_i, with within-study standard deviations σ_i. At the between-studies level, the observed relative treatment effects follow a normal distribution centred around summary treatment effect μ and between-studies standard deviation τ:

$$\begin{aligned} y_i &\sim Normal\left(\mu_i, \sigma_i^2\right) \\ \mu_i &\sim Normal\left(\mu, \tau^2\right) \end{aligned} \tag{2.3}$$

The data, $D_i = (y_i, \sigma_i)$, forms the bottom of the hierarchy structure and informs study-specific parameters μ_i. These in turn interact through a common dependency on the top level of the hierarchy, the summary parameters $\phi = (\mu, \tau)$. This interac-

tion allows the observed data to inform all μ_i, not just their immediate parent.[56] However, as the i elements at the bottom of the hierarchy are dependent on the summary parameters ϕ at the top (see Figure 2.1), a small change in the summary parameters results in a large change in the posterior density, particularly when data are sparse. This creates a 'funnel' shaped posterior density, with the sampler struggling to generate samples from the neck of the funnel.

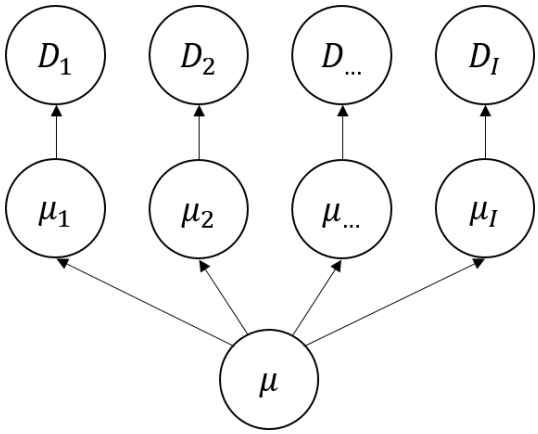


Figure 2.1: Structure of a one-level hierarchical model with a centred-parameterisation, adapted from a diagram by Betancourt and Girolami (2015).[56]

Dependencies between the layers of the hierarchical model can be reduced through a non-centred parameterisation, resulting in a posterior distribution that is easier for the sampler to explore. The between-studies level of the model in Equation 2.2.3.4 is re-expressed as:

$$\mu_i = \mu + \tau z_i$$

$$z_i \sim Normal(0, 1) \tag{2.4}$$

In Equation 2.4, the model layers are independent, conditional on z, and the sampler can more efficiently explore the posterior distribution. To demonstrate the increase in sampler efficiency gained through a non-centred parameterisation, data from 15 studies were simulated from the random effects meta-analysis model (described in Equation 2.2.3.4) assuming $\mu = 3$, $\tau = 3$ and $\sigma_i \sim Uniform(9, 10)$. Centred and non-centred parameterisations of this model were fit to the simulated data in Stan [53] and the values of $\log(\tau)$ and μ_1 for each parameterisation were plotted (see model code in Appendix A.1). Figure 2.2a and 2.2b show that Stan samples more efficiently from the posterior distribution 'funnel' neck using the non-centred parameterisation compared to the centred parameterisation.

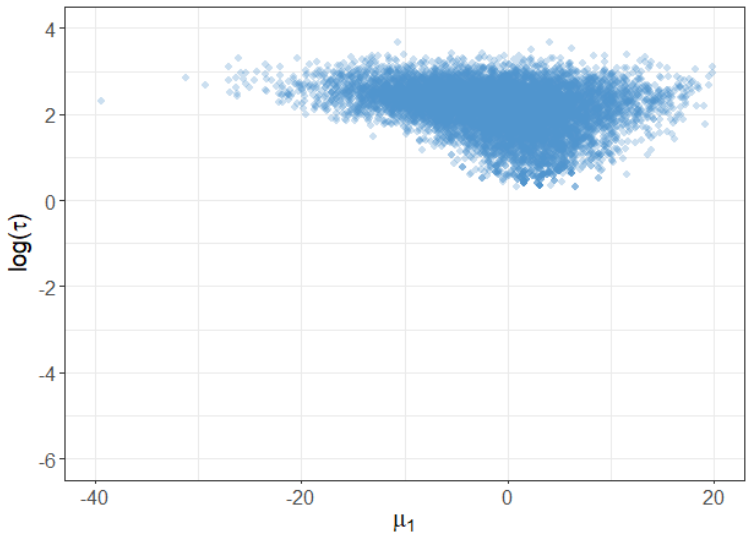


Figure 2.2a: Posterior distribution of centred parameterisation of the hierarchical normal model

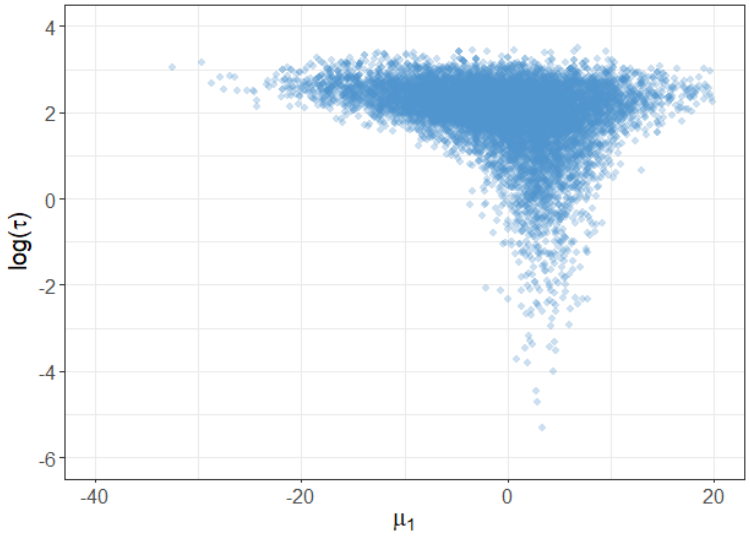


Figure 2.2b: Posterior distribution of non-centred parameterisation of the hierarchical normal model

2.2.4 Convergence diagnostics

A vital step in the Bayesian model fitting process is assessing the convergence of the parameters of interest to the target posterior distribution. This is done through the inspection of several diagnostic plots, which examine different properties of the Markov chains. The plots can only be used to detect non-convergence, and not as proof that convergence has been reached.[57] Highly complex models can require estimation of a large number of parameters, for which it may not be feasible to assess the convergence of each. In this case, the convergence of key parameters of interest should be evaluated, along with a random sample of the remaining parameters.[57] Running multiple Markov chains from different initial values enables additional convergence checks to be performed.

2.2.4.1 Trace plots

One such diagnostic plot is a trace plot (sometimes known as a history plot), which shows the value of a given parameter at each iteration of the sample. If multiple chains were ran, the plots can further be used to assess their mixing with one another. Where convergence has been reached, the plots should show random scatter around a stable mean, with chains that are similar in appearance and oscillate around similar point estimates. Visible trends in the trace plots may be indicative of slow convergence.[16]

2.2.4.2 Density plots

Density plots are smoothed histograms of the sample for a given parameter, used to examine its posterior distribution. The peak of the density plot indicates the mode of the distribution.[16] Unexpected shapes within the density plot, such as a bimodal distribution (characterised by two peaks) where a unimodal distribution is expected, may be a sign of poor convergence. Density plots for multiple chains can be overlaid to assess their mixing.

2.2.4.3 Autocorrelation plots

Due to the sequential sampling property of Markov chains, the MCMC sampler, whether Gibbs or HMC, will exhibit correlations between subsequent iterations. The

extent of this correlation is assessed through autocorrelation plots, which show the correlation between sampled values a specified number of iterations apart (known as the lag). As the lag increases, the correlation between samples should decrease; where this is not the case, slow convergence is indicated.[16]

2.2.5 Model fit and comparison

Model fit is assessed throughout this thesis by calculating the widely applicable information criterion (WAIC), also known as the Watanabe-Akaike information criterion.[58] The WAIC is a generalised form of the Akaike information criterion (AIC), and is preferable to the AIC or deviance information criterion (DIC) in a Bayesian setting.[59] The WAIC will be used for model comparison in later chapters, with a smaller WAIC indicating better model fit. The WAIC is estimated by computing the pointwise log-likelihood of the model, which is then penalised by the effective number of parameters in the model to adjust for overfitting.[59]

2.3 Diagnostic test accuracy

Diagnostic tests are used to identify the presence or absence of disease underlying a patient's signs and symptoms. Some tests are capable of measuring the extent of disease and providing information on prognosis, or may be used to screen for asymptomatic disease and aid in the monitoring and management of existing conditions. For a test to be used in clinical practice, it must first undergo evaluation to establish its accuracy and impact on clinical decision-making. Diagnostic accuracy is the ability of a test to correctly classify a person as with or without a target condition, and should be assessed within the context of its intended use, target population, and similar, existing tests for the same condition.

2.3.1 Phases of test evaluation

Diagnostic accuracy can be assessed through a variety of study designs, each of which addresses a different aspect of test performance. Test evaluation tends to run through a sequence of such studies, reflecting the increasing cost and resource use of each study design. Sackett and Haynes describe a framework for test evaluation formed of four phases.[60]

Phase I studies ask whether test results differ in those with the target condition compared to healthy participants. This is assessed through a case-control study, where a group of participants known to have the target condition (cases) and a group known not to have the target condition (controls) are recruited and the index test applied to each group. The ability of the test to distinguish between cases and controls is measured, for example by calculating the median and range of (continuous) test results in each group, or by comparing imaging test results between groups.

Phase II studies seek to establish whether patients with certain test results are more likely to have the target condition than patients with other test results. Also a case-control design, groups of participants known to have the target condition with differing degrees of severity are recruited, as well as healthy controls, and the index test applied to each group. At this phase a threshold for test positivity can be determined that best distinguishes between the groups. Case-control designs have been shown to inflate test accuracy estimates,[61] hence the need for more robust study designs further along the test evaluation process.

Phase III studies assess whether a test distinguishes between participants with and without the target condition in a clinically relevant population. A group of patients suspected to have the target condition are given both the index test and reference standard. The accuracy of the index test is compared to the reference standard, which is assumed to be perfectly accurate. This thesis focusses on studies of this design, as described in Section 2.3.2.

Phase IV studies measure the clinical effects of introducing a new test into practice, such as evaluating adverse events and the effect on clinical decision-making in terms of potential intervention. This is assessed using long-term follow up of randomised studies, where participants suspected to have the target condition are randomly assigned to either the index test or usual care.

Lijmer et al performed a systematic review of the phased test evaluation models proposed between 1978 and 2007,[62] including that of Sackett and Haynes,[60] concluding that evaluations are likely to be a repetitive, cyclic process rather than a linear one. While these models can be useful to classify different study designs and identify knowledge gaps, Lijmer et al argued that the reality of test evaluation is more complex and that a circular process of concurrent test development, assessment, and implementation is more appropriate.[62] Hovarth et al described a cyclical test evaluation framework outlining the pathway of a biomarker to becoming a clinically useful test.[63] The proposed test evaluation process consists of five components: analytical

and clinical performances, clinical and cost-effectiveness, and the broader impact of testing. The framework combines the five inter-related components in a dynamic, cyclical evaluation framework driven by the test's purpose and role within a specified clinical pathway, with the aim to increase collaboration between laboratory professionals, experts in evidence-based medicine, industry, policy-makers, and regulatory bodies.[63]

2.3.2 Diagnostic accuracy studies of a single test compared to a reference standard

The diagnostic accuracy of a test of interest, known as an index test, is evaluated through a diagnostic test accuracy study. The performance of the index test is compared to an existing test for the same condition assumed to perfectly classify participants as with or without the target condition, known as a reference standard. Results of such studies are typically reported in the form of four counts that measure the degree of misclassification between the index test and the reference standard, known as 2×2 data: the number of true positives (tp), false negatives (fn), true negatives (tn), and false positives (fp). The total number of participants who have and do not have the target condition are denoted as N^D and $N^{\bar{D}}$, respectively (Table 2.1).

Table 2.1: 2×2 data from a diagnostic accuracy study of a single test compared to a reference standard.

	Disease	**No-disease**
Test +	tp	fp
Test −	fn	tn
Total	N^D	$N^{\bar{D}}$

tp, true positives; fn, false negatives; tn, true negatives; fp, false positives; N^D, total diseased; $N^{\bar{D}}$, total non-diseased

2.3.3 Measures of test accuracy

Diagnostic accuracy can be quantified using several different measures calculated from the counts described in Section 2.3. Test accuracy is often summarised by a pair of parameters, which are estimated using distinct patient populations and represent two desirable qualities of a test: its ability to identify patients with the target

condition, and its ability to exclude patients without the target condition. These qualities are most commonly assessed by estimating sensitivity (se), the proportion of patients who have the target condition that have a positive test result, and specificity (sp), the proportion of patients who do not have the target condition that have a negative test result (Equation 2.5).[64, 6] Highly sensitive tests are unlikely to produce false negative results and can confidently rule out the target condition; on the other hand, highly specific tests produce few false positive results and can reliably rule in the target condition.[65]

$$\begin{aligned} se &= \frac{tp}{tp + fn} \\ sp &= \frac{tn}{tn + fp} \end{aligned} \tag{2.5}$$

An alternative approach for summarising test accuracy is using predictive values. A positive predictive value (ppv) is the proportion of patients with positive test results who have the target condition, while a negative predictive value (npv) is the proportion of patients with negative test results who do not have the target condition (Equation 2.6).[64] Predictive values are thought to have greater clinical utility than sensitivity and specificity, as when true disease status is unknown it can be more valuable to make probability statements concerning the known test result, but are sensitive to changes in the underlying prevalence of disease and therefore should be interpreted with caution.

$$\begin{aligned} ppv &= \frac{tp}{tp + fp} \\ npv &= \frac{tn}{tn + fn} \end{aligned} \tag{2.6}$$

Another summary of diagnostic accuracy is the likelihood ratio, which is the ratio of the proportion of patients with a specific test result who have the target condition to the proportion of patients who do not. Positive likelihood ratios (LR^+) summarise how many times more likely patients with disease are to have a positive test result compared to those without disease. Negative likelihood ratios (LR^-) summarise how many times less likely patients with disease are to have a negative test result compared to those without disease (Equation 2.7).[64] Unlike predictive values and, to an extent, sensitivity and specificity, likelihood ratios are not affected by the prevalence of the target condition. A likelihood ratio equal to 1 indicates that patients with and without the target condition are equally likely to receive a particular test result

(positive for LR^+ and negative for LR^-), meaning the result has no distinguishing ability.

$$LR^+ = \frac{se}{1 - sp}$$
$$LR^- = \frac{1 - se}{sp} \tag{2.7}$$

Alternatively, test accuracy can be summarised by a single measure, such as the diagnostic odds ratio (DOR) (Equation 2.8).[66] The DOR summarises the performance of a test using a single indicator ranging from zero to infinity, and has the advantage of allowing the formal ranking of tests, although the ability to distinguish between a test with high sensitivity and low specificity and vice versa is lost. Higher values of the DOR indicate a better discriminatory test performance. A value of 1 suggests that a test does not discriminate between patients with and without the target condition.

$$DOR = \frac{tp/fn}{fp/tn} = \frac{LR^+}{LR^-} \tag{2.8}$$

2.3.4 Diagnostic accuracy studies of two tests compared to a reference standard

To assess the clinical effectiveness of a novel diagnostic test, its performance should be compared to existing, relevant tests for the same condition.[12] Comparative diagnostic accuracy studies evaluate the accuracy of two tests against true disease status, inferred through a common reference standard, allowing direct comparison of their sensitivity and specificity in a single group of participants.[8] Comparative accuracy studies may use a paired design, in which all patients undergo both index tests plus a reference standard, or, less frequently, a randomised design, in which patients undergo the reference standard and are randomly assigned to one index test. The remainder of this thesis will focus on paired studies only, as associations between tests are not present within randomised studies.

Results of studies that utilise a paired design are most commonly reported as aggregate, 2×2 tables of the results of each test compared to the reference standard, containing the number of true positives (tp_j), false negatives (fn_j), true negatives (tn_j), and false positives (fp_j) for each of the $j = 1, 2$ tests. The number of patients with and without the target condition are denoted N^D and $N^{\bar{D}}$, respectively (see Table 2.2).

Table 2.2: 2×2 diagnostic accuracy data on two tests for a single study. All individuals undergo both tests plus a reference standard.

	Disease	**No disease**
Test 1 +	tp_1	fp_1
Test 1 −	fn_1	tn_1
Total	N^D	$N^{\bar{D}}$

	Disease	**No disease**
Test 2 +	tp_2	fp_2
Test 2 −	fn_2	tn_2
Total	N^D	$N^{\bar{D}}$

tp_j, true positives; fn_j, false negatives; tn_j, true negatives; fp_j, false positives; $j = 1, 2$ tests; N^D, total diseased; $N_{\bar{D}}$, total non-diseased

Less frequently, joint cross-classifications of test results compared to a common reference standard may be reported, producing a 2×4 table (Table 2.3). IPD for a study can be reconstructed using cross-classified data by recreating each participants' test results (dichotomised at their respective thresholds for positivity) and true disease status from the counts. The number of patients with each combination of test results with the target condition (x^D_{kl}) and without the target condition ($x^{\bar{D}}_{kl}$) are presented. $k, l = 0, 1$ denote the first and second test, respectively, where 0 indicates a negative test result and 1 indicates a positive test result. Cross-classified data allow inference on the agreement between tests, enabling estimation of within-study dependencies and evaluation of the accuracy of the tests in combination (see Section 2.3.5).

Table 2.3: Fully cross-classified diagnostic accuracy data for two tests for a single study. All individuals undergo both tests plus a reference standard.

	Disease			**No disease**		
	Test 2 +	**Test 2 −**	Total	**Test 2 +**	**Test 2 −**	Total
Test 1 +	x^D_{11}	x^D_{10}	tp_1	$x^{\bar{D}}_{11}$	$x^{\bar{D}}_{10}$	fp_1
Test 1 −	x^D_{01}	x^D_{00}	fn_1	$x^{\bar{D}}_{01}$	$x^{\bar{D}}_{00}$	tn_1
Total	tp_2	fn_2	N^D	fp_2	tn_2	$N^{\bar{D}}$

tp_j, true positives; fn_j, false negatives; tn_j, true negatives; fp_j, false positives; $j = 1, 2$ tests; x^D_{kl}, number of participants with the target condition with each combination of test results; $x^{\bar{D}}_{kl}$, number of participants without the target condition with each combination of test results; $k, l = 0, 1$, denote the first and second test, respectively, where 0 indicates a negative test result and 1 indicates a positive test result; N^D, total diseased; $N^{\bar{D}}$, total non-diseased

While cross-classified data are advantageous, they are often not reported in practice.[9, 10, 11, 4] Methods to obtain cross-classifications where they are not reported are accompanied by their own challenges and limitations. It may be possible to contact study authors for the additional data, however they may not respond or there may be ethical issues with sharing data, including identifiability

and permissions. Gabelica et al analysed the data availability statements of 3,416 articles in open-access journals, finding that 93% of authors that indicated their data were available upon request did not respond or declined to share their data.[67] Where IPD, equivalent to cross-classified data, are stored in electronic health records there may be issues associated with access and cost, with the records themselves subject to challenges including missing data, the potential for poor data quality, and the applicability of the data to the research question.[68] It may be possible to impute the missing data from studies in which full cross-classifications are reported, but the imputation method must be valid and reliable to avoid introducing bias.

2.3.5 Measures of comparative or combined test accuracy

Where only 2×2 data are reported, test accuracy estimates are limited to the individual (or marginal) sensitivity and specificity of each test (as in Equation 2.5), enabling their comparison but not evaluation of their performance in combination. Comparative accuracy measures such as the differences in sensitivity and specificity (Equation 2.9) or relative sensitivity and specificity (Equation 2.10) and their corresponding confidence or credible intervals can also be estimated from the marginal sensitivities and specificities.

Difference in sensitivity represents the change in the proportion of patients with the target condition detected by test 1 compared to test 2 (se_{diff}). Difference in specificity represents the change in the proportion of patients without the target condition detected by test 1 compared to test 2 (sp_{diff}).[4] Relative sensitivity (se_{rel}) is a ratio of two sensitivities; if relative sensitivity is 1 then the sensitivity of the two tests is the same (similarly for relative specificity, sp_{rel}).

$$se_{diff} = se_1 - se_2 \qquad sp_{diff} = sp_1 - sp_2 \tag{2.9}$$

$$se_{rel} = \frac{se_1}{se_2} \qquad sp_{rel} = \frac{sp_1}{sp_2} \tag{2.10}$$

Where full cross-classifications are reported, we can estimate additional measures that capture the joint diagnostic accuracy of the two tests. Joint accuracy measures, such as joint sensitivity and joint specificity, measure the concordance between the tests (Equation 2.11). Joint sensitivity (se_{joint}) is the proportion of patients who have the target condition that have a positive result on both tests, while joint specificity (sp_{joint}) is the proportion of patients who do not have the target condition that have

a negative result on both tests.

$$se_{joint} = \frac{x_{11}^{D}}{N^{D}} \qquad sp_{joint} = \frac{x_{00}^{\bar{D}}}{N^{\bar{D}}} \tag{2.11}$$

Using the joint accuracy measures, it is possible to derive summary estimates of test accuracy for test combinations (Equation 2.12). Tests may be combined under an 'AND' rule, also known as a 'both positive' rule, where both component tests must be positive for the final result to be positive. The 'AND' rule increases the specificity of the combined tests compared to the marginal specificities of each test considered in isolation, and is analogous to joint sensitivity. Alternatively, tests may be combined under an 'OR' rule, also known as an 'either positive' rule, where the final result is positive if either of the two component tests are positive. The 'OR' rule increases the sensitivity of the combined tests compared to the marginal sensitivities of each test considered in isolation.[69]

$$AND = se_{joint} = \frac{x_{11}^{D}}{N^{D}} \qquad OR = \frac{x_{11}^{D} + x_{10}^{D} + x_{01}^{D}}{N^{D}} \tag{2.12}$$

Where full cross-classifications are not reported, and only 2×2 data on each test is available, joint accuracy measures may be estimated by assuming independence between the results of the first and second test (Equation 2.13). Ignoring within-study associations between tests may introduce bias to the estimates of joint accuracy.[15]

$$se_{joint} = se_1 \times se_2 \qquad sp_{joint} = se_1 \times se_2 \tag{2.13}$$

2.4 Meta-analysis

Meta-analysis is a method of integrating the findings of two or more studies that answer the same research question by calculating a quantitative combination of their results, typically based on a weighted average.[5] It is a two-stage process. In the first, a parameter of interest is calculated for each study using its raw data, such as a treatment effect or test accuracy measure. In the second stage, a summary estimate of the parameter of interest is produced as a weighted average of the study-specific effects. Fixed effect meta-analysis assumes that each study is estimating the same parameter of interest. Random effects meta-analysis, assumes that studies are not estimating the same parameter but rather parameters that follow a distribution

across studies, can be performed (see Section 2.4.1). The latter allows for variation between studies, known as heterogeneity, due to differences in patient and study characteristics.

2.4.1 Meta-analysis of a single outcome

2.4.1.1 Continuous outcome

The majority of meta-analysis models in interventional research synthesise data on a single outcome. For example, a study that evaluates the efficacy of a treatment may report a mean difference, representing the difference in a continuous outcome of interest between treatment and control groups. Consider a meta-analysis of $i = 1, ..., I$ studies that evaluate such an outcome. Under random effects, it is assumed that the true treatment effect varies between studies, but according to a pre-specified distribution (e.g. normal distribution). The observed treatment effects in study i, y_i, are assumed to follow a normal distribution with study-specific mean true treatment effect, μ_i, with within-study variance σ_i^2.[70] At the between-studies level, the true treatment effects follow a normal distribution centred around summary treatment effect μ and between-studies variance τ^2:

$$y_i \sim Normal\left(\mu_i, \sigma_i^2\right)$$

$$\mu_i \sim Normal\left(\mu, \tau^2\right) \tag{2.14}$$

In a Bayesian framework, prior distributions are placed on the unknown parameters in the model: namely, μ and τ^2. In the absence of external evidence, minimally informative prior distributions may be specified. For a continuous treatment effect, a normal distribution centred at 0 (no effect) with a large variance relative to the scale of the outcome of interest may be used, such as $\mu \sim Normal\left(0, 10^2\right)$. Variance τ^2 is restricted to positive values and a prior distribution must be chosen that reflects this. A uniform distribution, (e.g. $\tau^2 \sim Uniform\left(0, 2\right)$) or a half-normal distribution (e.g. $\tau^2 \sim HalfNormal\left(2^2\right)$) may be suitable choices. As discussed in Section 2.2, minimally informative prior distributions may have a large impact on the posterior distribution and should be explored through sensitivity analyses.[44]

Under fixed effects, the true treatment effect, μ, is assumed to be the same across studies (analogous to τ =0 in a random effects model). The observed treatment

effects, y_i, are assumed to be normally distributed with a common summary true treatment effect μ and within-study standard deviations σ_i.[71]

$$y_i \sim Normal\left(\mu, \sigma_i^2\right) \tag{2.15}$$

In a Bayesian setting, a prior distribution must be provided for μ. Similar to the random effects model, a normal distribution with large variance that spans the range of plausible values for the outcome of interest, such as $\mu \sim Normal\left(0, 10^2\right)$, may be specified.

2.4.1.2 Dichotomous outcome

Where the data are dichotomous, studies may instead report odds ratios, representing the odds of an outcome of interest in the treatment group to the odds of that outcome in the control group. In this case, it is common to transform observed treatment effects to work on the log odds ratio scale. DerSimonian and Laird proposed that the logs odds ratios can be assumed to follow a normal distribution, similar to Equation 2.14.[72] However, it has been demonstrated that using a normal approximation can produce biased summary estimates, particularly when the outcome of interest is rare or data are sparse,[73] and lead to computational issues when there are no events in one or more study arms. An alternative approach is to assume that the number of events in the control arm, $r_{A,i}$, and the treatment arm, $r_{B,i}$, in the i^{th} study follow independent binomial distributions:

$$r_{i,A} \sim Binomial\left(p_{i,A}, N_{i,A}\right), \;\; r_{i,B} \sim Binomial\left(p_{i,B}, N_{i,B}\right), \tag{2.16}$$

where $p_{i,A}$ and $p_{i,B}$ are the study-specific true probabilities of an event and $N_{i,A}$ and $N_{i,B}$ are the number of patients in the control and treatment arms, respectively.[73] $p_{i,A}$ and $p_{i,B}$ are *logit*-transformed to calculate the baseline and treatment effects (Equation 2.17). The *logit*-transformation maps $p_{i,A}$ and $p_{i,B}$, which are bound between $[0, 1]$, to log-odds, which are real numbers in $(-\infty, \infty)$. This transformation overcomes the restricted range of probabilities. At the between-studies level, the observed relative treatment effects δ_i are modelled using a normal distribution, similar to Equation 2.14:

$$logit\left(p_{i,A}\right) = \mu_i, \;\; logit\left(p_{i,B}\right) = \mu_i + \delta_i$$

$$\delta_i \sim Normal\left(d, \tau^2\right) \tag{2.17}$$

To fit the model in a Bayesian framework, prior distributions must be placed on μ_i (for each of the i studies), d, and τ^2. Log-transformed data can be assumed to be approximately normally distributed; therefore, in the absence of external information, a minimally informative normal distribution with large variance may be specified, e.g. $\mu_i \sim Normal\,(0, 10^2)$ and $d \sim Normal\,(0, 10^2)$.[16] For between-study variance τ^2, a restricted distribution such as a uniform ($\tau^2 \sim Uniform\,(0, 2)$) or half-normal (e.g. $\tau^2 \sim HalfNormal\,(2^2)$) distribution ensures only positive values are sampled.

2.4.1.3 Heterogeneity

While random effects incorporate between-studies heterogeneity in the true treatment effects across studies, they do not account for all sources of variability. Random effects meta-analysis models make distributional assumptions about the effects being estimated in the different studies, addressing heterogeneity that cannot be explained by other, observed factors.[5] It is important to explore heterogeneity in the study pool through techniques such as subgroup analyses, where subsets of studies or patients are split into subgroups to aid their comparison, or meta-regression, where the effect of study or patient characteristics on the outcome are investigated through their incorporation as covariates in the model.[4]

2.4.2 Meta-analysis of a single diagnostic test compared to a reference standard

Meta-analysis of diagnostic accuracy studies differs from a typical interventional meta-analysis due to the presence of two, correlated outcomes within diagnostic accuracy data, e.g. sensitivity and specificity. Between-studies correlation arises between sensitivity and specificity, driven by heterogeneity in both parameters across studies. Test accuracy may be underestimated if such correlation is not adequately accounted for.[7] Univariate pooling of sensitivity and specificity using standard methods for meta-analysing proportions is not usually appropriate, except when there are few studies or sparse data.[74]

2.4.2.1 Bivariate random effects model

Reitsma et al proposed a bivariate meta-analysis approach to summarise the accuracy of a single test at a common threshold, assuming *logit*-transformed sensitivity and

specificity are normally distributed around a mean value.[75] Chu and Cole suggested that using exact binomial likelihoods are a more natural choice to model within-study variability in sensitivity and specificity,[76] a correction that is widely accepted and is referred to throughout this thesis as the bivariate random effects meta-analysis (BRMA) model.

At the within-study level, the number of true positive (tp_i) and true negative (tn_i) results for study $i = 1, ..., I$ are assumed to follow independent binomial distributions from a sample of disease positive and disease negative individuals, respectively. These counts are modelled independently as the populations of diseased and non-diseased are mutually exclusive, i.e. an individual cannot contribute to both the estimation of sensitivity and specificity:

$$tp_i \sim Binomial\left(se_i, N_i^D\right) \qquad tn_i \sim Binomial\left(sp_i, N_i^{\bar{D}}\right) \tag{2.18}$$

where se_i and sp_i denote the sensitivity and specificity in the i^{th} study, respectively, and N_i^D and $N_i^{\bar{D}}$ the number of patients with and without the target condition, respectively. At the between-studies level, *logit*-transformed study-specific sensitivities ($\mu_{i,se}$) and specificities ($\mu_{i,sp}$) are jointly modelled using a bivariate normal distribution centred around *logit*-transformed summary sensitivity and specificity, μ_{se} and μ_{sp}, accounting for between-studies correlation arising between sensitivity and specificity due to differences in study characteristics. Between-studies variances are denoted σ_{se}^2 and σ_{sp}^2, respectively, while ρ_b represents the between-studies correlation parameter.

$$logit\left(se_i\right) = \mu_{i,se}, \;\; logit\left(sp_i\right) = \mu_{i,sp}$$

$$\begin{pmatrix} \mu_{i,se} \\ \mu_{i,sp} \end{pmatrix} \sim Normal\left(\begin{pmatrix} \mu_{se} \\ \mu_{sp} \end{pmatrix}, \begin{pmatrix} \sigma_{se}^2 & \rho_b\sigma_{se}\sigma_{sp} \\ \rho_b\sigma_{se}\sigma_{sp} & \sigma_{sp}^2 \end{pmatrix} \right) \tag{2.19}$$

Prior distributions are placed on the unknown parameters in the model, namely μ_{se}, μ_{sp}, σ_{se}^2, σ_{sp}^2, and ρ_b. A normal distribution centred at 0 with a large variance is a suitable minimally informative prior for $\mu_{se}, \mu_{sp} \sim Normal\left(0, 10^2\right)$. For the between-studies variance parameters, a restricted distribution such as a uniform (e.g. $\sigma_{se}^2, \sigma_{sp}^2 \sim Uniform\left(0, 2\right)$) or half-normal (e.g. $\sigma_{se}^2, \sigma_{sp}^2 \sim HalfNormal\left(2^2\right)$) distribution is an appropriate choice. The between-studies correlation parameter is bound between $[-1, 1]$ and a prior must be chosen to reflect this. Examples of suitable priors include: a uniform distribution ($\rho \sim Uniform\left(0, 1\right)$) which assign

equal weighting to all possible values, positive or negative, of correlation; a transformed beta distribution (e.g. $\frac{\rho+1}{2} \sim Beta\left(1.5, 1.5\right)$), which is relatively flat across the range of values with exception of the extreme values (close to -1 or 1), which are highly unlikely; or the Fisher z-transformation ($\rho = \tanh\left(z\right), z \sim Normal\left(0, 1\right)$), which produces an approximately normal distribution with appropriate bounds. As discussed in Section 2.2, variance parameters are particularly sensitive to choice of prior and their impact upon the posterior should be explore through sensitivity analyses. Stan code for the BRMA model implemented in a Bayesian framework is presented in Appendix A.2.

2.4.2.2 Hierarchical summary receiver operating characteristic model

Another widely accepted method of meta-analysing diagnostic accuracy studies is the hierarchical summary receiver operating characteristic (HSROC) model, described by Rutter and Gatsonis.[77] A fixed effects summary receiver operating characteristic (SROC) model was first proposed by Moses et al, accounting for variation due to threshold for test positivity and chance only.[78, 79] The HSROC model extends the SROC model, accounting for both within-study and between-studies variability in sensitivity and specificity.[77]

For each of the $i = 1, ..., I$ studies, the number of patients testing positive in each of the diseased and non-diseased groups are denoted y_{i1} and y_{i0}, respectively. The number testing positive in each of the $j = 0, 1$ disease groups are assumed to follow independent binomial distributions from a sample of patients with (N_{i1}) and without (N_{i0}) the target condition individuals, respectively (Equation 2.20). π_{ij} denotes the probability of a positive test result in the diseased and non-diseased groups, representing the true positive and true negative rate in the i^{th} study, respectively.

$$y_{ij} \sim Binomial\left(\pi_{ij}, N_{ij}\right) \tag{2.20}$$

At the within-study level, the HSROC model takes the form:

$$logit\left(\pi_{ij}\right) = \left(\vartheta_i + \alpha_i X_{ij}\right) e^{-\beta X_{ij}} \tag{2.21}$$

X_{ij} is a dummy variable for true disease status, coded as 0.5 for the diseased group and -0.5 for the non-diseased group in the i^{th} study. ϑ_i is a threshold parameter and α_i is a measure of diagnostic accuracy that incorporates sensitivity and specificity

for the i^{th} study.[4] β is a shape or scale parameter that incorporates asymmetry in the SROC curve by allowing diagnostic accuracy to vary with threshold. ϑ_i and α_i are allowed to vary between studies and modelled as random effects; β is modelled as a fixed effect.

At the between-studies level, ϑ_i and α_i are modelled using independent normal distributions with means Θ and A and standard deviations σ_ϑ and σ_α, respectively:

$$\vartheta_i \sim N\left(\Theta, \sigma_\vartheta^2\right) \tag{2.22}$$

$$\alpha_i \sim N\left(A, \sigma_\alpha^2\right) \tag{2.23}$$

$$\tag{2.24}$$

The HSROC model produces an SROC curve summarising sensitivity and specificity across a number of different thresholds for test positivity. Using a range of values for 1 - specificity (false positive rate, FPR), corresponding average values for sensitivity (true positive rate, TPR) are calculated from the estimated average location (A) and scale parameter (β) [4]:

$$TPR = \frac{1}{1 + exp\left(-\left(Ae^{-0.5\beta} + logit\left(FPR\right)e^{-\beta}\right)\right)} \tag{2.25}$$

In a Bayesian setting, prior distributions are specified for parameters Θ, A, β, σ_α^2, and σ_ϑ^2. In the absence of prior beliefs, a minimally informative normal distribution centred around 0 for threshold parameter $\Theta \sim Normal\left(0, 10^2\right)$, accuracy parameter $A \sim Normal\left(0, 10^2\right)$, and slope parameter $\beta \sim Normal\left(0, 10^2\right)$ are suitable choices. For between-study variances, a restricted distribution such as a uniform ($\sigma_\alpha^2, \sigma_\vartheta^2 \sim Uniform\left(0, 2\right)$) or half-normal (e.g. $\sigma_\alpha^2, \sigma_\vartheta^2 \sim HalfNormal\left(2^2\right)$) distribution ensures only positive values are sampled.

2.4.2.3 Choice of model

While the BRMA and HSROC models differ in their parametrisations, they have been shown to be mathematically equivalent in the absence of study-level covariates.[80] Hierarchical models have been shown to outperform univariate meta-analysis methods for estimating pooled sensitivity and specificity, due to a 'borrowing of strength' across endpoints arising from the inclusion of between-studies correlation.[81, 82, 83]

The bivariate parametrisation models sensitivity, specificity, and the correlation between them directly, leading to summary point estimates (e.g. summary sensitivity and specificity) plus corresponding confidence intervals at a common threshold. The HSROC parameterisation, however, models functions of sensitivity and specificity, naturally giving rise to an SROC curve summarising the trade off between the two measures across different cut-off points.[4] The latter may be more appropriate where threshold varies across studies. As such, while the two models are equivalent under certain circumstances, choice of parametrisation is driven by the focus of inference (summary points or SROC curves), data availability (single threshold or multiple thresholds), and whether covariates will be included in the model to explore heterogeneity.[84] Neither the BRMA or HSROC approach explicitly incorporates information on varying thresholds for test positivity across studies.

As the majority of diagnostic meta-analysis models for evaluating two tests are extensions to the BRMA rather than the HSROC parameterisation (see Chapter 3), the remainder of this thesis will focus on evaluating and building upon the former.

2.4.3 Meta-analysis of two diagnostic tests

Systematic reviews comparing the accuracy of two (or more) diagnostic tests are of greater relevance to healthcare decision-makers than reviews assessing a single test.[84] Such reviews can adopt two approaches. In the first, analysis is restricted to studies that have evaluated both tests in the same group of participants. In the second, all studies that evaluated one or both tests are included in the analysis.[4] The latter approach may introduce bias to the study pool where test comparisons are potentially based on indirect evidence.[8]

Additional correlation structures are present when synthesising data on two diagnostic tests in a meta-analysis framework. Comparative diagnostic accuracy studies, unlike parallel studies comparing the efficacy of multiple interventions, often evaluate both index tests in the same patients using a paired design. If a study estimates the sensitivity and specificity, se_j and sp_j, of $j = 1, 2$ tests in the same individuals, then within-study dependence arises between se_1 and se_2, and between sp_1 and sp_2.[15] The impact of account for within-study associations on test accuracy parameters is, so far, unknown.

Current guidelines recommend a meta-regression approach to compare the accuracy of two or more diagnostic tests, where test type is included as a covariate in either

the BRMA or HSROC model.[4] However, this strategy does not account for within-study dependencies between tests. A number of novel models have recently been proposed to synthesise data on multiple diagnostic tests. Reimbursement agencies such as the NICE Decision Support Unit have highlighted that further evaluation of these methods is required before they are adopted into routine use.[15]

2.5 Chapter summary

This chapter has introduced key statistical theories and methodologies that are used and further developed throughout this thesis. The following chapter builds upon this theoretical knowledge, describing a methodological review undertaken to identify models for synthesising evidence on two or more diagnostic tests. Through the critical appraisal of existing methodologies, motivation for novel model development in Chapters 6 and 7 is established.

Chapter 3

Methodological review of meta-analysis models for evaluating two diagnostic tests

3.1 Chapter overview

Chapter 3 presents a methodological review of models for jointly synthesising data on two or more diagnostic tests. This chapter serves to summarise and critically appraise the existing literature base, providing motivation for methodological development undertaken in later chapters. Section 3.2 describes the methods used to identify articles describing such methods, including the search strategy and selection process. Section 3.3 presents the results of the methodological review, highlighting the key features, strengths and weaknesses of the existing models. Whether and how each methods accounts for within-study associations, present when more than one test is evaluated in the same patient group through a paired diagnostic accuracy study, is considered. Section 3.4 concludes the chapter with a discussion, including a summary of the limitations of the methods that Chapters 6 and 7 aim to address.

3.2 Methods

A methodological review was conducted to identify meta-analysis models to jointly synthesise data on two or more diagnostic tests.

3.2.1 Literature search

PubMed was searched for relevant articles on the 20th June 2023, combining terms for 'meta-analysis', 'diagnostic test' and 'comparative accuracy' in any field. Forward citation searching was used to identify articles that cited key papers. Database searching was supplemented through internet searches, using keywords such as 'test comparison', 'multiple tests' and 'meta-analysis' as a source of potentially relevant papers.

3.2.2 Inclusion criteria

Inclusion was restricted to articles published from 1st January 2014 onwards, the year that the first joint meta-analysis model for two diagnostic tests compared in the same participants using a paired design was published.[85] Meta-analysis methods for comparing diagnostic test accuracy published up to and including July 2014 have been summarised previously.[9] Articles that proposed a novel method for meta-analysis of two or more diagnostic tests were included.

3.2.3 Study selection

Titles and abstracts identified through electronic database and web searching were uploaded to Rayyan and screened by one reviewer to identify potentially relevant studies.[86] Papers were considered potentially relevant where they appeared to propose a novel meta-analysis model for evaluating diagnostic test accuracy. Following this initial assessment, full-text articles were obtained and assessed by one reviewer for inclusion in the review.

3.3 Results

A total of 1,309 articles published between January 1986 and June 2023 were identified through PubMed, most of which described applied systematic reviews and meta-analyses comparing the accuracy of multiple diagnostic tests. Figure 3.1 shows the number of PubMed articles meeting the search criteria published each year, demonstrating proliferating interest in performing comparative accuracy reviews. This is further supported by a study by Veroniki et al, which analysed applied network

meta-analyses of diagnostic test accuracy studies, finding a substantial increase in publications since 2010.[87]

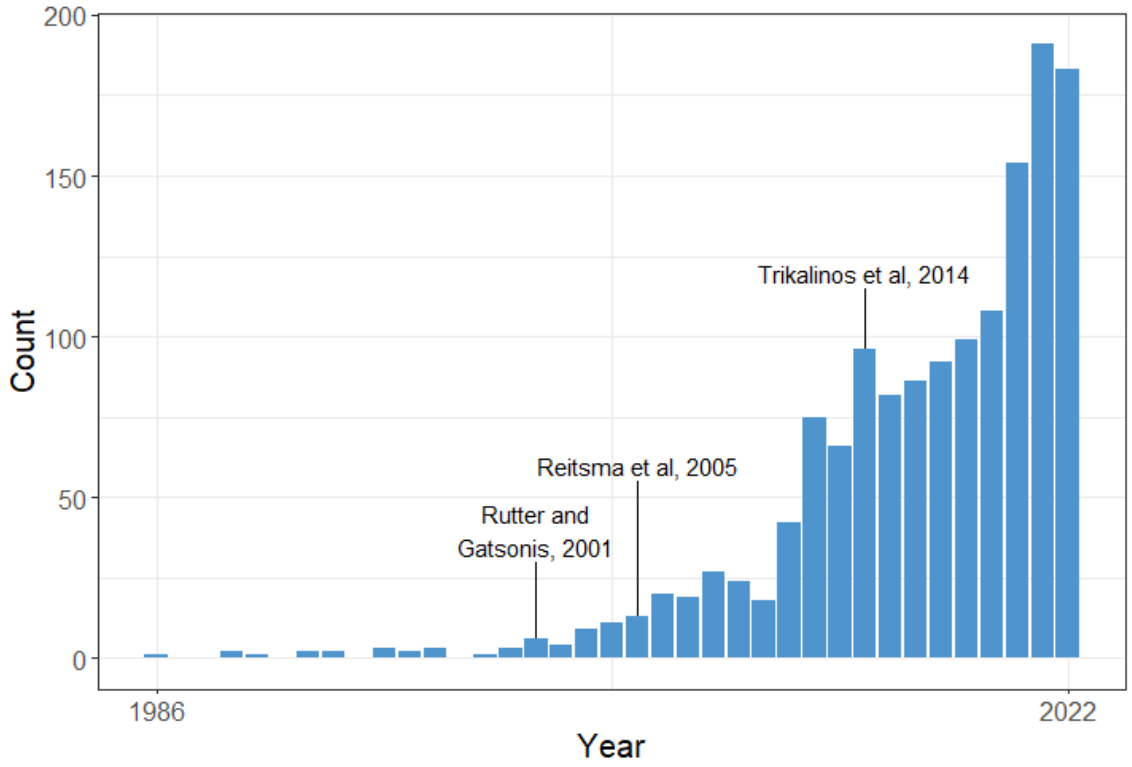


Figure 3.1: Number of PubMed articles containing the phrases 'meta-analysis', 'diagnostic test', and 'comparative accuracy' published per year. The publication of key methodological developments, discussed in detail within this chapter, are highlighted.[77, 75, 85]

After restricting search results to articles published from 1st January 2014, 1,024 potentially relevant articles were identified and underwent screening. Of these, a total of 12 relevant articles that met the inclusion criteria were identified.[85, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98] Within these papers, 15 methods for synthesising diagnostic accuracy data on two tests were described. Table 3.1, adapted from a similar table in Owen et al,[95] summarises the key features of each method.

Table 3.1: Summary of statistical meta-analysis models for evaluating the accuracy of two diagnostic tests, adapted from Owen et al [95]

Model type	Model description	Data required	Multiple thresholds per test	Software
Trikalinos et al, 2014.[85]				
Arm-based	Multinomial distributions jointly model the cross-classified results of the two tests. A six-dimensional normal distribution accounts for between-studies correlations.	Fully cross-classified data	No	JAGS 3.1.0 implemented through R package *rjags*. R code available in supplementary material.
Menten and Lesaffre, 2015.[88]				
Contrast-based	Hierarchical latent class model. Incorporates direct and indirect evidence on test accuracy and adjusts for the use of imperfect reference standards.	Fully cross-classified data	No	OpenBUGS 3.0.3 called from within R 3.0.1 using BRugs library. OpenBUGS code available in supplementary material.
Nyaga et al, 2016a.[89]				
Arm-based	Two-stage hierarchical model analogous to single factor ANOVA method with repeated measures. Shared random effects induce study-level correlations. Model allows borrowing of information where single-arm studies are included.	2×2 data for each test	No	Stan programming language implemented within R 3.3.0 using *rstan* 2.9.0 package. Stan code provided in supplementary material.

Model type	Model description	Data required	Multiple thresholds per test	Software
Nyaga et al, 2016b.[90]				
Arm-based	One-stage approach models directly on the probability scale (without-*logit* transformation), using marginal beta-binomial distributions for sensitivity and specificity linked by a copula density at the between-studies level.	2×2 data for each test	No	Stan implemented within R 3.3.1 using *rstan* 2.11.1 package.
Dimou et al, 2016.[91]				
Arm-based	Four-dimensional normal distribution with closed-form expression for the within-study covariance matrix (requires full cross-classifications). When only 2×2 data is available, full reporting used to impute the correlation between tests.	Fully cross-classified or 2×2 data for each test	No	Stata, illustrative code given in supplementary material.

Model type	Model description	Data required	Multiple thresholds per test	Software
Cheng, 2016.[92]				
Arm-based	All proposed models are capable of synthesising single-arm studies and accounting for partial cross-classifications. a. Multinomial model with decomposition of test- and study-specific effects. b. Multivariate extension of the HSROC model. c. Beta-binomial marginals are linked by multivariate Gaussian copulas at the between-studies level.	Fully cross-classified or 2×2 data for each test	a. No b. Yes c. No	JAGS called from R through package *R2jags*.
Ma et al, 2018.[93]				
Arm-based	Hierarchical model incorporating single-arm and randomised studies, studies with and without a gold standard, studies that did not evaluate all tests, and disease prevalence.	Fully cross-classified data	No	JAGS software via *rjags* package in R.

Model type	Model description	Data required	Multiple thresholds per test	Software
Hoyer and Kuss, 2018a.[94]				
Arm-based	Four-dimensional Gaussian and vine copula models. Allows for flexible modelling of correlation structures, while avoiding complex numerical approximation required by incorporating random effects.	2×2 data for each test	No	SAS PROC NLMIXED procedure with default options.
Owen et al, 2018.[95]				
Arm-based	Extension of the BRMA model, incorporating constraints on increasing test threshold while accounting for within-study correlations between sensitivities and specificities using random effects.	2×2 data for each test	Yes	WinBUGS 1.4.3 software. Example code given in supplementary material.
Hoyer and Kuss, 2018b.[96]				
Arm-based	Generalised linear mixed model, similar to Trikalinos et al.[85] A four-dimensional normal distribution models marginal sensitivities and specificities. Does not account for within-study dependencies.	2×2 data for each test	Yes	GLIMMIX procedure using SAS 9.3, available in supplementary material.

Model type	Model description	Data required	Multiple thresholds per test	Software
Lian et al, 2019.[97]				
Arm-based	Extension of HSROC model, incorporating single-arm and randomised studies, studies with and without a gold standard, studies that did not evaluate all tests, and disease prevalence.	Fully cross-classified data	Yes	JAGS 4.2.0 via *rjags* package in R 3.3.2.
Nikoloulopoulos, 2019.[98]				
Arm-based	Extension to generalised linear mixed model propsed by Hoyer and Kuss.[96] A vine copula representation of the random effects allows for between-studies dependencies.	2×2 data for each test	No	R functions to implement D-vine copula mixed model are part of the R package *CopulaREMADA*.

ANOVA, analysis of variance; BRMA, bivariate random effects meta-analysis; HSROC, hierarchical receiver operating characteristic

3.3.1 Trikalinos et al 2014

Trikalinos et al proposed a joint model for synthesising data on two diagnostic tests evaluated against a common reference standard under a paired design.[85] The arm-based model parametrises test accuracy in terms of the absolute sensitivity and specificity of each test.[88] At the within-study level, four-dimensional multinomial distributions jointly model cross-classified data for the diseased and non-diseased groups, capturing within-study associations between two tests assessed using a paired design. At the between-studies level, a six-dimensional normal distribution accounts for between-studies correlations. By utilising fully-cross classified data for each study, the model is able to directly estimate both the marginal and joint sensitivities and specificities while incorporating within-study dependencies. The model is implemented in a Bayesian framework using JAGS via the R package *rjags*.

While the authors assert that the incorporation of randomised or non-comparative studies is possible through minor modifications to the model, these adaptations are not explicitly described. It is also stated that the model can be extended to evaluate three or more tests in theory, although this is not demonstrated in the paper. With each additional test, the number of model parameter will increase rapidly, however, which may lead to convergence issues. Indeed, fitting this model to simulated data on the diagnostic accuracy of two diagnostic tests compared to a common reference standard using a paired study design was explored as part of the development of this thesis. The model was found to be highly computationally intensive, even for this relatively simple example, and convergence issues were common. This is further supported in the existing literature.[15]

3.3.2 Menten and Lesaffre 2015

where (*logit*-transformed) sensitivities and specificities are modelled relative to a baseline or comparator test.[95] Similar to network meta-analysis, both the

Menten and Lesaffre described a Bayesian approach to evaluating two or more diagnostic tests.[88] Their contrast-based method models the relative accuracy of two tests compared to a baseline test through their head-to-head comparison. Similar to network meta-analysis, both the direct comparisons, evaluated using comparative diagnostic accuracy studies, and indirect comparisons, assessed through a common diagnostic test, are modelled.[95] The model utilises latent class analysis to adjust

for the use of imperfect reference tests. Within this framework, the true disease status of study participants is treated as an unobserved, or latent, variable with two mutually exclusive categories: those with the target condition, and those without. The model also incorporates external information on the diagnostic accuracy of the reference standard through the prior distribution. The posterior distribution is estimated through an MCMC sampler via OpenBUGS.

The model requires cross-classified data from each study to inform the latent class analysis, but does not account for within-study associations present between two tests evaluated under a paired design. The sensitivities and specificities of three or more diagnostic tests, including those of the reference standard(s), may be modelled through separate bivariate normal distributions.

3.3.3 Nyaga et al 2016a

Nyaga et al proposed a two-way analysis of variance (ANOVA) method for synthesising data on multiple diagnostic tests for the same condition.[89] At the within-study level of their arm-based model, the number of true positive and true negative results are modelled using independent binomial distributions. At the between-studies level, random effects of study and fixed effects of test induce between-studies correlation between sensitivity and specificity. Within-study associations between two tests assessed in the same study participants using a paired design are not accounted for. The model is implemented in a Bayesian framework in Stan via the *rstan* package.

The model requires 2×2 data for each study only. The diagnostic accuracy of three or more index tests can be modelled without creating high computational demands. Using this model, it is possible to include studies that evaluated one, some, or all of the index tests. Missing test results within studies are assumed to be missing at random and tests must form a connected network, i.e. there must exist at least one study evaluating a test of interest together with at least one of the remaining index tests.

3.3.4 Nyaga et al 2016b

Nyaga et al proposed a second Bayesian approach for the meta-analysis of two or more diagnostic tests, utilising a copula approach for capturing associations present across studies.[90] In their arm-based method, independent binomial likelihoods

model the number of true positives and true negatives for each test at the within-study level. At the between-studies level, a bivariate copula captures the associations between the sensitivities and specificities of the tests. The authors consider one type of bivariate copula only: the Frank copula (see Section 6.4.1 for a description of copula families). The model is implemented in Stan through the *rstan* package.

Two-by-two data comparing the diagnostic accuracy of each test of interest to a common reference standard is required to fit the model, and within-study associations present between tests assessed using a paired design are not accounted for. The diagnostic accuracy of more than two index tests can be evaluated using the method, as is demonstrated in the paper through a motivating example comparing the accuracy of 11 index tests. It is also possible to include studies that evaluated one, some, or all of the index tests in the analysis.

3.3.5 Dimou et al 2016

Dimou et al outlined an arm-based frequentist approach for modelling the diagnostic accuracy of two tests for the same condition.[91] In their paper, the authors describe using a multivariate normal distribution to model *logit*-transformed true positive rates and true negative rates of the two tests. By assuming that data are missing at random, studies that evaluated only one of the two tests may be incorporated within the analysis. The model is implemented in Stata using the *mvmeta* command.

Where cross-classified data are available, the within-study covariances may be calculated through a closed-form expression. Where only 2×2 data are available for a subset of studies, the missing cross-classified counts are derived using those studies with complete data. Random effects meta-analysis is used to estimate summary conditional odds ratios for diseased and non-diseased groups, representing the association between two tests, for the complete studies. These summaries, combined with the available 2×2 data, are used to derive cross-classified for the remaining studies. While the authors state that the model can be extended to evaluate more than two diagnostic tests, this is not elaborated upon in the paper. Similar to the multivariate meta-analysis model proposed by Trikalinos et al [85], convergence issues may arise due to a high number of model parameters associated with additional tests.

3.3.6 Cheng 2016

Cheng proposed three Bayesian meta-analysis models for evaluating multiple diagnostic tests.[92] In their thesis, Cheng aimed to extend three meta-analysis models for single diagnostic tests to the case of three tests: the bivariate normal or BRMA model, the HSROC model, and the beta-binomial model with bivariate copulas.[99, 100] For each method, a shared-parameter modelling framework enables the synthesis of mixed study types, including studies that assessed a single test or a subset of the tests of interest. All three models are arm-based, and can synthesise cross-classified and/or 2×2 data, where available. The three models are implemented in JAGS via the R package *R2jags*.

The first model extends the BRMA model for single tests (described in Section 2.4.2.1) to three tests. At the within-studies level, cross-classified counts for the diseased and non-diseased groups are assumed to follow multinomial distributions. The *logit*-transformed true positive rates and false positive rates are decomposed into test- and study-specific effects, allowing for a full range of study types. Where 2×2 data only are available, the true positive and false positive counts for each test are modelled using independent binomial distributions. The second model extends the HSROC model for single tests (described in Section 2.4.2.2) to three tests. The model may be simplified where cross-classified data are not reported. As in the case for single tests (see Section 2.4.2.3), in the absence of study-level covariates the extended HSROC model for 2×2 data is equivalent to the extended BRMA model for 2×2 data. The third model extends the beta-binomial model with bivariate copulas for single tests [99, 100] to three tests. Beta-binomial marginal distributions capture the observed number of true positive and true negative test results. Multivariate Gaussian copulas link the pairs of marginals, modelling dependencies between and within studies.

3.3.7 Ma et al 2018

Ma et al described an arm-based Bayesian hierarchical meta-analysis model for two or more diagnostic tests.[93] Using this model, it is possible to synthesise data from paired design studies, randomised studies and studies that evaluated a single test using a missing data framework. By considering all studies as if participants underwent all index tests of interest plus a reference standard, and treating outcomes from non-evaluated tests as missing data, pooling of studies that assessed different tests of

interest is enabled. Studies without a reference standard may also be incorporated. The model is implemented in JAGS via *rjags*.

The model accounts for within-study associations present between multiple tests evaluated in the same study participants using a paired design, requiring cross-classified data for each study. Similar to network meta-analysis of multiple treatments, the model assumes that indirect evidence is consistent with direct evidence. In practice this assumption may not be upheld, and a formal test for inconsistency in network meta-analysis of diagnostic tests has yet to be developed.

3.3.8 Hoyer and Kuss 2018a

Hoyer and Kuss proposed an arm-based approach to modelling the sensitivities and specificities of two diagnostic tests compared to a common reference standard through the application of quadrivariate copulas.[94] A quadrivariate Gaussian copula model and a quadrivariate vine copula model are proposed to capture between-studies dependence between sensitivities and specificities. A closed-form likelihood function avoids the more complex numerical approximation methods required by random effects models. The vine copula was found to be the more robust of the two methods through a simulation study presented in the paper. The models are implemented in a frequentist framework using SAS software.

The model requires 2×2 data from each study, and therefore does not account for within-study associations present between two tests evaluated using a paired study design. It is not possible to incorporate single-arm studies using this approach, and the models are limited to two diagnostic tests.

3.3.9 Owen et al 2018

Owen et al outlined an arm-based method for synthesising data on two or more diagnostic tests where 2×2 data are reported at multiple thresholds within a study.[95] At the within-study level, independent binomial likelihoods model the number of true positive and true negative results. At the between-studies level, a bivariate normal distribution captures across-study dependence between sensitivities and specificities. Random effects of sensitivity and specificity due to study and study-specific test and fixed effects of sensitivity and specificity due to test-threshold combinations allow the incorporation of multiple tests and multiple thresholds of the same test. Con-

straints are specified on increasing test thresholds, such that higher thresholds have an increased specificity but decreased sensitivity compared to lower thresholds of the same test.

Implemented in a Bayesian framework, WinBUGS code for the model is provided. While the model accounts for within-study associations present due to multiple thresholds, it does not account for within-study associations present due to paired underlying study design. It is straightforward to extend this method to more than two diagnostic tests and it is possible to include single-arm studies, which evaluated a single index test only, in the analysis.

3.3.10 Hoyer and Kuss 2018b

Hoyer and Kuss proposed an arm-based quadrivariate extension to the BRMA model for single diagnostic tests (see Section 2.4.2.1).[96] Where two diagnostic tests are evaluated against a common reference standard, the true positive true negative results for each test are assumed to follow independent binomial distributions. At the between-studies level, a four-dimensional normal distribution captures dependencies between *logit*-transformed sensitivities and specificities. The model is implemented through a frequentist approach using SAS software.

Where studies report results at multiple thresholds for test positivity, this information is incorporated into the model as a covariate to the linear predictor. The model requires 2×2 data for each study, and does not account for within-study associations present in paired studies of two tests. The model is restricted to two diagnostic tests and cannot incorporate studies that evaluated a single test only.

3.3.11 Lian et al 2019

Lian et al described an arm-based extension to the HSROC model for single diagnostic tests (see Section 2.4.2.2) to two or more tests.[97] By implementing a missing data framework, it is possible to incorporate paired design studies, randomised studies and single arm studies, as well as studies without a reference standard. Disease prevalence, and its correlation with test accuracy measures such as sensitivity and specificity, is taken into account alongside within-study associations between multiple tests evaluated using a paired study design. The model requires cross-classified data from each study and is estimated using a Bayesian sampler in JAGS, via the

rjags package.

Similar to the model proposed by Ma et al [93], the HSROC network meta-analysis model makes several assumptions about the underlying evidence. The model assumes consistency across direct and indirect evidence on test accuracy, and where a study did not evaluate all tests of interest assumes this data is missing at random. The authors discuss the limitations of these assumptions, in light of the lack of methodology to detect inconsistency in network meta-analysis of diagnostic tests. While the model may be applied to three or more diagnostic tests, the number of parameters and therefore the computational burden may hinder the assessment of additional tests. Estimation of high-dimensional covariance matrices may be particularly challenging when the number of studies is small.

3.3.12 Nikoloulopoulos 2019

Nikoloulopoulos proposed an arm-based, frequentist approach to assessing the accuracy of two diagnostic tests compared to a common reference standard, capturing between-studies dependencies between sensitivities and specificities through the application of copula methodology.[98] At the within-study level, four independent binomial distributions model the number of true positive, false negative, true negative and false positive results. At the between-studies level, a drawable vine (D-vine) copula models the associations the sensitivities and specificities of the two tests. Vine copulas enable the extension of parametric bivariate copula families to more than two dimensions by decomposing high-dimensional probability density functions into bivariate copulas and marginal densities.[101] D-vine refers to a specific structure of vine copula, where each level (or tree) of the copula forms a path that determines both the dependence structure and the structure of the level(s) below.

The model requires 2×2 data for each study, and does not account for within-study associations between two tests. The model is limited to evaluating the diagnostic accuracy two index tests compared to a common reference standard, using paired comparative accuracy studies. The model may be implemented through the R package *CopulaREMADA*, increasing its usability.

3.4 Discussion

Fifteen meta-analysis methods for synthesising two or more diagnostic tests were identified from 12 papers included in the methodological review. A review by Veroniki et al, published in 2022, which aimed to identify novel methodological development in network meta-analysis of three or more diagnostic tests demonstrated significant overlap with this review. Nine relevant articles were included, all of which were also identified within this review. Veroniki searched three electronic databases for published and unpublished articles, including grey literature searches, and conducted paired, independent screening of all results. While the searches in this chapter could have been further augmented, for example through outreach to experts in the field to identify unpublished or potentially relevant published work that may have been missed, it seems likely that the search was comprehensive.

All proposed methods extend the BRMA model for meta-analysis of a single diagnostic test (described in Section 2.4.2.1), except for models suggested by Cheng [92] and Lian et al [97], which build upon the HSROC parameterisation (described in Section 2.4.2.2). All but one of the models [88] use an arm-based approach. Trikalinos et al [85], Hoyer and Kuss [96], and Cheng [92] propose using multinomial distributions to model within-study variation between multiple tests. Dimou et al [91] suggest using multivariate normal distributions to account for within-study covariances. Trikalinos et al [85], Dimou et al [91], and Hoyer and Kuss [96] make use of full cross-classifications (i.e. counts of all possible combinations of test results for patient groups both with and without the target condition), meaning within-study correlations are accounted for. Cheng [92] and Lian et al [97] extend the HSROC model to compare multiple tests, with the latter borrowing strength from indirect comparisons through a common diagnostic test. Menten and Lesaffre [88] also make use of indirect evidence in their model, which can be further extended to account for imperfect reference standards. Nyaga et al [89], Ma et al [93], and Owen et al [95] proposed hierarchical (two-stage) approaches based on *logit*-transformed sensitivity and specificity. Nyaga et al [90], Hoyer and Kuss [94], Cheng [92], and Nikoloulopoulos [98] propose copula approaches (discussed in more detail in Chapter 6 and 7) as a flexible method to capture between-study dependencies. Four of the models were performed within a frequentist framework [91, 94, 96, 98]; the remainder were developed in a Bayesian setting.

There are common themes that are present across many of the identified models.

The majority of methods do not account for within-study associations, while those that do typically require cross-classified data from each study, make restrictive assumptions about the underlying distributions of the variables, or are implemented using a Bayesian sampler that is less computationally efficient for parameters that are high correlated in the posterior (e.g. diagnostic accuracy data). There is a need for further model development to address these limitations. In particular, the development of models that can flexibly account for within-study associations, applied using an efficient Bayesian sampler. In practice, cross-classified data may be available for all comparative diagnostic accuracy studies included in an analysis, a subset of studies or one studies, or no studies. A series of models that make optimal use of the different levels of data would help to address this knowledge gap and contribute to the understanding of the comparative and combined accuracy of diagnostic tests in an evidence synthesis framework.

3.5 Chapter summary

This chapter described a methodological review of meta-analysis models for evaluating the accuracy of two diagnostic tests. Motivated by the strengths and weaknesses of the existing methods, Chapters 6 and 7 will aim to address these limitations through novel model development. The following chapter describes a literature review undertaken to identify comparative diagnostic accuracy studies in Alzheimer's disease dementia, on which methodological development in later chapters is based.

Chapter 4

Literature review of comparative diagnostic test accuracy studies in Alzheimer's disease dementia

4.1 Chapter overview

To meet the aims of this thesis, a motivating clinical example must be identified on which to base the methodological development described in later chapters. Driven by the need to assess the comparative and combined performance of diagnostic tests and testing strategies for Alzheimer's disease dementia, this chapter describes a literature review undertaken to source comparative diagnostic accuracy studies evaluating two or more tests for the condition. The studies are used in Chapter 5 to inform the data generating mechanism for a simulation study evaluating the performance of an existing meta-analytic model for jointly synthesising accuracy data on two diagnostic tests. Building on these models, novel evidence synthesis methods are proposed in Chapters 6 and 7. The models' utility as healthcare decision-making aids is explored through their application to comparative test accuracy data acquired in this chapter.

In Section 4.2, an overview of the types of diagnostic tests for Alzheimer's disease dementia, as well as the currently recommended testing pathway in the UK, is given. Section 4.3 outlines the literature review methodology. Methods of literature searching, study selection, and data extraction are described. Section 4.4 summarises key characteristics of the included studies. Comparative test accuracy data extracted from the studies, used throughout the remainder of this thesis, are presented. Sec-

tion 4.5 concludes the chapter with a discussion.

4.2 Diagnostic tests for Alzheimer's disease dementia

Current estimates from NHS England indicate that 38% of people living with dementia do not have a formal diagnosis as of September 2022.[102] An early and accurate dementia diagnosis grants access to additional support services and emerging treatments to manage symptoms and disease progression, as well as providing an explanation for potentially distressing symptoms. Prompt diagnosis also aids with planning both in terms of the health and social management of the individual living with dementia, but also for their families and carers. In accordance with current NICE guidelines in the UK,[36] patients with suspected Alzheimer's disease dementia are assessed using a combination of structured cognitive testing, imaging tests and/or by examining patients' CSF for biomarkers known to be associated with the development of Alzheimer's disease. People with suspected dementia typically present to primary care, where they undergo blood and urine tests to exclude reversible causes of cognitive decline (such as conditions of the liver, kidney, or thyroid, diabetes, vitamin B12 or folate deficiencies, or infection), and a structured cognitive questionnaire (Section 4.2.1). Where dementia is still suspected, patients are referred to a specialist dementia service such as a memory clinic or community old age psychiatry service. Dementia subtype is determined through additional cognitive testing, with further tests for Alzheimer's disease dementia including imaging (Section 4.2.2) and biomarker tests (Section 4.2.3).

4.2.1 Cognitive tests

Cognitive tests may be performed in primary or secondary care settings, including specialist dementia services such as memory clinics, and are a common starting point for investigations in patients presenting with cognitive complaint. Patients' short- and long-term memory, language and communication skills, concentration and orientation (awareness of time and place) are evaluated using structured, validated questionnaires. Common cognitive tests for Alzheimer's disease dementia include the Mini Mental State Examination (MMSE), Montreal Cognitive Assessment (MoCA), and Mini-cog. While cognitive questionnaires are advantageous due to their minimal

time and resource demands, and low risk of adverse events, there is a lack of evidence supporting their role as a stand-alone test for diagnosing dementia or distinguishing one dementia subtype from another.[103, 104, 105, 106, 107, 108] Cognitive questionnaires are often used for screening or triage purposes, i.e. identifying patients that require further testing.

4.2.2 Imaging tests

Brain imaging tests are often used as part of the diagnostic pathway for Alzheimer's disease dementia after referral to a specialist dementia service, aiding the identification of dementia subtype and ruling out of other conditions. Imaging tests for Alzheimer's disease dementia include structural magnetic resonance imaging (MRI), computerised tomography (CT), and positron emission tomography (PET).

4.2.3 Cerebrospinal fluid and plasma biomarkers

CSF is found in the tissue surrounding the brain and spinal cord and is sampled using a fine needle inserted between two vertebrae, known as a lumbar puncture. Changes in levels of CSF biomarkers such as Aβ, t-tau and phosphorylated tau (p-tau) are known to be associated with the development of Alzheimer's disease.

Plasma biomarkers, measured in the blood, are emerging as novel, highly sensitive tests for Alzheimer's disease dementia. Blood concentrations of Aβ and tau proteins have been shown to correspond to CSF levels as well as amyloid- and tau-based PET scans.[109] Blood-based biomarkers have the potential to revolutionise the diagnosis of Alzheimer's disease dementia due to their accessibility, affordability and minimally-invasive nature, although further validation is required before they are implemented into routine clinical practice.[33]

4.2.3.1 Aβ

Aβ plaques are one of the two hallmark pathologies of Alzheimer's disease. There is evidence that amyloid deposition can precede structural changes to the brain and clinically recognisable disease by decades.[1, 2] Aβ_{42} is the most commonly measured form of Aβ. Lower levels of Aβ_{42} measured in the CSF are associated with the accumulation of amyloid plaques in the brain and resulting cognitive impairment.

While another form of Aβ, Aβ_{40}, has been reported to have little ability to detect the condition on its own, it is thought that the ratio of Aβ_{42} to Aβ_{40} (A$\beta_{42/40}$) is superior to Aβ_{42} alone for predicting progression to Alzheimer's disease dementia from MCI.[110] CSF sampling is invasive compared to other diagnostic methods for Alzheimer's disease dementia, and is subject to potential adverse effects including headache, back pain, and swelling at the puncture site. Historically, CSF biomarkers suffered from a lack of standardisation in their conduct and interpretation, leading to variability between measurements across laboratories. Recently, however, a uniform protocol has been developed to homogenise testing procedures and enable their use in routine clinical practice.[111]

4.2.3.2 Tau

Tau protein tangles are the second hallmark pathology of Alzheimer's disease, with elevated levels of t-tau and p-tau associated with the development of the condition. There is a growing body of evidence suggesting that Aβ and tau work together, independently of their respective accumulation, with each marker exacerbating the toxic properties of the other.[112] Ratios of t-tau or p-tau to Aβ_{42} (t-tau/Aβ_{42} or p-tau/Aβ_{42}) have been shown to outperform any of the markers evaluated alone.[113]

4.3 Methods

The literature review was performed in two stages. In the first, systematic reviews of the accuracy of diagnostic tests for Alzheimer's disease dementia were identified through electronic database searching. In the second, the primary studies included within the systematic reviews were screened for comparative diagnostic accuracy data on two or more tests for Alzheimer's disease dementia. Where reported, these data were extracted and compiled in a data set of comparative diagnostic accuracy data. This data set forms the basis of model development in later chapters.

A full systematic review was not an appropriate use of time and resources for this thesis. As the focus of this thesis is methodological development, the aim of the literature review was to identify a clinically relevant example on which to base this methodology rather than the application of such methods to answer questions on the clinical utility of diagnostic tests for Alzheimer's disease dementia.

4.3.1 Literature search

Cochrane is an independent, international organisation that aims to produce trusted, accessible evidence on the effectiveness of healthcare.[4] Cochrane reviews are of consistently high quality and considered the gold standard of systematic reviews. Stringent reporting guidelines enable straightforward extraction of meta-analytic data collected as part of the systematic reviews, making them a valuable resource for data on which to base model evaluation and development. The Cochrane Library was searched on the 10$^{\text{th}}$ August 2023 for articles containing 'Alzheimer's disease' in the title, abstract or among the keywords. Results were filtered by diagnostic test accuracy reviews.

4.3.2 Inclusion criteria

Systematic reviews of diagnostic accuracy studies for tests for Alzheimer's disease dementia were identified from the pool of potentially relevant articles. Systematic reviews of other dementia subtypes (e.g. vascular, frontotemporal) were not included. Reviews that included only studies in which the target diagnosis was all-cause dementia that did not report separate accuracy data for Alzheimer's disease dementia were excluded. Reviews of non-numerical tests, such as clinical judgement as a diagnostic tool, were further excluded. Longitudinal cohort (index test administered to participants at baseline, reference standard of clinical diagnosis obtained through follow-up) or cross-sectional (index test and reference standard administered within a short, specified timespan, e.g. six months) study designs were included. Case-control designs were excluded as they have been shown to overestimate test accuracy.[61]

Primary studies identified within the systematic reviews that reported comparative test accuracy data on two or more diagnostic tests were included. Studies that used a paired or randomised design were eligible for inclusion. Studies that reported diagnostic test accuracy data on a single test only were excluded from further analyses, as the inclusion of non-comparative test accuracy studies in meta-analyses comparing multiple tests has been shown to lead to bias in summary accuracy estimates.[8] Furthermore, it is not possible to incorporate studies evaluating a single test within the models that will be fit to the extracted data in later chapters.

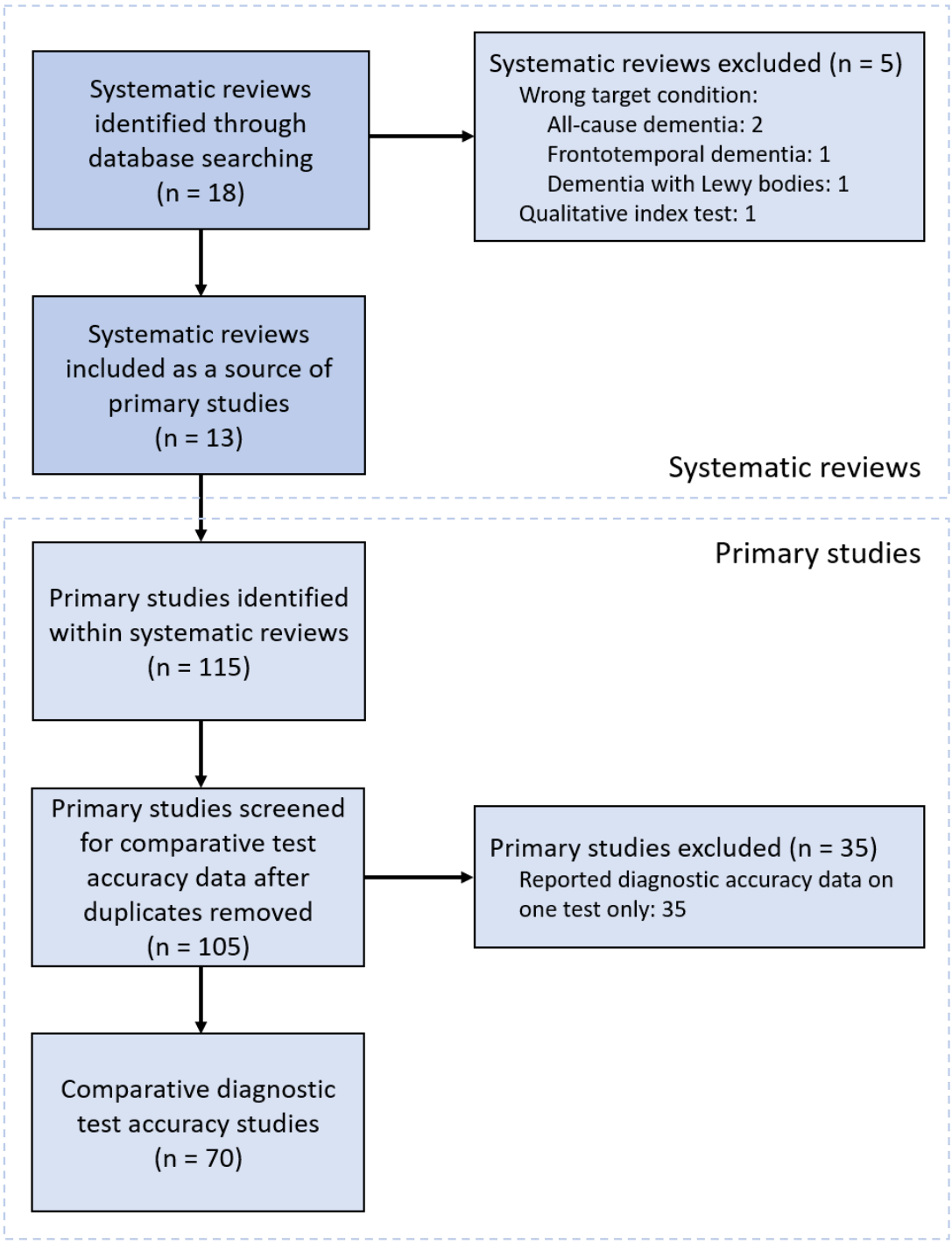


Figure 4.1: Adapted Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) diagram showing the flow of studies through the literature review.[114]

4.3.3 Study selection

In the first stage of study selection, titles and abstracts of systematic reviews identified through searching The Cochrane Library were screened by one reviewer. Articles considered potentially relevant were obtained and assessed by one reviewer for inclusion. Systematic reviews that met the inclusion criteria were used as a source of primary studies in the second stage of the review.

In the second stage of study selection, primary studies included in the systematic reviews were obtained and screened for comparative diagnostic accuracy data on two or more tests by one reviewer.

4.3.4 Data extraction

Data from each systematic review and primary study were extracted by one reviewer. Summary data on participant characteristics, index tests, and reference standard of included studies were extracted from each systematic review.

Two-by-two and/or 2×4 data (fully cross-classified data, see Section 2.3.4) evaluating the accuracy of two or more diagnostic tests for Alzheimer's disease dementia in the same patient group was extracted from each of the identified comparative diagnostic test accuracy studies. The threshold for positivity for each test was also extracted for each 2×2 or 2×4 table. Where diagnostic accuracy data were reported for all-cause dementia with Alzheimer's disease dementia as a subgroup, data were extracted for this subgroup only.

4.4 Results

4.4.1 Systematic review characteristics

A total of 18 potentially relevant systematic reviews were identified through a search of The Cochrane Library (see the adapted Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) study flow diagram in Figure 4.1).[114] Thirteen systematic reviews fulfilled the inclusion criteria. Two reviews for which the target condition was all cause dementia were excluded; studies included in the reviews were examined but none reported diagnostic test accuracy data on Alzheimer's disease dementia only.[105, 108] Two reviews were not considered relevant as the

target condition was a different dementia subtype (frontotemporal dementia and dementia with Lewy bodies, respectively),[115, 116] while another was excluded because the index test was qualitative in nature (clinical judgement by primary care physicians).[117] Key features of the included reviews, including participant characteristics, study type, index test and reference standard, are summarised in Table 4.1.

Table 4.1: Summary of Cochrane reviews of diagnostic tests for Alzheimer's disease dementia. The number of studies that were included after excluding studies with a target condition of all-cause dementia or other dementia subtype is indicated.

Test type	Studies (participants), n	Review characteristics
Arevalo-Rodriguez et al, 2021[1].[103]		
Cognitive	11 (1569) Included: 8 (1128)	**Date of search:** May 2014. **Participant characteristics and setting:** Participants with a diagnosis of MCI recruited from community, primary care and secondary care settings. **Exclusions:** Studies on participants with a secondary cause of cognitive impairment, including alcohol or drug abuse, central nervous system trauma, tumour and infection. **Study type:** Longitudinal cohort. **Index test(s):** MMSE. **Reference standard:** Clinical diagnosis using recognised diagnostic criteria.
Chan et al, 2021[1].[104]		
Cognitive	3 (1415) Included: 1 (279)	**Date of search:** March 2019. **Participant characteristics and setting:** Participants recruited from secondary care settings, including inpatient and outpatient hospital populations. **Exclusions:** Studies on participants with a developmental disability that prevented them from undergoing cognitive testing. **Study type:** Cross-sectional. **Index test(s):** Mini-cog. **Reference standard:** Clinical diagnosis using recognised diagnostic criteria, including additional confirmatory neuroimaging procedures.

Test type	Studies (participants), n	Review characteristics
Davis et al, 2021[1].[106]		
Cognitive	7 (9422) Included: 3 (6788)	**Date of search:** August 2012. **Participant characteristics and setting:** Participants recruited from community, primary care, memory clinic and other secondary care settings. **Exclusions:** Studies on participants with a secondary cause of cognitive impairment, including alcohol or drug abuse, central nervous system trauma, tumour and infection, and in specific clinical groups, e.g. participants with Parkinson's disease or MCI. **Study type:** Cross-sectional. **Index test(s):** MoCA. **Reference standard:** Clinical diagnosis using recognised diagnostic criteria.
Fage et al, 2021[1].[107]		
Cognitive	3 (1620) Included: 2 (464)	**Date of search:** March 2013. **Participant characteristics and setting:** Participants recruited from community settings. **Exclusions:** Studies on participants with a developmental disability that prevented them from undergoing cognitive testing. **Study type:** Cross-sectional. **Index test(s):** Mini-cog. **Reference standard:** Clinical diagnosis using recognised diagnostic criteria, including additional confirmatory neuroimaging procedures.

Test type	Studies (participants), n	Review characteristics
Kokkinou et al, 2021.[118]		
Biomarker	39 (5000) Included: 13 (1704)	**Date of search:** February 2020. **Participant characteristics and setting:** Participants with a clinical diagnosis of any form of dementia recruited from specialist dementia services, either inpatient or outpatient. **Exclusions:** Studies on participants with MCI. **Study type:** Cross-sectional. **Index test(s):** CSF Aβ_{42}. **Reference standard:** Clinical diagnosis using recognised diagnostic criteria.
Lombardi et al, 2020.[119]		
Imaging	33 (3935) Included: 22 (3150)	**Date of search:** January 2019. **Participant characteristics and setting:** Participants with a diagnosis of MCI. **Exclusions:** Studies on healthy participants, participants with subjective cognitive decline in the absence of objective cognitive dysfunction, and studies that based MCI definition on biomarker results. **Study type:** Longitudinal cohort. **Index test(s):** Structural MRI. **Reference standard:** Clinical diagnosis using recognised diagnostic criteria.

Test type	Studies (participants), n	Review characteristics
Martínez et al, 2017(a).[120]		
Imaging	1 (45) Included: 1 (45)	**Date of search:** May 2017. **Participant characteristics and setting:** Participants with a diagnosis of MCI. **Exclusions:** Studies on participants with a secondary cause of cognitive impairment, including alcohol or drug abuse, central nervous system trauma, and other neurological conditions, e.g. Parkinson's or Huntington's diseases. **Study type:** Longitudinal cohort. **Index test(s):** ^{18}F-florbetaben PET. **Reference standard:** Clinical diagnosis using recognised diagnostic criteria.
Martínez et al, 2017(b).[121]		
Imaging	2 (243) Included: 2 (243)	**Date of search:** May 2017. **Participant characteristics and setting:** Participants with a diagnosis of MCI. **Exclusions:** Studies on participants with a secondary cause of cognitive impairment, including alcohol or drug abuse, central nervous system trauma, and other neurological conditions, e.g. Parkinson's or Huntington's diseases. **Study type:** Longitudinal cohort. **Index test(s):** ^{18}F-flutemetamol PET. **Reference standard:** Clinical diagnosis using recognised diagnostic criteria.

Test type	Studies (participants), n	Review characteristics
Martínez et al, 2017(c).[122]		
Imaging	3 (453) Included: 2 (448)	**Date of search:** May 2017. **Participant characteristics and setting:** Participants with a diagnosis of MCI. **Exclusions:** Studies on participants with a secondary cause of cognitive impairment, including alcohol or drug abuse, central nervous system trauma, and other neurological conditions, e.g. Parkinson's or Huntington's diseases. **Study type:** Longitudinal cohort. **Index test(s):** ^{18}F-florbetapir PET. **Reference standard:** Clinical diagnosis using recognised diagnostic criteria.
Ritchie et al, 2014.[39]		
Biomarker	17 (2228) Included: 16 (1967)	**Date of search:** December 2012. **Participant characteristics and setting:** Participants with a diagnosis of MCI. **Exclusions:** Studies on participants with a secondary cause of cognitive impairment, including alcohol or drug abuse, central nervous system trauma, and other neurological conditions (e.g. Parkinson's or Huntington's diseases); participants with psychiatric, neurological, metabolic, immunological, hormonal, or cerebrovascular disorders; participants with other dementia co-morbidity (e.g. vascular or frontotemporal dementias) or potential genetic cause for their dementia; and participants under 50 years old. **Study type:** Longitudinal cohort. **Index test(s):** CSF Aβ42, CSF Aβ_{40}, CSF A$\beta_{42/40}$ ratio. **Reference standard:** Clinical diagnosis using recognised diagnostic criteria.

Test type	Studies (participants), n	Review characteristics
Ritchie et al, 2017.[40]		
Biomarker	15 (1282) Included[2]: 11 (987)	**Date of search:** January 2013. **Participant characteristics and setting:** Participants with a diagnosis of MCI. **Exclusions:** Studies on participants with a secondary cause of cognitive impairment, including alcohol or drug abuse, central nervous system trauma, and other neurological conditions, e.g. Parkinson's or Huntington's diseases. **Study type:** Longitudinal cohort. **Index test(s):** CSF t-tau, CSF p-tau, CSF t-tau/ Aβ ratio, CSF p-tau/Aβ ratio. **Reference standard:** Clinical diagnosis using recognised diagnostic criteria.
Smailagic et al, 2015.[123]		
Imaging	14 (421) Included: 14 (421)	**Date of search:** January 2013. **Participant characteristics and setting:** Participants with a diagnosis of MCI. **Exclusions:** Studies on participants with a secondary cause of cognitive impairment, including alcohol or drug abuse, central nervous system trauma, tumour, infection, and other neurological conditions, e.g. Parkinson's or Huntington's diseases. **Study type:** Longitudinal cohort. **Index test(s):** ^{18}F-fludeoxyglucose (^{18}F-FDG) PET. **Reference standard:** Clinical diagnosis using recognised diagnostic criteria.

Test type	Studies (participants), n	Review characteristics
Zhang et al, 2014.[124]		
Imaging	9 (274) Included: 9 (274)	**Date of search:** January 2013. **Participant characteristics and setting:** Participants with a diagnosis of MCI. **Exclusions:** Studies on participants with a secondary cause of cognitive impairment, including alcohol or drug abuse, central nervous system trauma, tumour, infection, and other neurological conditions, e.g. Parkinson's or Huntington's diseases. **Study type:** Longitudinal cohort. **Index test(s):** ^{11}C-labelled Pittsburgh Compound-B (^{11}C-PIB) PET. **Reference standard:** Clinical diagnosis using recognised diagnostic criteria.

Aβ, amyloid beta; CSF, cerebrospinal fluid; MCI, mild cognitive impairment; MMSE, Mini Mental State Examination; MoCA, Montreal Cognitive Assessment; MRI, magnetic resonance imaging; PET, positron emission tomography; p-tau, phosphorylated tau; t-tau, total tau

[1]The title and objectives of this review were revised in 2021 to clarify that short screening tests alone cannot make a diagnosis of a dementia subtype such as Alzheimer's disease dementia, in line with feedback from a group of dementia researchers. There were no other changes to the review, but the citation was updated to reflect the revision date. The date of the literature search is included to make the cut-off point for publication inclusion clear.
[2]Two studies involved the same participants, only one was included.

Participants were recruited from the community (population-based screening), primary care (where patients may present to their general practitioner with cognitive issues, or be screened opportunistically), and secondary care (specialist care services, including hospitals and memory clinics). Due to their expense and complexity, imaging and CSF biomarker tests were assessed in secondary care settings only, whereas cognitive tests were evaluated across all settings. Most reviews specifically targeted participants diagnosed with mild cognitive impairment (MCI), also known as prodromal dementia, based on cognitive decline beyond what would be expected as part of the normal ageing process that does not meet the criteria for clinically probable dementia.[125] MCI differs from a dementia diagnosis in that cognitive impairment is not severe enough to significantly impact daily activities, although people with MCI are more likely to go on to develop dementia than people without MCI. One study found that 46% of participants with MCI at baseline developed dementia within the 3 years of follow-up, compared with 3% of age-matched participants without impairment.[126] Prevalence of Alzheimer's disease dementia across studies exhibited a high pf heterogeneity, ranging from 12.3-77.1%.

The reviews assessed the diagnostic accuracy of cognitive, imaging, and CSF and plasma biomarker tests for Alzheimer's disease dementia. Two reviews evaluated the accuracy of multiple diagnostic tests.[39, 40] Most reviews restricted their inclusion to longitudinal cohort studies.

In all reviews the reference standard for confirming the presence or absence of Alzheimer's disease dementia was clinical diagnosis using recognised diagnostic criteria, such as the Diagnostic and Statistical Manual of Mental Disorders (DSM), the International Classification of Diseases (ICD), or the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA). All criteria require an individual to experience impairment in multiple areas of cognitive function that impact daily living to meet the threshold for diagnosis. These criteria are considered the gold standard for ante-mortem diagnosis of Alzheimer's disease dementia, although a definitive diagnosis can only be made post-mortem via autopsy. Some reviews included studies that used additional neuroimaging procedures, such as MRI or CT scans, to further clarify the diagnosis.[104, 107]

Table 4.2: Details of the included comparative diagnostic test accuracy studies.

Study, year	Study design	Study population			Index tests								
		Population	N	ADD, %	Test type	Test	Threshold	*tp*	*fn*	*tn*	*fp*	*se*, %	*sp*, %
Anchisi et al, 2005.[127]	Longitudinal cohort	MCI	48	29.2	Imaging	^{18}F-FDG PET	rCGMr≤1.14	13	1	28	6	92.9	82.4
					Cognitive	CVLT-LDFR	6/16	13	1	20	14	92.9	58.8
Bjerke et al, 2009.[128]	Longitudinal cohort	MCI	162	12.3	Biomarker	CSF Aβ_{42}	512 ng/L	18	2	99	43	90.0	69.7
					Biomarker	CSF t-tau	613 ng/L	12	8	130	12	60.0	91.5
					Biomarker	CSF p-tau	66 ng/L	17	3	109	33	85.0	76.8
					Biomarker	CSF Aβ_{42}/t-tau	1.2	19	1	101	41	95.0	71.1
Blom et al, 2009.[129]	Longitudinal cohort	MCI	28	50.0	Biomarker	CSF Aβ_{42}	82 pM	9	5	5	9	64.3	35.7
					Biomarker	CSF Aβ_{40}	1615 pM	9	5	12	2	64.3	85.7
					Biomarker	CSF t-tau	428 ng/L	7	7	11	3	50.0	78.6
					Biomarker	CSF p-tau	51 ng/L	8	6	11	3	57.1	78.6
Borson et al, 2005.[130]	Cross-sectional	Community	252	44.4	Cognitive	MMSE	23/30	106	6	118	22	94.6	84.3
					Cognitive	Mini-cog	2/5	111	1	116	24	99.1	82.9
Brettschneider et al, 2006.[131]	Cross-sectional	All-cause dementia	165	66.1	Biomarker	CSF Aβ_{42}	612 pg/mL	89	20	26	30	81.7	46.4
					Biomarker	CSF A$\beta_{42/40}$	1.15	101	8	24	32	92.7	42.9
					Biomarker	CSF t-tau	389 pg/mL	74	35	36	20	67.9	64.3
					Biomarker	CSF p-tau	62.5 pg/mL	72	37	40	16	66.1	71.4
					Biomarker	CSF NfH^{SMI35}	15 pg/mL	85	24	16	40	78.0	28.6

Study, year	Study design	Study population			Index tests								
		Population	N	ADD, %	Test type	Test	Threshold	*tp*	*fn*	*tn*	*fp*	*se*, %	*sp*, %
Brys et al, 2009.[132]	Longitudinal cohort	MCI	65	33.8	Biomarker	CSF A$\beta_{42/40}$	0.099	19	3	26	17	86.4	60.5
					Biomarker	CSF p-tau	38.6 pg/mL	16	6	38	5	72.7	88.4
					Biomarker	CSF t-tau	605.5 pg/mL	15	7	39	4	68.2	90.7
					Biomarker	CSF t-tau/A$\beta_{42/40}$	6127	17	5	35	8	77.3	81.4
					Biomarker	CSF p-tau/A$\beta_{42/40}$	445.9	16	6	37	6	72.7	86.0
					Biomarker	CSF IP	47.5 pg/mL	12	10	39	4	54.5	90.7
Buchhave et al, 2008.[133]	Longitudinal cohort	MCI	147	42.9	Cognitive	MMSE	28/30	56	7	28	56	88.9	33.3
					Cognitive	CDT	3/5	33	30	57	27	52.4	67.9
					Cognitive	CCT	12/20	34	29	60	24	54.0	71.4
Carmichael et al, 2007.[134]	Longitudinal cohort	MCI	29	41.4	Imaging	MRI-LV	NR	9	3	8	9	75.0	47.1
					Imaging	MRI-WB	NR	11	1	7	10	91.7	41.2
Caroli et al, 2007.[135]	Longitudinal cohort	MCI	23	39.1	Imaging	MRI-HC, total	NR	9	0	6	8	100.0	42.9
					Imaging	MRI-HC, left	NR	8	1	9	5	88.9	64.3
					Imaging	MRI-HC, right	NR	9	0	6	8	100.0	42.9
					Imaging	MRI-MTL, total	NR	6	3	11	3	66.7	78.6

Study, year	Study design	Study population			Index tests								
		Population	N	ADD, %	Test type	Test	Threshold	*tp*	*fn*	*tn*	*fp*	*se*, %	*sp*, %
Chiasserini et al, 2010.[136]	Longitudinal cohort	MCI	41	56.1	Biomarker	CSF Aβ_{42}	741 pg/mL	17	6	16	2	73.9	88.9
					Biomarker	CSF t-tau	363 pg/mL	12	11	16	2	52.2	88.9
					Biomarker	CSF p-tau	51 pg/mL	20	3	13	5	87.0	72.2
					Biomarker	CSF HFABP	451 pg/mL	19	4	9	9	82.6	50.0
					Biomarker	CSF HFABP/Aβ_{42}	0.7	18	5	18	0	78.3	100.0
					Biomarker	CSF HFABP/t-tau	1.3	10	13	16	2	43.5	88.9
					Biomarker	CSF HFABP/p-tau	8.6	8	15	14	4	34.8	77.8
Clerx et al, 2013.[137]	Longitudinal cohort	MCI	328	27.7	Imaging	MRI-HC, total	NR	71	20	154	83	78.0	65.0
					Imaging	MRI-MTL, total	NR	60	31	152	85	65.9	64.1
					Imaging	MRI-LV	NR	48	43	161	76	52.7	67.9
deToledo-Morrell et al, 2004.[138]	Longitudinal cohort	MCI	27	37.0	Imaging	MRI-HC, total	NR	9	1	10	7	90.0	58.8
					Imaging	MRI-ERC, total	NR	5	5	17	0	50.0	100.0
Devanand et al, 2007.[139]	Longitudinal cohort	MCI	139[1]	25.2	Imaging	MRI-HC, total	NR	20	15	87	17	57.1	83.7
					Imaging	MRI-HC, left	NR	24	11	71	33	68.6	68.3
					Imaging	MRI-HC, right	NR	29	6	62	42	82.9	59.6
					Imaging	MRI-ERC, total	NR	22	12	89	15	64.7	85.6
					Imaging	MRI-ERC, left	NR	22	12	88	16	64.7	84.6
					Imaging	MRI-ERC, right	NR	20	14	92	12	58.8	88.5

Study, year	Study design	Study population			Index tests								
		Population	N	ADD, %	Test type	Test	Threshold	*tp*	*fn*	*tn*	*fp*	*se*, %	*sp*, %
Devanand et al, 2008.[140]	Longitudinal cohort	MCI	125[1]	26.4	Cognitive	MMSE	NR	9	24	83	9	27.3	90.2
					Cognitive	SRT	NR	16	16	84	9	50.0	90.3
					Cognitive	UPSIT	NR	16	17	82	9	48.5	90.1
					Cognitive	FAQ	NR	11	22	77	8	33.3	90.6
					Imaging	MRI-HC, total	NR	12	17	80	9	41.4	89.9
					Imaging	MRI-ERC, total	NR	14	14	80	9	50.0	89.9
Eckerström et al, 2013.[141]	Longitudinal cohort	MCI	34	38.2	Imaging	MRI-HC, left	NR	11	2	16	5	84.6	76.2
					Imaging	MRI-HC, right	NR	9	4	17	4	69.2	81.0
Erten-Lyons et al, 2006.[142]	Longitudinal cohort	MCI	37	59.5	Imaging	MRI-HC, total	NR	14	8	11	4	63.6	73.3
					Imaging	MRI-LV	NR	13	9	11	4	59.1	73.3
					Imaging	MRI-WB	NR	19	3	7	8	86.4	46.7
Fei et al, 2011.[143]	Longitudinal cohort	MCI	565[1]	43.4	Biomarker	Plasma Aβ_{42}	0.64 ng/mL	210	35	159	161	85.7	49.7
					Biomarker	Plasma A$\beta_{42/40}$	0.95 ng/mL	210	35	221	96	85.7	69.7
Fellgiebel et al, 2007.[144]	Longitudinal cohort	MCI	16	25.0	Biomarker	CSF p-tau	50 ng/mL	4	0	4	8	100.0	33.3
					Imaging	^{18}F-FDG PET	NR	4	0	9	3	100.0	75.0
Frölich et al, 2017.[145]	Longitudinal cohort	MCI	115	24.3	Imaging	MRI-HC, total	NR	18	10	71	16	64.3	81.6
					Biomarker	CSF Aβ_{42}	600 pg/mL	17	11	65	22	60.7	74.7
					Biomarker	CSF t-tau	300 pg/mL	24	4	46	41	85.7	52.9
					Biomarker	CSF p-tau	60 pg/mL	17	11	60	27	60.7	69.0
Galluzzi et al, 2010.[146]	Longitudinal cohort	MCI	86[1]	25.6	Imaging	^{18}F-FDG PET	t sum>11.09	11	3	7	17	78.6	29.2
					Biomarker	CSF Aβ_{42}	500 pg/mL	18	4	25	17	81.8	59.5
					Imaging	MRI-MTL, total	NR	12	10	51	13	54.5	79.7

Study, year	Study design	Study population			Index tests								
		Population	N	ADD, %	Test type	Test	Threshold	*tp*	*fn*	*tn*	*fp*	*se*, %	*sp*, %
Galton et al, 2005.[147]	Longitudinal cohort	MCI	29	37.9	Imaging	MRI-LTL, right	NR	9	2	9	9	81.8	50.0
					Cognitive	ADAS	NR	7	4	16	2	63.6	88.9
					Cognitive	ACE	NR	8	3	18	0	72.7	100.0
Gaser et al, 2013.[148]	Longitudinal cohort	MCI	195[1]	68.2	Cognitive	MMSE	NR	94	39	38	24	70.7	61.3
					Cognitive	CDR-SB	NR	85	48	48	14	63.9	77.4
					Cognitive	ADAS	NR	86	47	50	12	64.7	80.6
					Imaging	MRI-HC, left	NR	70	63	50	12	52.6	80.6
					Biomarker	CSF Aβ_{42}	NR	59	7	12	21	89.4	36.4
					Biomarker	CSF t-tau	NR	58	8	13	20	87.9	39.4
					Biomarker	CSF p-tau	NR	45	21	19	14	68.2	57.6
					Biomarker	CSF Aβ_{42}/p-tau	NR	61	5	14	19	92.4	42.4
Hampel et al, 2004.[149]	Longitudinal cohort	MCI	52	55.8	Biomarker	CSF t-tau	479 ng/L	26	3	11	12	89.7	47.8
					Biomarker	CSF Aβ_{42}	679 ng/L	24	5	13	10	82.8	56.5
Hansson et al, 2006.[150]	Longitudinal cohort	MCI	134	42.5	Biomarker	CSF t-tau	350 pg/mL	55	2	47	30	96.5	61.0
					Biomarker	CSF p-tau	60 pg/mL	54	3	45	32	94.7	58.4
					Biomarker	CSF Aβ_{42}/p-tau	6.5 pg/mL	55	2	61	16	96.5	79.2
					Biomarker	CSF Aβ_{42}	530 ng/L	56	1	50	27	98.2	64.9
Hansson et al, 2007.[151]	Longitudinal cohort	MCI	134	42.5	Biomarker	CSF Aβ_{42}	0.64 ng/L	53	4	41	36	93.0	53.2
					Biomarker	CSF A$\beta_{42/40}$	0.95 ng/L	50	7	60	17	87.7	77.9

Study, year	Study design	Study population			Index tests								
		Population	N	ADD, %	Test type	Test	Threshold	*tp*	*fn*	*tn*	*fp*	*se*, %	*sp*, %
Hertze et al, 2010.[152]	Longitudinal cohort	MCI	159	32.7	Biomarker	CSF Aβ_{42}	209 pg/mL	47	5	75	32	90.4	70.1
					Biomarker	CSF t-tau	100 pg/mL	38	14	82	25	73.1	76.6
					Biomarker	CSF p-tau	51 pg/mL	22	30	96	11	42.3	89.7
					Biomarker	CSF Aβ_{42}/t-tau	2.5	47	5	75	32	90.4	70.1
					Biomarker	CSF Aβ_{42}/p-tau	6.6	45	7	77	30	86.5	72.0
					Biomarker	CSF t-tau/Aβ_{40}	0.010	43	9	78	29	82.7	72.9
					Biomarker	CSF A$\beta_{42/40}$	0.069	45	7	70	37	86.5	65.4
Herukka et al, 2008.[153]	Longitudinal cohort	MCI	21	38.1	Imaging	MRI-HC, total	NR	8	0	8	5	100.0	61.5
					Imaging	MRI-HC, left	NR	6	2	8	5	75.0	61.5
					Imaging	MRI-HC, right	NR	7	1	9	4	87.5	69.2
					Imaging	MRI-ERC, total	NR	7	1	10	3	87.5	76.9
					Imaging	MRI-ERC, left	NR	7	1	9	4	87.5	69.2
					Imaging	MRI-ERC, right	NR	7	1	9	4	87.5	69.2
					Biomarker	CSF Aβ_{42}	450 pg/mL	6	2	11	2	75.0	84.6
					Biomarker	CSF t-tau	400 pg/mL	6	2	7	6	75.0	53.8
					Biomarker	CSF p-tau	70 pg/mL	7	1	7	6	87.5	53.8
Jang et al, 2018.[154]	Longitudinal cohort	MCI	340	20.3	Imaging	MRI-LV	NR	44	25	161	110	63.8	59.4
					Imaging	^{18}F-florbetapir PET	SUVR≥1.10	61	8	143	128	88.4	52.8
Kapaki et al, 2003.[155]	Cross-sectional	All-cause dementia	64	76.7	Biomarker	CSF Aβ_{42}	435 pg/mL	35	14	12	3	71.4	80.0
					Biomarker	CSF t-tau	437 pg/mL	35	14	14	1	71.4	93.3
					Biomarker	CSF t-tau/Aβ_{42}	0.94	35	14	15	0	71.4	100.0

Study, year	Study design	Study population			Index tests								
		Population	N	ADD, %	Test type	Test	Threshold	*tp*	*fn*	*tn*	*fp*	*se*, %	*sp*, %
Kester et al, 2011.[156]	Longitudinal cohort	MCI	100	42.0	Biomarker	CSF t-tau	356 pg/mL	35	7	29	29	83.3	50.0
					Biomarker	CSF Aβ_{42}	495 pg/mL	32	10	42	16	76.2	72.4
Knapskog et al, 2019.[157]	Cross-sectional	All-cause dementia	205	67.3	Biomarker	CSF Aβ_{42}	550 pg/mL	59	79	55	12	42.8	82.1
					Biomarker	CSF t-tau	300 ng/L	90	48	52	15	65.2	77.6
					Biomarker	CSF p-tau	79 ng/L	65	73	60	7	47.1	89.6
Koivunen et al, 2008.[158]	Longitudinal cohort	MCI	14	35.7	Biomarker	CSF p-tau	70 pg/mL	2	3	2	7	40.0	22.2
					Biomarker	CSF Aβ_{42}/p-tau	6.5 pg/mL	4	1	3	6	80.0	33.3
Lee et al, 2008.[159]	Cross-sectional	Secondary care	159	27.7	Cognitive	MoCA	25/30	44	0	59	56	100.0	51.3
					Cognitive	MMSE	25/30	38	6	81	34	86.4	70.4
Ledig et al, 2018.[160]	Longitudinal cohort	MCI	343	51.6	Imaging	MRI-HC, right	NR	108	69	106	60	61.0	63.9
					Imaging	MRI-ERC, total	NR	108	69	100	66	61.0	60.2
					Imaging	MRI-LV	NR	90	87	116	50	50.8	69.9
					Imaging	MRI-WB	NR	92	85	81	85	52.0	48.8
					Imaging	MRI-MTG	NR	104	73	100	66	58.8	60.2
					Imaging	MRI-amygdala, total	NR	112	65	113	53	63.3	68.1
					Imaging	MRI-amygdala, left	NR	108	69	106	60	61.0	63.9
					Imaging	MRI-amygdala, right	NR	112	65	111	55	63.3	66.9
					Imaging	MRI-CGM	NR	97	80	93	73	54.8	56.0

Study, year	Study design	Study population			Index tests								
		Population	N	ADD, %	Test type	Test	Threshold	*tp*	*fn*	*tn*	*fp*	*se*, %	*sp*, %
Lewczuk et al, 2004.[161]	Cross-sectional	All-cause dementia	33[1]	66.7	Biomarker	CSF Aβ_{42}	500 pg/mL	19	3	9	2	86.4	81.8
					Biomarker	CSF A$\beta_{42/40}$	11.0	20	1	8	2	95.2	80.0
					Biomarker	CSF Aβ40	4800 pg/mL	8	13	9	1	38.1	90.0
					Biomarker	CSF t-tau	400 pg/mL	18	4	8	3	81.8	72.7
Maddalena et al, 2003.[162]	Cross-sectional	All-cause dementia	81	63.0	Biomarker	CSF Aβ_{42}	490 pg/mL	40	11	21	9	78.4	70.0
					Biomarker	CSF p-tau	35 pg/mL	37	14	19	11	72.5	63.3
					Biomarker	CSF p-tau/Aβ_{42}	83	41	10	22	8	80.4	73.3
Milian et al, 2012.[163]	Cross-sectional	Secondary care	279	77.1	Cognitive	Mini-cog	NR	195	20	64	0	90.7	100.0
					Cognitive	CDT	3/6	174	41	62	2	80.9	96.9
					Cognitive	MMSE	24/30	175	40	64	0	81.4	100.0
Modrego et al, 2005.[164]	Longitudinal cohort	MCI	53	54.7	Cognitive	MMSE	29/35	22	7	16	8	75.9	66.7
					Imaging	MRI-HC, left	NR	18	11	13	11	62.1	54.2
					Imaging	MRI-PC, right	NR	18	11	17	7	62.1	70.8
					Imaging	MRI-OC, left	NR	29	0	18	6	100.0	75.0
Modrego et al, 2013.[165]	Longitudinal cohort	MCI	105	54.3	Cognitive	MMSE	30/35	26	31	42	6	45.6	87.5
					Cognitive	MIS	2/8	39	18	35	13	68.4	72.9
Monge-Argilés et al, 2011.[166]	Longitudinal cohort	MCI	37	29.7	Biomarker	CSF t-tau	77.5 pg/mL	8	3	18	8	72.7	69.2
					Biomarker	CSF p-tau	54.5 pg/mL	9	2	15	11	81.8	57.7
					Biomarker	CSF t-tau/Aβ_{42}	0.18 pg/mL	10	1	13	13	90.9	50.0
					Biomarker	CSF p-tau/Aβ_{42}	0.17 pg/mL	9	2	17	9	81.8	65.4
					Biomarker	CSF Aβ_{42}	320 pg/mL	9	2	16	10	81.8	61.5

Study, year	Study design	Study population			Index tests								
		Population	N	ADD, %	Test type	Test	Threshold	*tp*	*fn*	*tn*	*fp*	*se*, %	*sp*, %
Montine et al, 2001.[167]	Cross-sectional	All-cause dementia	27	70.4	Biomarker	CSF Aβ_{42}	1125 pg/mL	19	0	2	6	100.0	25.0
					Biomarker	CSF t-tau	300 pg/mL	15	4	5	3	78.9	62.5
					Biomarker	CSF IP	25 pg/mL	17	2	3	5	89.5	37.5
Mosconi et al, 2004.[168]	Longitudinal cohort	MCI	37	21.6	Imaging	^{18}F-FDG PET	NR	3	5	28	1	37.5	96.6
					Cognitive	MMSE	NR	2	6	28	1	25.0	96.6
Nesteruk et al, 2016.[169]	Longitudinal cohort	MCI	40	22.5	Imaging	MRI-HC, left	NR	6	3	24	7	66.7	77.4
					Imaging	MRI-HC, right	NR	6	3	23	8	66.7	74.2
					Imaging	MRI-ERC, left	NR	5	4	21	10	55.6	67.7
					Biomarker	CSF Aβ_{42}	609.5 pg/mL	7	2	18	13	77.8	58.1
					Biomarker	CSF t-tau	277.0 pg/mL	6	3	20	11	66.7	64.5
					Biomarker	CSF p-tau	55.1 pg/mL	3	6	24	7	33.3	77.4
Nobili et al, 2008.[170]	Longitudinal cohort	MCI	33	33.3	Imaging	^{18}F-FDG PET	NR	9	2	20	2	81.8	90.9
					Cognitive	SRT-IR	NR	7	4	14	8	63.6	63.6
					Cognitive	SRT-DR	NR	6	5	11	11	54.5	50.0
Ong et al, 2015.[171]	Longitudinal cohort	MCI	45	44.4	Imaging	^{18}F-florbetaben PET	SUVR>1.45	18	2	19	6	90.0	76.0
					Imaging	MRI-HC, total	NR	10	10	16	9	50.0	64.0
Ossenkoppele et al, 2012a.[172]	Longitudinal cohort	MCI	12	33.3	Imaging	^{18}F-FDG PET	NR	3	1	7	1	75.0	87.5
					Imaging	^{11}C-PIB PET	NR	4	0	7	1	100.0	87.5
Ossenkoppele et al, 2012b.[173]	Longitudinal cohort	MCI	12	50.0	Imaging	^{18}F-FDG PET	NR	5	1	6	0	83.3	100.0
					Imaging	^{11}C-PIB PET	NR	6	0	5	1	100.0	83.3

Study, year	Study design	Study population			Index tests								
		Population	N	ADD, %	Test type	Test	Threshold	*tp*	*fn*	*tn*	*fp*	*se*, %	*sp*, %
Palmqvist et al, 2012.[174]	Longitudinal cohort	MCI	133	39.1	Biomarker	CSF t-tau	87 pg/mL	42	10	58	23	80.8	71.1
					Biomarker	CSF p-tau	39 pg/mL	35	17	70	11	67.3	86.4
					Biomarker	CSF Aβ_{42}	208 pg/mL	47	5	56	25	90.4	69.1
					Cognitive	MMSE	26/30	32	20	68	13	61.5	84.0
					Cognitive	CDT	3/5	23	29	70	11	44.2	86.4
Parnetti et al, 2006.[175]	Longitudinal cohort	MCI	44	25.0	Biomarker	CSF Aβ_{42}	500 pg/L	4	7	30	3	36.4	90.9
					Biomarker	CSF t-tau	$\leq$50 years: 300 pg/mL 51-70 years: 450 pg/mL $\geq$71 years: 500 pg/mL	5	6	32	1	45.5	97.0
					Biomarker	CSF p-tau	80 pg/mL	9	2	32	1	81.8	97.0
Parnetti et al, 2012.[176]	Longitudinal cohort	MCI	90	35.6	Biomarker	CSF Aβ_{42}	421 pg/mL	18	14	56	2	56.3	96.6
					Biomarker	CSF Aβ_{40}	12282 pg/mL	20	12	31	27	62.5	53.4
					Biomarker	CSF A$\beta_{42/40}$	5	23	9	53	5	71.9	91.4
					Biomarker	CSF t-tau	406 pg/mL	20	12	51	7	62.5	87.9
					Biomarker	CSF p-tau	62 pg/mL	26	6	52	6	81.3	89.7
					Biomarker	CSF Aβ_{42}/t-tau	348	30	2	38	20	93.8	65.5
					Biomarker	CSF Aβ_{42}/p-tau	1074	26	6	55	3	81.3	94.8

Study, year	Study design	Study population			Index tests								
		Population	N	ADD, %	Test type	Test	Threshold	*tp*	*fn*	*tn*	*fp*	*se*, %	*sp*, %
Perani et al, 2016.[177]	Cross-sectional	All-cause dementia	75[1]	62.7	Biomarker	CSF Aβ_{42}	500 ng/L	40	7	13	15	85.1	46.4
					Biomarker	CSF t-tau	350 ng/L	18	29	24	4	38.3	85.7
					Biomarker	CSF p-tau	61 ng/L	33	14	18	10	70.2	64.3
					Biomarker	CSF t-tau/Aβ_{42}	NR	37	10	19	9	78.7	67.9
					Biomarker	CSF p-tau/Aβ_{42}	NR	39	8	18	10	83.0	64.3
					Imaging	MRI-WB	NR	9	10	8	8	47.4	50.0
					Imaging	^{18}F-FDG PET	NR	44	3	24	4	93.6	85.7
Pozueta et al, 2011.[178]	Longitudinal cohort	MCI	105	47.6	Cognitive	MMSE	26/30	32	18	44	11	64.0	80.0
					Cognitive	CVLT-LDFR	4/16	37	13	39	16	74.0	70.9
Prestia et al, 2013a.[179]	Longitudinal cohort	MCI	73	39.7	Imaging	MRI-HC, total	W-score < -2.90	16	13	34	10	55.2	77.3
					Imaging	^{18}F-FDG PET	t sum>13.48	23	6	32	12	79.3	72.7
					Biomarker	CSF Aβ_{42}	500 pg/mL	21	8	29	15	72.4	65.9
Prestia et al, 2013b.[180]	Longitudinal cohort	MCI	36	50.0	Imaging	MRI-HC, total	W-score < -2.76	5	13	17	1	27.8	94.4
					Biomarker	CSF Aβ_{42}	500 pg/mL	17	1	9	9	94.4	50.0
					Biomarker	CSF t-tau	≤70 years: 450 pg/mL ≥71 years: 500 pg/mL	11	7	15	3	61.1	83.3

Study, year	Study design	Study population			Index tests								
		Population	N	ADD, %	Test type	Test	Threshold	*tp*	*fn*	*tn*	*fp*	*se*, %	*sp*, %
Rhodius-Meester et al, 2016.[181]	Longitudinal cohort	MCI	171[1]	60.8	Imaging	MRI-MTL, total	NR	42	62	57	10	40.4	85.1
					Biomarker	CSF Aβ_{42}	550 pg/mL	58	27	38	14	68.2	73.1
					Biomarker	CSF t-tau	375 pg/mL	72	13	35	17	84.7	67.3
					Biomarker	CSF p-tau	52 pg/mL	75	10	30	22	88.2	57.7
Rosler et al, 2001.[182]	Cross-sectional	All-cause dementia	51	52.9	Biomarker	CSF Aβ_{42}	375 pg/mL	21	6	14	10	77.8	58.3
					Biomarker	CSF t-tau	440 pg/mL	24	3	16	8	88.9	66.7
Schreiber et al, 2015.[183]	Longitudinal cohort	MCI	401[1]	15.2	Imaging	^{18}F-florbetapir PET	SUVR>1.11	53	8	172	168	86.9	50.6
					Biomarker	CSF Aβ_{42}	192 pg/mL	33	4	86	133	89.2	39.3
Smach et al, 2008.[184]	Cross-sectional	All-cause dementia	108	67.6	Biomarker	CSF Aβ_{42}	505 pg/mL	60	13	25	10	82.2	71.4
					Biomarker	CSF t-tau	355 pg/mL	59	14	24	11	80.8	68.6
Spies et al, 2010.[185]	Cross-sectional	All-cause dementia	138	50.0	Biomarker	CSF Aβ_{42}	NR	57	12	51	18	82.6	73.9
					Biomarker	CSF A$\beta_{42/40}$	NR	59	10	59	10	85.5	85.5
Tapiola et al, 2000.[186]	Cross-sectional	All-cause dementia	107	74.8	Biomarker	CSF Aβ_{42}	340 pg/mL	55	25	16	11	68.8	59.3
					Biomarker	CSF t-tau	380 pg/mL	50	30	18	9	62.5	66.7
Tariciotti et al, 2018.[187]	Cross-sectional	All-cause dementia	749	35.2	Biomarker	CSF Aβ_{42}	500 pg/mL	214	50	262	223	81.1	54.0
					Biomarker	CSF t-tau	350 pg/mL	203	61	165	320	76.9	34.0
					Biomarker	CSF p-tau	61 pg/mL	193	71	97	388	73.1	20.0
Thurfjell et al, 2012.[188]	Longitudinal cohort	MCI	19	47.4	Imaging	^{18}F-flutemetamol PET	SUVR>1.5	8	1	8	2	88.9	80.0
					Imaging	MRI-HC, total	NR	7	2	3	7	77.8	30.0

Study, year	Study design	Study population			Index tests								
		Population	N	ADD, %	Test type	Test	Threshold	*tp*	*fn*	*tn*	*fp*	*se*, %	*sp*, %
van der Flier et al, 2005.[189]	Longitudinal cohort	MCI	15	60.0	Imaging	MRI-HC, total	NR	7	2	4	2	77.8	66.7
					Imaging	MRI-HC, left	NR	4	5	6	0	44.4	100.0
					Imaging	MRI-HC, right	NR	8	1	4	2	88.9	66.7
					Imaging	MRI-MTL, total	NR	7	2	3	3	77.8	50.0
					Imaging	MRI-MTL, left	NR	8	1	2	4	88.9	33.3
					Imaging	MRI-MTL, right	NR	2	7	6	0	22.2	100.0
					Imaging	MRI-WB	NR	3	6	6	0	33.3	100.0
Visser et al, 1999.[190]	Longitudinal cohort	MCI	13	69.2	Imaging	MRI-HC, total	NR	9	0	3	1	100.0	75.0
					Imaging	MRI-LTL, total	NR	5	4	3	1	55.6	75.0
Visser et al, 2002.[191]	Longitudinal cohort	MCI	30[1]	23.3	Imaging	MRI-HC, total	NR	6	1	17	5	85.7	77.3
					Imaging	MRI-MTL, total	NR	6	1	17	6	85.7	73.9
Visser et al, 2009.[192]	Longitudinal cohort	MCI	158	22.2	Biomarker	CSF p-tau	51 pg/mL	31	4	46	77	88.6	37.4
					Biomarker	CSF Aβ_{42}/t-tau	1	35	0	53	70	100.0	43.1
					Biomarker	CSF Aβ_{42}/p-tau	9.92	28	7	74	49	80.0	60.2
Vos et al, 2013.[193]	Longitudinal cohort	MCI	214	42.5	Biomarker	CSF Aβ_{42}	500 pg/mL	62	29	82	41	68.1	66.7
					Biomarker	CSF t-tau	<70 years: 450 pg/mL ≥70 years: 500 pg/mL	65	26	95	28	71.4	77.2
					Biomarker	CSF Aβ_{42}/t-tau	1	87	4	63	60	95.6	51.2

Study, year	Study design	Study population			Index tests								
		Population	N	ADD, %	Test type	Test	Threshold	*tp*	*fn*	*tn*	*fp*	*se*, %	*sp*, %
Wang et al, 2006.[194]	Longitudinal cohort	MCI	58	32.8	Imaging	MRI-HC, total	NR	15	4	32	7	78.9	82.1
					Imaging	MRI-HC, left	NR	15	4	31	8	78.9	79.5
					Imaging	MRI-HC, right	NR	14	5	30	9	73.7	76.9
					Imaging	MRI-amygdala, total	NR	11	8	32	7	57.9	82.1
					Imaging	MRI-amygdala, left	NR	15	4	24	15	78.9	61.5
					Imaging	MRI-amygdala, right	NR	12	7	28	11	63.2	71.8
Wood et al, 2016.[195]	Longitudinal cohort	MCI	15[1]	60.0	Imaging	MRI-HC, total	NR	6	1	4	2	85.7	66.7
					Biomarker	CSF t-tau/Aβ_{42}	0.821	6	1	6	0	85.7	100.0
					Cognitive	MMSE	27/30	4	5	2	4	44.4	33.3
					Cognitive	4 Mountains Test	8/15	9	0	5	1	100.0	83.3
					Cognitive	RAVLT	3	6	2	3	3	75.0	50.0
					Cognitive	Trail Making Test-B	102.6 seconds	5	3	6	0	62.5	100.0
Xu et al, 2002.[196]	Longitudinal cohort	MCI	351	13.4	Cognitive	MMSE	26/30	29	18	252	52	61.7	82.9
					Cognitive	CCSE	25/30	35	12	258	46	74.5	84.9

[1]Not all participants underwent all index tests.

Abbreviations:
^{11}C-PIB, ^{11}C-labelled Pittsburgh Compound B; ^{18}F-FDG, ^{18}flourine-fluorodeoxyglucose; Aβ, amyloid beta; ACE, Addenbrooke's Cognitive Examination; ADAS, Alzheimer's Disease Assessment Scale; ADD, Alzheimer's disease dementia; CCSE, Cognitive Capacity Screening Examination; CCT, cube copying test; CDR-SB, Clinical Dementia Rating 'sum of boxes'; CDT, clock drawing test; CSF, cerebrospinal fluid; CVLT-LDFR, California Verbal Learning Test - Long Delay Free Recall; FAQ, Functional Activities Questionnaire; fn, false negatives; fp, false positives; HFABP, heart fatty acid binding protein; IP, isoprostane; MCI, mild cognitive impairment; MIS, Memory Impairment Screen; MMSE, Mini Mental State Examination; MoCA, Montreal Cognitive Assessment; MRI-CGM, magnetic resonance imaging - cortical grey matter; MRI-ERC, magnetic resonance imaging - entorhinal cortex; MRI-HC, magnetic resonance imaging - hippocampus; MRI-LTL, magnetic resonance imaging - lateral temporal lobe; MRI-LV, magnetic resonance imaging - lateral ventricles; MRI-MTG, magnetic resonance imaging - medial temporal gyrus; MRI-MTL, magnetic resonance imaging - medial temporal lobe; MRI-OC, magnetic resonance imaging - occipital cortex; MRI-PC, magnetic resonance imaging - parietal cortex; MRI-WB, magnetic resonance imaging - whole brain; N, number of participants; NfH, neurofilament heavy chain isoform; NR, not reported; PET, positron emission tomography; pM, picomole; p-tau, phosphorylated tau; RAVLT, Rey Auditory Verbal Learning Test; rCGMr, regional cerebral glucose metabolism ratio; se, sensitivity; sp, specificity; SRT, Selective Reminding Test; SRT-DR, Selective Reminding Test - Delayed Recall; SRT-IR, Selective Reminding Test - Immediate Recall; SUVR, Standardised Uptake Volume ratio; tn, true negatives; tp, true positives; t-tau, total tau; UPSIT, University of Pennsylvania Smell Identification Test

4.4.2 Diagnostic accuracy data comparing two or more tests

One hundred and fifteen diagnostic accuracy studies for tests for Alzheimer's disease dementia were identified from the 13 systematic reviews (see the adapted PRISMA study flow diagram in Figure 4.1). After removing duplicates, 105 diagnostic accuracy studies were screened for the presence of comparative test accuracy data.

Of the 105 studies identified within the reviews, 70 compared the accuracy of two or more tests using the same patient group. No studies randomised patients to different testing arms. Table 4.2 summarises the study and test characteristics. Two hundred and fifty-one 2×2 tables, reported in Table 4.2, were extracted from the comparative studies. The majority (126 [50%]) of the 2×2 tables contained data on the accuracy of biomarkers; 89 (36%) contained imaging data and 36 (14%) contained cognitive data. The comparative diagnostic test accuracy studies evaluated a minimum of two and a maximum of nine tests.

Of the 70 comparative studies, 14 (20%) reported cross-classified data, extracted as 2×4 tables, on their combined test performance (see Table 4.3).

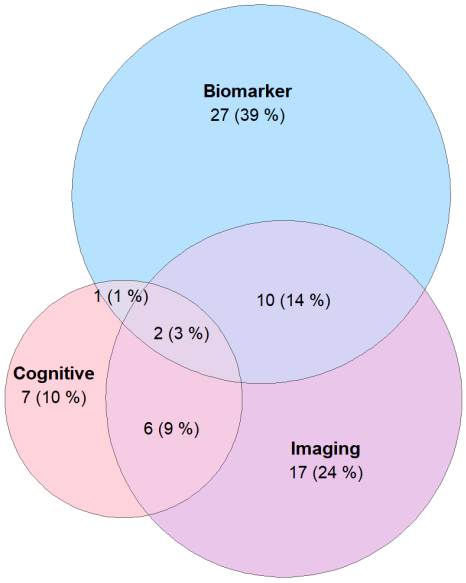


Figure 4.2: Diagram showing the frequencies of different test type comparisons.

Table 4.3: Fully cross-classified data. Column headings correspond with notation in Table 2.2 and 2.3.

Study	**Test 1**	**Test 2**	**Alzheimer's disease dementia**				**No Alzheimer's disease dementia**				se_1	se_2	se_{12}	sp_1	sp_2	sp_{12}
			x_{11}^D	x_{10}^D	x_{01}^D	x_{00}^D	$x_{11}^{\bar{D}}$	$x_{10}^{\bar{D}}$	$x_{01}^{\bar{D}}$	$x_{00}^{\bar{D}}$						
Anchisi et al, 2005.[127]	^{18}F-FDG PET	CVLT-LDTR	12	1	1	0	1	5	13	15	92.9	92.9	85.7	82.4	58.8	44.1
Hansson et al, 2006.[150]	CSF Aβ_{42}	CSF t-tau	54	2	1	0	13	14	17	33	98.2	96.5	94.7	64.9	61.0	42.9
	CSF Aβ_{42}	CSF p-tau	54	2	0	1	15	12	18	32	98.2	94.7	94.7	64.9	57.1	41.6
	CSF Aβ_{42}/p-tau	CSF t-tau	54	1	1	1	11	5	20	41	96.5	96.5	94.7	79.2	59.7	53.2
Hertze et al, 2010.[152]	CSF Aβ_{42}	CSF t-tau	46	1	4	1	20	12	46	29	90.4	96.2	88.5	70.1	38.3	27.1
Kapaki et al, 2003.[155]	CSF Aβ_{42}	CSF t-tau	27	8	8	6	0	3	1	11	71.4	71.4	55.1	80.0	93.3	73.3
Knapskog et al, 2019.[157]	CSF Aβ_{42}	CSF t-tau	40	19	50	29	5	7	10	45	42.8	65.2	29.0	82.1	77.6	67.2
	CSF Aβ_{42}	CSF p-tau	31	28	34	45	3	9	4	51	42.8	47.1	22.5	82.0	89.6	76.1
Lewczuk et al, 2004.[161]	CSF Aβ_{42}	CSF t-tau	15	4	3	0	0	2	3	6	86.4	81.8	68.2	81.8	72.7	54.5
	CSF Aβ_{40}	CSF t-tau	6	2	11	2	0	1	3	6	38.1	81.0	28.6	90.0	70.0	60.0
	CSF A$\beta_{42/40}$	CSF t-tau	16	4	1	0	2	0	1	7	95.2	81.0	76.2	80.0	70.0	70.0
Milan et al, 2012.[163]	MMSE	CDT	152	23	22	18	0	0	2	62	81.4	80.9	70.7	100.0	96.9	96.9
Montine et al, 2001.[167]	CSF Aβ_{42}	CSF IP	17	2	0	0	3	3	2	0	100.0	89.5	89.5	25.0	37.5	0.0
Ossenkoppele et al, 2012b.[173]	^{18}F-FDG PET	^{11}C-PIB PET	5	0	1	0	0	0	1	5	83.3	100.0	83.3	100.0	83.3	83.3
Pozueta et al, 2011.[178]	MMSE	CVLT-LDTR	22	10	15	3	6	5	10	34	64.0	74.0	44.0	80.0	70.9	61.8

Study	Test 1	Test 2	Alzheimer's disease dementia				No Alzheimer's disease dementia				se_1	se_2	se_{12}	sp_1	sp_2	sp_{12}
			x_{11}^D	x_{10}^D	x_{01}^D	x_{00}^D	$x_{11}^{\bar{D}}$	$x_{10}^{\bar{D}}$	$x_{01}^{\bar{D}}$	$x_{00}^{\bar{D}}$						
Prestia et al, 2013a.[179]	MRI-HC, total	^{18}F-FDG PET	15	1	8	5	7	3	5	29	55.2	79.3	51.7	77.3	72.7	65.9
	MRI-HC, total	CSF Aβ_{42}	11	5	10	3	3	7	12	22	55.2	72.4	37.9	77.3	65.9	50.0
	^{18}F-FDG PET	CSF Aβ_{42}	17	6	4	2	4	8	11	21	79.3	72.4	58.6	72.7	65.9	47.7
Rosler et al, 2001.[182]	CSF Aβ_{42}	CSF t-tau	20	2	4	1	4	6	4	10	81.5	88.9	74.1	58.3	66.7	41.7
Tapiola et al, 2000.[186]	CSF Aβ_{42}	CSF t-tau	42	13	8	17	4	7	5	11	68.8	62.5	52.5	59.3	66.7	40.7
Thurfjell et al, 2012.[188]	^{18}F-flutemetamol PET	MRI-HC, total	6	2	1	0	0	2	7	1	88.9	77.8	66.7	80.0	30.0	10.0

se_j, sensitivities; sp_j, specificities; $j = 1, 2$ tests; se_{12}, joint sensitivity; sp_{12}, joint specificity; x_{kl}^D, number of participants with the target condition with each combination of test results; $x_{kl}^{\bar{D}}$, number of participants without the target condition with each combination of test results; $k, l = 0, 1$, denote the first and second test, respectively, where 0 indicates a negative test result and 1 indicates a positive test result

^{11}C-PIB, ^{11}C-labelled Pittsburgh Compound B; ^{18}F-FDG, ^{18}flourine-fluorodeoxyglucose; Aβ, amyloid beta; CDT, clock drawing test; CSF, cerebrospinal fluid; CVLT-LDFR, California Verbal Learning Test - Long Delay Free Recall; IP, isoprostane; MMSE, Mini Mental State Examination; MRI-HC, magnetic resonance imaging - hippocampus; PET, positron emission tomography; p-tau, phosphorylated tau; t-tau, total tau

4.5 Discussion

In this literature review, 13 systematic reviews of diagnostic accuracy studies for tests for Alzheimer's disease dementia were identified, which included 105 primary studies that were further examined to identify those that evaluated the accuracy of multiple tests. Comparative accuracy data, where two or more tests are evaluated in the same patient group, were reported in 70 studies. In total, 251 2×2 tables of comparative accuracy data were extracted. The rate of reporting for cross-classified data was low, with 20% of included comparative diagnostic accuracy studies providing data on the concordance between tests in the form of 2×4 tables.

This review identified relevant systematic reviews through electronic searching of The Cochrane Library, considered to be of high methodological rigour. Cochrane reviews are a high quality source of evidence on healthcare interventions and diagnostic tests, requiring authors to undertake a thorough literature search, independent screening and data extraction of all studies performed in duplicate. The reviews covered a broad range of tests, including cognitive, imaging and CSF biomarker tests, primarily in populations with MCI. This literature review is novel in its objective to identify and extract cross-classified data from comparative test accuracy studies in Alzheimer's disease dementia. This greater level of granularity enables the estimation of within-study associations between multiple tests evaluated using a paired design. The data can be used to inform models that account for these dependencies, either by feeding the data into the model directly or by fixing a dependence parameter (e.g. Pearson's correlation, Kendall's tau, copula dependence parameter) used in the model.

The data collection methods used are not without limitations. By design, this literature review will not identify every diagnostic test accuracy study in Alzheimer's disease dementia. The literature review lacked the rigorous methodology of a systematic review, such as an exhaustive search strategy including grey literature sources, or the use of established guidelines for the performing (e.g. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy [4]) or reporting (e.g. PRISMA for Diagnostic Test Accuracy [114]) of systematic reviews of diagnostic test accuracy. Screening and data extraction were performed by one reviewer (single screening), rather than independently and in duplicate by two reviewers (double screening). Single screening of titles and abstracts has been shown to lead to false exclusions, with a methodological review finding a median proportion of missed studies of 5% for sin-

gle compared to double screening.[197] Some of the reviews were recently published, while the oldest review was published in 2014. It is likely that additional studies have been published since, which were potentially not included in the literature review. Data on novel blood-based biomarkers for Alzheimer's disease dementia were sparse; plasma Aβ_{42} and A$\beta_{42/40}$ were evaluated in a single comparative test accuracy study only. No studies were identified assessing the diagnostic accuracy of some cognitive tests recommended in the current NICE guidelines for Alzheimer's disease dementia, including the 10-point cognitive screener (10-CS), the 6-item cognitive impairment test (6CIT), the 6-item screener, and Test Your Memory (TYM).[36]

Diagnostic reviews in particular have been shown to entail a high reviewer workload,[198] and reporting quality of diagnostic studies is often poor,[199] impeding study screening, data extraction, and quality assessment. A comprehensive systematic review of this breadth is beyond the scope of this thesis. This literature review was undertaken to identify motivational data sets on which to base model evaluation and development; it will not, and indeed should not, be used to inform clinical decision-making in Alzheimer's disease dementia. Novel diagnostic tests for Alzheimer's disease dementia, such as plasma biomarkers, continue to emerge, and as such there is an ongoing need to evaluate both their accuracy compared to and combined with existing tests. Future research could include a full systematic review of comparative diagnostic studies in Alzheimer's disease dementia, using meta-analysis models proposed in Chapters 6 and 7 to synthesise the extracted evidence to model diagnostic pathways utilising novel plasma and CSF biomarkers.

4.6 Chapter summary

In this chapter, a literature review of comparative diagnostic accuracy studies in Alzheimer's disease dementia was performed. The literature review was not intended to be exhaustive; rather, the purpose of the review was to identify motivational data sets on which to base model evaluation and development in later chapters. A review of systematic reviews was undertaken to identify comparative accuracy studies containing study- or individual-level data on two or more tests assessed in the same patients. In the following chapter, the impact of accounting for within-study dependencies between two tests evaluated under a paired design in a meta-analysis framework is assessed, using simulations based on a motivational data set identified in this chapter. Novel meta-analysis models for jointly synthesising comparative ac-

curacy data are proposed in Chapters 6 and 7, using bivariate and trivariate copulas, respectively, to account for within-study dependencies. The models are fit to motivational examples in Alzheimer's disease dementia identified through this literature review to illustrate their use.

Chapter 5

Meta-analysis of two diagnostic tests: the effect of ignoring within-study association

5.1 Chapter overview

Chapter 3 included a methodological review of meta-analysis models for jointly synthesising accuracy data on two diagnostic tests evaluated on the same individuals. A number of these methods focussed on modelling 2×2 data for each test, meaning that within-study associations present between the two tests cannot be accounted for. The impact of modelling these associations on key test accuracy parameters has not yet been evaluated. In this chapter, the performance of a meta-regression model currently recommended in evidence synthesis guidelines is assessed through a simulation study. This method uses 2×2 data for each index test against a common reference standard, treating the two tests as independent and thus ignoring within-study associations. By fitting the model to simulated data with known within-study dependencies, based on a motivating example extracted as part of the literature review conducted in the previous chapter, the impact of ignoring these associations is measured. The findings of this study will further motivate novel model development for meta-analysing data on two tests in Chapters 6 and 7.

5.2 Introduction

In a healthcare decision-making context, we are often interested in simultaneously evaluating the accuracy of two or more diagnostic tests. Joint meta-analysis models for two or more tests answer clinically relevant questions that assessing tests in isolation do not, enabling comparison of the accuracy of multiple tests for the same condition, as well as evaluation of testing strategies or diagnostic pathways comprising of multiple tests. Several methods for the synthesis of diagnostic accuracy data for two or more tests in a meta-analysis framework have been proposed in recent years, as summarised in Section 2.4.3. However, despite calls for their evaluation,[12, 4, 10] there has been no assessment of these approaches so far.

Evaluating test accuracy within a meta-analysis framework makes the best use of the available evidence by combining all relevant published studies in a single analysis. Meta-analyses of diagnostic accuracy studies presents different challenges compared to meta-analyses of randomised controlled trials (RCTs), in part due to additional dependence structures present in the data. As introduced in Section 2.4.2 and 2.4.3, there are two sources of association present when meta-analysing diagnostic accuracy data on two tests assessed using a within-subject design: between-study and within-study dependencies. Diagnostic accuracy studies that compare two tests under a paired design may report either 2×2 data on each test (Table 2.2), where the overlap between results on the first and second test is unknown, or cross-classified data (Table 2.3), which contain all possible combinations of test results compared to true disease status. The latter enables inference on the agreement between the tests, through estimation of measures such as joint sensitivity and specificity.

In the recently published Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy,[4] a meta-regression approach is recommended to compare the accuracy of two diagnostic tests compared to a common reference standard, extending meta-analysis models for single tests to incorporate test type as a covariate. Results of two tests evaluated under a paired design are treated as independent, meaning the model fails to account for within-study dependencies between tests. Further development of guidance for analysing data from paired diagnostic accuracy studies is needed, along with investigation into the impact of ignoring within-study associations on test accuracy parameters.[15]. Based on literature in the field of multivariate meta-analysis,[81, 85] summary estimates of sensitivity and specificity are likely to be relatively robust to the assumption of test independence. Measures of joint test

accuracy (such as joint sensitivity and specificity) may be subject to substantial bias, however.[15]

In this chapter, a meta-regression model for synthesising diagnostic accuracy studies that compare two tests, assessed in the same individuals, against a common reference standard is evaluated. The method treats the tests as independent, using binomial likelihoods to model 2×2 data on each test, and does not account for within-study associations present between them. The effect of ignoring within-study dependencies on estimates of marginal and joint test accuracy measures is assessed through a simulation study. Diagnostic accuracy data on two tests are generated under various realistic scenarios with known within-study associations, based on a motivational example in Alzheimer's disease dementia. The model is fit to the simulated data sets and key performance characteristics are calculated to estimate bias and variability. Section 5.3 describes the meta-regression model for combining comparative diagnostic accuracy studies that will be used to assess the impact of accounting for dependencies between two tests evaluated using a paired design. The simulation study methodology is also outlined. The results of the simulation study are presented in Section 5.4. Section 5.5 concludes the chapter with a discussion.

5.3 Methods

In this section, the methods for the simulation study are described. The meta-regression model for synthesising diagnostic accuracy studies comparing two tests to a common reference standard is described in Section 5.3.1. The meta-analysis model for the joint synthesis of two diagnostic test, capturing within-study dependencies using multinomial likelihoods, from which the diagnostic accuracy data are simulated is outlined in Section 5.3.2. The motivating example in Alzheimer's disease dementia, on which the simulated data are based, is presented in Section 5.3.3, followed by the simulation methodology, including the data generating mechanism and calculation of performance statistics, in Section 5.3.4.

5.3.1 Meta-regression model with test type as a covariate

The BRMA model, described in Section 2.4.2.1, synthesises diagnostic accuracy data on a single test compared to a common reference standard.[75, 76] The number of true positive and true negative results are assumed to follow independent binomial

distributions; at the between-studies level, *logit*-transformed sensitivities and specificities are jointly modelled by a bivariate normal distribution. To evaluate the diagnostic accuracy of two tests assessed against a common reference standard using a paired design, current evidence synthesis guidelines advise extending the BRMA model to incorporate test type as a binary covariate (i.e. meta-regression), allowing the comparison of summary sensitivities and specificities of two tests.[4]

This model requires study-level 2×2 data on each test compared to a common reference standard (Table 2.2). Consider a meta-analysis of $i = 1, \ldots, I$ comparative diagnostic accuracy studies in which all patients undergo both tests. In a meta-regression with test type as a covariate, the tests are treated as independent, and, as such, within-study dependencies between tests are ignored. The number of true positives ($tp_{i,j}$) and true negatives ($tn_{i,j}$) for the $j = 1, 2$ tests follow independent binomial distributions:

$$tp_{i,1} \sim Binomial(se_{i,1}, N_i^D), \quad tp_{i,2} \sim Binomial(se_{i,2}, N_i^D), \tag{5.1}$$

$$tn_{i,1} \sim Binomial(sp_{i,1}, N_i^{\bar{D}}), \quad tn_{i,2} \sim Binomial(sp_{i,2}, N_i^{\bar{D}})$$

$se_{i,j}$ and $sp_{i,j}$ are the sensitivity and specificity of the j^{th} test in the i^{th} study, and N_i^D and $N_i^{\bar{D}}$ the number of patients with and without the target condition, respectively. To assess the accuracy of two diagnostic tests, two pairs of *logit*-transformed study-specific sensitivities ($\mu_{i,sej}$) and specificities ($\mu_{i,spj}$) are modelled. Therefore, the bivariate normal distribution at the between-studies level of the BRMA model for single tests (Equation 2.19, Section 2.4.2.1) is replaced by a four-dimensional multivariate normal distribution. By jointly modelling the pairs of sensitivities and specificities, between-studies correlation between the measures induced by differences in study characteristics is accounted for:

$$logit\,(se_{i,1}) = \mu_{i,se1}, \quad logit\,(se_{i,2}) = \mu_{i,se2}, \tag{5.2}$$

$$logit\,(sp_{i,1}) = \mu_{i,sp1}, \quad logit\,(sp_{i,2}) = \mu_{i,sp2},$$

$$\begin{pmatrix} \mu_{i,se1} \\ \mu_{i,se2} \\ \mu_{i,sp1} \\ \mu_{i,sp2} \end{pmatrix} \sim Normal \left(\begin{pmatrix} \mu_{se1} \\ \mu_{se2} \\ \mu_{sp1} \\ \mu_{sp2} \end{pmatrix}, \quad \boldsymbol{\Sigma} \right)$$

$\boldsymbol{\Sigma}$ is the unstructured between-study covariance matrix. A common between-studies correlation parameter, ρ_b, may be estimated across pairs of sensitivities and specificities to minimise the number of model parameters and reduce the likelihood of non-convergence:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2_{se1} & \rho_b\sigma_{se1}\sigma_{se2} & \rho_b\sigma_{se1}\sigma_{sp1} & \rho_b\sigma_{se1}\sigma_{sp2} \\ & \sigma^2_{se2} & \rho_b\sigma_{se2}\sigma_{sp1} & \rho_b\sigma_{se2}\sigma_{sp2} \\ & & \sigma^2_{sp1} & \rho_b\sigma_{sp1}\sigma_{sp2} \\ & & & \sigma^2_{sp2} \end{pmatrix} \tag{5.3}$$

Alternatively, a separate between-studies correlation parameter for each test may be specified, enabling estimation of between-study and between-test variability [4]:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2_{se1} & \rho_{se1se2}\sigma_{se1}\sigma_{se2} & \rho_{se1sp1}\sigma_{se1}\sigma_{sp1} & \rho_{se1sp2}\sigma_{se1}\sigma_{sp2} \\ & \sigma^2_{se2} & \rho_{se2sp1}\sigma_{se2}\sigma_{sp1} & \rho_{se2sp2}\sigma_{se2}\sigma_{sp2} \\ & & \sigma^2_{sp1} & \rho_{sp1sp2}\sigma_{sp1}\sigma_{sp2} \\ & & & \sigma^2_{sp2} \end{pmatrix} \tag{5.4}$$

Post-sampling, summary sensitivities and specificities are derived through an inverse-*logit* transformation:

$$\nu_{se1} = logit^{-1}\left(\mu_{se1}\right), \quad \nu_{se2} = logit^{-1}\left(\mu_{se2}\right), \tag{5.5}$$

$$\nu_{sp1} = logit^{-1}\left(\mu_{sp1}\right), \quad \nu_{sp2} = logit^{-1}\left(\mu_{sp2}\right)$$

The meta-regression model assumes independence between tests. Therefore, joint sensitivity (ν_{se12}) and specificity (ν_{sp12}) are estimated by multiplying the marginal sensitivities and specificities:

$$\nu_{se12} = \nu_{se1} \times \nu_{se2} \tag{5.6}$$

$$\nu_{sp12} = \nu_{sp1} \times \nu_{sp2}$$

Prior distributions were placed on the unknown parameters in the model. *Logit*-transformed summary sensitivities and specificities were assumed to follow a minimally informative $Normal\left(0, 10^2\right)$ prior distribution. The between-studies variance parameters were restricted to positive values and were assumed to have a uniform $Half - Normal\left(0, 2.5^2\right)$ prior distribution. For the between-studies correlation parameter, the Fisher z-transformation was used, i.e $\rho_b = \tanh\left(z\right), z \sim Normal\left(0, 0.8\right)$). This transformation produces an approximately normal distribution bound between $[-1, 1]$.

5.3.2 Meta-analysis model with multinomial likelihoods

Trikalinos et al [85] proposed an alternative extension to the BRMA model, using multinomial likelihoods to jointly model the accuracy of two diagnostic tests evaluated against a common reference standard using a within-subject design. The method requires individual-level, cross-classified data (Table 2.3) from all studies and captures within-study dependencies between sensitivities and specificities through estimating joint sensitivity and specificity parameters.

As before, let there be $i = 1, ..., I$ comparative diagnostic accuracy studies in which all patients undergo $j = 1, 2$ tests. Unlike the meta-regression model (described in Section 5.3.1), cross-classified data are required for each study. Notation for the probability of each possible combination of test results for populations with and without the target condition for study i are presented in Table 5.1.

Multinomial likelihoods are used to model the observed cross-classified results of the two tests. Notation for the cross-classified counts is presented in Table 2.2. $\left(\pi^D_{11}, \pi^D_{10}, \pi^D_{01}, \pi^D_{00}\right)^T$ and $\left(\pi^{\bar{D}}_{11}, \pi^{\bar{D}}_{10}, \pi^{\bar{D}}_{01}, \pi^{\bar{D}}_{00}\right)^T$ denote the vectors containing the probability of each combination of test results for groups with and without the target condition, respectively, where T is the transpose of the vector. Notation for the probabilities of possible combinations of test results against the reference standard

Table 5.1: Notation for the probability of fully cross-classified diagnostic accuracy data for two tests for a single study i. All individuals undergo both tests plus a reference standard. Adapted from Trikalinos et al [85]

	Disease			**No disease**		
	Test 2 +	**Test 2 −**	Total	**Test 2 +**	**Test 2 −**	Total
Test 1 +	$\pi^{D}_{i,11}$	$\pi^{D}_{i,10}$	$tp_{i,1}$	$\pi^{\bar{D}}_{i,11}$	$\pi^{\bar{D}}_{i,10}$	$fp_{i,1}$
Test 1 −	$\pi^{D}_{i,01}$	$\pi^{D}_{i,00}$	$fn_{i,1}$	$\pi^{\bar{D}}_{i,01}$	$\pi^{\bar{D}}_{i,00}$	$tn_{i,1}$
Total	$tp_{i,2}$	$fn_{i,2}$	N^{D}_{i}	$fp_{i,2}$	$tn_{i,2}$	$N^{\bar{D}}_{i}$

$tp_{i,j}$, true positives; $fn_{i,j}$, false negatives; $tn_{i,j}$, true negatives; $fp_{i,j}$, false positives; $i = 1, ..., I$, $j = 1, 2$ tests; $x^{D}_{i,kl}$, number of participants with the target condition with each combination of test results; $x^{\bar{D}}_{i,kl}$, number of participants without the target condition with each combination of test results; $k, l = 0, 1$, denote the first and second test, respectively, where 0 indicates a negative test result and 1 indicates a positive test result; N^{D}_{i}, total diseased; $N^{\bar{D}}_{i}$, total non-diseased

is displayed in Table 5.1.

$$\begin{pmatrix} x^{D}_{i,11} \\ x^{D}_{i,10} \\ x^{D}_{i,01} \\ x^{D}_{i,00} \end{pmatrix} \sim Multinomial \left(\begin{pmatrix} \pi^{D}_{i,11} \\ \pi^{D}_{i,10} \\ \pi^{D}_{i,01} \\ \pi^{D}_{i,00} \end{pmatrix}, \quad N^{D}_{i} \right) \tag{5.7}$$

$$\begin{pmatrix} x^{\bar{D}}_{i,11} \\ x^{\bar{D}}_{i,10} \\ x^{\bar{D}}_{i,01} \\ x^{\bar{D}}_{i,00} \end{pmatrix} \sim Multinomial \left(\begin{pmatrix} \pi^{\bar{D}}_{i,11} \\ \pi^{\bar{D}}_{i,10} \\ \pi^{\bar{D}}_{i,01} \\ \pi^{\bar{D}}_{i,00} \end{pmatrix}, \quad N^{\bar{D}}_{i} \right)$$

The marginal sensitivity and specificity of each test can be expressed in terms of the following probabilities, where $\cdot$ indicates summation over the respective subscript:

$$se_{i,1} = \pi^{D}_{i,1\cdot} = \pi^{D}_{i,10} + \pi^{D}_{i,11} \tag{5.8}$$

$$se_{i,2} = \pi^{D}_{i,\cdot 1} = \pi^{D}_{i,01} + \pi^{D}_{i,11} \tag{5.9}$$

$$sp_{i,1} = \pi^{\bar{D}}_{i,0\cdot} = \pi^{\bar{D}}_{i,01} + \pi^{\bar{D}}_{i,00} \tag{5.10}$$

$$sp_{i,2} = \pi^{\bar{D}}_{i,\cdot 0} = \pi^{\bar{D}}_{i,10} + \pi^{\bar{D}}_{i,00} \tag{5.11}$$

The joint sensitivity and joint specificity are also incorporated into the model to measure the agreement between tests:

$$se_{i,12} = \pi^{D}_{i,11} \tag{5.12}$$

$$sp_{i,12} = \pi^{\bar{D}}_{i,00} \tag{5.13}$$

At the between-studies level, the *logit*-transformed marginal and joint test accuracy measures are modelled by a six-dimensional normal distribution with mean equal to the summary estimates of the marginal and joint sensitivities and specificities, and between-study covariance matrix $\mathbf{\Phi}$:

$$logit\left(se_{i,1}\right) = \mu_{i,se1}, \quad logit\left(se_{i,2}\right) = \mu_{i,se2}, \quad logit\left(se_{i,12}\right) = \mu_{i,se,12}, \tag{5.14}$$

$$logit\left(sp_{i,1}\right) = \mu_{i,sp1}, \quad logit\left(sp_{i,2}\right) = \mu_{i,sp2}, \quad logit\left(sp_{i,12}\right) = \mu_{i,sp,12},$$

$$\begin{pmatrix} \mu_{i,se1} \\ \mu_{i,se2} \\ \mu_{i,se12} \\ \mu_{i,sp1} \\ \mu_{i,sp2} \\ \mu_{i,sp12} \end{pmatrix} \sim Normal \left(\begin{pmatrix} \mu_{se1} \\ \mu_{se2} \\ \mu_{se12} \\ \mu_{sp1} \\ \mu_{sp2} \\ \mu_{sp12} \end{pmatrix}, \quad \mathbf{\Phi} \right)$$

$\mathbf{\Phi}$ is the unstructured between-study covariance matrix. $\mu_{i,se1}$, $\mu_{i,se2}$, $\mu_{i,sp1}$ and $\mu_{i,sp2}$ denote the *logit*-transformed study-specific marginal sensitivities and specificities, while $\mu_{i,se12}$ and $\mu_{i,sp12}$ denote the *logit*-transformed joint sensitivity and joint specificity. $\mathbf{\Phi}$ is parametrised as follows, where σ_{se1}, σ_{se2}, σ_{se12}, σ_{sp1}, σ_{sp2}, and σ_{sp12} are the standard deviations in *logit*-transformed sensitivities and specificities and ρ_b is

the between-studies correlation parameter:

$$\mathbf{\Phi} = \begin{pmatrix} \sigma^2_{se1} & \rho_b\sigma_{se1}\sigma_{se2} & \rho_b\sigma_{se1}\sigma_{se12} & \rho_b\sigma_{se1}\sigma_{sp1} & \rho_b\sigma_{se1}\sigma_{sp2} & \rho_b\sigma_{se1}\sigma_{sp12} \\ & \sigma^2_{se2} & \rho_b\sigma_{se2}\sigma_{se12} & \rho_b\sigma_{se2}\sigma_{sp1} & \rho_b\sigma_{se2}\sigma_{sp2} & \rho_b\sigma_{se2}\sigma_{sp12} \\ & & \sigma^2_{se12} & \rho_b\sigma_{se12}\sigma_{sp1} & \rho_b\sigma_{se12}\sigma_{sp2} & \rho_b\sigma_{se12}\sigma_{sp12} \\ & & & \sigma^2_{sp1} & \rho_b\sigma_{sp1}\sigma_{sp2} & \rho_b\sigma_{sp1}\sigma_{sp12} \\ & & & & \sigma^2_{sp2} & \rho_b\sigma_{sp2}\sigma_{sp12} \\ & & & & & \sigma^2_{sp12} \end{pmatrix} \tag{5.15}$$

Post-sampling, summary test accuracy measures are derived through an inverse-*logit* transformation. Both marginal and joint sensitivities and specificities are estimated directly within the model.

$$\nu_{se1} = logit^{-1}\left(\mu_{se1}\right), \quad \nu_{se2} = logit^{-1}\left(\mu_{se2}\right), \quad \nu_{se12} = logit^{-1}\left(\mu_{se12}\right), \tag{5.16}$$

$$\nu_{sp1} = logit^{-1}\left(\mu_{sp1}\right), \quad \nu_{sp2} = logit^{-1}\left(\mu_{sp2}\right), \quad \nu_{sp12} = logit^{-1}\left(\mu_{sp12}\right)$$

Prior distributions are placed on the unknown parameters in the model. *Logit*-transformed summary marginal and joint sensitivities and specificities were assumed to follow a minimally informative $Normal\left(0, 10^2\right)$ prior distribution. The between-studies variance parameters were restricted to positive values and were assumed to have a uniform $Half - Normal\left(0, 2.5^2\right)$ prior distribution. For the between-studies correlation parameter, the Fisher z-transformation was used, i.e $\rho_b = \tanh\left(z\right), z \sim Normal\left(0, 0.8\right)$). This transformation produces an approximately normal distribution bound between $[-1, 1]$.

5.3.3 Motivational example

Data collection methods are described in detail in Chapter 4. From the 70 comparative diagnostic accuracy studies identified through the literature review, a motivating example was selected that maximised the availability of comparative data. Priority was given to test comparisons supported by the largest number of studies and greatest number of total participants, as well as ensuring cross-classified data were reported for at least one study. The selected data set compares Aβ42 and t-tau, measured in the CSF, on which 36 2×2 tables from 18 studies were extracted. Two studies contained cross-classified data and the remaining 16 reported 2×2 data only

(Table 5.2).

Table 5.2: Data on the accuracy of amyloid-β 42 (Aβ_{42}, test 1) and total tau (t-tau, test 2) for diagnosing Alzheimer's disease dementia. The table contains 2×2 data for each test within each study, and fully-cross classified data where available.

Study	**Alzheimer's disease dementia**									**No Alzheimer's disease dementia**								
	tp_1	tp_2	fn_1	fn_2	x_{11}^D	x_{10}^D	x_{01}^D	x_{00}^D	N^D	tn_1	tn_2	fp_1	fp_2	$x_{11}^{\bar{D}}$	$x_{10}^{\bar{D}}$	$x_{01}^{\bar{D}}$	$x_{00}^{\bar{D}}$	$N^{\bar{D}}$
Bjerke et al, 2009.[128]	18	12	2	8	-	-	-	-	20	99	130	43	12	-	-	-	-	142
Blom et al, 2009.[129]	9	7	5	7	-	-	-	-	14	5	11	9	3	-	-	-	-	14
Chiasserini et al, 2010.[136]	17	12	6	11	-	-	-	-	23	16	16	2	2	-	-	-	-	18
Frölich et al, 2017.[145]	17	24	11	4	-	-	-	-	28	65	46	22	41	-	-	-	-	87
Gaser et al, 2013.[148]	59	58	7	8	-	-	-	-	66	12	13	21	20	-	-	-	-	33
Hampel et al, 2004.[149]	24	26	5	3	-	-	-	-	29	13	11	10	12	-	-	-	-	23
Hansson et al, 2006.[150]	56	55	1	2	54	2	1	0	57	50	47	27	30	13	14	17	33	77
Hertze et al, 2010.[152]	47	38	5	14	46	1	4	1	52	75	82	32	25	20	12	46	29	107
Herukka et al, 2008.[153]	6	6	2	2	-	-	-	-	8	11	7	2	6	-	-	-	-	13
Kester et al, 2011.[156]	32	35	10	7	-	-	-	-	42	42	29	16	29	-	-	-	-	58
Monge-Argilés et al, 2011.[166]	9	8	2	3	-	-	-	-	11	16	18	10	8	-	-	-	-	26
Nesteruk et al, 2016.[169]	7	6	2	3	-	-	-	-	9	18	20	13	11	-	-	-	-	31
Palmqvist et al, 2012.[174]	47	42	5	10	-	-	-	-	52	56	58	25	23	-	-	-	-	81
Parnetti et al, 2006.[175]	4	5	7	6	-	-	-	-	11	30	32	3	1	-	-	-	-	33
Parnetti et al, 2012.[176]	18	20	14	12	-	-	-	-	32	56	51	2	7	-	-	-	-	58
Prestia et al, 2013b.[180]	17	11	1	7	-	-	-	-	18	9	15	9	3	-	-	-	-	18
Rhodius-Meester et al, 2016.[181]	58	72	27	13	-	-	-	-	85	38	35	14	17	-	-	-	-	52
Vos et al, 2013.[193]	62	65	29	26	-	-	-	-	91	82	95	41	28	-	-	-	-	123

tp_j is the number of true positives, fn_j false negatives, tn_j true negatives, and fp_j false positives for each of the $j = 1, 2$ tests.
x_{kl}^D is the number of participants with the target condition and $x_{kl}^{\bar{D}}$ is the number of participants without the target condition with each combination of test results. $k, l = 0, 1$ denote the first and second test, respectively, where 0 indicates a negative test result and 1 indicates a positive test result.

5.3.4 Simulation study

5.3.4.1 Data generating mechanism

Data were simulated under nine scenarios. Based on the meta-analysis model for two diagnostic tests developed by Trikalinos et al,[85] fully cross-classified data were simulated under a multinomial distribution to jointly model the number of patients with each possible combination of test results (see Step 6 of the data generation process below). Separate data sets were generated for the distinct groups of patients with and without the target condition. The within-study associations were varied through the joint sensitivity and specificity parameters. The impact of the within-study dependencies on model performance is likely to depend on the strength of the associations. Therefore, data were simulated assuming strong, moderate and weak within-study associations (Step 6). To explore the effect of the magnitude of the marginal sensitivities and specificities on model performance, three sets of scenarios were further considered for each within-study association: one where sensitivities and specificities were both high (0.8; chosen from the maximum mean sensitivities and specificities of Aβ_{42} and t-tau in the motivating example), a second where sensitivities and specificities were both low (0.4; chosen as half of the maximum mean), and a third where sensitivities were high (0.8) and specificities were low (0.4). Strong associations represented perfect dependence between the tests (e.g. $se_{12} = se_1 = se_2 = 0.8$), weak associations represented no dependence ($se_{12} = se_1 \times se_2 = 0.64$), and moderate association was chosen as the midpoint between strong and weak association. Macaskill et al demonstrated that joint sensitivity of the combined tests will be no more than the lower sensitivity of the component tests.[200] Therefore, joint sensitivity and specificity are bounded above by the smallest of the marginal sensitivities and specificities, i.e. $se_{12} \leq min\,(se_1, se_2)$ and $sp_{12} \leq min\,(sp_1, sp_2)$, respectively. Additionally, a lower bound was chosen that represents independence between the two tests, i.e. $se_1 \times se_2 = se_{12}$ and $sp_1 \times sp_2 = sp_{12}$. Within-study associations were selected with respect to these constraints (Table 5.3).

Table 5.3: Nine scenarios from which data were simulated from. 1000 data sets were simulated per scenario

Scenarios	Within-study associations		
	Strong	**Moderate**	**Weak**
High sensitivities, high specificities	$se_j = se_{12} = 0.8$ $sp_j = sp_{12} = 0.8$	$se_j = 0.8, se_{12} = 0.7$ $sp_j = 0.8, sp_{12} = 0.7$	$se_j = 0.8, se_{12} = 0.64$ $sp_j = 0.8, sp_{12} = 0.64$
Low sensitivities, low specificities	$se_j = se_{12} = 0.4$ $sp_j = sp_{12} = 0.4$	$se_j = 0.4, se_{12} = 0.3$ $sp_j = 0.4, sp_{12} = 0.3$	$se_j = 0.4, se_{12} = 0.16$ $sp_j = 0.4, sp_{12} = 0.16$
High sensitivities, low specificities	$se_j = se_{12} = 0.8$ $sp_j = sp_{12} = 0.4$	$se_j = 0.8, se_{12} = 0.7$ $sp_j = 0.4, sp_{12} = 0.3$	$se_j = 0.8, se_{12} = 0.64$ $sp_j = 0.4, sp_{12} = 0.16$

se_j and sp_j are the marginal sensitivities and specificities of the $j = 1, 2$ tests. $se_{1,2}$ and sp_{12} are the joint sensitivity and specificity, respectively.

The data generation process was as follows:

1. Set the number of studies ($N = 18$), based the motivating example comparing CSF Aβ_{42} and t-tau (Table 5.2).

2. Simulate the prevalence of the target condition using a uniform distribution for each of the i studies. A $Uniform\,(0.30, 0.50)$ distribution was used, with the upper and lower bounds corresponding to the upper 75^{th} and lower 25^{th} quartiles estimated from the prevalence of Alzheimer's disease dementia in the studies included in the motivational example.

3. Simulate the number of patients in each study using a uniform distribution, rounding to the nearest integer. A $Uniform\,(40, 134)$ distribution was used, corresponding to the upper and lower quartiles estimated from the number of patients per study in the motivational example.

4. Calculate the number of patients with and without the target condition in each study based on the prevalence and the total number of patients in the study.

5. At the between-studies level, simulate *logit*-transformed marginal and joint sensitivities and specificities in each study using a six-dimensional normal distribution (Equation 5.15). Between-studies correlation was fixed as -0.15, estimated using the motivating example. Between-study standard deviations were fixed as 0.2; heterogeneity was reduced compared to the motivating example to prevent sampling at extreme values.

6. At the within-study level, simulate the number of patients with each possible combination of results across the two tests using a multinomial distribution for patients with and without the target condition (Equation 5.7). Marginal and joint accuracy measures estimated in the previous step were used to populate the probabilities of combinations of test results in the diseased and non-diseased patient groups.

7. For the independent binomial likelihoods model, sum across the full cross-classifications to obtain 2×2 tables of diagnostic accuracy data for each test.

To investigate the impact of ignoring within-study dependencies, the meta-regression model with a common between-studies correlation parameter (described in Sections 5.3.1) was fit to each of the simulated data sets across the nine scenarios. The posterior median and 95% CrI of key test accuracy parameters, such as marginal sensitivities and specificities, joint sensitivity and specificity, were monitored across the 1000 simulations. Using these estimates and the known truths from which the data were simulated, summary performance measures (detailed in Section 5.3.4.2) were produced.

5.3.4.2 Performance measures

Model performance was assessed on coverage, bias, percentage bias and root mean square error (RMSE). The measures were calculated for the marginal and joint sensitivities and specificities, which were monitored during the simulation, to assess the impact of ignoring within-study dependencies on key test accuracy parameters used in clinical decision-making.

The coverage of a confidence interval is the proportion of simulations for which the estimated confidence interval contains the true parameter value:

$$\text{Coverage} = \hat{\theta}_m \pm Z_{1-\alpha/2} SE\left(\hat{\theta}_m\right) \tag{5.17}$$

$\hat{\theta_m}$ is the summary estimate of a parameter drawn from the m^{th} simulation, $SE\left(\hat{\theta_m}\right)$ is its standard error and $Z_{1-\alpha/2}$ is the $1-\alpha/2^{th}$ quantile of the standard normal distribution. Assuming the samples are approximately normally distributed, the coverage should approximately equal the nominal coverage percentage, 95%. Where coverage rates are greater than 95%, known as over-coverage, this indicates variability in the parameter estimates is too large and the estimates are too conservative. Where

coverage rates are less than 95%, known as under-coverage, this suggests variability is too small and there is overconfidence in the estimates.[201] Coverage can be difficult to interpret if the parameters are biased, as the true value is less likely to be contained in the confidence interval. Poor coverage may arise due to bias in the estimates drawn from the simulation study, bias in the model variability parameters, poor distributional properties or a combination of these factors.

Bias measures the deviation in the estimate from the true parameter value, allowing inference on model performance. For $m = 1, ..., M$ simulations, bias is calculated as the average difference between summary estimate drawn from the simulation study, $\hat{\theta}$, and the true parameter value, θ,[201]:

$$\begin{aligned} \text{Bias} &= \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i - \theta \\ &= \bar{\hat{\theta}} - \theta \end{aligned} \tag{5.18}$$

Provided that the true parameter value does not equal 0, percentage bias is further calculated as:

$$\%\ \text{Bias} = \left(\frac{\bar{\hat{\theta}} - \theta}{\theta}\right) \times 100 \tag{5.19}$$

The RMSE assesses the accuracy, or variability of the estimates by incorporating both bias and variability into a single measure [201]:

$$\text{RMSE} = \sqrt{\left(\bar{\hat{\theta}} - \theta\right)^2 + SE\left(\hat{\theta}\right)^2} \tag{5.20}$$

where $SE\left(\hat{\theta}\right)$ is the standard error of the estimate $\hat{\theta}$.

5.3.4.3 Sample size

To ensure the simulation study is adequately powered, a sample size must be chosen that achieves an acceptable Monte Carlo standard error (a measure of uncertainty due to using a finite number of repetitions) for key performance measures.[202] Based on bias being the performance measure of interest, and assuming a variance of 0.04, 1000 simulations per scenario lead to a Monte Carlo standard error of 0.006. As this is below 0.01, it is deemed to be satisfactory and the simulation study proceeds generating 1000 simulated data sets for each scenario.

5.3.4.4 Estimation

The model was implemented in a Bayesian framework using Stan version 2.32.2 [53] within R version 4.3.1 [203] via the rstan version 2.32.5 package.[54] Stan, a Bayesian sampler for performing Hamiltonian Monte Carlo simulation, is particularly computationally efficient and stable when parameters are highly correlated in the posterior distribution – as in diagnostic meta-analysis. A non-centred parameterisation of the model was used to reduce dependencies between successive levels of the hierarchical structure, further increasing the efficiency of the sampler (see Section 2.2.3.4).[55] After discarding 1,000 burn-in iterations, posterior estimates were obtained using three chains initialised at different starting values, consisting of 5,000 iterations each. The three chains were initialised at different starting values to aid the evaluation of model convergence. 95% CrIs were computed as highest posterior density intervals.

Model convergence was assessed using diagnostics described in Section 2.2.4, including trace plots, density plots, and autocorrelation plots. Stan code for the meta-regression model is provided in Appendix B.1; Stan code for the multinomial model - from which the data are simulated, is also provided in B.2 for reference. R code for the simulation study is available in Appendix B.3.

5.4 Results

5.4.1 Marginal sensitivities and specificities

Table 5.4, 5.5 and 5.6 display the performance measures for marginal and joint test accuracy parameters across the 9 scenarios, averaged over 1000 simulations per scenario. The meta-regression model performed well regardless of the strength of the within-study associations, achieving 0 bias, approximately 95% coverage probabilities and low RMSE across all scenarios. The magnitude of the marginal sensitivities and specificities did not impact model performance, with consistently accurate estimates and acceptable coverage probabilities across all combinations. Overall, there was no observable effect of ignoring within-study associations between the two tests on the marginal sensitivities and specificities, with the model achieving high performance across all scenarios.

5.4.2 Joint sensitivity and specificity

Figure 6.2 and 6.3 present the coverage, bias and RMSEs of joint sensitivity and specificity estimates across the nine scenarios. In the scenarios where within-study associations were low, equivalent to independence between tests, the meta-regression model accurate estimates of joint sensitivity and specificity, with 0 bias, approximately 95% coverage and low RMSE. There was no difference in model performance across the three scenarios simulated using different marginal sensitivities and specificities.

When within-study associations were moderate or strong, the meta-regression model resulted in downwardly biased estimates of joint sensitivity and specificity, with higher RMSEs and under-coverage. The magnitude of the marginal sensitivities and specificities impacted model performance, with increased bias and RMSEs but lower coverage for joint test accuracy measures when marginal accuracies low compared to when marginal accuracies were high. When sensitivities and specificities were low and within-study associations were strong or moderate coverage was 0% for both joint sensitivity and specificity, indicating that none of the estimated confidence intervals across the 1000 simulated data sets per scenario contained the true parameter value. When sensitivities and specificities were high and within-study associations were strong coverage was 0% and 0% for joint sensitivity and specificity, respectively, rising to 32% and 14% when within-study associations were moderate. The difference in coverage of joint sensitivity is likely explained by the prevalence of the target condition, which leads to a greater number of patients in the non-diseased than in the diseased group. The larger sample of patients without the target condition increases the precision of the joint test accuracy estimate, leading to a narrower confidence interval and reduced coverage.

Percentage bias, presented in Table 5.4-5.6, reflects these findings. Stronger within-study associations lead to increased bias in the joint sensitivity and specificity estimates. The lower the marginal sensitivities and specificities, the greater the bias. When sensitivities and specificities were high and within-study associations were moderate, percentage bias for joint sensitivity and specificity were 8% and 9%, respectively, increasing to 20% and 20% when within-study associations were strong. When sensitivities and specificities were low this effect was exacerbated, with a percentage bias for of 47% for both joint accuracy measures when within-study associations were moderate, rising to 60% when within-study associations were strong.

Table 5.4: Performance measures for key test accuracy parameters across the 3 scenarios where sensitivities and specificities are both high, averaged over 1000 simulations per scenario.

Parameter	% Coverage	Bias	% Bias	RMSE
Scenario 1: high sensitivities and specificities with strong within-study associations, $se_j = sp_j = 0.8, se_{12} = sp_{12} = 0.8$				
se_1	95.9	0.002	0.2	0.019
se_2	95.9	0.002	0.2	0.019
se_{12}	0.1	-0.157	-19.6	0.160
sp_1	96.2	0.000	0.0	0.016
sp_2	96.1	0.000	0.0	0.016
sp_{12}	0.0	-0.160	-20.0	0.161
Scenario 2: high sensitivities and specificities with moderate within-study associations, $se_j = sp_j = 0.8, se_{12} = sp_{12} = 0.7$				
se_1	95.7	0.001	0.1	0.019
se_2	96.3	0.001	0.1	0.019
se_{12}	32.2	-0.059	-8.4	0.063
sp_1	95.4	0.000	0.0	0.016
sp_2	96.0	-0.001	-0.1	0.015
sp_{12}	13.6	-0.061	-8.7	0.064
Scenario 3: high sensitivities and specificities with weak within-study associations, $se_j = sp_j = 0.8, se_{12} = sp_{12} = 0.64$				
se_1	96.0	-0.006	-0.8	0.019
se_2	95.7	-0.008	-0.9	0.020
se_{12}	95.7	-0.011	-1.7	0.023
sp_1	96.3	-0.007	-0.8	0.016
sp_2	95.2	-0.007	-0.8	0.017
sp_{12}	94.1	-0.011	-1.7	0.020

se_j and sp_j are the marginal sensitivities and specificities of the $j = 1, 2$ tests. $se_{1,2}$ and sp_{12} are the joint sensitivity and specificity, respectively. RMSE, root mean square error

Table 5.5: Performance measures for key test accuracy parameters across the 3 scenarios where sensitivities and specificities are both low, averaged over 1000 simulations per scenario.

Parameter	% Coverage	Bias	% Bias	RMSE
Scenario 4: low sensitivities and specificities with strong within-study associations, $se_j = sp_j = 0.4, se_{12} = sp_{12} = 0.4$				
se_1	96.3	-0.001	-0.2	0.023
se_2	95.8	-0.001	-0.2	0.023
se_{12}	0.0	-0.240	-60.0	0.241
sp_1	95.7	-0.001	-0.2	0.020
sp_2	95.7	0.000	-0.1	0.019
sp_{12}	0.0	-0.240	-60.0	0.241
Scenario 5: low sensitivities and specificities with moderate within-study associations, $se_j = sp_j = 0.4, se_{12} = sp_{12} = 0.3$				
se_1	96.6	0.000	-0.1	0.024
se_2	96.7	0.000	0.0	0.022
se_{12}	0.0	-0.140	-46.6	0.141
sp_1	94.5	0.000	-0.1	0.021
sp_2	95.0	0.001	0.1	0.020
sp_{12}	0.0	-0.140	-46.6	0.140
Scenario 6: low sensitivities and specificities with weak within-study associations, $se_j = sp_j = 0.4, se_{12} = sp_{12} = 0.16$				
se_1	96.9	0.000	0.01	0.023
se_2	95.0	0.001	0.2	0.024
se_{12}	97.0	0.000	0.3	0.013
sp_1	95.0	0.000	-0.1	0.020
sp_2	95.3	0.000	-0.1	0.019
sp_{12}	96.3	0.000	-0.1	0.011

se_j and sp_j are the marginal sensitivities and specificities of the $j = 1, 2$ tests. $se_{1,2}$ and sp_{12} are the joint sensitivity and specificity, respectively. RMSE, root mean square error

Table 5.6: Performance measures for key test accuracy parameters across the 3 scenarios where sensitivities are high and specificities are low, averaged over 1000 simulations per scenario.

Parameter	% Coverage	Bias	% Bias	RMSE
Scenario 7: high sensitivities and low specificities with strong within-study associations, $se_j = 0.8, se_{12} = 0.8, sp_j = 0.4, sp_{12} = 0.4$				
se_1	97.2	0.002	0.3	0.019
se_2	96.2	0.002	0.2	0.019
se_{12}	0.1	-0.157	-19.6	0.159
sp_1	96.5	-0.001	-0.2	0.020
sp_2	96.1	0.000	0.0	0.020
sp_{12}	0.0	-0.240	-60.0	0.240
Scenario 8: high sensitivities and low specificities with moderate within-study associations, $se_j = 0.8, se_{12} = 0.7, sp_j = 0.4, sp_{12} = 0.3$				
se_1	97.1	-0.001	-0.1	0.018
se_2	96.8	0.000	0.0	0.018
se_{12}	28.5	-0.060	-8.6	0.065
sp_1	95.7	0.000	0.0	0.020
sp_2	94.9	0.000	0.1	0.021
sp_{12}	0.0	-0.140	-46.6	0.140
Scenario 9: high sensitivities and low specificities with weak within-study associations, $se_j = 0.8, se_{12} = 0.64, sp_j = 0.4, sp_{12} = 0.16$				
se_1	95.7	-0.007	-0.8	0.020
se_2	96.1	-0.006	-0.7	0.019
se_{12}	95.7	-0.010	-1.5	0.023
sp_1	94.3	0.000	-0.1	0.020
sp_2	96.3	0.001	0.1	0.020
sp_{12}	96.1	0.000	0.1	0.011

se_j and sp_j are the marginal sensitivities and specificities of the $j = 1, 2$ tests. $se_{1,2}$ and sp_{12} are the joint sensitivity and specificity, respectively. RMSE, root mean square error

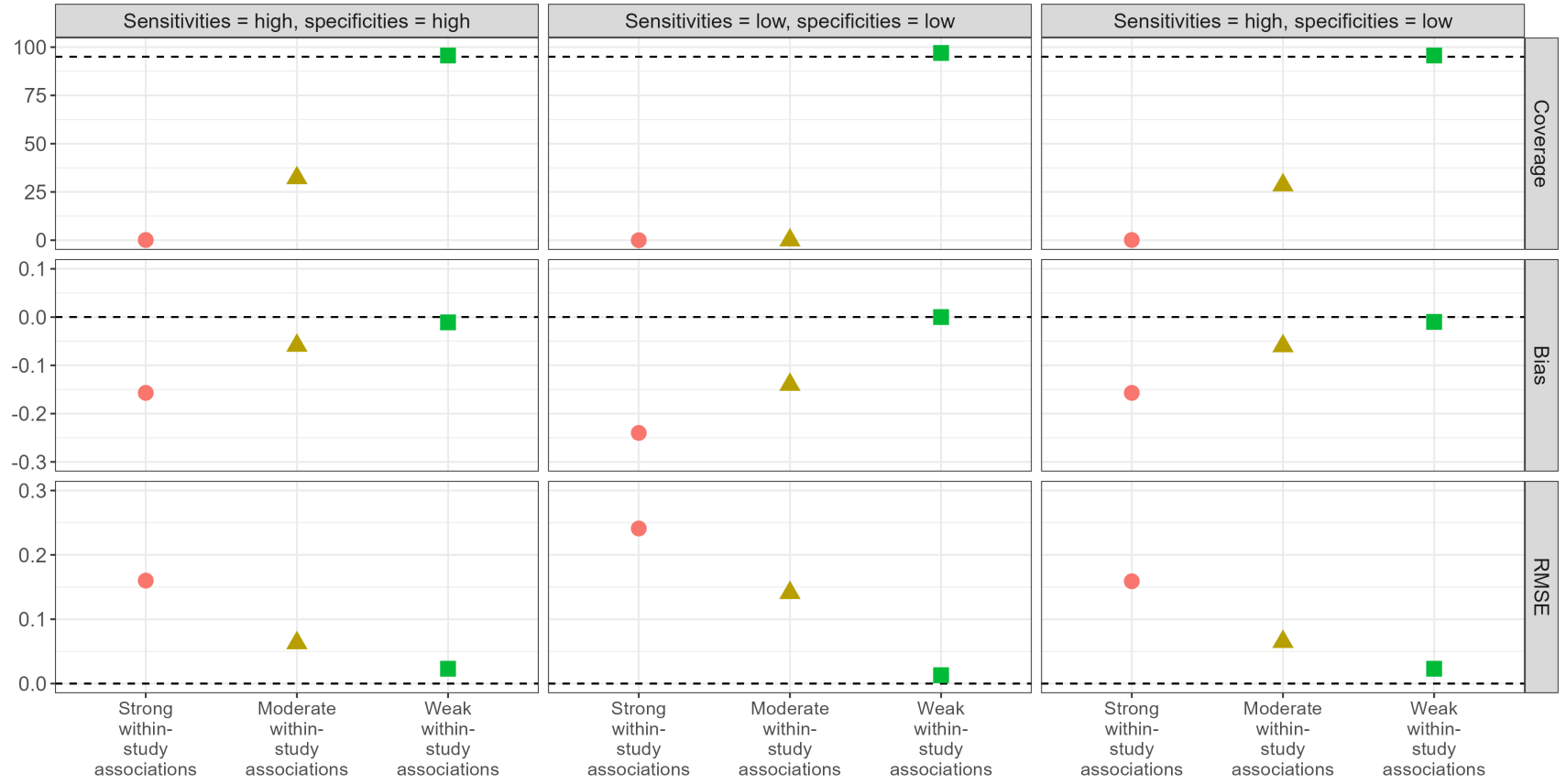


Figure 5.1: Bias, coverage and root mean square error (RMSE) of joint sensitivity averaged over the 1000 simulations across the 9 scenarios.

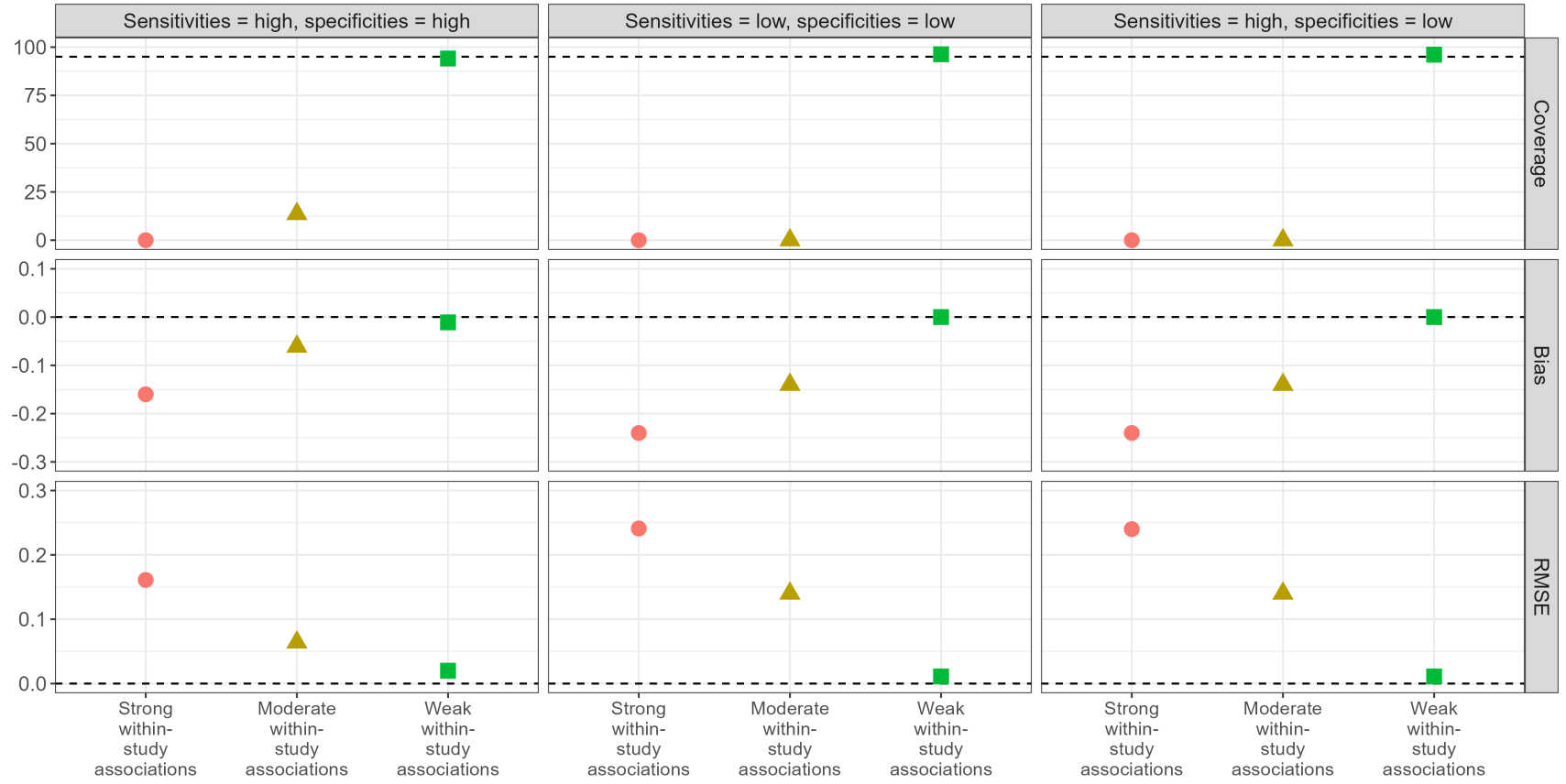


Figure 5.2: Bias, coverage and root mean square error (RMSE) of joint specificity averaged over the 1000 simulations across the 9 scenarios.

5.4.3 Convergence diagnostics

Convergence diagnostic plots for key model parameters are given in Appendix B.4. Interpretation of the diagnostic plots was described in Section 2.2.4. Due to the large number of simulated data sets and scenarios, a selection of diagnostic plots are presented for a single simulated data set; however, a random sample of plots for each scenario were inspected for non-convergence. Trace plots showed random oscillation around a stable mean and MCMC chains initiated at different starting values appeared to mix well. Density plots indicated no issues with convergence or mixing of chains. Autocorrelation between samples decreases sharply, indicating rapid model convergence. Overall, no issues with convergence or mixing of chains were detected for any of the scenarios.

5.5 Discussion

In this chapter, the impact of ignoring within-study associations between sensitivities and specificities in a meta-analysis of comparative diagnostic accuracy studies was assessed through a simulation study. A meta-regression approach, which does not account for dependencies between two tests assessed under a within-subject design, was fit to simulated data with known within-study associations. The model gave downwardly biased estimates of joint sensitivity and specificity when within-study associations were strong or moderate, underestimating the joint test accuracy measures by as much as 60%. As the strength of the within-study associations increased, model performance declined. The effect was also associated with the magnitude of the marginal test accuracy measures, with lower sensitivities and specificities leading to increased bias. When within-study associations were weak, the model performed well across all simulated scenarios. The model produced accurate estimates of marginal sensitivities and specificities regardless of the strength of the within-study associations.

Empirically testing whether ignoring dependencies between multiple tests impacts upon test accuracy measures is an important and timely addition to the existing literature on diagnostic meta-analysis models. Joint test accuracy measures enable estimation of the combined accuracy of two tests through 'AND'/'OR' rules (see Section 2.3.5). In the context of healthcare decision-making, combined test accuracy measures are used to evaluate the accuracy of different testing strategies and diagnos-

tic pathways, which often comprise of multiple components. Using a meta-analysis model that does not appropriately account for within-study dependencies present between multiple tests may lead to underestimation of joint accuracy, particularly where associations are strong, potentially resulting in suboptimal decision-making about which test combinations to use in clinical practice.

Based on the findings of the simulation study, the meta-regression approach is more suitable for a test comparison framework. When marginal test accuracy is of interest, for example when the aim of an analysis is to compare test performance with the view to use the most sensitive or specific, this simpler model appears to be sufficient. By utilising binomial likelihoods at the within-study level, the model requires only 2×2 data on each test, reflecting current reporting practices for comparative diagnostic accuracy studies. A method that accounts for within-study dependencies between multiple tests would be more appropriate to model the combined accuracy of two tests, particularly when the dependencies are strong. Any such method is likely to require cross-classified data from some, if not all, of the included studies in order to estimate within-study dependencies between tests. This level of granularity is often not reported in diagnostic studies; cross-classifications were available for only 20% of the comparative studies identified in the literature review in Chapter 4. This is further supported in the existing literature.[9, 10, 11, 4]

Choice of meta-analysis model to combine comparative diagnostic accuracy studies requires consideration of a number of factors. Firstly, data availability; where cross-classified data are reported for a subset of studies or no studies, the models that can be fit to the meta-analytic data set are limited. As evidenced in this chapter, choosing a model that requires only 2×2 data on each test may impact on joint test accuracy estimates, particularly where within-study associations are strong. Where combined test accuracy is of interest, it is advisable to choose model that accounts for these dependencies using cross-classified data where available. The number of index tests is another consideration. A number of meta-analysis models for two diagnostic tests can readily be extended to three or more tests (see Chapter 3), while for others this extension is non-trivial and requires further methodological development. Some models, extended beyond two index tests compared to a reference standard, may encounter convergence issues due to the large number of parameters that need to be estimated. Another factor when deciding which model to use is the number of test thresholds; neither the meta-regression or multinomial likelihoods model discussed in this chapter accommodate estimates of sensitivity and specificity at more than

one diagnostic threshold per study, however models developed by Cheng [92], Owen et al [95], Hoyer and Kuss [96] and Lian et al [97] do. As demonstrated through the simulation study conducted in this chapter, the test accuracy measures of interest, particularly in regards to whether they are comparative (e.g. marginal sensitivities and specificities, differences in sensitivities and specificities) or combined (e.g. joint sensitivity and specificity, combined accuracy measures based on 'AND'/'OR' rules) in nature, form a important component of model choice. In selecting an appropriate meta-analysis model, consideration should also be given to model fit, using measures such as the AIC or DIC for frequentist methods or the WAIC for Bayesian methods.

The simulation study is subject to a number of potential limitations. Due to the small number of studies reporting cross-classified data identified through the literature review in the previous chapter, it was not possible to simulate the diagnostic accuracy data for the study from the joint sensitivity and specificity estimated using the motivating example. Instead, plausible values of marginal and joint sensitivities and specificities were selected based on comparative diagnostic accuracy data extracted as part of the literature review. The motivating example was used to inform the study sample sizes and disease prevalence. A literature review by Bachmann et al estimated the median sample size of diagnostic accuracy studies as 118 (interquartile range (IQR) 71-350) and median prevalence of the target condition as 43% (IQR 27-61%).[204]. The median sample size (95, IQR 40-134) and prevalence (41, IQR 31-50) from which the data were simulated was reflective of the wider evidence base of diagnostic accuracy studies, increasing the generalisability of the simulation study findings. The simulation study could be extended to consider additional scenarios that may introduce bias to test accuracy estimates and impact model performance. Riley et al suggested that within-study correlation in multivariate meta-analysis exerts the greatest influence on results when within-study variation in the study estimates (i.e. sampling error) is large relative to between-study variation in the true underlying study values.[205] Further scenarios that vary between-study heterogeneity, modelled using between-study variance parameters, could be simulated to explore this hypothesis in the context of diagnostic meta-analysis of multiple tests. Diagnostic meta-analyses are often limited to a small number of studies; a study by Rosenberger et al summarised 1,379 meta-analyses of diagnostic test accuracy published in The Cochrane Library, finding that a median of 4 (IQR 2-9) studies were synthesised per analysis.[206] The number of studies in this simulation, which were chosen based on the motivating example, was relatively high at 18. It was not possible to explore the impact of few studies on test accuracy, with sparse data leading to

convergence issues. However, a similar study exploring the impact of within-study associations on surrogate endpoint evaluation found that smaller study size led to increased bias compared to larger study size.[207]

The need for comparison of existing meta-analysis methods, including whether more complex models outperform meta-regression, has been highlighted.[12, 4, 10] The simulation study could be extended to evaluate and compare the performance of novel meta-analysis models for jointly analysing diagnostic accuracy data on two tests, such as the multinomial likelihoods model developed by Trikalinos et al,[85] to the currently recommended meta-regression approach. There is some uncertainty over what information would be gained from such a simulation study, however. Data would need to be simulated from the multinomial model to capture strong tail dependencies, as it is in this chapter, therefore multinomial model would be the best fit for the data, potentially leading to a circular argument. The question of how to simulate diagnostic accuracy data on two tests, and the complex dependence structures that underlie it, holds true for any joint meta-analysis model (such as those summarised in Chapter 3) to which the meta-regression approach might be compared.

Fitting the multinomial likelihoods model and comparing the results to those of the meta-regression model was explored as part of thesis development. While the multinomial model was originally formulated in JAGS,[85] it has been observed to exhibit high dimensionality and computational demands, even for the case of two index tests compared to a common reference standard.[15] These findings were corroborated by exploratory analyses conducted in this chapter. Despite running the model for hundreds of thousands - and then millions - of iterations, it failed to achieve satisfactory convergence across any of the simulated scenarios. Thus, there arises a pressing need for further model development to capture the complexities of diagnostic accuracy data on two tests evaluated using a paired design; specifically, the need for computationally efficient and flexible methods to model dependencies between multiple tests. It was not possible to compare the meta-regression model with separate between-studies correlation (Equation 5.4) to the model with common between-studies correlation (Equation 5.3), due to the additional parameters leading to convergence issues when trying to implement the former as part of the simulation study. In the following chapter, the two meta-regression models are compared to a novel method that accounts for within-study associations through their application to two motivating examples. The analyses demonstrate that, although there may be a small gain in model performance through estimating a separate between-studies

correlation parameter for certain data sets, both meta-regression methods fall short of the models that incorporate within-study associations. Similarly, if it were feasible to include the separate between-studies correlation model in the simulation study then bias would likely still be observed, albeit a reduction compared to the common between-studies correlation model.

Given the substantial impact of ignoring within-study dependencies between multiple tests, particularly when associations are strong, we recommend that cross-classified data are reported wherever possible. This enables the application of models that appropriately account for within-study associations, a number of which were identified in the methodological review in Chapter 3, and adds value to healthcare decision-making about testing pathways. Technical Support Documents (TSDs), produced by the NICE Decision Support Unit, make recommendations on the implementation of methods for technology appraisal. TSDs developed for multivariate meta-analysis of interventional research could be extended to further aid in the translation of diagnostic meta-analysis models for multiple tests into healthcare decision-making.[209]

5.6 Chapter summary

In this chapter, a simulation study explored the impact of ignoring within-study dependencies between two tests in a meta-analysis framework on key test accuracy parameters. When within-study associations are weak, or when inference is focussed on comparative test accuracy measures, a meta-regression approach requiring 2×2 data on each test is sufficient. However, it does not make optimal use of the available evidence base and uncertainty persists around how to incorporate cross-classified data in comparative meta-analysis models when marginal sensitivities and specificities are the primary measures of interest. When within-study associations are present, a model that accounts for these dependencies using cross-classified data is required to capture the joint accuracy of multiple tests. Presently, many meta-analysis models for synthesising diagnostic accuracy data on two tests evaluated under a paired design ignore within-study dependencies, while models that incorporate these associations often result in slow convergence and long computation times. Motivated by the findings of the simulation study, the following two chapters introduce novel model development for meta-analysing diagnostic accuracy data on two tests evaluated in the same individuals, using copula methodology to flexibly capture within-study dependencies between sensitivities and specificities.

Chapter 6

Joint meta-analysis of two diagnostic tests using a bivariate copula to model comparative test accuracy

6.1 Chapter overview

Chapter 5 highlighted the importance of accounting for dependencies between two tests evaluated using a within-subject design, demonstrating their impact upon key test accuracy parameters through a simulation study. In this chapter, novel Bayesian meta-analysis models for evaluating the accuracy of two diagnostic tests are developed. Building on model development undertaken in the field of surrogate endpoint evaluation and diagnostic test accuracy, the bivariate random effects meta-analysis model for a single test is extended to jointly synthesise accuracy data on two tests, using bivariate copulas to flexibly capture within-study dependencies between sensitivities and specificities. The proposed models address limitations of existing methods, described in Chapter 3, by making better use of the available evidence base through the synthesis of full or partial cross-classifications, where available.

6.2 Introduction

There is an increasing availability of diagnostic tests in contemporary healthcare, with clinicians wishing to know which of the available options is 'best'. While the majority of diagnostic accuracy studies evaluate the performance a single test against that of a reference standard, most clinically relevant questions are comparative in nature.[210] Whether a novel test is sufficiently accurate to use in practice depends on the accuracy of similar, existing tests for the same condition.[12] Furthermore, joint analysis of two or more diagnostic tests is vital to assess the accuracy of different testing strategies and diagnostic pathways, which often comprise of multiple components.

In the context of healthcare decision-making, meta-analysis is a powerful tool for informing evidence-based practice.[3, 16] Several meta-analysis models for evaluating two or more diagnostic tests have been proposed in recent years, as summarised in Chapter 3. Whilst the existing models adequately capture between-studies association, within-study association is often ignored. As illustrated in the previous chapter, this assumption can have a substantial impact on test accuracy measures when within-study associations are moderate or strong. Models that account for within-study dependencies often make restrictive assumptions about the distributions of the marginal variables, and result in slow convergence and long computation times.[15] Furthermore, their use is often hindered by the limited availability of cross-classified data (see Section 2.3.4).

Copula theory is used to flexibly capture dependence between two (or more) random variables.[211] Copulas have previously been used to model between-studies correlation between sensitivity and specificity in a meta-analytic framework. For a single test, Kuss et al [99] and Nyaga et al [212] proposed using beta-binomial marginal distributions, linked by various coupling functions, as a natural choice to model sensitivity and specificity, which are bound between 0 and 1. Nyaga et al [90] extended their previous work to evaluate the accuracy of two tests, while Cheng [92] suggested a similar extension to Kuss et al's model. Hoyer and Kuss,[94] and later Nikoloulopoulos,[98] described four-dimensional copula models to capture between-studies dependence between the sensitivities and specificities. Within the context of surrogate endpoint meta-analysis, copulas have been used to model within-study association; either between the surrogate endpoint and the final clinical outcome in a meta-analysis of RCTs with time-to-event IPD by Burzykowski et al [213] or, as

proposed by Papanikos et al, between the treatment effects on the surrogate endpoint and the final clinical outcome within each study in an aggregate-level meta-analysis of RCTs with binary outcomes.[207] The application of copula models offers a potentially novel solution to capture within-study dependencies arising between two tests assessed using a paired design.

Building on the model proposed by Papanikos et al in the field of surrogate endpoint evaluation and meta-analytic models for diagnostic test accuracy,[212, 207] the following chapter proposes Bayesian meta-analysis models for evaluating the accuracy of two diagnostic tests, using copulas to capture within-study dependencies between the tests. Models are developed for five types of bivariate copula: Gaussian, Frank, Gumbel, Clayton and Clayton 180°. Section 6.3 describes the motivational data sets, extracted as part of Chapter 4, that the developed methodology is applied to. Section 6.4 describes the bivariate copula methodology. The application of the methods is demonstrated in Section 6.5, where the results of fitting both the new models and the meta-regression approach to the motivating examples are presented and compared. Section 6.6 concludes the chapter with a discussion.

6.3 Motivational examples

Data collection methods are described in detail in Chapter 4. From the 70 comparative diagnostic accuracy studies identified through the literature review, two motivating examples were selected that maximised the availability of comparative data. Priority was given to test comparisons supported by the largest number of studies and greatest number of total participants, as well as ensuring cross-classified data were reported for at least one study. The first example compares Aβ42 and t-tau, measured in the CSF, on which 36 2×2 tables from 18 studies were extracted. Two studies contained cross-classified data and the remaining 16 reported 2×2 data only (Table 5.2, presented in the previous chapter). The second compares the diagnostic accuracy of CSF Aβ42 and CSF p-tau in 14 studies, of which one reported cross-classified data (Table 6.1).

Table 6.1: Data on the accuracy of amyloid-β 42 (Aβ_{42}, test 1) and phosphorylated tau (p-tau, test 2) for diagnosing Alzheimer's disease dementia. The table contains 2×2 data for each test within each study, and fully-cross classified data where available.

Study	**Alzheimer's disease dementia**									**No Alzheimer's disease dementia**								
	tp_1	tp_2	fn_1	fn_2	x_{11}^D	x_{10}^D	x_{01}^D	x_{00}^D	N^D	tn_1	tn_2	fp_1	fp_2	$x_{11}^{\bar{D}}$	$x_{10}^{\bar{D}}$	$x_{01}^{\bar{D}}$	$x_{00}^{\bar{D}}$	$N^{\bar{D}}$
Bjerke et al, 2009.[128]	18	17	2	3	-	-	-	-	20	99	109	43	33	-	-	-	-	142
Blom et al, 2009[129].	9	8	5	6	-	-	-	-	14	5	11	9	3	-	-	-	-	14
Chiasserini et al, 2010.[136]	17	20	6	3	-	-	-	-	23	16	13	2	5	-	-	-	-	18
Frölich et al, 2017.[145]	17	17	11	11	-	-	-	-	28	65	60	22	27	-	-	-	-	87
Gaser et al, 2013.[148]	59	45	7	21	-	-	-	-	66	12	19	21	14	-	-	-	-	33
Hansson et al, 2006.[150]	56	54	1	3	54	2	0	1	57	50	45	27	32	15	12	18	32	77
Herukka et al, 2008.[153]	6	7	2	1	-	-	-	-	8	11	7	2	6	-	-	-	-	13
Hertze et al, 2010.[152]	47	22	5	30	-	-	-	-	52	75	96	32	11	-	-	-	-	107
Monge-Argilés et al, 2011.[166]	9	9	2	2	-	-	-	-	11	16	15	10	11	-	-	-	-	26
Nesteruk et al, 2016.[169]	7	3	2	6	-	-	-	-	9	18	24	13	7	-	-	-	-	31
Palmqvist et al, 2012.[174]	47	35	5	17	-	-	-	-	52	56	70	25	11	-	-	-	-	81
Parnetti et al, 2006.[175]	4	9	7	2	-	-	-	-	11	30	32	3	1	-	-	-	-	33
Parnetti et al, 2012.[176]	18	26	14	6	-	-	-	-	32	56	52	2	6	-	-	-	-	58
Rhodius-Meester et al, 2016.[181]	58	75	27	10	-	-	-	-	85	38	30	14	22	-	-	-	-	52

tp_j is the number of true positives, fn_j false negatives, tn_j true negatives, and fp_j false positives for each of the $j = 1, 2$ tests.
x_{kl}^D is the number of participants with the target condition and $x_{kl}^{\bar{D}}$ is the number of participants without the target condition with each combination of test results. $k, l = 0, 1$ denote the first and second test, respectively, where 0 indicates a negative test result and 1 indicates a positive test result.

6.4 Methods

In this section, the bivariate copula methodology is described. Section 6.4.1, 6.4.2 and 6.4.3 introduce bivariate copula theory and the various classes of copula that are explored in this chapter. Section 6.4.4 outlines the proposed methodology for the meta-analysis of two diagnostic tests, using bivariate copulas to model dependencies between tests. Section 6.4.5 describes the bootstrap method used to estimate within-study associations. Section 6.4.6 describes the meta-regression models which serves as comparators for the bivariate copula models and Section 6.4.7 details the estimation methods.

6.4.1 Bivariate copula theory

Copula theory is used to flexibly capture dependence between two or more random variables.[211] Standard multivariate distributions for modelling correlated variables can make strong assumptions about the marginal distributions of each variable. For example, the multivariate normal distribution assumes each individual component (and linear combinations of these components) follows a univariate normal distribution, which in practice may be violated. Diagnostic accuracy data, in particular, are often skewed and non-normal; assuming normality may lead to poor model fit and impact test accuracy estimates.

Copulas enable the separation of the marginals from the dependence structure of a multivariate distribution, allowing separate specification of both the random variables and the relationship between them. Copulas are often used to model high-dimensional data due to this property.[214] As discussed in Section 2.4.3, within-study dependencies arise between sensitivities and between specificities when two tests are assessed under a paired design. Diagnostic accuracy data on two tests are likely to exhibit strong tail dependencies (when sensitivity or specificity is close to 1). The multivariate normal distribution assumes weak association at extreme ends of the distribution and may not adequately capture the relationship between tests. There are a number of different copula families and types, discussed in more detail in Section 6.4.2. These allow for a variety of dependence structures, including strong tail dependencies and asymmetry, which may better represent the relationship between two tests.

A bivariate copula is a bivariate cumulative distribution function (CDF) with uniform

marginal distributions on the interval [0,1].[214, 211] Let there be two continuous, correlated random variables, u_1 and u_2. In accordance with Sklar's theorem,[215] any bivariate distribution, H, can be expressed in terms of univariate marginal distribution functions, F_1 and F_2, and a copula, C, which describes their relationship to one another. θ is the copula dependence parameter, which captures the association between the two random variables. A copula is a coupling function which links the univariate marginal distribution functions to their bivariate distribution function, allowing separate specification of their dependence structure:

$$H\left(u_1, u_2, \theta\right) = C\left(F_1\left(u_1\right), F_2\left(u_2\right), \theta\right). \tag{6.1}$$

The derivation of the joint distribution is possible through partial derivatives when the random variables are continuous. For discrete random variables, such as those used to model diagnostic accuracy data, which are summarised by counts, the joint probability mass function is derived through finite differences:

$$\begin{aligned} h(u_1, u_2, \theta) = {} & C(F_1(u_1), F_2(u_2), \theta) - C(F_1(u_1 - 1), F_2(u_2), \theta) \\ & - C(F_1(u_1), F_2(u_2 - 1), \theta) + C(F_1(u_1 - 1), F_2(u_2 - 1), \theta) \end{aligned} \tag{6.2}$$

6.4.2 Families of copulas

Several families of copulas have been described, which classify different copulas based on their properties. The properties of these different classes of copulas make them more appropriate for modelling certain types of relationships between variables, such as strong dependencies at the extreme (tail) ends of their distributions or asymmetric dependencies. Five types of bivariate copulas belonging to two families are implemented within the diagnostic meta-analytic framework in this chapter. These are the Gaussian copula, belonging to the elliptical family, and the Frank, Gumbel, Clayton, and Clayton 180°copulas, belonging to the Archimedean family. The Gaussian copula was chosen due to its similarities to the multivariate normal distribution, which has previously been used to model dependencies between multiple tests.[85, 91, 96] The four Archimedean copulas were specified to explore different approaches to modelling within-study associations present between two tests, such as assuming strong, tail dependencies, weak tail dependencies or asymmetry. In order to visualise the dependence structures modelled by the five bivariate copula types, 4000 samples were simulated from each within R version 4.3.1 using the *copula* version 1.1-2 package

(Figure 6.1).

6.4.2.1 Elliptical copulas

Elliptical copulas join univariate marginals through an elliptical distribution, the most commonly used of which is the Gaussian copula. Elliptical copulas model the full range of correlation between the marginal distributions, meaning they are not limited to positive or negative dependence only. However, they cannot be expressed in closed form and are restricted to radial symmetry.

6.4.2.1.1 Gaussian copula The Gaussian copula is a symmetric copula with weak dependence in the tails of the distribution (see Figure 6.1). For a given correlation matrix $R \in [-1,1]^{2\times 2}$, the Gaussian copula modelling the association between two random variables, u_1 and u_2, can be written as:

$$C^R_{Gauss}(u_1, u_2) = \phi_R\left[\phi^{-1}(u_1), \phi^{-1}(u_2)\right], \quad \theta \in [-1,1] \tag{6.3}$$

where ϕ_R is the joint CDF of the bivariate normal distribution with mean $(0,0)^T$ and correlation matrix R. ϕ^{-1} is the inverse CDF of the standard normal distribution. θ is the copula dependence parameter, which measures the association between the variables (discussed in more detail in Section 6.4.3). In the case of the Gaussian copula, θ is equivalent to the Pearson's correlation coefficient.[216]

6.4.2.2 Archimedean copulas

Archimedean copulas are expressed as an explicit formula and depend on a single parameter that dictates the strength of the dependence. For random variables u_1 and u_2, bivariate Archimedean copulas are defined by:

$$C(u_1, u_2) = \psi^{-1}\left[\psi(u_1) + \psi(u_2)\right], \quad u_1, u_2 \in [0,1] \tag{6.4}$$

where ψ and ψ^{-1} are the generator function and inverse generator function of the copula, respectively. Generator functions for each copula type used in this thesis are given in Table 6.2); the bivariate copulas can be constructed by applying these functions to Equation 6.4.

Table 6.2: Bivariate Archimedean copula cumulative distribution functions, $C\left(u_2 \mid u_1; \theta\right)$, generator functions, ψ, and inverse generator functions, ψ^{-1}.

Copula	$\boldsymbol{C\left(u_2 \mid u_1; \theta\right)}$	$\boldsymbol{\theta}$	**Generator** $\boldsymbol{\psi\left(t\right)}$	**Inverse generator** $\boldsymbol{\psi^{-1}\left(t\right)}$
Frank	$e^{-\theta u_1}\left[\frac{1-e^{-\theta}}{1-e^{-\theta u_2}-\left(1-e^{-\theta u_1}\right)}\right]^{-1}$	$\theta \in \mathbb{R} \backslash \{0\}$	-log $\left(\frac{e^{-\theta t}-1}{e^{-\theta}-1}\right)$	$\frac{\log\left(1+e^{-t}\left(e^{-\theta}-1\right)\right)}{\theta}$
Gumbel	$e^{\left[-\left((-\log(u_1))^{\theta}+(-\log(u_2))^{\theta}\right)^{\frac{1}{\theta}}\right]}$	$\theta \in [1, \infty)$	$\left(-\log\left(t\right)\right)^{\theta}$	$e^{-t^{\frac{1}{\theta}}}$
Clayton	$\left[1+u_1^{\theta}\left(u_2^{-\theta}-1\right)\right]^{-1-\frac{1}{\theta}}$	$\theta \in (0, \infty)$	$\frac{1}{\theta}\left(t^{-\theta}-1\right)$	$\left(1+\theta t\right)^{-\frac{1}{\theta}}$
Clayton 180°	$1-\left[1+\left(1-u_1\right)^{\theta}\left(\left(1-u_2\right)^{-\theta}-1\right)\right]^{-1-\frac{1}{\theta}}$	$\theta \in (0, \infty)$	$\frac{1}{\theta}\left(t^{-\theta}-1\right)$	$\left(1+\theta t\right)^{-\frac{1}{\theta}}$

u_j, where $j = 1, 2$, are random variables. u_j is substituted for t.
θ is the copula dependence parameter.

6.4.2.2.1 Frank copula The Frank copula is a symmetric Archimedean copula with weak tail dependence (see Figure 6.1). As copula dependence parameter $\theta \in (-\inf, \inf)$, the Frank copula can model both positive and negative dependence. Given two random variables, u_1, u_2, the Frank copula is expressed as:

$$C_{Frank}(u_1, u_2) = -\frac{1}{\theta}\log\left[1 + \frac{\left(e^{-\theta u_1} - 1\right)\left(e^{-\theta u_2} - 1\right)}{\left(e^{-\theta} - 1\right)}\right], \quad \theta \in \mathbb{R}\backslash\{0\} \tag{6.5}$$

6.4.2.2.2 Gumbel copula The Gumbel copula is an asymmetric copula exhibiting positive dependence, with stronger dependence in the right-hand tail than in the left-hand tail (see Figure 6.1). Negative (left-hand tail) dependence can be induced by rotating the copula function by 90°or 270°. For random variables u_1 and u_2, and copula dependence parameter θ, the Gumbel copula is defined as:

$$C_{Gumbel}(u_1, u_2) = e^{\left[-\left[(-\log(u_1))^{-\theta} + (-\log(u_2))^{-\theta}\right]^{-\frac{1}{\theta}}\right]}, \quad \theta \in [1, \infty) \tag{6.6}$$

6.4.2.2.3 Clayton copula The Clayton copula is another asymmetric Archimedean copula exhibiting positive dependence, and can be rotated in a similar fashion to a Gumbel copula. It forms a funnel shaped distribution, with strong, left-hand tail dependence and weak right-hand tail dependence (see Figure 6.1). For random variables u_1 and u_2 and copula dependence parameter θ, the Clayton copula is expressed as follows:

$$C_{Clayton}(u_1, u_2) = \left[u_1^{-\theta} + u_2^{-\theta} - 1\right]^{-\frac{1}{\theta}}, \quad \theta \in (0, \infty) \tag{6.7}$$

6.4.2.2.4 Clayton 180°copula When evaluating the accuracy of two diagnostic tests, it is likely that there will be stronger dependence in the right-hand tail (i.e. when sensitivities and specificities are close to 1) than in the left (when sensitivities and specificities are low). This can be induced by rotating the copula function by 180°(see Figure 6.1). For two random variables u_1 and u_2 and copula dependence parameter θ, the Clayton 180°copula is defined as:

$$\begin{aligned} C_{Clayton180}(u_1, u_2) &= u_1 + u_2 - 1 + C_{Clayton}\left((1 - u_1), (1 - u_2)\right) \\ &= u_1 + u_2 - 1 + \left[(1 - u_1)^{-\theta} + (1 - u_2)^{-\theta} - 1\right]^{-\frac{1}{\theta}}, \quad \theta \in (0, \infty) \end{aligned} \tag{6.8}$$

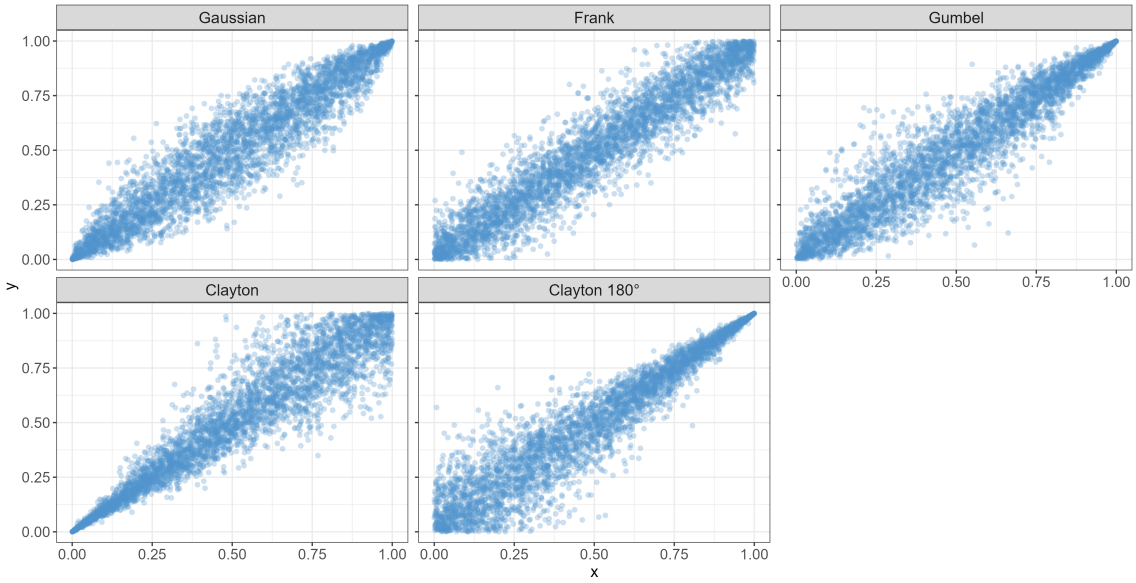


Figure 6.1: Simulated samples from bivariate a) Gaussian, b) Frank, c) Gumbel, d) Clayton, and e) Clayton 180°copulas. 4000 samples were simulated for each copula type, with fixed Spearman's correlation coefficient $\rho_s = 0.95$.

Negative dependence can also be induced by rotating the copula function by 90° or 270°.

6.4.3 Copula dependence parameter

The copula dependence parameter, θ, captures the association between two random variables modelled by a bivariate copula. Where fully cross-classified results (see Table 2.3) are available for each study, a separate copula dependence parameter can be estimated for each, making full use of the available data. Where full cross-classifications are not available across all studies, a common dependence parameter can be assumed. Where only 2×2 tables for each test is available, an informative prior distribution for the dependence parameter can be constructed using external sources of evidence.[3, 207]

Interpretation of θ differs from measures of correlation such as a Kendall's tau or Spearman's rank correlation coefficient, which are bounded between -1 (perfect negative association) and 1 (perfect positive association). Generally, a negative or positive θ indicates a negative or positive association between the variables, respectively. $\theta = 0$ is interpreted as no association. θ is restricted to positive values for certain copula types, such as the Gumbel and Clayton copulas, however the copula can

be rotated to induce negative association (Section 6.4.2). θ can be converted to Kendall's tau or Spearman's rho; the transformation depends on the copula type (see Table 6.3). Both Kendall's tau and Spearman's rank depend only on the copula and not the marginal densities.[214]

Table 6.3: Kendall's tau and Spearman's rank correlation coefficient as functions of their corresponding copula dependence parameter, θ for each bivariate copula. Adapted from Kojadinovic and Yan [217] and Nikoloulopoulos [98].

Copula	$\boldsymbol{\theta}$	Kendall's tau	Spearman's rank
Gaussian	$\theta \in [-1, 1]$	$\frac{2}{\pi} \arcsin(\theta)$	$\frac{6}{\pi} \arcsin\left(\frac{\theta}{2}\right)$
Frank	$\theta \in \mathbb{R} \backslash \{0\}$	$1 - \frac{4}{\theta}\left[D_1(-\theta) - 1\right]$	$1 - \frac{12}{\theta}\left[D_2(-\theta) - D_1(-\theta)\right]$
Gumbel	$\theta \in [1, \infty)$	$1 - \frac{1}{\theta}$	Numerical approximation
Clayton	$\theta \in (0, \infty)$	$\frac{\theta}{(\theta+2)}$	Numerical approximation
Clayton 180°	$\theta \in (0, \infty)$	$\frac{\theta}{(\theta+2)}$	Numerical approximation

θ is the copula dependence parameter.
D_k is the Debye function, $D_k(\theta) = \frac{k}{\theta^k} \int_\theta^0 \frac{t^k}{e^t - 1} dt$, for $k = 1, 2$.[218]

6.4.4 Bivariate copula model

An extension of the BRMA model for single tests described in Section 2.4.2.1 to jointly synthesise data on two diagnostic tests evaluated using a within-subject design is proposed. This builds on method development for surrogate endpoint evaluation introduced by Papanikos et al,[207] for which the number of events in each of the treatment and placebo arms of a randomised controlled trial follow bivariate distributions with binomial marginal distributions. A bivariate copula jointly models the number of events for the surrogate endpoint and final clinical outcome in each arm, accounting for within-study dependencies between the two outcomes, while correlation between treatment effects on the surrogate and final outcome (on the *log*-odds ratio scale) is incorporated through a bivariate normal distribution at the between-studies level.

There are two outcomes of interest in the surrogate endpoint model described above: treatment effect on the surrogate endpoint, and treatment effect on the final clinical outcomes. To jointly synthesise data on two diagnostic tests, four outcomes of interest are modelled: sensitivity and specificity of first test, and sensitivity and specificity of the second test. Bivariate copulas capture within-study dependencies between the sensitivities and specificities for the two tests. For study $i = 1, \ldots, I$,

the number of true positive and true negative events for each test follow bivariate distributions:

$$\begin{pmatrix} tp_{i,1} \\ tp_{i,2} \end{pmatrix} \sim h\left(se_{i,1}, se_{i,2}, N_i^D, \theta_{i,se}\right), \quad \begin{pmatrix} tn_{i,1} \\ tn_{i,2} \end{pmatrix} \sim h\left(sp_{i,1}, sp_{i,2}, N_i^{\bar{D}}, \theta_{i,sp}\right) \quad (6.9)$$

with binomial marginal distributions. $se_{i,j}$ and $sp_{i,j}$ denote the true study-specific sensitivities and specificities of the two tests, respectively, N_i^D and $N_i^{\bar{D}}$ the number of patients with and without the target condition as determined by the reference standard and $\theta_{i,se}$ and $\theta_{i,sp}$ are the copula dependence parameters, representing the within-study dependencies between the sensitivities and specificities, respectively. A separate dependence parameter can be estimated for each study i, allowing the within-study dependencies to vary across studies. In practice, full cross-classifications required to estimate the dependence parameters, θ_{se} and θ_{sp} may not be available for all studies. In this case, a pair of common dependence parameters can be assumed across studies (see Section 6.4.5). Where only 2×2 data are available, informative prior distributions for the dependence parameters can be constructed using external sources of evidence.[3, 207]

The joint probability mass functions are derived through finite differences:

$$\begin{aligned} h\left(tp_{i,1}, tp_{i,2} | se_{i,1}, se_{i,2}, N_i^D, \theta_{i,se}\right) = & \\ & C\left(F_1\left(tp_{i,1}\right), F_2\left(tp_{i,2}\right), \theta_{i,se}\right) - \\ & C\left(F_1\left(tp_{i,1}-1\right), F_2\left(tp_{i,2}\right), \theta_{i,se}\right) - \\ & C\left(F_1\left(tp_{i,1}\right), F_2\left(tp_{i,2}-1\right), \theta_{i,se}\right) + \\ & C\left(F_1\left(tp_{i,1}-1\right), F_2\left(tp_{i,2}-1\right), \theta_{i,se}\right) \end{aligned}$$

$$\begin{aligned} h\left(tn_{i,1}, tn_{i,2} | sp_{i,1}, sp_{i,2}, N_i^{\bar{D}}, \theta_{i,sp}\right) = & \\ & C\left(F_1\left(tn_{i,1}\right), F_2\left(tn_{i,2}\right), \theta_{i,sp}\right) - \\ & C\left(F_1\left(tn_{i,1}-1\right), F_2\left(tn_{i,2}\right), \theta_{i,sp}\right) - \\ & C\left(F_1\left(tn_{i,1}\right), F_2\left(tn_{i,2}-1\right), \theta_{i,sp}\right) + \\ & C\left(F_1\left(tn_{i,1}-1\right), F_2\left(tn_{i,2}-1\right), \theta_{i,sp}\right) \end{aligned} \quad (6.10)$$

$F_1\left(tp_{i,1}\right)$, $F_2\left(tp_{i,2}\right)$, $F_1\left(tn_{i,1}\right)$ and $F_2\left(tn_{i,2}\right)$ are the CDFs of the binomial marginal distributions on the number of true positives and true negatives, and $C\left(\cdot,\cdot\right)$ is the bivariate copula.

At the between-studies level, the *logit*-transformed study-specific sensitivities and specificities follow a four-dimensional multivariate normal distribution with means μ_{sej} and μ_{spj}, between-study variances σ^2_{sej} and σ^2_{spj}, and $j = 1, 2$. A common between-studies correlation parameter, ρ_b, is estimated across pairs of sensitivities and specificities to minimise the number of model parameters and reduce the likelihood of non-convergence.

$$logit\left(se_{i,1}\right) = \mu_{i,se1}, \quad logit\left(se_{i,2}\right) = \mu_{i,se2}, \tag{6.11}$$

$$logit\left(sp_{i,1}\right) = \mu_{i,sp1}, \quad logit\left(sp_{i,2}\right) = \mu_{i,sp2},$$

$$\begin{pmatrix} \mu_{i,se1} \\ \mu_{i,se2} \\ \mu_{i,sp1} \\ \mu_{i,sp2} \end{pmatrix} \sim Normal \left(\begin{pmatrix} \mu_{se1} \\ \mu_{se2} \\ \mu_{sp1} \\ \mu_{sp2} \end{pmatrix}, \begin{pmatrix} \sigma^2_{se1} & \rho_b\sigma_{se1}\sigma_{se2} & \rho_b\sigma_{se1}\sigma_{sp1} & \rho_b\sigma_{se1}\sigma_{sp2} \\ & \sigma^2_{se2} & \rho_b\sigma_{se2}\sigma_{sp1} & \rho_b\sigma_{se2}\sigma_{sp2} \\ & & \sigma^2_{sp1} & \rho_b\sigma_{sp1}\sigma_{sp2} \\ & & & \sigma^2_{sp2} \end{pmatrix} \right)$$

Post-sampling, summary sensitivities and specificities are derived through an inverse-*logit* transformation:

$$\eta_{se1} = logit^{-1}\left(\mu_{se1}\right), \quad \eta_{se2} = logit^{-1}\left(\mu_{se2}\right), \tag{6.12}$$

$$\eta_{sp1} = logit^{-1}\left(\mu_{sp1}\right), \quad \eta_{sp2} = logit^{-1}\left(\mu_{sp2}\right)$$

Prior distributions were placed on the unknown parameters in the model. *Logit*-transformed summary sensitivities and specificities were assumed to follow a minimally informative $Normal\left(0, 10^2\right)$ prior distribution. The between-studies variance parameters were restricted to positive values and were assumed to have a uniform $Half - Normal\left(0, 2.5^2\right)$ prior distribution. For the between-studies correlation parameter, the Fisher z-transformation was used, i.e $\rho_b = \tanh\left(z\right), z \sim Normal\left(0, 0.8\right)$). This transformation produces an approximately normal distribution bound between $[-1, 1]$.

6.4.5 Bootstrapping methods to obtain copula dependence parameter

The copula dependence parameters were estimated from available cross-classifications, which allow the reconstruction of IPD, using a double bootstrap method.[219, 207] Double bootstrapping allows the estimation of the association between outcomes with uncertainty. Prior to bootstrapping, IPD was recreated for each study with cross-classified data by transforming the counts into a data set of zeros and ones indicating each patients' test results and disease status. Bootstrapping involves repeatedly sampling the IPD (with replacement) from a single study to create many simulated data sets. Sensitivities and specificities were estimated for each simulated data set, then the association between them estimated across the multiple bootstrap samples. Where cross-classifications are available for each of the $i = 1, ..., I$ studies, the simulated data sets are used to estimate $\theta_{i,se}$ and $\theta_{i,sp}$ for each study using maximum likelihood estimation. In the motivating example used in this chapter, cross-classified data were available for two of the 18 studies.[150, 152] A common pair of dependence parameters, θ_{se} and θ_{sp}, were assumed across studies by applying the double bootstrap method to each study and estimating the mean of the two for each dependence parameter. Code for the bootstrapping method is provided in Appendix C.1.

6.4.6 Meta-regression with test type as a covariate

The BRMA model, described in Section 2.4.2.1, can be extended to incorporate covariates, allowing the comparison of summary sensitivity and specificity between groups in a meta-regression framework.[4] The addition of test type as a covariate enables the synthesis of comparative diagnostic accuracy studies that evaluate two tests against a common reference standard. The model requires study-level 2×2 data on each test (see Table 2.2), meaning the two tests are treated as independent and within-study dependencies between tests are not accounted for. Two meta-regression models, applying a common or separate between-studies correlation parameter, are used as comparators for the novel meta-analytic models introduced in this chapter, which account for associations between the two tests using bivariate copulas. Both the meta-regression and bivariate copula methods are fit to the motivational data sets, described in Section 6.3. Full specification of the meta-regression models are given in the previous chapter, see Section 5.3.1.

6.4.7 Estimation

All models were implemented in a Bayesian framework using Stan version 2.32.2 [53] within R version 4.3.1 [203] via the rstan version 2.32.5 package.[54] A non-centred parameterisation was used for all models to reduce dependencies between successive levels of the hierarchical structure, further increasing the efficiency of the sampler (see Section 2.2.3.4).[55] Copula dependence parameters were estimated using the bootstrap method described in Section 6.4.5 using R version 4.2.1.[54] After discarding 1,000 burn-in iterations, posterior estimates were obtained using three chains initialised at different starting values, consisting of 5,000 iterations each. The three chains were initialised at different starting values to aid the evaluation of model convergence. 95% CrIs were computed as HPD intervals.

Model convergence was assessed using diagnostics described in Section 2.2.4, including trace plots, density plots, and autocorrelation plots. Model fit was compared by calculating the WAIC using the R package *loo*,[220] with a smaller WAIC indicating better model fit.[58] Stan code for the two meta-regression models and the five bivariate copula models is available in the Appendix B.1 and C.2, respectively.

6.5 Results

6.5.1 Comparison of model fit

Table 6.4 presents the values of the WAIC across the seven fitted models for the two motivational examples. The meta-regression model with a common between-studies correlation parameter was the poorest fit for the meta-analytic data set comparing CSF Aβ_{42} to t-tau, corresponding to the largest WAIC of 180.6. There is strong evidence that the bivariate copula models resulted in a better fit compared to either meta-regression approach, with reductions in WAIC ranging from 21.0-29.3 for the common between-studies correlation model and 12.5-20.8 for the separate between-studies correlation model. Of the five bivariate copula models, the Gumbel copula was the best fit for the data with the smallest WAIC of 151.3. The differences in WAIC between the Gaussian (WAIC = 159.6), Frank (WAIC = 155.4), Clayton (WAIC = 151.6) and Clayton 180°(WAIC = 156.9) copulas were small, however, and the evidence to support one copula over another is marginal.

The meta-regression model with a common between-studies correlation parameter

was also the poorest fit for the 14 studies comparing the diagnostic accuracy of CSF Aβ_{42} to p-tau, corresponding to the largest WAIC of 135.0. Estimating a separate between-studies correlation parameter did not lead to improved model fit (WAIC = 134.3). Greater differences in model fit were observed between the bivariate copula models for this data set, with reductions in WAIC ranging between 8.5-28.3. The Gumbel copula model was again the best fit for the data, with the smallest WAIC of 106.7. The differences between the Gumbel, Frank (WAIC = 112.8) and Clayton 180°(WAIC = 113.2) copulas were marginal. There was evidence that the Gaussian (WAIC = 126.4) and Clayton (124.8) copulas led to a poorer fit compared to the other bivariate copulas, however this is still a notable improvement over the meta-regression approach.

Table 6.4: Values of the widely applicable information criterion (WAIC) across the meta-regression and bivariate copula models for each motivational example.

Model	**WAIC**	**Change in WAIC compared to the meta-regression model with common BSC**
Aβ_{42} vs t-tau		
Meta-regression (common BSC)	180.6	-
Meta-regression (separate BSC)	172.1	-8.5
Bivariate copula (Gaussian)	159.6	-21.0
Bivariate copula (Frank)	155.4	-25.2
Bivariate copula (Gumbel)	151.3	-29.3
Bivariate copula (Clayton)	151.6	-29.0
Bivariate copula (Clayton 180°)	156.9	-22.8
Aβ_{42} vs p-tau		
Meta-regression (common BSC)	135.0	-
Meta-regression (separate BSC)	134.3	-0.7
Bivariate copula (Gaussian)	126.4	-8.6
Bivariate copula (Frank)	112.8	-22.2
Bivariate copula (Gumbel)	106.7	-28.3
Bivariate copula (Clayton)	124.8	-10.2
Bivariate copula (Clayton 180°)	113.2	-21.8

Aβ, amyloid-β; BSC, between-studies correlation; p-tau, phosphorylated tau; t-tau, total tau; WAIC, widely applicable information criterion

6.5.2 Summary sensitivities and specificities η

The results of fitting the two meta-regression and five bivariate copula models to the two motivational examples in Alzheimer's disease dementia are presented in Table 6.5 and 6.6. Figure 6.2 and 6.3 display the posterior medians and 95% CrIs for key test accuracy parameters across each of the models for the two data sets.

The first example compares the diagnostic accuracy of CSF Aβ_{42} and t-tau (Table 6.5). Based on the best fitting Gumbel copula model, CSF Aβ_{42} and t-tau demonstrated 80.9% (95% CrI: 73.4, 87.5) and 76.4% (95% CrI: 69.4, 83.1) sensitivity to differentiate Alzheimer's disease dementia from MCI, respectively. Summary specificity was 70.3% (95% CrI: 61.3, 78.4) and 72.5% (95% CrI: 63.7, 81.3), respectively. The second compares the diagnostic accuracy of CSF Aβ_{42} and p-tau (Table 6.6). Using the same model, sensitivity of CSF Aβ_{42} and p-tau were 80.2% (95% CrI: 68.8, 89.4) and 76.8% (95% CrI: 66.1, 87.2), respectively. Summary specificity was 72.5% (95% CrI: 60.8, 82.7) and 75.7% (95% Cri: 65.9, 84.1), respectively. The clinical role of biomarkers for dementia is evolving. At present, UK guidelines recommend using CSF markers if diagnostic subtype of dementia is uncertain. Where there is already a clinical suspicion of Alzheimer's dementia, test accuracy may be high in this scenario. Overall, similar inferences about the sensitivities and specificities of CSF Aβ_{42}, t-tau and p-tau were drawn regardless of the model used.

Posterior median summary sensitivities and specificities were similar across all the models; however, the bivariate copula models yielded narrower 95% CrIs for these parameters than the meta-regression models. The Gumbel copula produced the narrowest CrIs for sensitivity and specificity estimates corresponding to both data sets, with a 16% and 10% decrease in CrI width for sensitivity of Aβ_{42} and t-tau, respectively, compared to the meta-regression model with common between-studies correlation.

6.5.3 Between-studies standard deviations σ

Posterior median between-studies standard deviation estimates of *logit*-transformed sensitivities and specificities of Aβ_{42} and t-tau from the Gumbel copula model were 0.80 (95% CrI: 0.44, 1.23), 0.64 (95% CrI: 0.36, 1.00), 0.73 (95% CrI: 0.38, 1.19) and 0.83 (95% CrI: 0.50, 1.27), respectively. For the data set comparing Aβ_{42} and p-tau, estimates were 1.02 (95% CrI: 0.57, 1.66), 0.94 (95% CrI: 0.53, 1.52), 0.90 (95%

CrI: 0.46, 1.50) and 0.78 (95% CrI: 0.40, 1.27), respectively. Across both motivating examples, estimates were larger for the meta-regression than the bivariate copula models. 95% CrIs corresponding to the standard deviations were also wider for the meta-regression models compared to the bivariate copula models, with the Gumbel copula model resulting in a 16% and 18% reduction in CrI width for between-studies standard deviations in sensitivity of Aβ_{42} and t-tau, respectively, compared to the meta-regression model with common between-studies correlation.

6.5.4 Between-studies correlation ρ_b

Posterior median between-studies correlation for Aβ_{42} compared to t-tau was estimated as -0.21 (-0.33, 0.03) using the Gumbel copula model, indicating negative association between sensitivities and specificities. For the data set comparing Aβ_{42} and p-tau, estimated between-studies correlation was -0.07 (-0.30, 0.27). Estimates were consistently lower, indicating stronger negative association between sensitivities and specificities across studies, for the bivariate copula models compared to the meta-regression model with common between-studies correlation. The 95% CrIs for the bivariate copula models were wide, and all spanned 0. When a separate between-studies correlation parameter was estimated for all possible combination of sensitivities and specificities across the two tests, the association ranged from -0.77 to 0.49 for the example comparing Aβ_{42} to t-tau and -0.48 to 0.50 for the example comparing Aβ_{42} and p-tau.

6.5.5 Convergence diagnostics

Convergence diagnostic plots for key model parameters are given in Appendix C.3-C.8. Interpretation of the diagnostic plots was described in Section 2.2.4. Due to the large number of model parameters, and as each model was fit to two motivational examples, a sample of diagnostic plots are presented; however, plots for each parameter, model, and example were checked to detect non-convergence. Trace plots did not show any systematic trends and MCMC chains initiated at different starting values appeared to mix well. Density plots display unimodal distributions for each parameter and no issues were detected with mixing of chains. Autocorrelation between samples decreases sharply, indicating rapid model convergence. Overall, no issues with convergence or mixing of chains were detected for any of the models or data sets.

Table 6.5: Posterior medians and 95% credible intervals estimated by fitting the meta-regression and bivariate copula models to data comparing the diagnostic accuracy of amyloid-β (Aβ_{42}, test 1) and total tau (t-tau, test 2) for detecting Alzheimer's disease dementia.

Parameter	Meta-regression, median (95% CrI)		Bivariate copula, median (95% CrI)	
	Common BSC	Separate BSC	Gaussian	Frank
η_{se1}	80.24 (71.35, 87.89)	79.45 (69.66, 87.63)	80.82 (73.11, 87.47)	81.03 (73.23, 87.75)
η_{se2}	76.06 (67.64, 82.79)	76.26 (68.26, 83.36)	76.36 (69.17, 82.94)	76.53 (68.86, 83.00)
η_{sp1}	70.39 (61.34, 79.38)	70.64 (61.32, 79.05)	70.33 (61.86, 78.72)	70.40 (61.36, 78.77)
η_{sp2}	72.90 (63.34, 82.14)	73.37 (63.77, 81.52)	72.87 (63.44, 81.08)	72.97 (63.89, 81.96)
σ_{se1}	0.962 (0.508, 1.437)	0.955 (0.526, 1.447)	0.802 (0.453, 1.261)	0.829 (0.472, 1.279)
σ_{se2}	0.721 (0.356, 1.131)	0.760 (0.429, 1.165)	0.640 (0.356, 1.013)	0.674 (0.374, 1.021)
σ_{sp1}	0.806 (0.444, 1.284)	0.808 (0.448, 1.238)	0.750 (0.406, 1.212)	0.762 (0.429, 1.224)
σ_{sp2}	0.881 (0.525, 1.316)	0.879 (0.551, 1.305)	0.846 (0.516, 1.265)	0.853 (0.528, 1.281)
ρ_b	-0.174 (-0.308, 0.044)	range: -0.772, 0.490	-0.212 (-0.333, 0.014)	-0.194 (-0.325, 0.038)

Parameter	Bivariate copula, median (95% CrI)		
	Gumbel	Clayton	Clayton 180°
η_{se1}	80.88 (73.43, 87.52)	81.12 (73.02, 88.12)	80.67 (73.09, 87.61)
η_{se2}	76.39 (69.43, 83.13)	76.60 (68.90, 83.47)	76.29 (68.92, 83.00)
η_{sp1}	70.26 (61.34, 78.36)	70.52 (61.28, 79.19)	70.27 (61.72, 78.56)
η_{sp2}	72.48 (63.66, 81.27)	73.02 (63.90, 82.18)	72.50 (63.61, 81.39)
σ_{se1}	0.798 (0.439, 1.231)	0.842 (0.486, 1.329)	0.803 (0.451, 1.264)
σ_{se2}	0.642 (0.355, 1.002)	0.681 (0.370, 1.051)	0.654 (0.367, 1.018)
σ_{sp1}	0.733 (0.375, 1.188)	0.772 (0.406, 1.232)	0.742 (0.383, 1.202)
σ_{sp2}	0.831 (0.502, 1.268)	0.861 (0.532, 1.325)	0.831 (0.505, 1.278)
ρ_b	-0.205 (-0.333, 0.030)	-0.178 (-0.328, 0.069)	-0.207 (-0.333, 0.022)

η_{se1}, η_{se2}, η_{sp1}, and η_{sp2} denote summary sensitivities and specificities of test 1 and test 2, respectively.
σ_{se1}, σ_{se2}, σ_{sp1}, and σ_{sp2} denote between-studies standard deviation in *logit*-transformed sensitivities and specificities of test 1 and test 2, respectively.
ρ_b denotes the between-studies correlation parameter.
BSC, between-studies correlation; CrI, credible interval

Table 6.6: Posterior medians and 95% credible intervals estimated by fitting the meta-regression and bivariate copula models to data comparing the diagnostic accuracy of amyloid-β (Aβ_{42}, test 1) and phosphorylated tau (p-tau, test 2) for detecting Alzheimer's disease dementia.

Parameter	Meta-regression, median (95% CrI)		Bivariate copula, median (95% CrI)	
	Common BSC	Separate BSC	Gaussian	Frank
η_{se1}	79.85 (69.11, 89.60)	79.84 (67.96, 89.21)	80.25 (68.92, 89.32)	80.42 (70.12, 90.04)
η_{se2}	75.85 (63.81, 86.66)	76.04 (63.83, 87.02)	76.60 (65.60, 86.63)	76.93 (66.13, 87.26)
η_{sp1}	72.64 (60.73, 84.12)	73.13 (61.50, 83.62)	72.74 (60.80, 82.99)	72.88 (61.16, 83.13)
η_{sp2}	75.95 (66.08, 85.33)	76.06 (66.08, 84.88)	75.90 (66.45, 84.79)	76.02 (65.88, 84.66)
σ_{se1}	1.038 (0.539, 1.693)	1.064 (0.592, 1.752)	1.018 (0.559, 1.675)	1.030 (0.578, 1.676)
σ_{se2}	0.965 (0.508, 1.571)	1.017 (0.562, 1.655)	0.936 (0.489, 1.531)	0.943 (0.517, 1.527)
σ_{sp1}	0.965 (0.503, 1.598)	0.939 (0.516, 1.519)	0.916 (0.470, 1.525)	0.923 (0.459, 1.511)
σ_{sp2}	0.829 (0.454, 1.375)	0.819 (0.453, 1.313)	0.785 (0.414, 1.288)	0.801 (0.416, 1.309)
ρ_b	-0.032 (-0.261, 0.286)	range: -0.479, 0.499	-0.071 (-0.297, 0.263)	-0.077 (-0.313, 0.265)

Parameter	Bivariate copula, median (95% CrI)		
	Gumbel	Clayton	Clayton 180°
η_{se1}	80.16 (68.76, 89.38)	80.20 (69.08, 89.40)	80.27 (69.36, 89.70)
η_{se2}	76.79 (66.06, 87.18)	76.45 (64.89, 86.67)	76.77 (65.56, 86.10)
η_{sp1}	72.45 (60.80, 82.73)	72.82 (61.27, 84.10)	72.28 (60.79, 83.22)
η_{sp2}	75.71 (65.91, 84.12)	76.18 (66.99, 84.97)	75.76 (66.04, 83.88)
σ_{se1}	1.024 (0.572, 1.664)	1.051 (0.585, 1.736)	1.023 (0.533, 1.633)
σ_{se2}	0.939 (0.528, 1.519)	0.964 (0.515, 1.564)	0.930 (0.491, 1.511)
σ_{sp1}	0.899 (0.455, 1.503)	0.948 (0.482, 1.584)	0.902 (0.454, 1.517)
σ_{sp2}	0.775 (0.397, 1.270)	0.813 (0.414, 1.337)	0.774 (0.403, 1.268)
ρ_b	-0.074 (-0.296, 0.268)	-0.059 (-0.283, 0.284)	-0.069 (-0.306, 0.265)

η_{se1}, η_{se2}, η_{sp1}, and η_{sp2} denote summary sensitivities and specificities of test 1 and test 2, respectively.
σ_{se1}, σ_{se2}, σ_{sp1}, and σ_{sp2} denote between-studies standard deviation in *logit*-transformed sensitivities and specificities of test 1 and test 2, respectively.
ρ_b denotes the between-studies correlation parameter.
BSC, between-studies correlation; CrI, credible interval

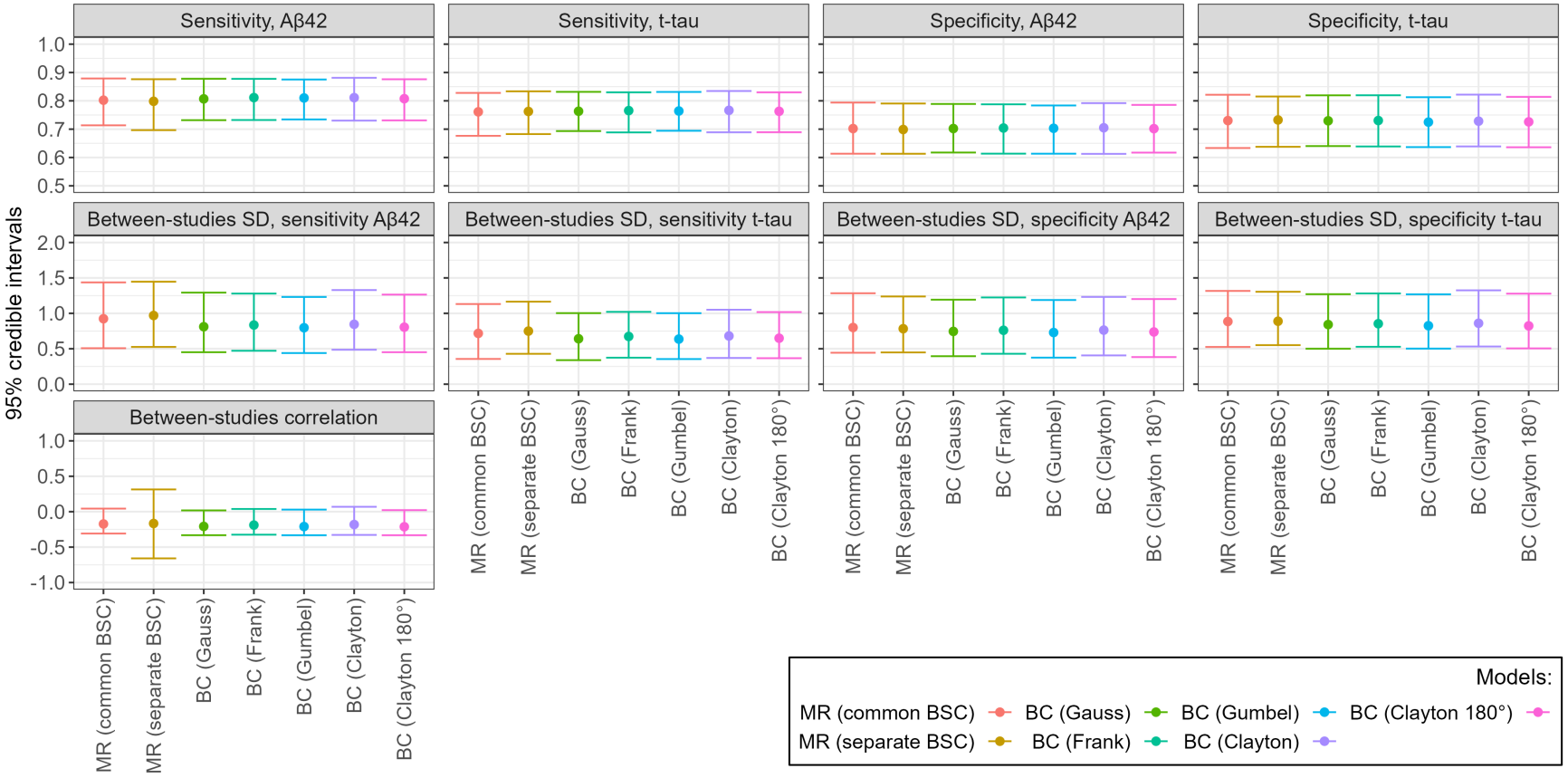


Figure 6.2: Posterior medians (solid dots) and 95% credible intervals (solid bars) of key test accuracy parameters across the meta-regression and bivariate copula (BC) models for amyloid-β (Aβ_{42}) and total tau (t-tau) data.

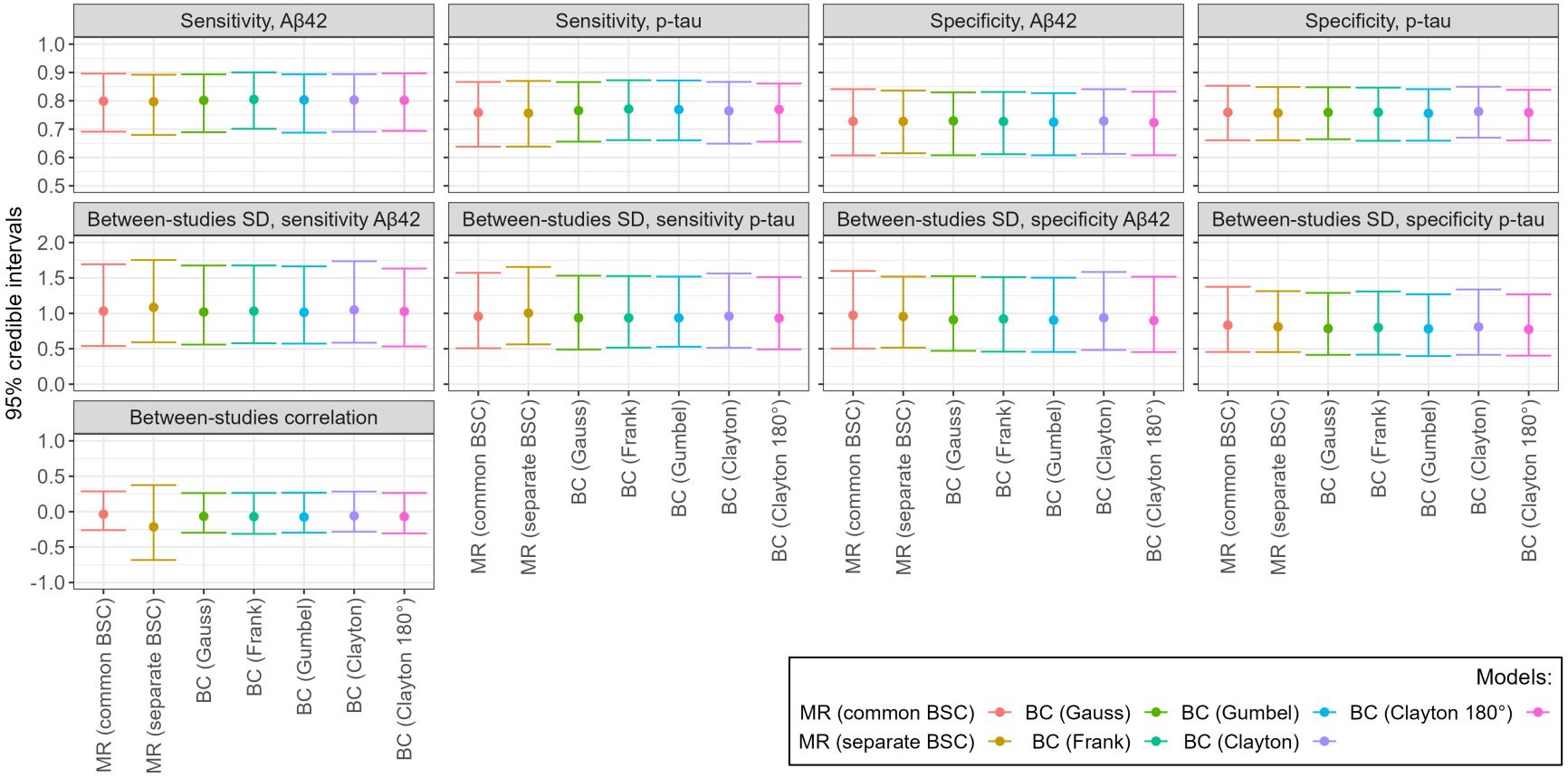


Figure 6.3: Posterior medians (solid dots) and 95% credible intervals (solid bars) of key test accuracy parameters across the meta-regression and bivariate copula (BC) models for amyloid-β (Aβ_{42}) and phosphorylated tau (p-tau) data.

6.6 Discussion

In this chapter, novel bivariate copula models for synthesising evidence on the accuracy of two diagnostic tests were developed, accounting for associations between and within studies present between two tests evaluated using a paired design. The new models offer a robust yet flexible approach to modelling comparative test accuracy data, maximising the available evidence base by making use of both study- and individual-level data where available. The bivariate copula method resulted in improved model fit compared to the currently recommended meta-regression approach.

When applied to the motivating data sets, the bivariate copula models resulted in lower point estimates of the between-studies standard deviation and correlation parameters. It has been hypothesised that when within-study dependencies are not taken into account, the 'excess' of the association manifests itself as an upwardly biased estimate of the between-studies heterogeneity parameters.[221, 207] This may explain the differences observed between the meta-regression and bivariate copula approach. The bivariate copula models yielded narrower CrIs for summary sensitivity and specificity parameters, likely due to the additional evidence on within-study associations utilised by the method. In a HTA context, sensitivity and specificity estimates are used in decision-making to determine whether a novel diagnostic test should be used in clinical practice over other technologies. Increased precision in these estimates aids the evaluation of clinical and cost-effectiveness, enabling more precise decisions on the most efficient use of health resources. Where the estimates are used to populate a health economic model, it is vital to capture the uncertainty in the parameter estimates to provide appropriate recommendations for reimbursement.[222]

There are a number of existing meta-analysis models for jointly synthesising data on multiple diagnostic tests, as summarised in Table 3.1. Trikalinos et al [85] proposed extending the BRMA model (Section 2.4.2.1), using multinomial likelihoods to capture within-study dependencies present between two tests evaluated under a within-subject design. Multivariate methods such as this make strong distributional assumptions about the marginal variables, and require cross-classified data from all studies. A scoping review of meta-analysis models for three or more diagnostic tests by Veroniki et al found that the majority of existing methods require cross-classified test results,[87] although publications of diagnostic accuracy studies regularly do not facilitate access to this data.[9, 10, 11, 4] Indeed, in the review undertaken to identify

a motivating example for this chapter, full cross-classifications were reported in only 20% of comparative studies. Meta-analysis of multiple outcomes is, in fact, often hindered when within study correlations are not reported in primary studies.[223, 224] Copula methodology relaxes the requirement for cross-classifications across all studies, reflecting current reporting standards and making its application in systematic reviews more generalisable.

The flexibility of copulas, of which a number of types have been defined and explored, make them a natural approach to model diagnostic accuracy data, which are often non-normal and likely to exhibit strong tail dependence when marginal sensitivities (and specificities) are high. The adaptable nature of copulas increases their suitability for a range of diagnostic data sets, or indeed any type of data in which multiple, correlated outcomes are present. An advantage of a copula approach to modelling related variables is the estimation of a single association parameter, θ, in contrast to the need to estimate both standard deviation and correlation parameters when fitting a multivariate normal distribution. If these parameters conflict with one another, for example when extreme values are sampled from the prior distribution, it can cause the Bayesian sampler to fail to converge. Indeed, the heterogeneity parameters estimated using a four-dimensional normal distribution at the between-studies level in the bivariate copula models were noted to be highly sensitive to choice of prior distribution and starting values.[44, 225]

The proposed methodology is subject to a number of potential limitations. Inference focussed on sensitivity and specificity, but alternative test accuracy measures may be of greater clinical utility in practice when true disease status is unknown. Nonetheless, it is straightforward to derive estimates of predictive values, likelihood ratios and DORs from posterior estimates of sensitivity and specificity. Meta-analysis of these measures directly is subject to limitations. Predictive values are often more heterogeneous than sensitivity and specificity due to increased variation with disease prevalence, which leads to reduced goodness of fit for meta-analysis models.[226] Bivariate meta-analysis of likelihood ratios can lead to implausible corresponding values of sensitivity and specificity (i.e. <0).[227] Synthesising DORs rather than paired test accuracy measures results in the loss of ability to distinguish between tests with high sensitivity and high specificity.[66] Conversely, paired test accuracy measures, including sensitivity and specificity, hinder the ranking of tests. Trade-off between sensitivity and specificity should be considered in the context of the clinical use of the test; the potential consequences of a false positive or false negative result

may not be equal.[11]

It can be challenging to interpret copula dependence parameters directly, which are not as intuitive as correlation between two variables. While conversion between dependence parameters and correlation coefficients are possible (Table 6.3), this may be less clinically applicable than joint or combined test accuracy measures, and only makes use of studies where IPD are available to inform estimates, rather than pooling estimates across studies accounting for study weight and variability. The bivariate copula models do not explicitly capture joint accuracy measures, which would allow the comprehensive evaluation of diagnostic pathways in which multiple tests are applied. The models' focus primarily lies in test comparison; extension to a trivariate copula in the following chapter enables further assessment of combined test accuracy. In selecting an appropriate copula model, consideration should be given to model fit, using measures such as WAIC. The Gumbel copula was the best fit for the motivational example; however, underlying dependence structures may vary by patient and test characteristics and different copula types should be compared as part of the model fitting process.

Alternative methods for synthesising data on two diagnostic tests incorporate other, desirable features that could be considered for future copula model development. Menten and Lesaffre [88] and Lian et al [97] described meta-analysis models that account for imperfect reference standards, the former using latent class analysis and the latter by comparing tests within a missing data framework. Methods introduced by Owen et al [95] and Hoyer and Kuss [96] allow for multiple thresholds per test by including threshold information as a covariate, making use of a greater proportion of the available literature. Several models allow the addition of single arm studies.[89, 93, 97] Sensitivity and specificity are known to vary with disease prevalence.[228] Hoyer and Kuss [100] and Nikoloulopoulos [101] suggested meta-analysis models for a single diagnostic test, accounting for disease prevalence using a trivariate copula. The models could also be adapted to include a copula at the between-studies level,[90, 92, 94, 98] removing the need to estimate between-studies heterogeneity parameters that may be sensitive to prior distribution choices.

Several models allow the addition of single test studies,[89, 93, 97] although caution is advisable as non-comparative studies are known to introduce bias to test accuracy estimates,[8] as, unlike network meta-analysis of interventional research - which includes indirect evidence on comparative treatment accuracy through a common comparator - there is no equivalent control test through which adjustment for differences

in accuracy between studies is possible.[229, 230] A compromise may be to downgrade the single test studies within the meta-analysis model. While meta-analysis models have been described for combining real world evidence from single arm studies with randomised controlled trials, where single arm data are down-weighted through prior specification or variance inflation,[231] further methodological development is needed to incorporate these techniques in diagnostic meta-analysis models. Recommendations for incorporating real-world evidence into meta-analyses of treatment effects highlight the importance of quality assessment of the real-world studies and the need for sensitivity analyses to understand the uncertainty around meta-analytic estimates.[232] Leeflang et al suggested including only single test studies that appear to be at lower risk of bias as another possible solution to combining comparative with non-comparative diagnostic test accuracy studies.[12] There is also the potential to incorporate IPD obtained from linked electronic health records or cohorts such as Dementias Platform UK, which facilitates access to cohort data on over 3 million participants to enable research into the detection and treatment of dementia. IPD could be incorporated to the bivariate copula models to maximise the use of the available evidence base through the inclusion of patient-level covariates that measure the effect of patient characteristics on summary accuracy measures.[233]

6.7 Chapter summary

In this chapter, novel Bayesian meta-analysis models were developed for synthesising data on two diagnostic tests compared to a common reference standard under a paired design, using bivariate copulas to capture within-study dependencies between multiple tests. The models were compared to a meta-regression approach that ignores within-study associations using a motivational example in Alzheimer's disease dementia, as described in Chapter 5. Differences were observed in estimates of the between-studies heterogeneity parameters produced by the two methods; this is reflective of findings in the previous chapter, which demonstrated the impact of accounting for within-study dependencies on key test accuracy measures. The bivariate copula methodology introduced in this chapter makes use of a broader evidence base by utilising both study-level and IPD, where available. The models are applicable to a wide range of disease areas, aiding evidence synthesis of commonly reported data items from diagnostic test accuracy studies to improve healthcare policy and decision-making. The following chapter extends these models to incorporate

a trivariate copula to fully account for within-study associations between two tests when cross-classified data are available for each study, allowing the direct estimation of joint accuracy measures.

Chapter 7

Joint meta-analysis of two diagnostic tests using a trivariate copula to model combined test accuracy

7.1 Chapter overview

In the previous chapter, bivariate copulas were used to model the associations between the true positive and true negative results of each test, allowing the evaluation of comparative test accuracy in a meta-analysis framework. Cross-classified data, where available, are incorporated into the model through the copula dependence parameter. In this chapter, the bivariate copula methodology is extended to a trivariate copula model, developed to directly estimate both joint and marginal test results where cross-classified data are available for each study. A simulation study in Chapter 5 demonstrated the impact of accounting for within-study dependencies present between multiple tests evaluated using a within-subject design on joint accuracy measures. Furthermore, a methodological review in Chapter 3 found that present methods that capture these associations often suffer from computational difficulties and inflexible distributional assumptions, thus motivating the need for further model development. The novel Bayesian meta-analysis models introduced in this chapter, which account for within-study dependencies between sensitivities and specificities using a flexible trivariate copula approach, aim to address limitations of

existing meta-analytic methods for synthesising cross-classified data.

7.2 Introduction

In Chapter 5, ignoring within-study dependencies was shown to lead to underestimation of joint sensitivity and specificity, particularly when associations were strong. This emphasises the importance of selecting an appropriate meta-analysis model to estimate the joint accuracy of two more diagnostic tests; in particular, one which fully captures within-study associations present between multiple tests evaluated under a paired design. In the previous chapter, bivariate copulas were used to model within-study associations between sensitivities and specificities of two tests assessed using a paired design. While the bivariate copula models provide a flexible framework for test comparison, able to make use of the entire study pool where cross-classifications reported for a subset of studies only, they do not directly estimate joint accuracy parameters used to inform summary estimates of combined test accuracy. Existing methods for evaluating the joint accuracy of two diagnostic tests often make restrictive assumptions about the distributions of the marginal variables, and can run into computational difficulties.

Building on novel model development in Chapter 6, the following chapter describes an extension to the bivariate copula models for evaluating the accuracy of two diagnostic tests in a meta-analysis framework. Bayesian meta-analysis models for synthesising comparative diagnostic accuracy studies are described, using trivariate copulas to account for within-study dependencies between two tests evaluated using a paired design. Joint sensitivity and specificity are directly estimated in addition to marginal sensitivities and specificities, requiring cross-classified data from each study to inform the joint accuracy measures. Models are developed for three types of trivariate copula: Frank, Gumbel and Clayton. Section 7.3 describes the motivating example, simulated as part of Chapter 5, that the developed methodology is applied to. Section 7.4 describes the trivariate copula methodology, extending the bivariate copula framework introduced in the previous chapter to a trivariate copula. The application of the methods is demonstrated in Section 7.5, where the results of fitting the trivariate copula models to the simulated example are presented. Section 7.6 concludes the chapter with a discussion.

7.3 Motivational examples

A data generating mechanism for simulating full cross-classifications based on a motivating example was developed in Chapter 5. The process is briefly summarised here; for a full explanation see Section 5.3.4.1. The motivational data set, provided in Table 5.2, compares the diagnostic accuracy of CSF Aβ42 and t-tau for diagnosing Alzheimer's disease dementia. The number of studies were fixed, while the prevalence of the target condition and the number of patients in each study was simulated based on descriptive statistics estimated using the motivating example. The marginal and joint sensitivities and specificities were simulated from a six-dimensional normal distribution at the between-studies level, with fixed between-studies standard deviation and correlation. At the within-study level, the number of patients with each possible combination of test results in the diseased and non-diseased groups, populated by plausible values of marginal and joint sensitivities and specificities based on comparative diagnostic accuracy data extracted as part of the literature review, were simulated from a multinomial distribution. Two data sets were simulated with moderate within-study associations: in the first, sensitivities and specificities were high (0.8); in the second, sensitivities and specificities were low (0.4). The simulated data sets that the trivariate model is fit to are provided in Table 7.1 and 7.2.

Table 7.1: Simulated data on the accuracy of two diagnostic tests evaluated using a paired design. Sensitivities and specificities are high and within-study associations are moderate.

Study	Disease									No disease								
	tp_1	tp_2	fn_1	fn_2	x_{11}^D	x_{10}^D	x_{01}^D	x_{00}^D	N^D	tn_1	tn_2	fp_1	fp_2	$x_{11}^{\bar{D}}$	$x_{10}^{\bar{D}}$	$x_{01}^{\bar{D}}$	$x_{00}^{\bar{D}}$	$N^{\bar{D}}$
Study 1	26	29	6	3	25	1	4	2	32	51	55	15	11	0	15	11	40	66
Study 2	21	16	3	8	16	5	0	3	24	36	42	9	3	3	6	0	36	45
Study 3	36	34	6	8	31	5	3	3	42	42	40	8	10	2	6	8	34	50
Study 4	49	44	8	13	40	9	4	4	57	49	55	15	9	9	6	0	49	64
Study 5	48	47	7	8	40	8	7	0	55	57	59	15	13	12	3	1	56	72
Study 6	16	15	3	4	15	1	0	3	19	16	19	8	5	2	6	3	13	24
Study 7	38	41	12	9	36	2	5	7	50	46	41	13	18	6	7	12	34	59
Study 8	33	25	8	16	22	11	3	5	41	33	39	13	7	4	9	3	30	46
Study 9	26	25	7	8	23	3	2	5	33	39	37	8	10	6	2	4	35	47
Study 10	22	22	9	9	17	5	5	4	31	42	43	12	11	0	12	11	31	54
Study 11	41	41	9	9	37	4	4	5	50	39	36	11	14	6	5	8	31	50
Study 12	39	39	8	8	33	6	6	2	47	61	61	11	11	1	10	10	51	72
Study 13	43	43	10	10	33	10	10	0	53	68	67	12	13	11	1	2	66	80
Study 14	21	20	8	9	20	1	0	8	29	48	58	18	8	5	13	3	45	66
Study 15	35	29	7	13	25	10	4	3	42	40	41	9	8	7	2	1	39	49
Study 16	16	16	6	6	14	2	2	4	22	21	22	7	6	5	2	1	20	28
Study 17	18	17	8	9	13	5	4	4	26	43	46	9	6	3	6	3	40	52
Study 18	14	13	2	3	13	1	0	2	16	24	23	2	3	0	2	3	21	26

tp_j is the number of true positives, fn_j false negatives, tn_j true negatives, and fp_j false positives for each of the $j = 1, 2$ tests.
x_{kl}^D is the number of participants with the target condition and $x_{kl}^{\bar{D}}$ is the number of participants without the target condition with each combination of test results. $k, l = 0, 1$ denote the first and second test, respectively, where 0 indicates a negative test result and 1 indicates a positive test result.

Table 7.2: Simulated data on the accuracy of two diagnostic tests evaluated using a paired design. Sensitivities and specificities are low and within-study associations are moderate.

Study	Disease									No disease								
	tp_1	tp_2	fn_1	fn_2	x^D_{11}	x^D_{10}	x^D_{01}	x^D_{00}	N^D	tn_1	tn_2	fp_1	fp_2	$x^{\bar{D}}_{11}$	$x^{\bar{D}}_{10}$	$x^{\bar{D}}_{01}$	$x^{\bar{D}}_{00}$	$N^{\bar{D}}$
Study 1	13	17	19	15	12	1	5	14	32	22	23	44	43	32	12	11	11	66
Study 2	12	13	13	12	9	3	4	9	25	16	18	27	25	20	7	5	11	43
Study 3	13	12	28	29	12	1	0	28	41	24	29	35	30	29	6	1	23	59
Study 4	4	4	16	16	4	0	0	16	20	6	9	17	14	14	3	0	6	23
Study 5	28	28	27	27	19	9	9	18	55	24	25	48	47	46	2	1	23	72
Study 6	4	4	15	15	4	0	0	15	19	14	11	10	13	9	1	4	10	24
Study 7	12	14	25	23	12	0	2	23	37	38	25	39	52	37	2	15	23	77
Study 8	22	31	30	21	14	8	17	13	52	35	44	34	25	17	17	8	27	69
Study 9	13	15	22	20	11	2	4	18	35	24	28	33	29	22	11	7	17	57
Study 10	11	10	11	12	6	5	4	7	22	22	19	24	27	24	0	3	19	46
Study 11	31	23	28	36	19	12	4	24	59	35	29	38	44	34	4	10	25	73
Study 12	17	11	20	26	10	7	1	19	37	20	25	28	23	17	11	6	14	48
Study 13	14	12	17	19	10	4	2	15	31	26	16	29	39	25	4	14	12	55
Study 14	12	11	17	18	11	1	0	17	29	27	37	39	29	27	12	2	25	66
Study 15	18	12	24	30	8	10	4	20	42	16	17	33	32	31	2	1	15	49
Study 16	7	6	15	16	5	2	1	14	22	9	10	19	18	17	2	1	8	28
Study 17	10	9	16	17	5	5	4	12	26	18	20	34	32	27	7	5	13	52
Study 18	5	4	11	12	4	1	0	11	16	11	9	15	17	11	4	6	5	26

tp_j is the number of true positives, fn_j false negatives, tn_j true negatives, and fp_j false positives for each of the $j = 1, 2$ tests.
x^D_{kl} is the number of participants with the target condition and $x^{\bar{D}}_{kl}$ is the number of participants without the target condition with each combination of test results. $k, l = 0, 1$ denote the first and second test, respectively, where 0 indicates a negative test result and 1 indicates a positive test result.

7.4 Methods

In this section, novel methods for jointly analysing data on two diagnostic tests are introduced, using a trivariate copula to model within-study dependencies between sensitivities and specificities. A brief overview of trivariate copula theory and the copula types explored in this chapter are presented in Section 7.4.1 and 7.4.2. The novel trivariate copula models are described in Section 7.4.3. Section 7.4.4 describes the bootstrap method used to estimate within-study dependencies and Section 7.4.5 details the estimation methods.

7.4.1 Trivariate copula theory

Bivariate copula methodology, introduced in Section 6.4.1, can be extended from two- to three-dimensions, known as a trivariate copula. A trivariate copula is a trivariate CDF with uniform marginal distributions on the interval [0,1].[214, 211] Let there be three continuous, correlated random variables, u_1, u_2 and u_4. Similarly to the bivariate case, Sklar's theorem states that any trivariate distribution, H, can be expressed in terms of univariate marginal distribution functions, F_1, F_2 and F_3, and a copula, C, which describes their relationship to one another.[215] θ is the copula dependence parameter, which captures the association between the three random variables. The copula links the univariate marginal distribution functions to their trivariate distribution function, allowing separate specification of their dependence structure:

$$H\left(u_1, u_2, , u_3 \theta\right) = C\left(F_1\left(u_1\right), F_2\left(u_2\right), , F_3\left(u_3\right) \theta\right). \tag{7.1}$$

For discrete random variables, Equation 6.2 is extended to derive the joint probability

mass function for the trivariate case via finite differences:

$$\begin{aligned} h(u_1, u_2, u_3\theta) = {} & C(F_1(u_1), F_2(u_2), F_3(u_3), \theta) - \\ & C\left(F_1(u_1 - 1), F_2(u_2), F_3(u_3), \theta\right) - \\ & C\left(F_1(u_1), F_2(u_2 - 1), F_3(u_3), \theta\right) - \\ & C\left(F_1(u_1), F_2(u_2), F_3(u_3 - 1), \theta\right) + \\ & C\left(F_1(u_1 - 1), F_2(u_2 - 1), F_3(u_3), \theta\right) + \\ & C\left(F_1(u_1 - 1), F_2(u_2), F_3(u_3 - 1), \theta\right) + \\ & C\left(F_1(u_1 - 1), F_2(u_2), F_3(u_3 - 1), \theta\right) - \\ & C\left(F_1(u_1 - 1), F_2(u_2 - 1), F_3(u_3 - 1), \theta\right) \end{aligned} \tag{7.2}$$

7.4.2 Families of copulas

As described in Section 6.4.2, different classes of copulas possess various properties that make them more appropriate for modelling certain types of relationships between variables, such as strong tail dependencies or asymmetry. The bivariate copula families described in the previous chapter can be extended to the trivariate case. Three types of trivariate copulas belonging to the Archimedean family are implemented within the diagnostic meta-analytic framework in this chapter. The Frank, Gumbel and Clayton copulas were specified to explore different assumptions about within-study associations present between two tests, such as weak tail dependencies or tail dependencies that are stronger at one end of the distribution than the other (asymmetry). In order to visualise the dependence structures modelled by the three trivariate copula types, 2000 samples were simulated from each within R version 4.3.1 using the *copula* version 1.1-2 package (Figure 7.1).

7.4.2.1 Archimedean copulas

The bivariate Archimedean copulas described in the previous chapter can be extended to trivariate copulas, capturing the relationship between three random variables with varying dependence structures. For random variables u_1, u_2 and u_3, trivariate Archimedean copulas are given by:

$$C\left(u_1, u_2, u_3\right) = \phi^{-1}\left[\phi\left(u_1\right) + \phi\left(u_2\right) + \phi\left(u_3\right)\right] \tag{7.3}$$

where ψ and ψ^{-1} are the generator function and inverse generator function of the

copula, respectively. The generator functions presented in Table 6.3 can be applied to Equation 7.3 to implement different types of trivariate Archimedean copulas.

7.4.2.1.1 Frank copula Introduced in Section 6.4.2, the Frank copula is a symmetric Archimedean copula with weak tail dependence that can model both positive and negative dependence (see Figure 7.1). Given three random variables, u_1, u_2 and u_3, and a copula dependence parameter, θ, the Frank copula is expressed as:

$$C_{Frank}\left(u_1, u_2, u_3\right) = -\frac{1}{\theta}\log\left[1 + \frac{\left(e^{-\theta u_1} - 1\right)\left(e^{-\theta u_2} - 1\right)\left(e^{-\theta u_3} - 1\right)}{\left(e^{-\theta} - 1\right)^2}\right] \tag{7.4}$$

7.4.2.1.2 Gumbel copula Recall that Gumbel copula is an asymmetric copula exhibiting positive, right-hand tail dependence (see Figure 7.1). For random variables, u_1, u_2 and u_3, and copula dependence parameter θ, the Gumbel copula is defined as:

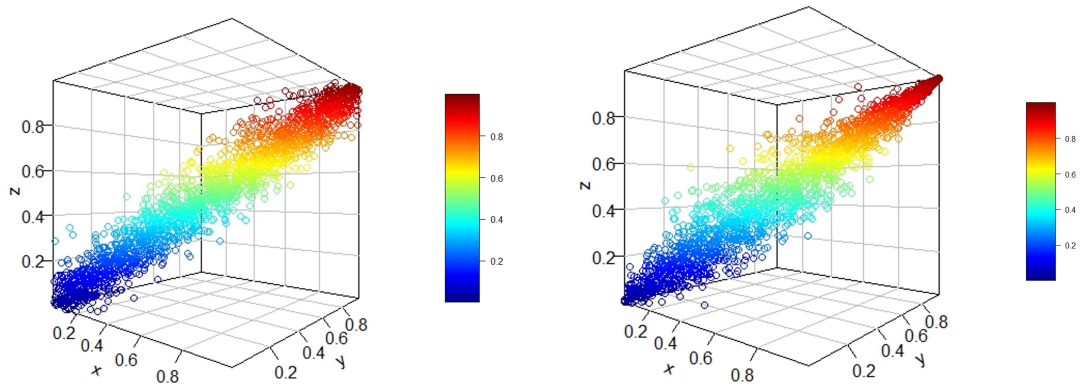
$$C_{Gumbel}\left(u_1, u_2, u_3\right) = e^{\left[-\left[(-\log(u_1))^{-\theta} + (-\log(u_2))^{-\theta} + (-\log(u_3))^{-\theta}\right]^{-\frac{1}{\theta}}\right]} \tag{7.5}$$

7.4.2.1.3 Clayton copula As described in Section 6.4.2, the Clayton copula is an asymmetric Archimedean copula exhibiting positive, left-hand tail dependence (see Figure 7.1). For three random variables, u_1, u_2 and u_3, and a copula dependence parameter, θ, the Clayton copula is expressed as follows:

$$C_{Clayton}\left(u_1, u_2, u_3\right) = \left[u_1^{-\theta} + u_2^{-\theta} + u_3^{-\theta} - 2\right]^{-\frac{1}{\theta}} \tag{7.6}$$

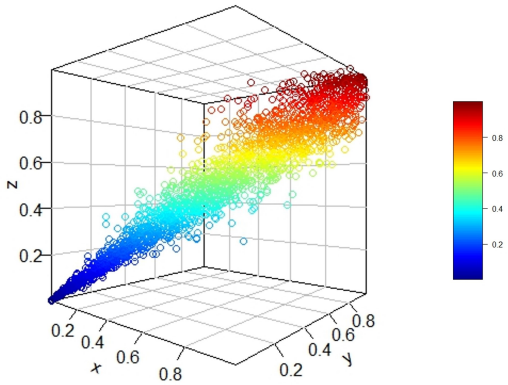
7.4.3 Trivariate copula model

The bivariate copula model proposed in Chapter 6 (Section 6.4.4) is extended to a trivariate copula model. Cross-classified data from each study, containing counts of combined test results in patient groups with and without the target condition, are assumed to follow trivariate distributions with binomial marginal distributions. Correlation between joint and marginal test accuracy measures (on the *logit* scale) is incorporated through a six-dimensional normal distribution at the between-studies level.



(a) Frank copula

(b) Gumbel copula



(c) Clayton copula

Figure 7.1: Simulated samples from trivariate a) Frank, b) Gumbel, and c) Clayton copulas. 2000 samples were simulated for each copula type, with fixed Spearman's correlation coefficient $\rho_s = 0.95$.

A trivariate copula captures within-study dependencies between the two tests. For the diseased patient group in each study, three counts are used to estimate three test accuracy measures: the marginal sensitivities of the two tests, $se_{i,1}$ and $se_{i,2}$, and their joint sensitivity, $se_{i,*}$. As in the bivariate copula model, $se_{i,1}$ and $se_{i,2}$ are estimated from the number of true positive results for each test, denoted $x^D_{i,1\cdot}$ for the first test and $x^D_{i,\cdot 1}$ for the second for consistency with notation for cross-classified data (Equation 7.7). Joint sensitivity, $se_{i,*}$, is estimated from the number of patients positive on both tests, denoted $x^D_{i,11}$ (corresponding to notation in Table 2.2). Similarly, the marginal specificities, $sp_{i,1}$ and $sp_{i,2}$, and joint specificity, $sp_{i,*}$ are estimated from counts $x^{\bar{D}}_{i,0\cdot}$, $x^{\bar{D}}_{i,\cdot 0}$ and $x^{\bar{D}}_{i,00}$ for the non-diseased patient group (Equation 7.8).

$$\begin{aligned} x^D_{i,1\cdot} &= x^D_{i,10} + x^D_{i,11} \\ &= tp_{i,1} \\ \\ x^D_{i,\cdot 1} &= x^D_{i,01} + x^D_{i,11} \\ &= tp_{i,2} \end{aligned} \tag{7.7}$$

$$\begin{aligned} x^{\bar{D}}_{i,0\cdot} &= x^{\bar{D}}_{i,01} + x^{\bar{D}}_{i,00} \\ &= tn_{i,1} \\ x^{\bar{D}}_{i,\cdot 0} &= x^{\bar{D}}_{i,10} + x^{\bar{D}}_{i,00} \\ &= tn_{i,2} \end{aligned} \tag{7.8}$$

For study $i = 1, \ldots, I$, the joint and marginal counts follow trivariate distributions $h(se_{i,1}, se_{i,2}, se_{i,*}, N^D_i, \theta_{i,se})$ and $h(sp_{i,1}, sp_{i,2}, sp_{i,*}, N^{\bar{D}}_i, \theta_{i,sp})$, respectively, with binomial marginal distributions. N^D_i and $N^{\bar{D}}_i$ denote the number of patients with and without the target condition, respectively, as determined by the reference standard.

$$\begin{pmatrix} x^D_{i,1\cdot} \\ x^D_{i,\cdot 1} \\ x^D_{i,11} \end{pmatrix} \sim h\left(se_{i,1}, se_{i,2}, se_{i,*}N^D_i, \theta_{i,se*}\right) \tag{7.9}$$

$$\begin{pmatrix} x^{\bar{D}}_{i,0\cdot} \\ x^{\bar{D}}_{i,\cdot 0} \\ x^{\bar{D}}_{i,00} \end{pmatrix} \sim h\left(sp_{i,1}, sp_{i,2}, sp_{i,*}N^D_i, \theta_{i,sp*}\right) \tag{7.10}$$

At the between-studies level, the true marginal and joint sensitivities and specificities

are transformed using a logit link function. The true study-level effects follow a six-dimensional multivariate normal distribution with means μ_{se1}, μ_{se2}, μ_{se*}, μ_{sp1}, μ_{sp2}, and μ_{sp*}. Between-study variances are denoted σ^2_{se1}, σ^2_{se2}, σ^2_{se*}, σ^2_{sp1}, σ^2_{sp2}, and σ^2_{sp*}. ρ_b corresponds to the between-studies correlation parameter.

$$logit\left(se_{i,1}\right) = \mu_{i,se1}, \quad logit\left(se_{i,2}\right) = \mu_{i,se2}, \quad logit\left(se_{i,*}\right) = \mu_{i,se*}, \tag{7.11}$$

$$logit\left(sp_{i,1}\right) = \mu_{i,sp1}, \quad logit\left(sp_{i,2}\right) = \mu_{i,sp2}, \quad logit\left(sp_{i,*}\right) = \mu_{i,sp*},$$

$$\begin{pmatrix} \mu_{i,se1} \\ \mu_{i,se2} \\ \mu_{i,se*} \\ \mu_{i,sp1} \\ \mu_{i,sp2} \\ \mu_{i,sp*} \end{pmatrix} \sim Normal \left(\begin{pmatrix} \mu_{se1} \\ \mu_{se2} \\ \mu_{se*} \\ \mu_{sp1} \\ \mu_{sp2} \\ \mu_{sp*} \end{pmatrix}, \quad \boldsymbol{\xi} \right),$$

$\boldsymbol{\xi}$ is parametrised as follows:

$$\boldsymbol{\xi} = \begin{pmatrix} \sigma^2_{se1} & \rho_b\sigma_{se1}\sigma_{se2} & \rho_b\sigma_{se1}\sigma_{se*} & \rho_b\sigma_{se1}\sigma_{sp1} & \rho_b\sigma_{se1}\sigma_{sp2} & \rho_b\sigma_{se1}\sigma_{sp*} \\ & \sigma^2_{se2} & \rho_b\sigma_{se2}\sigma_{se*} & \rho_b\sigma_{se2}\sigma_{sp1} & \rho_b\sigma_{se2}\sigma_{sp2} & \rho_b\sigma_{se2}\sigma_{sp*} \\ & & \sigma^2_{se*} & \rho_b\sigma_{se*}\sigma_{sp1} & \rho_b\sigma_{se*}\sigma_{sp2} & \rho_b\sigma_{se*}\sigma_{sp*} \\ & & & \sigma^2_{sp1} & \rho_b\sigma_{sp1}\sigma_{sp2} & \rho_b\sigma_{sp1}\sigma_{sp*} \\ & & & & \sigma^2_{sp2} & \rho_b\sigma_{sp2}\sigma_{sp*} \\ & & & & & \sigma^2_{sp*} \end{pmatrix} \tag{7.12}$$

Post-sampling, summary test accuracy measures are derived through an inverse-*logit* transformation. Both marginal and joint sensitivities and specificities are estimated directly within the model.

$$\upsilon_{se1} = logit^{-1}\left(\mu_{se1}\right), \quad \upsilon_{se2} = logit^{-1}\left(\mu_{se2}\right), \quad \upsilon_{se*} = logit^{-1}\left(\mu_{se*}\right), \tag{7.13}$$

$$\upsilon_{sp1} = logit^{-1}\left(\mu_{sp1}\right), \quad \upsilon_{sp2} = logit^{-1}\left(\mu_{sp2}\right), \quad \upsilon_{sp*} = logit^{-1}\left(\mu_{sp*}\right)$$

Using the marginal and joint accuracy measures, it is also possible to derive summary estimates of combined test accuracy, based on the 'AND' ('both positive') rule or the 'OR' ('either positive') rule (see Section 2.3.5). Estimation of summary combined test accuracy based on the 'AND' rule, analogous to summary joint sensitivity, is

straightforward:

$$\upsilon_{AND} = \upsilon_{se*} \tag{7.14}$$

Summary combined test accuracy based on the 'OR' rule is estimated through the relationships described in Equation 2.12 and 7.7:

$$\begin{aligned} OR &= \frac{x_{10}^D + x_{01}^D + x_{11}^D}{N^D} \\ &= \frac{\left(x_{1\cdot}^D - x_{11}^D\right) + \left(x_{\cdot 1}^D - x_{11}^D\right) + x_{11}^D}{N^D} \\ &= \frac{x_{1\cdot}^D + x_{\cdot 1}^D - x_{11}^D}{N^D} \\ &= \frac{x_{1\cdot}^D}{N^D} + \frac{x_{\cdot 1}^D}{N^D} - \frac{x_{11}^D}{N^D} \end{aligned} \tag{7.15}$$

Therefore:

$$\upsilon_{OR} = \upsilon_{se1} + \upsilon_{se2} - \upsilon_{se*} \tag{7.16}$$

Prior distributions were placed on the unknown parameters in the model. *Logit*-transformed summary sensitivities and specificities were assumed to follow a minimally informative $Normal\left(0, 10^2\right)$ prior distribution. The between-studies variance parameters were restricted to positive values and were assumed to have a uniform $Half - Normal\left(0, 2.5^2\right)$ prior distribution. For the between-studies correlation parameter, the Fisher z-transformation was used, i.e $\rho_b = \tanh\left(z\right), z \sim Normal\left(0, 0.8\right)$). This transformation produces an approximately normal distribution bound between $[-1, 1]$.

7.4.4 Bootstrapping methods to obtain copula dependence parameter

The copula dependence parameters were estimated from the simulated cross-classifications, which allow the reconstruction of IPD, using a double bootstrap method.[219, 207] Prior to bootstrapping, IPD was recreated for each study by transforming the cross-classified counts into a data set of zeros and ones indicating each patients' test results and disease status. Bootstrapping involves repeatedly sampling the IPD (with replacement) from a single study to create many simulated data sets. Sensitivities and specificities were estimated for each simulated data set, then the association between them estimated across the multiple bootstrap

samples. For each of the $i = 1, ..., I$ studies, $\theta_{i,se}$ and $\theta_{i,sp}$ were estimated using maximum likelihood estimation. Code for the bootstrapping method is provided in Appendix D.1.

7.4.5 Estimation

All models were implemented in a Bayesian framework using Stan version 2.32.2 [53] within R version 4.3.1 [203] via the rstan version 2.32.5 package.[54] A non-centred parameterisation was used for all models to reduce dependencies between successive levels of the hierarchical structure, further increasing the efficiency of the sampler (see Section 2.2.3.4).[55] Copula dependence parameters were estimated using the bootstrap method described in Section 7.4.4 using R version 4.2.1.[54] After discarding 5,000 burn-in iterations, posterior estimates were obtained using three chains initialised at different starting values, consisting of 20,000 iterations each. The three chains were initialised at different starting values to aid the evaluation of model convergence. 95% CrIs were computed as HPD intervals.

Model convergence was assessed using diagnostics described in Section 2.2.4, including trace plots, density plots, and autocorrelation plots. Model fit was compared by calculating the WAIC using the R package *loo*,[220] with a smaller WAIC indicating better model fit.[58] Stan code for the three trivariate copula models is available in the Appendix D.2.

7.5 Results

7.5.1 Comparison of model fit

Table 7.3 presents the values of the WAIC across the three fitted models for two simulated examples. The Clayton copula model was the poorest fit for the data set simulated assuming high sensitivities and specificities and moderate within-study associations, corresponding to the largest WAIC of 238.3. Of the three trivariate copula models, the Gumbel copula was the best fit for the data with the smallest WAIC of 225.0. The differences in WAIC between the Gumbel and Frank (WAIC = 229.5) copulas were small and the evidence to support the Gumbel copula over the Frank copula is marginal. There was evidence that the Clayton copula (WAIC = 238.3) resulted in a poorer fit compared to the other trivariate copulas.

The Clayton copula was also the poorest fit for the data set with low sensitivities and specificities and moderate within-study associations, corresponding to the largest WAIC of 238.1. There was evidence that the Gumbel copula was a better fit for the data than the other trivariate copula models, corresponding to the smallest WAIC of 221.5. The difference between the Frank copula (WAIC = 236.9) and the Clayton copula (WAIC = 238.1) was marginal.

Table 7.3: Values of the widely applicable information criterion (WAIC) the meta-regression and trivariate copula models for each simulated example.

Model	WAIC	Change in WAIC compared to the model with the poorest fit
High sensitivities and specificities with moderate within-study associations		
Trivariate copula (Frank)	229.5	-8.8
Trivariate copula (Gumbel)	225.0	-13.3
Trivariate copula (Clayton)	238.3	-
Low sensitivities and specificities with moderate within-study associations		
Trivariate copula (Frank)	236.9	-1.2
Trivariate copula (Gumbel)	221.5	-16.6
Trivariate copula (Clayton)	238.1	-

WAIC, widely applicable information criterion

7.5.2 Summary sensitivities and specificities η

The results of fitting the three trivariate copula models to the two simulated data sets are presented in Table 7.4 and 7.5. Figure 7.2 and 7.3 display the posterior medians and 95% CrIs for key test accuracy parameters across each of the models for the two data sets.

Posterior median summary sensitivities and specificities were similar across all three models for both data sets, as were the corresponding 95% CrIs. Within-study associations underlying the data sets were fixed, meaning comparisons can be made between the known values of joint test accuracy and the estimates produced by the models. In the first simulated data set, marginal sensitivities and specificities were fixed at 80.0% and joint sensitivity and specificity at 70.0%, indicating moderate within-study associations. For the best fitting Gumbel copula model, joint sensitivity and specificity were 68.8% (95% CrI: 64.6, 72.9) and 71.5% (95% CrI: 67.1, 75.9), respectively. In the second simulated data set, marginal and joint sensitivities and

specificities were fixed at 40% and 30%, respectively. Using the same model, joint sensitivity was 30.5% (95% CrI: 26.3, 34.8) and joint specificity was 31.0% (95% CrI: 27.2, 34.8). The trivariate copula methodology appears to adequately capture within-study dependencies between the two tests, producing accurate estimates of joint test accuracy measures. Estimates of marginal accuracy measures across all models were also close to the known values.

7.5.3 Between-studies standard deviations $\boldsymbol{\sigma}$

Posterior median between-studies standard deviation estimates of *logit*-transformed sensitivities and specificities of test 1 and test 2 for the first data set were 0.32 (95% CrI: 0.07, 0.58), 0.27 (95% CrI: 0.00, 0.48), 0.22 (95% CrI: 0.00, 0.48) and 0.34 (95% CrI: 0.00, 60), respectively. Standard deviations estimates were 0.18 (95% CrI; 0.00, 0.36) for joint sensitivity and 0.32 (95% CrI: 0.15, 0.53) for joint specificity. For the second data set, between-studies standard deviation estimates for the marginal sensitivities and specificities of test 1 and test 2 were 0.29 (95% CrI: 0.11, 0.51), 0.34 (95% CrI: 0.16, 0.55), 0.25 (95% CrI: 0.06, 0.48) and 0.34 (95% CrI: 0.19, 0.55), respectively. Standard deviations estimates were 0.12 (95% CrI; 0.00, 0.32) for joint sensitivity and 0.19 (95% CrI: 0.00, 0.38) for joint specificity. 95% CrIs corresponding to the standard deviations were wide, with many spanning 0. Estimates were similar across the trivariate Frank, Gumbel and Clayton copula models.

7.5.4 Between-studies correlation $\boldsymbol{\rho_b}$

Using the Gumbel copula model, posterior median between-studies correlation for the data set with high sensitivities and specificities was estimated as 0.33 (95% CrI: -0.19, 0.74), indicating positive association across the six marginal and joint test accuracy measures. Findings were similar for the data set with low sensitivities and specificities, with between-studies correlation estimated as 0.34 (95% CrI: -0.11, 0.74). The Gumbel model, which was the best fit for both examples, produced narrower 95% CrIs for between-studies correlation than the Frank or Clayton copulas. The CrIs for between-studies correlation yielded by all three models across the two data sets were wide, and all spanned 0.

7.5.5 Convergence diagnostics

Convergence diagnostic plots for key model parameters are given in Appendix D.3, D.4 and D.5 (see Section 2.2.4 for interpretation). Diagnostic plots for key test accuracy parameters for each model are presented. Trace plots did not show any systematic trends and MCMC chains initiated at different starting values appeared to mix well. Density plots for the between-studies correlation parameter, rho_b, showed a wide distribution with no clear mode across all three models when sensitivities and specificities were high. Although between-studies correlation was fixed at -0.15 for the simulated data, the models estimated rho_b as approximately 0.3-0.4. This may be explained by the positive associations between sensitivities and specificities, specified through the joint accuracy parameters in the data generating mechanism, driving the between-studies correlation upwards. Autocorrelation between samples for between-studies correlation decreases less rapidly than other test accuracy parameters, indicating slower model convergence for this parameter. Consistent findings across the three models indicate that is likely to be a feature of the underlying simulated data, rather than an issue with model estimation. Furthermore, the simulated data with low sensitivities and specificities - and by extension, lower associations between the measures compared to the scenario with high sensitivities and specificities - appears to reach satisfactory convergence. Density plots for this data set show a unimodal, albeit wide, posterior distribution for rho_b.

Table 7.4: Posterior medians and 95% credible intervals estimated by fitting the trivariate copula models to simulated data comparing the diagnostic accuracy two tests against a common reference standard under a paired design. Data were simulated assuming high marginal test accuracy (sensitivities and specificities = 80%) and moderate within-study associations between tests.

Parameter	Trivariate copula, median (95% CrI)		
	Frank	**Gumbel**	**Clayton**
υ_{se1}	79.89 (76.23, 83.53)	80.29 (76.11, 84.02)	79.77 (76.04, 83.74)
υ_{se2}	76.70 (72.88, 80.43)	76.87 (72.64, 80.71)	76.02 (72.23, 80.02)
υ_{se*}	68.76 (64.80, 72.62)	68.81 (64.63, 72.90)	68.69 (64.69, 72.37)
υ_{sp1}	78.87 (75.29, 82.28)	79.34 (76.00, 82.60)	79.53 (76.25, 82.74)
υ_{sp2}	82.43 (78.99, 85.91)	82.61 (79.04, 86.01)	82.22 (79.01, 85.51)
υ_{sp*}	70.91 (66.23, 75.34)	71.46 (67.10, 75.91)	70.83 (66.59, 75.08)
σ_{se1}	0.286 (0.087, 0.511)	0.318 (0.072, 0.575)	0.260 (0.000, 0.520)
σ_{se2}	0.217 (0.000, 0.418)	0.270 (0.000, 0.476)	0.237 (0.000, 0.457)
σ_{se*}	0.170 (0.000, 0.350)	0.175 (0.000, 0.364)	0.182 (0.000, 0.359)
σ_{sp1}	0.250 (0.000, 0.501)	0.215 (0.000, 0.480)	0.220 (0.000, 0.459)
σ_{sp2}	0.311 (0.000, 0.587)	0.340 (0.000, 0.596)	0.265 (0.000, 0.521)
σ_{sp*}	0.331 (0.148, 0.567)	0.322 (0.149, 0.533)	0.313 (0.127, 0.542)
ρ_b	0.403 (-0.159, 0.814)	0.333 (-0.190, 0.742)	0.344 (-0.190, 0.769)

υ_{se1}, υ_{se2}, υ_{sp1}, and υ_{sp2} denote summary marginal sensitivities and specificities of test 1 and test 2, respectively. υ_{se*} and υ_{sp*} are summary joint sensitivity and specificity, respectively.

σ_{se1}, σ_{se2}, σ_{se*}, σ_{sp1}, σ_{sp2}, and σ_{sp*} denote between-studies standard deviation in *logit*-transformed marginal and joint sensitivities and specificities of test 1 and test 2, respectively.

ρ_b denotes the between-studies correlation parameter.

CrI, credible interval

Table 7.5: Posterior medians and 95% credible intervals estimated by fitting the trivariate copula models to simulated data comparing the diagnostic accuracy two tests against a common reference standard using a paired design. Data were simulated assuming low marginal test accuracy (sensitivities and specificities = 40%) and moderate within-study associations between tests.

Parameter	Trivariate copula, median (95% CrI)		
	Frank	**Gumbel**	**Clayton**
υ_{se1}	39.40 (34.27, 44.67)	39.19 (33.86, 44.33)	39.63 (34.89, 44.22)
υ_{se2}	37.36 (32.03, 43.21)	37.93 (32.36, 43.47)	37.66 (32.92, 42.46)
υ_{se*}	30.14 (26.33, 34.02)	30.48 (26.25, 34.81)	31.53 (27.66, 35.63)
υ_{sp1}	39.99 (35.95, 44.25)	40.27 (35.73, 44.63)	40.14 (35.97, 44.35)
υ_{sp2}	42.57 (37.85, 47.40)	41.37 (36.16, 46.60)	41.57 (36.41, 46.49)
υ_{sp*}	31.44 (27.63, 35.04)	31.03 (27.18, 34.77)	31.25 (27.18, 35.02)
σ_{se1}	0.283 (0.097, 0.514)	0.294 (0.112, 0.512)	0.207 (0.000, 0.417)
σ_{se2}	0.327 (0.116, 0.565)	0.336 (0.162, 0.554)	0.256 (0.000, 0.487)
σ_{se*}	0.105 (0.000, 0.296)	0.120 (0.000, 0.324)	0.161 (0.000, 0.359)
σ_{sp1}	0.224 (0.002, 0.405)	0.250 (0.059, 0.475)	0.217 (0.000, 0.412)
σ_{sp2}	0.283 (0.108, 0.498)	0.343 (0.185, 0.553)	0.321 (0.138, 0.538)
σ_{sp*}	0.187 (0.000, 0.374)	0.188 (0.001, 0.378)	0.226 (0.003, 0.418)
ρ_b	0.407 (-0.075, 0.818)	0.341 (-0.105, 0.744)	0.397 (-0.084, 0.811)

υ_{se1}, υ_{se2}, υ_{sp1}, and υ_{sp2} denote summary marginal sensitivities and specificities of test 1 and test 2, respectively. υ_{se*} and υ_{sp*} are summary joint sensitivity and specificity, respectively.

σ_{se1}, σ_{se2}, σ_{se*}, σ_{sp1}, σ_{sp2}, and σ_{sp*} denote between-studies standard deviation in *logit*-transformed marginal and joint sensitivities and specificities of test 1 and test 2, respectively.

ρ_b denotes the between-studies correlation parameter.

CrI, credible interval

Figure 7.2: Posterior medians (solid dots) and 95% credible intervals (solid bars) of key test accuracy parameters across the trivariate copula (TC) models for the simulated data set with high sensitivities and specificities and moderate within-study associations.

Figure 7.3: Posterior medians (solid dots) and 95% credible intervals (solid bars) of key test accuracy parameters across the trivariate copula (TC) models for the simulated data set with low sensitivities and specificities and moderate within-study associations.

7.6 Discussion

In this chapter, novel Bayesian meta-analysis models for jointly analysing accuracy data on two diagnostic tests were developed, using trivariate copulas to capture within-study dependencies between two tests evaluated using a within-subject design. The models use cross-classified data from each study to estimate joint sensitivity and specificity, enabling inference on the combined accuracy of the tests. Estimation of combined test accuracy measures permits the comparison of different testing strategies, as well as the evaluation of the accuracy of a novel test in the context of its intended diagnostic pathway, supporting healthcare decision-making.

The trivariate copula models were applied to a simulated data set with known within-study associations, allowing comparisons to be made between the estimates of the sensitivities and specificities and their true, underlying values. Both the marginal and joint accuracy measures estimated by the models were close to the values from which the data were simulated. The model appears to capture within-study associations adequately; the gain in accuracy for the joint sensitivity and specificity between the meta-regression model that was fit to the simulated data in Chapter 5 and the models considered in this chapter is notable. Three classes of trivariate copula are specified, offering a flexible approach to capturing associations between tests. The trivariate Gumbel copula - an asymmetric copula exhibiting strong positive dependence in the right-hand tail - was the best fit for both data sets.

The flexibility of the trivariate copula framework, for which several classes of copula were defined and explored in this chapter, make it possible to avoid strong distributional assumptions about the underlying variables. Diagnostic accuracy data are often skewed, with strong dependencies at the extreme ends of the distribution, and non-normal. Where this is the case, the assumption of normality may lead to poor model fit and impact test accuracy estimates. Joint meta-analysis models that account for within-study associations are often computationally intensive and slow to reach convergence. The trivariate copula models developed in this chapter, parameterised in Stan,[53] reach rapid convergence and are computationally feasible, taking just minutes to run.

The proposed methodology is subject to a number of potential limitations. As discussed for the bivariate copula models, inference in this chapter focussed on sensitivity and specificity over alternative test accuracy measures, such as predictive values, likelihood ratios and DORs. In some scenarios, predictive values may be more clin-

ically relevant than sensitivities and specificities. For example, when a clinician is screening an individual for a condition, predictive values allow inference on the test result, e.g. if the test returns a positive result, indicating that the condition of interest is present, the positive predictive value is the probability that the patient does indeed have the condition.[65] Sensitivity, on the other hand, enables inference on the true disease status of the patient, which is unknown at the time of testing, e.g. if patient has the condition of interest, the sensitivity is the probability that they will be correctly identified by the test. However, as in the previous chapter, it is straightforward to derive estimates of other test accuracy measures using posterior estimates of sensitivity and specificity; direct meta-analyses of predictive values, likelihood ratios or DORs may have disadvantageous properties,[226, 227, 66] as detailed in Chapter 6. Similar to the bivariate copula framework in Chapter 6, the between-studies heterogeneity parameters for the trivariate copula models were highly sensitive to prior distribution and initial values choices.[44, 225] Future model development could focus on including a copula at the between-studies level,[90, 92, 94, 98] removing the need to estimate between-studies heterogeneity parameters that may be sensitive to choices of prior distribution.

The trivariate copula model requires cross-classified data across all included studies in order to estimate the study-specific copula dependence parameters for the model. However, publications of studies that compare the accuracy of two or more diagnostic tests in the same individuals do not consistently report this data.[9, 10, 11, 4] The issues around requesting or imputing missing cross-classifications were discussed in Section 2.3.4. Model development in this chapter further emphasises the need for comparative diagnostic accuracy studies to report cross-classified data wherever possible.

7.7 Chapter summary

In this chapter, novel Bayesian meta-analysis models were developed for synthesising data on two diagnostic tests compared to a common reference standard under a paired design, using trivariate copulas to capture within-study dependencies between multiple tests. The models use cross-classified data from each study to directly estimate joint sensitivity and specificity, making full use of the available evidence base. By enabling the assessment of combined test accuracy, the trivariate copula framework allows the evaluation of realistic diagnostic pathways and testing strategies

to support healthcare decision-making. The following chapter concludes the thesis, summarising the findings and conclusions from previous chapters, together with a discussion of the limitations and the opportunities for further work.

Chapter 8

Discussion

8.1 Summary

This thesis considers a range of methodological challenges related to the synthesis of diagnostic accuracy studies that evaluate two or more tests using a within-subject design. The development of meta-analysis models that jointly synthesise diagnostic accuracy data on multiple tests is relatively novel, and guidance and best practice on their use are evolving. While meta-analysis guidelines for combining diagnostic accuracy studies focus primarily on the synthesis of data on individual tests,[14, 4] there are an increasing number of systematic reviews and meta-analyses that address comparative research questions.[87] A number of models for comparing the accuracy of two or more tests in a meta-analytic framework have been proposed in the last 10 years; however, there is a lack of clear guidance on their use.[15] Many of these models do not account for dependencies between tests, while those that do often result in slow convergence and long computation times. The use of these models can be restricted by data availability, with many models that capture within-study associations requiring cross-classified data on combined test performance that are not consistently reported in the literature.[9, 10, 11, 4] Existing meta-analysis models for multiple diagnostic tests should be assessed to inform recommendations for their use. Additionally, the development of novel methods that are less computationally demanding may help to address the limitations of current approaches. Within-study associations, present between the sensitivities and specificities of two or more tests assessed using the same individuals, represent another methodological challenge, necessitating the evaluation and refinement of existing models as well as the exploration

of novel approaches to adequately capture such dependencies. The methodological developments presented in this thesis were motivated by the challenges of diagnosing Alzheimer's disease dementia. However, the evaluation of the comparative or combined accuracy of multiple diagnostic tests is relevant to a wide range of disease areas. The novel meta-analysis models developed here are applicable to a variety of clinical settings, underscoring their potential for widespread impact in enhancing evidence-based practice.

Chapter 1 outlined the aims and structure of the thesis. A brief background to the thesis was provided, covering key concepts in meta-analysis of diagnostic test accuracy studies, Bayesian statistical methodology and Alzheimer's disease dementia. Current methodological challenges in the synthesis of evidence on the accuracy of multiple diagnostic tests were described. Chapter 2 summarised statistical theory relevant to the thesis. Chapter 3 presents a comprehensive review of existing meta-analysis models for two or more diagnostic tests, highlighting the limitations of existing methods that model development in later chapters sets out to address. Chapter 4 described the literature review undertaken to identify a motivating example in Alzheimer's disease dementia on which to base novel methodological development in later chapters. Cross-classified data, needed to account for within-study associations between multiple tests, were provided in only 20% of the included comparative diagnostic accuracy studies. This emphasises the importance of developing models that make realistic assumptions about data availability, utilising a combination of 2×2 and cross-classified data where reported.

Chapter 5 assessed the impact of ignoring within-study dependencies between sensitivities and specificities in a meta-analysis of comparative diagnostic accuracy studies through a simulation study. A meta-regression approach, which does not account for associations between two tests assessed under a paired design, was fit to simulated data with known within-study dependencies. When within-study associations were strong or moderate, joint sensitivity and specificity were underestimated and model performance declined. Lower marginal sensitivities and specificities led to increased bias of joint accuracy measures. When within-study associations were weak, or when marginal test accuracy is of interest (i.e. in a test comparison framework), the meta-regression model appears to be sufficient. To model the joint accuracy of two or more diagnostic tests evaluated in the same individuals, a meta-analysis model that accounts for within-study associations between multiple tests is more appropriate.

Chapter 6 described novel Bayesian model development for synthesising evidence

on the accuracy of two diagnostic tests, accounting for associations between and within studies present between two tests evaluated using a paired design. Using a novel application of bivariate copulas to capture within-study dependencies between sensitivities and specificities, the models offer a robust approach to modelling comparative test accuracy data. The models optimise the available evidence base by making use of both study- and individual-level data where available. The new models introduced in this chapter were compared to the currently recommended meta-regression approach that was formally evaluated in Chapter 5. Compared to the meta-regression model, which treats the two tests independent and ignores for within-study associations, the bivariate copula models resulted in improved model fit and increased precision in sensitivity and specificity estimates. In a HTA setting, this aids the evaluation of clinical and cost-effectiveness of diagnostic technologies, enabling more precise decisions on the most efficient use of health resources.

Chapter 7 extended the bivariate copula methodology in Chapter 6 to a trivariate copula model for jointly synthesising diagnostic test accuracy on two tests assessed using a within-subject design. The models use individual-level data from each study to directly estimate joint test accuracy, making full use of the available evidence base. The trivariate copula framework enables the assessment of realistic diagnostic pathways and testing strategies. Between the bivariate copula models described in Chapter 6, which synthesise study- and individual-level data where available, and the trivariate copula models proposed in Chapter 7, which make use of individual-level data from every study, novel methodological development in this thesis covers the full range of possible reporting formats for comparative diagnostic accuracy studies.

8.2 Strengths and limitations

This section discusses the strengths and limitations of the novel methodology for diagnostic test evaluation developed in this thesis.

In Chapter 5, the impact of accounting for within-study associations in a meta-analysis of two diagnostic tests evaluated using a paired design was assessed. A mechanism for simulating comparative diagnostic accuracy data, incorporating both between and within study associations, was developed. This represents novel methodological advancement, and may be used to assess a number of future models for the meta-analysis of comparative diagnostic accuracy. Uncertainties around the impact of within-study associations on test accuracy measures has been highlighted,[15] mak-

ing this simulation study an important and timely addition to the existing literature on diagnostic meta-analysis models. Based on the findings of the simulation study, recommendations were made on the use of joint meta-analysis models for multiple diagnostic tests in regards to their handling of within-study associations between sensitivities and specificities. While simpler models that assume independence between multiple tests assessed in the same individuals are sufficient when inference is focussed on comparative test accuracy, more complex models that account for within-study associations between tests are needed when joint test accuracy is of interest. Due to the small number of comparative diagnostic accuracy studies that reported cross-classified data identified through the literature review in Chapter 4, it was not possible to simulate the data using joint sensitivities and specificities from a real world example. Concern arises, therefore, whether findings based on the simulated data will be reflected in meta-analyses of data collected in a clinical setting. To mitigate this issue, several scenarios for which the strength of the within-study associations and the magnitude of the sensitivities and specificities were varied. The findings were consistent across a number of scenarios, increasing confidence in their robustness.

In Chapter 6, novel meta-analysis models for evaluating the accuracy of two diagnostic tests in a Bayesian framework were assessed, using bivariate copulas to account for within-study associations between two tests assessed using a paired design. The methodology led to improved model fit compared to the meta-regression approach, which does not account for within-study associations, and findings were similar across two motivational examples comparing different diagnostic tests for Alzheimer's disease dementia. The models also reduced the uncertainty around sensitivities and specificities, aiding evaluation of the clinical and cost-effectiveness of novel diagnostic tests. The model can be fit to 2×2 data on each test where cross-classified data are not available, incorporating external evidence on within-study associations through informative priors placed on copula dependence parameters. Where cross-classified data are available for a subset of studies, as in the motivating examples to which the models are applied in this chapter, a common copula dependence parameter can be assumed across studies. Bivariate copula methodology relaxes the requirement for cross-classifications needed for each study, reflecting current reporting standards and increasing the possibility of its application within systematic reviews. Diagnostic accuracy data are often non-normal and likely to exhibit strong tail dependence, particularly when marginal sensitivities and specificities are high; the flexible nature of copulas make them a natural approach for modelling associations between tests

in comparison to multivariate methods that make strong distributional assumptions about the marginal variables. A number of bivariate copulas were explored in Chapter 6, varying the strength and direction of the association between sensitivities and specificities, increasing the generalisability of the models for a range of diagnostic data sets. The bivariate copula framework allows estimation of the comparative accuracy of two tests, but does not enable direct estimation of joint or combined test accuracy measures.

In Chapter 7, the bivariate copula methodology developed in the previous chapter was extended to trivariate copula models for evaluating the combined accuracy of two diagnostic tests assessed against a common reference standard using a within-subject design. The model makes full use of the available evidence base, synthesising cross-classified data from all studies, allowing estimation of joint sensitivity and specificity. However, comparative diagnostic accuracy studies are known to rarely facilitate access to this data.[9, 10, 11, 4] Where these data are not reported, it will not be possible to include this study in the trivariate meta-analysis model. Advancement of imputation methods for comparative diagnostic accuracy data, discussed in more detail in Section 8.3, may allow the synthesis of such studies in the future. An advantage of a copula approach to modelling related variables is the estimation of a single association parameter, in contrast to the need to estimate standard deviation and correlation parameters when fitting a multivariate normal distribution. Where these parameters conflict with one another, for example when prior distributions are misspecified, it can lead to long computation times or non-convergence. The trivariate copula method is highly computationally efficient compared to other meta-analysis models for joint synthesis of two diagnostic tests. The advancement of HMC simulation techniques within the freely available Stan software,[53] and the development of novel methodology for capturing within-study associations between tests, has made joint meta-analysis of two diagnostic tests computationally feasible.

Hamiltonian Monte Carlo sampling is an efficient method for modelling related variables, and was used for all methodological development in this thesis. Nevertheless, the Stan programming language and model specification differs from other more commonly used Bayesian samplers in the context of HTA, such as JAGS and WinBUGS/OpenBUGS, resulting initially in a potential barrier to uptake of these methods. There are, however, a multitude of online resources available on Stan. R packages such as *shinystan* can also aid the visualisation of results and convergence assessment.[235] To support the implementation of these novel models into wider

practice, future research could involve the development of an R package or Shiny app to provide a web-based, user-friendly interface to the models.

8.3 Further work

In this thesis, novel meta-analysis methods for the synthesis of diagnostic accuracy data on two tests were developed. An area for future methodological development is the extension of the proposed bivariate and trivariate copula frameworks to three or more diagnostic tests compared to a common reference standard. Vine copulas, which enable the extension of parametric bivariate copula families to higher dimensions,[101] may be one approach to accomplish this generalisation. While a number of publications on joint meta-analysis methods for comparative diagnostic accuracy studies claim that it is possible to extend the respective models to three or more tests, many do not demonstrate their application beyond two tests.[12] Some methods, such as the meta-regression model that includes test type as a covariate, can be readily extended to three or more tests. For other methods, such as the multinomial likelihoods model described by Trikalinos et al [85], this extension is non-trivial and requires further methodological development. Existing models that account for within-study associations between multiple tests assessed using a paired design are known to encounter issues with convergence and computational difficulties.[15] As the number of tests increases, so too does the number of estimated parameters in the meta-analysis model. Models that are already computationally intensive may become infeasible beyond a certain number of diagnostic tests.

The need for the evaluation and comparison of existing meta-analysis methods before their adoption into health technology decision-making has been highlighted.[12, 4, 10, 15] The simulation study performed in Chapter 5 makes some progress towards this goal, evaluating the currently recommended meta-regression model and its assumption of independence between two tests assessed using a within-subject design. The performance of other, more complex meta-analysis models for two tests could be formally compared to the meta-regression method through further simulation studies. The issue remains, however, as to how to simulate comparative diagnostic accuracy data and the complex dependence structures that underlie it in a way that does not influence the conclusions of the study. To capture between and within study associations present in meta-analytic data sets on two or more diagnostic tests, a model that captures these dependencies must be used to simulate the data. This may

lead to a circular argument, however; whichever model the data were simulated from would be the best fitting model. The data generating mechanism and corresponding R code developed in Chapter 5 goes some way to addressing this question, enabling quantification of potential bias from selecting a more simplistic model.

Uncertainty remains around the appropriate method for combining comparative and non-comparative diagnostic accuracy studies. Restricting to head-to-head comparisons, while methodologically sound, does not make best use of available evidence. However, including all studies that evaluated one or a subset of the tests of interest in a meta-analysis of multiple diagnostic tests may introduce bias to the study pool where test comparisons are potentially based on indirect evidence.[8] Future model development could consider how to combine comparative and non-comparative data in single analysis while adjusting for potential bias that indirect comparisons may introduce.

A number of publications of joint meta-analysis models for synthesising two or more diagnostic tests that account for within-study associations repeatedly make use of data from a meta-analysis of the diagnostic accuracy of two second trimester ultrasound markers for trisomy 21, also known as Down syndrome, published in 2001.[236] Trikalinos et al first used this data set to demonstrate the application of a joint meta-analysis model for two diagnostic tests.[85] As full cross-classifications were not available for all studies, Trikalinos et al partially imputed the data set by drawing on associations within studies that reported complete data, although the precise imputation method is not specified. This partially imputed data set has been repeatedly reused to illustrate the application of other meta-analysis models for multiple diagnostic tests, presumably at least in part because cross-classified data are infrequently reported in the literature.[9, 10, 11, 4] Future versions of reporting checklists for diagnostic accuracy studies, such as STARD,[208] could include an item stipulating that cross tabulation of all index test results should be reported where appropriate. Cross tabulation of index test results was identified as an item in the dementia-specific extension to STARD, STARDdem.[234] This would improve reporting practices and increase awareness of the need for comparative data to evaluate combined test accuracy, as well as removing barriers to the implementation of models proposed in this thesis. Future research could also focus on the development of a validated method to impute partially missing cross-classifications, using associations from studies that report data at the individual-level to derive cross-classifications for studies that report 2×2 tables for each test only.

In terms of the clinical application of the methodological developments in this thesis, the proposed methodology is generalisable to a wide range of other disease areas. The diagnostic pathway for Alzheimer's disease dementia will undergo rapid evolution in the near future, with the emergence of blood-based biomarkers representing an scalable, minimally invasive and early diagnosis. These include Aβ markers, with plasma Aβ_{42} levels being shown to be negatively correlated with CSF Aβ_{42} levels.[237] Given the recent development of amyloid-based therapies for Alzheimer's disease dementia,[34, 35] this discovery could not be more timely. Assessment of the real world performance of blood-based biomarkers is urgently needed. Novel methodology proposed in this thesis could be applied to blood-based biomarkers to address this issue. With a number of blood-based amyloid and tau markers identified as potential novel diagnostic tests, the application of bivariate copula models developed in Chapter 6 could be used to compare their accuracy to one another, as well as existing technologies such as CSF biomarkers or imaging tests. To implement the blood-based biomarkers within various healthcare systems, an understanding is needed of the accuracy of the markers in combination with other diagnostic tests for Alzheimer's disease dementia. This could be evaluated through the application of trivariate copula models, developed in Chapter 7, to estimate joint test accuracy measures while accounting for within-study associations present between multiple tests.

If cross-classified data on blood-based biomarkers are not available in the published literature, IPD from electronic health records could be used to inform associations between sensitivities and specificities. For example, Dementias Platform UK brings together records of more than 3 million participants, and includes rich biomarker and imaging data not usually available through routinely collected sources.[238] The trivariate copula models developed in Chapter 7 could be extended to incorporate IPD through the inclusion of patient-level covariates that measure the effect of patient characteristics on summary accuracy measures.[233]

8.4 Conclusions

In conclusion, this thesis has demonstrated the impact of ignoring within-study dependencies present between two diagnostic tests evaluated using a paired design in a meta-analytic framework. This, in turn, motivated novel methodological developments to synthesise comparative diagnostic accuracy studies, using a new application

of copula models to capture within-study associations between tests. The models developed as part of this thesis are generalisable to a wide range of other research areas where comparing or combining multiple diagnostic tests is of interest. The models make use of a broader evidence base by utilising both study- and individual-level data, where available. The work presented in this thesis has the ability to guide the development of protocols for synthesising comparative diagnostic accuracy studies and facilitate informed decision-making regarding the use of diagnostic tests.

Appendix A

A.1 Centred and non-centred parameterisations of a hierarchical normal model

This section contains Stan code for implementing a hierarchical normal model for meta-analysing data on a continuous treatment effect. The difference between centred and non-centred parameterisations of the model is illustrated in Figure 2.2a and 2.2b.

Centred parameterisation of a hierarchical normal model

```
data {
  int<lower = 0> N;
  vector[N] y;
  vector<lower = 0>[N] sigma;}

parameters {
  real<lower = 0> tau;
  real m;
  vector[N] mu;}

model {
  m ~ normal(0,3);
  tau ~ cauchy(0,5);
  //centred parameterisation
  mu ~ normal(m,tau);
```

```
  y ~ normal(mu,sigma);}
```

Non-centred parameterisation of a hierarchical normal model

```
data {
  int<lower = 0> N;
  vector[N] y;
  vector<lower = 0>[N] sigma;}

parameters {
  real<lower = 0> tau;
  real m;
  vector[N] z;}

transformed parameters {
  vector[N] mu;
  for (i in 1:N) {
    //non-centred parameterisation
    mu[i] = m  + tau * z[i];}}

model {
  z ~ std_normal();
  m ~ normal(0,3);
  tau ~ cauchy(0,5);
  y ~ normal(mu, sigma);}
```

A.2 Stan code for bivariate random effects meta-analysis model for evaluating the diagnostic accuracy of a single test

```
data {
  int<lower = 0> Ns;
  int<lower = 0> tp[Ns];
  int<lower = 0> tn[Ns];
```

```
  int<lower = 0> disease[Ns];
  int<lower = 0> nodisease[Ns];}

parameters {
  real rr;
  vector[2] b;
  vector<lower = 0>[2] tau;
  vector[2] z[Ns];}

transformed parameters {
  matrix[2,2] Tau;
  matrix[2,2] L;
  vector[2] mu[Ns];
  real<lower = -1, upper = 1> rho;
  rho = tanh(rr);
  Tau[1,1] = tau[1]^2;
  Tau[1,2] = tau[1]*tau[2]*rho;
  Tau[2,1] = tau[1]*tau[2]*rho;
  Tau[2,2] = tau[2]^2;
  L = cholesky_decompose(Tau);
  for (i in 1:Ns) {
    mu[i] = b + Tau*z[i];}}

model {
  // priors
  rr ~ normal(0,0.8);
  b ~ normal(0,10);
  tau ~ normal(0,2.5);
  for (i in 1:Ns) {
    z[i] ~ std_normal();
    // likelihoods
    tp[i] ~ binomial_logit(disease[i], mu[i,1]);
    tn[i] ~ binomial_logit(nodisease[i], mu[i,2]);}}

generated quantities {
  real<lower = 0,upper = 1> sens;
```

```
real<lower = 0,upper = 1> spec;
sens = inv_logit(b[1]);
spec = inv_logit(b[2]);}
```

Appendix B

B.1 Meta-regression model for evaluating the accuracy of two diagnostic tests

Stan code for implementing the Bayesian meta-regression method for synthesising diagnostic accuracy data on two tests evaluated using a paired study design, in which all patients undergo both tests plus a reference standard. Two versions of the model are available: one that estimates a common between-studies correlation parameter, one test estimates a separate between-studies correlation parameter for each test. The model specification is described in Section 5.3.1.

```
data {
  int<lower = 0> Ns;
  int<lower = 0> tp[Ns,2]; //2x2 data on each test
  int<lower = 0> tn[Ns,2];
  int<lower = 0> disease[Ns];
  int<lower = 0> nodisease[Ns];}

parameters {
  real rr; //common between-studies correlation parameter
  //vector[6] rr; //separate between-studies correlation parameter
  vector[4] b;
  vector<lower = 0>[4] tau;
  vector[4] z[Ns];}

transformed parameters {
  matrix[4,4] Tau;
```

```
  matrix[4,4] L;
  vector[4] mu[Ns];
  real<lower = -1, upper = 1> rho; //common
  //vector<lower = -1, upper = 1>[6] rho; //separate
  rho = tanh(rr);
  //Tau[1,2] = tau[1]*tau[2]*rho[1]; //separate
  //Tau[1,3] = tau[1]*tau[3]*rho[2];
  //Tau[1,4] = tau[1]*tau[4]*rho[3];
  //Tau[2,3] = tau[2]*tau[3]*rho[4];
  //Tau[2,4] = tau[2]*tau[4]*rho[5];
  //Tau[3,4] = tau[3]*tau[4]*rho[6];
  for (j in 1:4) {
    Tau[j,j] = tau[j]^2;
    for (k in (j+1):4) {
      Tau[j,k] = tau[j]*tau[k]*rho; //common
      Tau[k,j] = Tau[j,k];}}
  L = cholesky_decompose(Tau);
  //non-centred parameterisation
  for (i in 1:Ns) {
    mu[i] = b + L*z[i];}}

model {
 //priors
 rr ~ normal(0,0.8);
 b ~ normal(0,10);
 tau ~ normal(0,2.5);
 for (i in 1:Ns) {
   z[i] ~ std_normal();
   //binomial likelihoods
   tp[i,1] ~ binomial_logit(disease[i], mu[i,1]);
   tp[i,2] ~ binomial_logit(disease[i], mu[i,2]);
   tn[i,1] ~ binomial_logit(nodisease[i], mu[i,3]);
   tn[i,2] ~ binomial_logit(nodisease[i], mu[i,4]);}}

generated quantities {
  real<lower = 0, upper = 1> sens1;
```

```
  real<lower = 0, upper = 1> sens2;
  real<lower = 0, upper = 1> spec1;
  real<lower = 0, upper = 1> spec2;
  real<lower = 0, upper = 1> jsens;
  real<lower = 0, upper = 1> jspec;
  real<lower = -1, upper = 1> diffsens;
  real<lower = -1, upper = 1> diffspec;
  vector[Ns] log_lik;
  sens1 = inv_logit(b[1]);
  sens2 = inv_logit(b[2]);
  spec1 = inv_logit(b[3]);
  spec2 = inv_logit(b[4]);
  jsens = sens1 * sens2;
  jspec = spec1 * spec2;
  diffsens = sens1 - sens2;
  diffspec = spec1 - spec2;
  for (n in 1:Ns) {
   //monitor log-likelihood to calculate WAIC post-estimation
   log_lik[n] = binomial_logit_lpmf(tp[n,1:2] | disease[n],
                 mu[n,1:2]);}}
```

B.2 Multinomial likelihoods model for joint meta-analysis of diagnostic accuracy data on two tests

Stan code for implementing the Bayesian multinomial likelihoods model for synthesising diagnostic accuracy data on two tests evaluated using a paired study design, in which all patients undergo both tests plus a reference standard. The model specification is described in Section 5.3.2.

```
data {
  int<lower = 0> Ns;
  int<lower = 0> x1[Ns,4]; //cross-classified data
  int<lower = 0> x0[Ns,4];
  int<lower = 0> disease[Ns];
```

```
  int<lower = 0> nodisease[Ns];}

parameters {
  real rr;
  vector[6] b;
  vector<lower = 0>[6] tau;
  vector[6] z[Ns];}

transformed parameters {
  matrix[6,6] Tau;
  matrix[6,6] L;
  vector<lower = 0, upper = 1>[4] p1[Ns];
  vector<lower = 0, upper = 1>[4] p0[Ns];
  real<lower = 0, upper = 1> p1_1dot[Ns];
  real<lower = 0, upper = 1> p1_dot1[Ns];
  real<lower = 0, upper = 1> p0_0dot[Ns];
  real<lower = 0, upper = 1> p0_dot0[Ns];
  vector[6] mu[Ns];
  real<lower = -1, upper = 1> rho;
  rho = tanh(rr);
  for (j in 1:6) {
    Tau[j,j] = tau[j]^2;
    for (k in (j+1):6) {
      Tau[j,k] = tau[j]*tau[k]*rho;
      Tau[k,j] = Tau[j,k];}}
  L = cholesky_decompose(Tau);
  //non-centred parameterisation
  for (i in 1:Ns) {
    //between-studies model
    mu[i] = b + L*z[i];
    //diseased group
    p1[i,4] = inv_logit(mu[i,3]);
    p1_1dot[i] = inv_logit(mu[i,1]);
    p1_dot1[i] = inv_logit(mu[i,2]);
    p1[i,2] = p1_dot1[i] - p1[i,4];
    p1[i,3] = p1_1dot[i] - p1[i,4];
```

```
    p1[i,1] = 1 - p1[i,2] - p1[i,3] - p1[i,4];
    //non-diseased group
    p0[i,1] = inv_logit(mu[i,6]);
    p0_0dot[i] = inv_logit(mu[i,4]);
    p0_dot0[i] = inv_logit(mu[i,5]);
    p0[i,2] = p0_0dot[i] - p0[i,1];
    p0[i,3] = p0_dot0[i] - p0[i,1];
    p0[i,4] = 1 - p0[i,1] - p0[i,2] - p0[i,3];}}

model {
  //priors
  rr ~ normal(0,0.8);
  b ~ normal(0,10);
  tau ~ normal(0,2.5);
  for (i in 1:Ns) {
    z[i] ~ std_normal();
    //multinomial likelihoods
    x1[i,1:4] ~ multinomial(p1[i,1:4]);
    x0[i,1:4] ~ multinomial(p0[i,1:4]);}}

generated quantities {
  real<lower = 0, upper = 1> sens1;
  real<lower = 0, upper = 1> sens2;
  real<lower = 0, upper = 1> spec1;
  real<lower = 0, upper = 1> spec2;
  real<lower = 0, upper = 1> jsens;
  real<lower = 0, upper = 1> jspec;
  sens1 = inv_logit(b[1]);
  sens2 = inv_logit(b[2]);
  jsens = inv_logit(b[3]);
  spec1 = inv_logit(b[4]);
  spec2 = inv_logit(b[5]);
  jspec = inv_logit(b[6]);}
```

B.3 Simulation study

R code for implementing simulation study evaluating the performance of the Bayesian meta-regression model. The model specification is described in Section 5.3.1 and the data generation mechanism in Section 5.3.4.1.

```
## Load packages
library(foreign)
library(boot)
library(MASS)
library(coda)
library(rstan)
library(purrr)
library(furrr)
library(parallel)

## Stan options
rstan_options(auto_write = TRUE)

## FUNCTION: Simulate a single data set ----------------------------
# Define a function to simulate a single data set
simulate_data <- function(extra, sens1, sens2, sens12, spec1, spec2,
                          spec12, nstudies) {
  disease <- list() # create list for diseased cohort
  nodisease <- list() # create list for non-diseased cohort
  for (i in 1:nstudies) {
    # Prevalence
    prev <- runif(1, 0.30, 0.50)
    # Total number of patients
    n <- floor(runif(1, 40, 134))
    # Number of patients with disease
    n1 <- round(n * prev) # number of patients with disease
    # Number of patients without disease
    n0 <- n - n1 # number of patients without disease
    # Between-studies model
    logit_sens1 <- logit(sens1) # logit transform sens & spec
    logit_sens2 <- logit(sens2)
```

```
logit_sens12 <- logit(sens12)
logit_spec1 <- logit(spec1)
logit_spec2 <- logit(spec2)
logit_spec12 <- logit(spec12)
theta <- c(logit_sens1, logit_sens2, logit_sens12,
             logit_spec1, logit_spec2, logit_spec12)
S <- matrix(c(0.2, 0, 0, 0, 0, 0,
              0, 0.2, 0, 0, 0, 0,
              0, 0, 0.2, 0, 0, 0,
              0, 0, 0, 0.2, 0, 0,
              0, 0, 0, 0, 0.2, 0,
              0, 0, 0, 0, 0, 0.2), byrow = TRUE, nrow = 6)
R <- matrix(c(1, -0.15, -0.15, -0.15, -0.15, -0.15,
              -0.15, 1, -0.15, -0.15, -0.15, -0.15,
              -0.15, -0.15, 1, -0.15, -0.15, -0.15,
              -0.15, -0.15, -0.15, 1, -0.15, -0.15,
              -0.15, -0.15, -0.15, -0.15, 1, -0.15,
              -0.15, -0.15, -0.15, -0.15, -0.15, 1),
                byrow = TRUE, nrow = 6)
tau <- S %*% R %*% S # fix matrix for all scenarios
eta <- mvrnorm(1, theta, tau)
# Upper bound for joint accuracy measures
# Joint sensitivity
if (eta[3] > eta[1] | eta[3] > eta[2]) {
  eta[3] <- min( eta[1], eta[2] )
}
# Joint specificity
if (eta[6] > eta[4] | eta[6] > eta[5]) {
  eta[6] <- min( eta[4], eta[5] )
}
# Within-study model
# Diseased cohort
# Proportion of patients with each combination of test results
p1_11 <- inv.logit(eta[3])
p1_10 <- inv.logit(eta[1]) - p1_11
p1_01 <- inv.logit(eta[2]) - p1_11
```

```
  p1_00 <- 1 - p1_11 - p1_10 - p1_01
  if (p1_11 + p1_10 + p1_01 > 1) {
    # condition to prevent negative probabilities
    a <- abs(1 - (p1_11 + p1_10 + p1_01))
    p1_11 <- p1_11 - a/3
    p1_10 <- p1_10 - a/3
    p1_01 <- p1_01 - a/3
    p1_00 <- 0
  }
  prob1 <- c(p1_00, p1_01, p1_10, p1_11)
  x1 <- rmultinom(1, n1, prob1) # multinomial distribution
  disease[[i]] <- x1 # add to data list
  # Non-diseased cohort
  # Proportion of patients with each combination of test results
  p0_00 <- inv.logit(eta[6])
  p0_10 <- inv.logit(eta[5]) - p0_00
  p0_01 <- inv.logit(eta[4]) - p0_00
  p0_11 <- 1 - p0_00 - p0_10 - p0_01
  if (p0_00 + p0_10 + p0_01 > 1) {
    # condition to prevent negative probabilities
    b <- abs(1 - (p0_00 + p0_10 + p0_01))
    p0_11 <- 0
    p0_10 <- p0_10 - b/3
    p0_01 <- p0_01 - b/3
    p0_00 <- p0_00 - b/3
  }
  prob0 <- c(p0_00, p0_01, p0_10, p0_11)
  x0 <- rmultinom(1, n0, prob0) # multinomial distribution
  nodisease[[i]] <- x0 # add to data list
}
# Transform lists to data frames
data1 <- do.call(cbind, disease) # transform list to data frame
data0 <- do.call(cbind, nodisease)
# Format data frames
# Transpose data frames
data1 <- t(data1)
```

```
  data0 <- t(data0)
  # Generate ID variable
  id <- 1:nrow(data1)
  data1 <- cbind(id = id, data1) # add id variable to data frames
  data0 <- cbind(id = id, data0)
  # Merge data frames using ID variable
  data <- merge(data1, data0, by = "id")
  # Rename columns
  colnames(data) <- c("id", "x1_00", "x1_01", "x1_10", "x1_11",
                      "x0_00", "x0_01", "x0_10", "x0_11")
  # Calculate number of diseased and non-diseased patients
  data$n1 <- data$x1_00 + data$x1_01 + data$x1_10 + data$x1_11
  data$n0 <- data$x0_00 + data$x0_01 + data$x0_10 + data$x0_11
  data$tp.1 <- data$x1_10 + data$x1_11
  data$tp.2 <- data$x1_01 + data$x1_11
  data$tn.1 <- data$x0_00 + data$x0_01
  data$tn.2 <- data$x0_00 + data$x0_10
  # Return data in list
  return(list(data = data))
}

## FUNCTION: Fit model to data set ---------------------------------
# Define a function to fit binomial model to data set
fit_model <- function(data, nstudies, warmup, iter) {
  # Extract data from list returned by simulate data function
  data <- data$data
  # Build data list for Stan
  data_binomial <- list(tp = cbind(data$tp.1, data$tp.2),
                        tn = cbind(data$tn.1, data$tn.2),
                        disease = data$n1, nodisease = data$n0,
                        Ns = nstudies)
  # Generate initial values from data
  data$se1 <- data$tp.1 / data$n1
  data$se2 <- data$tp.2 / data$n1
  data$sp1 <- data$tn.1 / data$n0
  data$sp2 <- data$tn.2 / data$n0
```

```
  b1 <- logit(mean(data$se1))
  b2 <- logit(mean(data$se2))
  b3 <- logit(mean(data$sp1))
  b4 <- logit(mean(data$sp2))
  # Build initial value list for Stan
  init_binomial1 <- list(z = structure(.Data = rep(0,(nstudies*4)),
                         .Dim = c(nstudies,4)),
                         tau = rep(0.2,4), rr = -0.15,
                         b = c(b1, b2, b3, b4))
  init_binomial2 <- list(z = structure(.Data = rep(0,(nstudies*4)),
                         .Dim = c(nstudies,4)),
                         tau = rep(0.1,4), rr = -0.20,
                         b = c(b1-0.1, b2-0.1, b3-0.1, b4-0.1))
  init_binomial3 <- list(z = structure(.Data = rep(0,(nstudies*4)),
                         .Dim = c(nstudies,4)),
                         tau = rep(0.3,4), rr = -0.10,
                         b = c(b1+0.1, b2+0.1, b3+0.1, b4+0.1))
  init_binomial <- list(init_binomial1, init_binomial2,
                          init_binomial3)
  # Sample from the posterior distribution
  binomial <- stan(
    file = 'model_binomial.stan',
    data = data_binomial,
    init = init_binomial,
    chains = 3,
    warmup = warmup,
    iter = iter,
    cores = 3,
    seed = 1234,
    control = list(adapt_delta = 0.99, stepsize = 0.1))
  # Return results in a list
  return(list(binomial = binomial))
}

## FUNCTION: Extract relevant results -----------------------------
# Define a function to extract relevant results at each
```

```
# iteration of simulation study
extract_results <- function(data, nstudies, warmup, iter) {
  # Fit model
  fit.model <- fit_model(data = data, nstudies = nstudies,
                         warmup = warmup, iter = iter)
  # Create data frame of results from each iteration of sampler
  all.chains <- data.frame(fit.model[["binomial"]]
                           @sim[["samples"]][[1]])
  # Calculate posterior summaries
  # Mean
  sens.1 <- mean(all.chains$sens1)
  sens.2 <- mean(all.chains$sens2)
  spec.1 <- mean(all.chains$spec1)
  spec.2 <- mean(all.chains$spec2)
  joint.sens <- mean(all.chains$jsens)
  joint.spec <- mean(all.chains$jspec)
  # Standard deviation
  sens.sd.1 <- sd(all.chains$sens1)
  sens.sd.2 <- sd(all.chains$sens2)
  spec.sd.1 <- sd(all.chains$spec1)
  spec.sd.2 <- sd(all.chains$spec2)
  joint.sens.sd <- sd(all.chains$jsens)
  joint.spec.sd <- sd(all.chains$jspec)
  # Median
  sens.med.1 <- median(all.chains$sens1)
  sens.med.2 <- median(all.chains$sens2)
  spec.med.1 <- median(all.chains$spec1)
  spec.med.2 <- median(all.chains$spec2)
  joint.sens.med <- median(all.chains$jsens)
  joint.spec.med <- median(all.chains$jspec)
  # Lower credible interval
  sens.lb.1 <- quantile(all.chains$sens1, .025)
  sens.lb.2 <- quantile(all.chains$sens2, .025)
  spec.lb.1 <- quantile(all.chains$spec1, .025)
  spec.lb.2 <- quantile(all.chains$spec1, .025)
  joint.sens.lb <- quantile(all.chains$jsens, .025)
```

```
  joint.spec.lb <- quantile(all.chains$jspec, .025)
  # Upper credible interval
  sens.ub.1 <- quantile(all.chains$sens1, .975)
  sens.ub.2 <- quantile(all.chains$sens2, .975)
  spec.ub.1 <- quantile(all.chains$spec1, .975)
  spec.ub.2 <- quantile(all.chains$spec2, .975)
  joint.sens.ub <- quantile(all.chains$jsens, .975)
  joint.spec.ub <- quantile(all.chains$jspec, .975)
  # Create data frame of posterior summaries
  extract.results <-
    data.frame(sens.1, sens.2, spec.1, spec.2,
               joint.sens, joint.spec,
               sens.sd.1, sens.sd.2, spec.sd.1, spec.sd.2,
               joint.sens.sd, joint.spec.sd,
               sens.med.1, sens.med.2, spec.med.1, spec.med.2,
               joint.sens.med, joint.spec.med,
               sens.lb.1, sens.lb.2, spec.lb.1, spec.lb.2,
               joint.sens.lb, joint.spec.lb,
               sens.ub.1, sens.ub.2, spec.ub.1, spec.ub.2,
               joint.sens.ub, joint.spec.ub)
}

## FUNCTION: Calculate study performance measures ------------------
# Calculate simulation study performance measures
summary_stats <- function(sim.results, sens1, sens2, sens12,
                          spec1, spec2, spec12, nSim) {
  # Calculate study performance measures
  # Sensitivity, test 1
  sens.1.bias <- sum(abs(sim.results$sens.1 - sens1 )) / nSim
  sens.1.rmse <- sqrt(sum((sim.results$sens.1 - sens1)^2) / nSim)
  sim.results$cov.1 <- (sim.results$sens.lb.1 <= sens1) *
                       (sens1 <= sim.results$sens.ub.1)
  sens.1.cov <- sum(sim.results$cov.1) / nSim
  # Sensitivity, test 2
  sens.2.bias <- sum(abs(sim.results$sens.2 - sens2)) / nSim
  sens.2.rmse <- sqrt(sum((sim.results$sens.2 - sens2)^2) / nSim)
```

```
sim.results$cov.2 <- (sim.results$sens.lb.2 <= sens2) *
                      (sens2 <= sim.results$sens.ub.2)
sens.2.cov <- sum(sim.results$cov.2) / nSim
# Specificity, test 1
spec.1.bias <- sum(abs(sim.results$spec.1 - spec1)) / nSim
spec.1.rmse <- sqrt(sum((sim.results$spec.1 - spec1)^2) / nSim)
sim.results$cov.3 <- (sim.results$spec.lb.1 <= spec1) *
                      (spec1 <= sim.results$spec.ub.1)
spec.1.cov <- sum(sim.results$cov.3) / nSim
# Specificity, test 2
spec.2.bias <- sum(abs(sim.results$spec.2 - spec2)) / nSim
spec.2.rmse <- sqrt(sum((sim.results$spec.2 - spec2)^2) / nSim)
sim.results$cov.4 <- (sim.results$spec.lb.2 <= spec2) *
                      (spec2 <= sim.results$spec.ub.2)
spec.2.cov <- sum(sim.results$cov.4) / nSim
# Joint sensitivity
joint.sens.bias <- sum(abs(sim.results$joint.sens - sens12))
                       / nSim
joint.sens.rmse <- sqrt(sum((sim.results$joint.sens - sens12)^2)
                        / nSim)
sim.results$cov.7 <- (sim.results$joint.sens.lb <= sens12) *
                      (sens12 <= sim.results$joint.sens.ub)
joint.sens.cov <- sum(sim.results$cov.7) / nSim
# Joint specificity
joint.spec.bias <- sum(abs(sim.results$joint.spec - spec12))
                    / nSim
joint.spec.rmse <- sqrt(sum((sim.results$joint.spec - spec12)^2)
                        / nSim)
sim.results$cov.8 <- (sim.results$joint.spec.lb <= spec12) *
                      (spec12 <= sim.results$joint.spec.ub)
joint.spec.cov <- sum(sim.results$cov.8) / nSim
# Return list of summaries of performance measures
sum.stats <-
  list(sens.1.bias = sens.1.bias, sens.1.rmse = sens.1.rmse,
       sens.1.cov = sens.1.cov, sens.2.bias = sens.2.bias,
       sens.2.rmse = sens.2.rmse, sens.2.cov = sens.2.cov,
```

```
          spec.1.bias = spec.1.bias, spec.1.rmse = spec.1.rmse,
          spec.1.cov = spec.1.cov, spec.2.bias = spec.2.bias,
          spec.2.rmse = spec.2.rmse, spec.2.cov = spec.2.cov,
          joint.sens.bias = joint.sens.bias,
          joint.sens.rmse = joint.sens.rmse,
          joint.sens.cov = joint.sens.cov,
          joint.spec.bias = joint.spec.bias,
          joint.spec.rmse = joint.spec.rmse,
          joint.spec.cov = joint.spec.cov)
  }

  ## FUNCTION: Run simulation study ----------------------------------
  # Define function to run simulation study: simulate nSim data
  # sets, fit model to each data set, extract relevant results at
  # each iteration and calculate performance measures
  run_simulation_study <- function(nSim, nstudies,sens1, sens2,
                                   sens12, spec1, spec2, spec12,
                                   warmup, iter) {
    # Simulate nSim data sets
    data <- lapply(X <- 1:nSim, FUN = simulate_data,
                   nstudies = nstudies, sens1 = sens1,
                   sens2 = sens2, sens12 = sens12, spec1 = spec1,
                   spec2 = spec2, spec12 = spec12)
    # Fit binomial model to each data set and extract results
    sim.results <- do.call(rbind.data.frame,
                           mclapply(X <- data,FUN = extract_results,
                                    nstudies = nstudies,
                                    warmup = warmup, iter = iter))
    # Calculate simulation study performance measures
sum.stats <-
  summary_stats(sim.results = sim.results, sens1 = sens1,
                sens2 = sens2, sens12 = sens12, spec1 = spec1,
                spec2 = spec2, spec12 = spec12, nSim = nSim)
# Return data set of relevant results and summary statistics
return(list(sim.results = sim.results, sum.stats = sum.stats))
  }
```

```
## ANALYSIS: Run simulation and summarise model performance --------
# Run simulation study for a specified number of iterations and
# summarise model performance
# Set seed for reproducibility
set.seed(1234)
# Run simulation study
# Strong within-study associations, high sens and high spec
ptm <- proc.time() # start the clock
simulation.1 <-
  run_simulation_study(nSim = 1000, nstudies = 18, sens1 = 0.8,
                       sens2 = 0.8, sens12 = 0.8, spec1 = 0.8,
                       spec2 = 0.8, spec12 = 0.8, warmup = 1000,
                       iter = 5000)
proc.time()-ptm # stop the clock
# Display performance measures
simulation.1[["sum.stats"]]
# Save simulation results as R data file
results.simulation <- simulation.1[["sim.results"]]
save(results.simulation, file = "simulation_1.RData")
```

B.4 Convergence diagnostics for the meta-regression model

Figure B.1: Trace plots obtained from the meta-regression analysis fit to a simulated data set assuming high sensitivities and specificities with strong within-study associations.

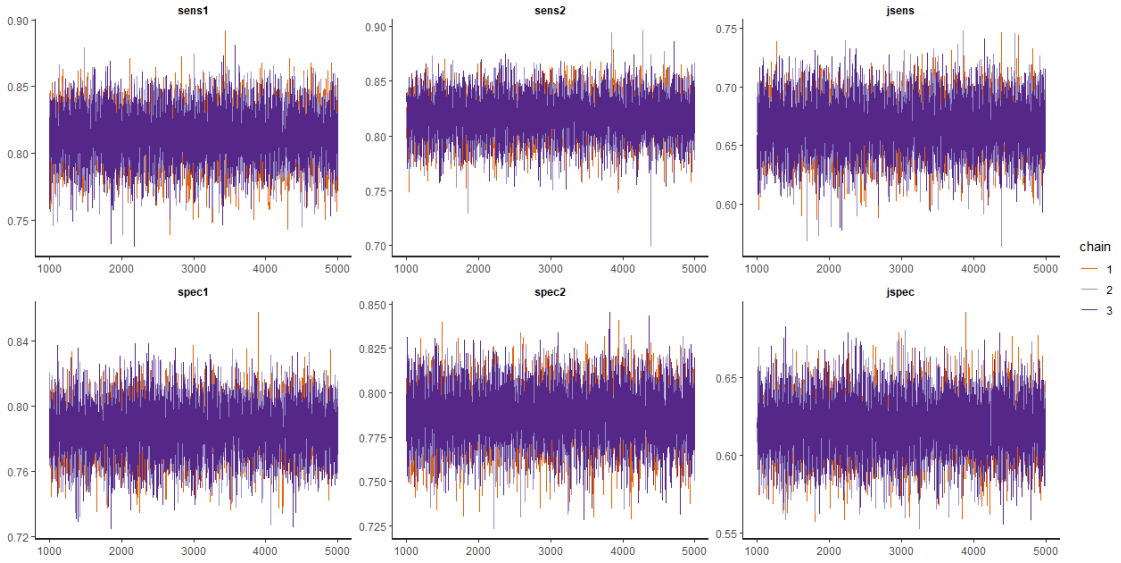


Figure B.2: Density plots obtained from the meta-regression analysis fit to a simulated data set assuming high sensitivities and specificities with strong within-study associations.

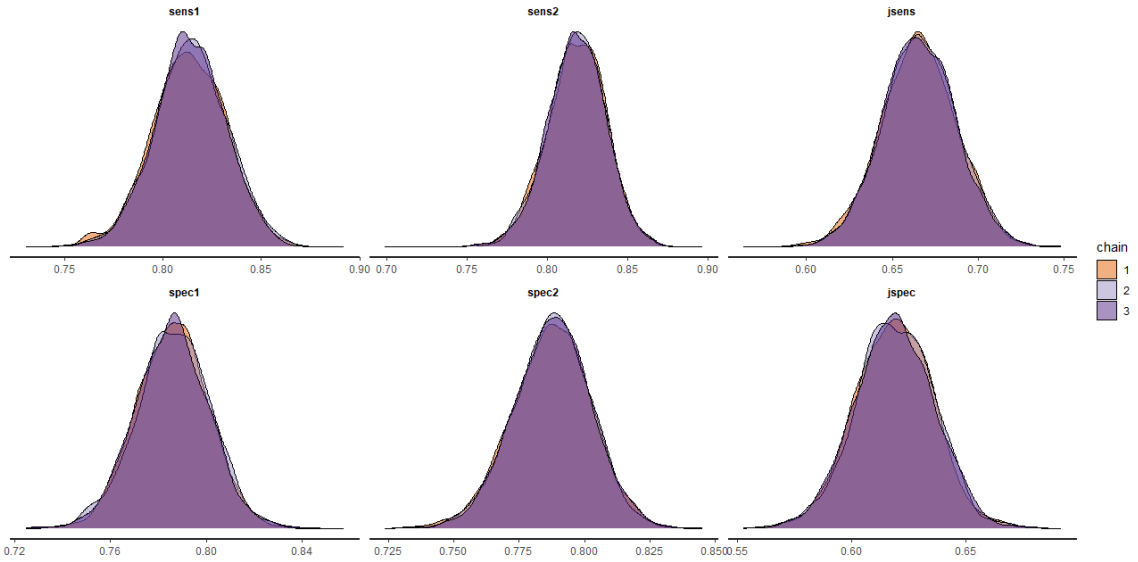
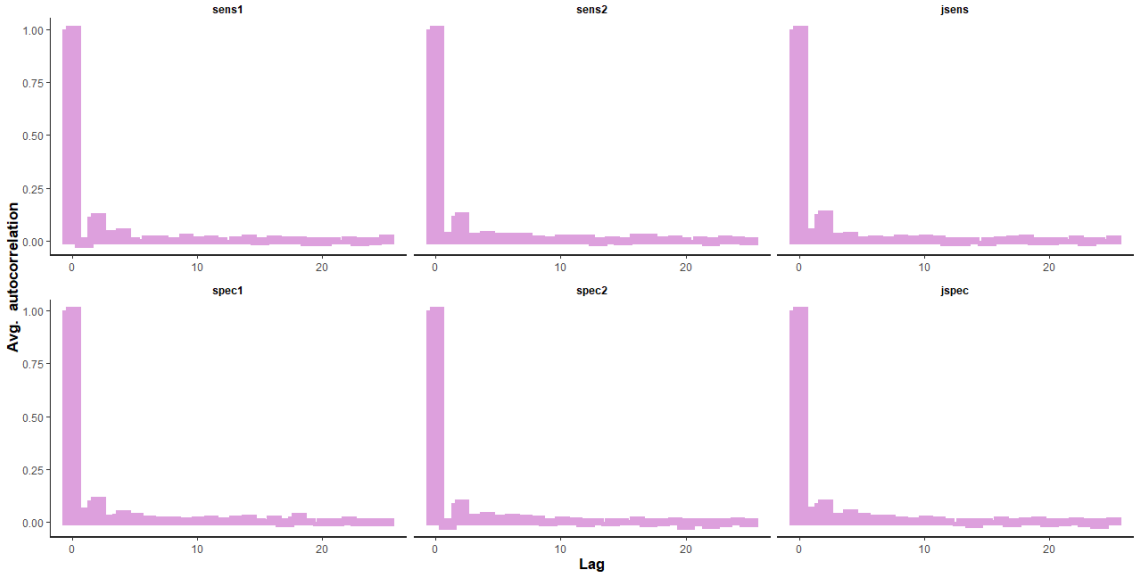


Figure B.3: Autocorrelation plots obtained from the meta-regression analysis fit to a simulated data set assuming high sensitivities and specificities with strong within-study associations. Autocorrelation is averaged over the three chains.



Appendix C

C.1 Bootstrapping method for bivariate copula models

R code for double bootstrap method, used to obtain estimates of the bivariate copula dependence parameters using individual participant data reconstructed from full cross-classifications from a single study. Where cross-classified data are available across all studies, the function below should be looped over each study to obtain a pair of copula dependence parameters per study. More detail on the double bootstrap method is available in Section 6.4.5.

```
bootstrap_copula <- function (df, Nb, copula) {
  # df = dataframe containing IPD, Nb = number of bootstrap samples
  names(df) <- paste(c('n1_00', 'n1_01', 'n1_10', 'n1_11',
                       'n0_00', 'n0_01', 'n0_10', 'n0_11'))
  s <- length(df$n1_00) # number of observations in the data
  tp1 <- tp2 <- tn1 <- tn2 <- array(1, Nb) # array of ones
  for (k in 1:Nb) {
    while ((tp1[k]==1 && tp2[k]==1) | (tn1[k]==1 && tn2[k]==1)) {
      sam <- sample(s, replace = TRUE)
      boot.1 <- df$n1_00[sam]
      boot.2 <- df$n1_01[sam]
      boot.3 <- df$n1_10[sam]
      boot.4 <- df$n1_11[sam]
      boot.5 <- df$n0_00[sam]
      boot.6 <- df$n0_01[sam]
      boot.7 <- df$n0_10[sam]
```

```
    boot.8 <- df$n0_11[sam]
    tp1[k] <- sum(boot.3 + boot.4)
    tp2[k] <- sum(boot.2 + boot.4)
    tn1[k] <- sum(boot.5 + boot.6)
    tn2[k] <- sum(boot.5 + boot.7)}}
llik1 <- function(p) - sum(dbinom(tp1, prob = p, size = s,
           log = TRUE))
llik2 <- function(p) - sum(dbinom(tp2, prob = p, size = s,
           log = TRUE))
llik3 <- function(p) - sum(dbinom(tn1, prob = p, size = s,
           log = TRUE))
llik4 <- function(p) - sum(dbinom(tn2, prob = p, size = s,
           log = TRUE))
p.tp1.hat <- optimize(llik1, c(0,1))$min
p.tp2.hat <- optimize(llik2, c(0,1))$min
p.tn1.hat <- optimize(llik3, c(0,1))$min
p.tn2.hat <- optimize(llik4, c(0,1))$min
u.tp   <- as.numeric(pbinom(tp1, s, p.tp1.hat))
u.tp.1 <- as.numeric(pbinom(tp1 - 1, s, p.tp1.hat))
v.tp   <- as.numeric(pbinom(tp2, s, p.tp2.hat))
v.tp.1 <- as.numeric(pbinom(tp2 - 1, s, p.tp2.hat))
u.tn   <- as.numeric(pbinom(tn1, s, p.tn1.hat))
u.tn.1 <- as.numeric(pbinom(tn1 - 1, s, p.tn1.hat))
v.tn   <- as.numeric(pbinom(tn2, s, p.tn2.hat))
v.tn.1 <- as.numeric(pbinom(tn2 - 1, s, p.tn2.hat))
if (copula == "gauss") {
  fA <- function(theta1) {-sum(log(cop.pmf(theta1, u.tp, v.tp,
                          u.tp.1, v.tp.1, "gauss")))}
  fB <- function(theta2) {-sum(log(cop.pmf(theta2, u.tn, v.tn,
                          u.tn.1, v.tn.1, "gauss")))}
  optf1 <- nlminb(c(.2), fA, lower = -.99, upper = .99,
             control = list(iter.max = 1000, eval.max = 1000))
  optf2 <- nlminb(c(.2), fB, lower = -.99, upper = .99,
             control = list(iter.max = 1000, eval.max = 1000))
} else if (copula == "frank") {
  fA <- function(theta1) {-sum(log(cop.pmf(theta1, u.tp, v.tp,
```

```
                          u.tp.1, v.tp.1, "frank")))}
  fB <- function(theta2) {-sum(log(cop.pmf(theta2, u.tn, v.tn,
                          u.tn.1, v.tn.1, "frank")))}
  optf1 <- nlminb(c(.2), fA,
            control = list(iter.max = 1000, eval.max = 1000))
  optf2 <- nlminb(c(.2), fB,
            control = list(iter.max = 1000, eval.max = 1000))
} else if (copula == "gumbel") {
  fA <- function(theta1) {-sum(log(cop.pmf(theta1, u.tp, v.tp,
                          u.tp.1, v.tp.1, "gumbel")))}
  fB <- function(theta2) {-sum(log(cop.pmf(theta2, u.tn, v.tn,
                          u.tn.1, v.tn.1, "gumbel")))}
  optf1 <- nlminb(c(.2), fA, lower = 1,
            control = list(iter.max = 1000, eval.max = 1000))
  optf2 <- nlminb(c(.2), fB, lower = 1,
            control = list(iter.max = 1000, eval.max = 1000))
} else if (copula == "clayton") {
  fA <- function(theta1) {-sum(log(cop.pmf(theta1, u.tp, v.tp,
                          u.tp.1, v.tp.1, "clayton")))}
  fB <- function(theta2) {-sum(log(cop.pmf(theta2, u.tn, v.tn,
                         u.tn.1, v.tn.1, "clayton")))}
  optf1 <- nlminb(c(.2), fA, lower = 0.01,
            control = list(iter.max = 1000, eval.max = 1000))
  optf2 <- nlminb(c(.2), fB, lower = 0.01,
            control = list(iter.max = 1000, eval.max = 1000))
} else if (copula == "clayton180") {
  fA <- function(theta1) {-sum(log(cop.pmf(theta1, u.tp, v.tp,
                          u.tp.1, v.tp.1, "clayton180")))}
  fB <- function(theta2) {-sum(log(cop.pmf(theta2, u.tn, v.tn,
                          u.tn.1, v.tn.1, "clayton180")))}
  optf1 <- nlminb(c(.2), fA, lower = 0.01,
            control = list(iter.max = 1000, eval.max = 1000))
  optf2 <- nlminb(c(.2), fB, lower = 0.01,
            control = list(iter.max = 1000, eval.max = 1000))
} else {
  print("Invalid choice of copula")
```

```
  }
  return(list(thetaf1 = optf1$par, thetaf2 = optf2$par,
              diag1 = optf1$convergence,
              diag2 = optf2$convergence))}

# Gaussian copula PDF
pbvncop <- function(u, v, cpar) {
  # Boundary corrections to prevent NaN
  u[1 - u < 1.e-9] <- 1-1.e-9
  v[1 - v < 1.e-9] <- 1-1.e-9
  u[u < 1.e-9] <- 1.e-9
  v[v < 1.e-9] <- 1.e-9
  t <- qnorm(u)
  s <- qnorm(v)
  cdf <- pbivnorm(t, s, cpar)
  cdf}

# Frank copula PDF
pfrank <- function(u, v, cpar) {
 cdf <- -1 / cpar * log1p((expm1((-cpar * u)) *
              (expm1(-cpar * v)) / (expm1(-cpar))))
 cdf}

# Gumbel copula PDF
pgumbel <- function(u, v, cpar) {
 cdf <- exp(-((-log(u))^cpar + (-log(v))^cpar)^(cpar^-1))
 cdf}

# Clayton copula PDF
pclayton <- function(u, v, cpar) {
  cdf <- (u^(-cpar) + v^(-cpar) - 1)^(-1 / cpar);
  cdf}

# Clayton180 PDF
pclayton180 <- function(u, v, cpar) {
 cdf <- u + v - 1 + ((1-u)^(-cpar) +
```

```
          (1-v)^(-cpar) - 1)^(-1 / cpar);
 cdf}

cop.pmf <- function(theta, u, v, u1, v1, copula) {
  if (copula == "gauss") {
    pmf <- pbvncop(u, v, theta) - pbvncop(u1, v, theta) {
             pbvncop(u, v1, theta) + pbvncop(u1, v1, theta)
  } else if (copula == "frank") {
    pmf <- pfrank(u, v, theta) - pfrank(u1, v, theta) {
             pfrank(u, v1, theta) + pfrank(u1, v1, theta)
  } else if (copula == "gumbel") {
    pmf <- pgumbel(u, v, theta) - pgumbel(u1, v, theta) {
             pgumbel(u, v1, theta) + pgumbel(u1, v1, theta)
  } else if (copula == "clayton") {
    pmf <- pclayton(u, v, theta) - pclayton(u1, v, theta) {
             pclayton(u, v1, theta) + pclayton(u1, v1, theta)
  } else if (copula == "clayton180") {
    pmf <- pclayton180(u, v, theta) - pclayton180(u1, v, theta) {
             pclayton180(u, v1, theta) + pclayton180(u1, v1, theta)
  } else {
    print("Invalid choice of copula")}}
```

C.2 Bivariate copula models for joint meta-analysis of diagnostic accuracy data on two tests

Stan code for implementing the novel Bayesian meta-analysis models for synthesising diagnostic accuracy data on two tests evaluated using a paired study design, in which all patients undergo both tests plus a reference standard. The model specification is described in Section 6.4.4. Five types of bivariate copula capture within-study dependencies between two tests: Gaussian, Frank, Gumbel, Clayton and Clayton 180°.

```
functions {
  // Gaussian copula CDF
```

```
real fcop(real theta, real u1, real u2) {
  real z1 = inv_Phi(u1);
  real z2 = inv_Phi(u2);
  if (z1 != 0 || z2 != 0) {
    real denom = fabs(theta) < 1.0 ? sqrt((1 + theta) *
                  (1 - theta)) : not_a_number();
    real a1 = (z2 / z1 - theta) / denom;
    real a2 = (z1 / z2 - theta) / denom;
    real product = z1 * z2;
    real delta = product < 0 || (product == 0 && (z1 + z2) < 0);
    return 0.5 * (u1 + u2 - delta) - owens_t(z1, a1) -
             owens_t(z2, a2); }
return 0.25 + asin(theta) / (2 * pi());}

// Frank copula CDF
real fcop2(real theta, real u1, real u2) {
  return -1 / theta * log1p((expm1(-theta * u1)) *
           (expm1(-theta * u2)) / (expm1(-theta)));}

// Gumbel copula CDF
real fcop3(real theta, real u1, real u2) {
  real neg_log_u1;
  real neg_log_u2;
  neg_log_u1 = -log(u1);
  neg_log_u2 = -log(u2);
  return exp(-(neg_log_u1^theta +
                neg_log_u2^theta)^(1 / theta));}

// Clayton copula CDF
real fcop4(real theta, real u1, real u2) {
  return (u1^(-theta) + u2^(-theta) - 1)^(-1 / theta);}

// Clayton180 copula CDF
real fcop5(real theta, real u1, real u2) {
  return u1 + u2 - 1 + ((1 - u1)^(-theta) +
           (1 - u2)^(-theta) - 1)^(-1 / theta);}
```

```
  real Bivfcop_lpmf(int[] r, int n, real theta, vector p) {
    vector[2] f;
    vector[2] f1;
    real prob;
    for(i in 1:2) {
      f1[i] = binomial_cdf(r[i] - 1| n, p[i]);
      if (f1[i] > 1-1.e-9) f1[i] = 1-1.e-9;
      if (f1[i] < 1.e-9) f1[i] = 1.e-9;
      f[i] = binomial_cdf(r[i] | n, p[i]);
      if (f[i] > 1-1.e-9) f[i] = 1-1.e-9;
      if (f[i] < 1.e-9) f[i] = 1.e-9; }
    prob = fcop(theta,f[1],f[2]) - fcop(theta,f[1],f1[2]) -
             fcop(theta,f1[1],f[2]) + fcop(theta,f1[1],f1[2]);
    return log(prob);}}

data {
 int<lower = 0> Ns;
 int<lower = 0> tp[Ns,2]; //2x2 data on each test
 int<lower = 0> tn[Ns,2];
 int<lower = 0> disease[Ns];
 int<lower = 0> nodisease[Ns];
 real theta1; //copula dependence parameters
 real theta2;}

parameters {
 real rr;
 vector[4] b;
 vector<lower = 0>[4] tau;
 vector[4] z[Ns];}

transformed parameters {
  cov_matrix[4] Tau;
  cholesky_factor_cov[4] L;
  vector[4] mu[Ns];
  vector<lower = 0, upper = 1>[4] p[Ns];
```

```
  real<lower = -1, upper = 1> rho;
  rho = tanh(rr);
  for (j in 1:4) {
    Tau[j,j] = tau[j]^2;
    for (k in (j+1):4) {
      Tau[j,k] = tau[j]*tau[k]*rho;
      Tau[k,j] = Tau[j,k];}}
  L = cholesky_decompose(Tau);
  //non-centred parameterisation
  for (i in 1:Ns) {
    mu[i] = b + (L*z[i]);
    p[i] = inv_logit(mu[i]);}}

model {
  //priors
  rr ~ normal(0,0.8);
  b ~ normal(0,10);
  tau ~ normal(0,2.5);
  for (i in 1:Ns) {
    z[i] ~ std_normal();
    //likelihoods
    target += Bivfcop_lpmf(tp[i,1:2] | disease[i],
                theta1, p[i,1:2]);
    target += Bivfcop_lpmf(tn[i,1:2] | nodisease[i],
                theta2, p[i,3:4]);}}

generated quantities {
 real<lower = 0, upper = 1> sens1;
 real<lower = 0, upper = 1> sens2;
 real<lower = 0, upper = 1> spec1;
 real<lower = 0, upper = 1> spec2;
 sens1 = inv_logit(b[1]);
 sens2 = inv_logit(b[2]);
 spec1 = inv_logit(b[3]);
 spec2 = inv_logit(b[4]);
 vector[Ns] log_lik;
```

```
for (n in 1:Ns) {
//monitor log-likelihood to calculate WAIC post-estimation
log_lik[n] = Bivfcop_lpmf(tp[n,1:2] | disease[n],
                theta1, p[n,1:2]);}}
```

C.3 Convergence diagnostics for the meta-regression model

Figure C.1: Trace plots obtained from the meta-regression analysis comparing the diagnostic accuracy of amyloid-β and phosphorylated tau.

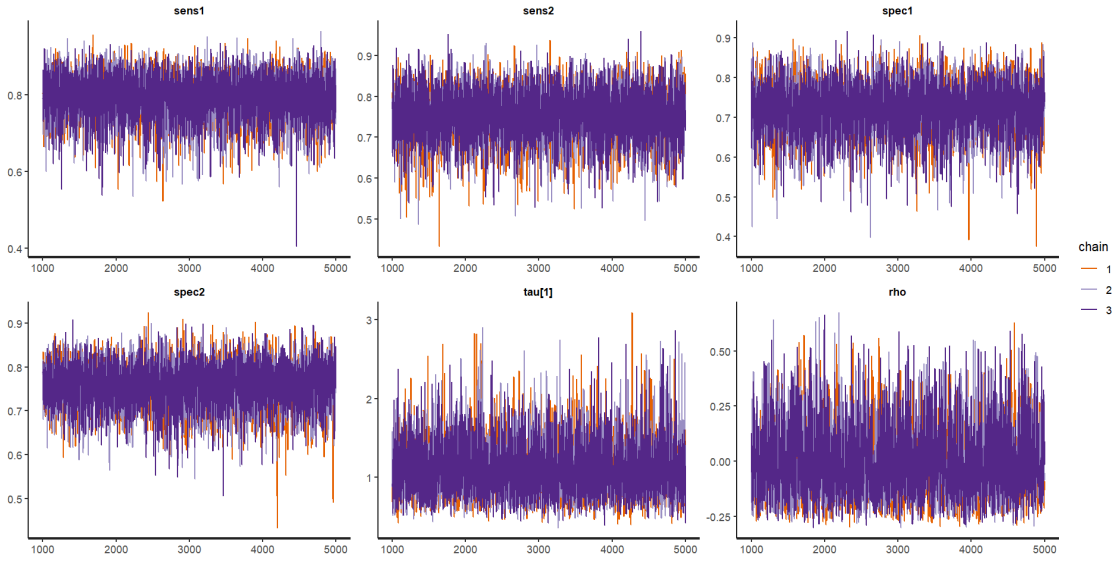


Figure C.2: Density plots obtained from the meta-regression analysis comparing the diagnostic accuracy of amyloid-β and phosphorylated tau.

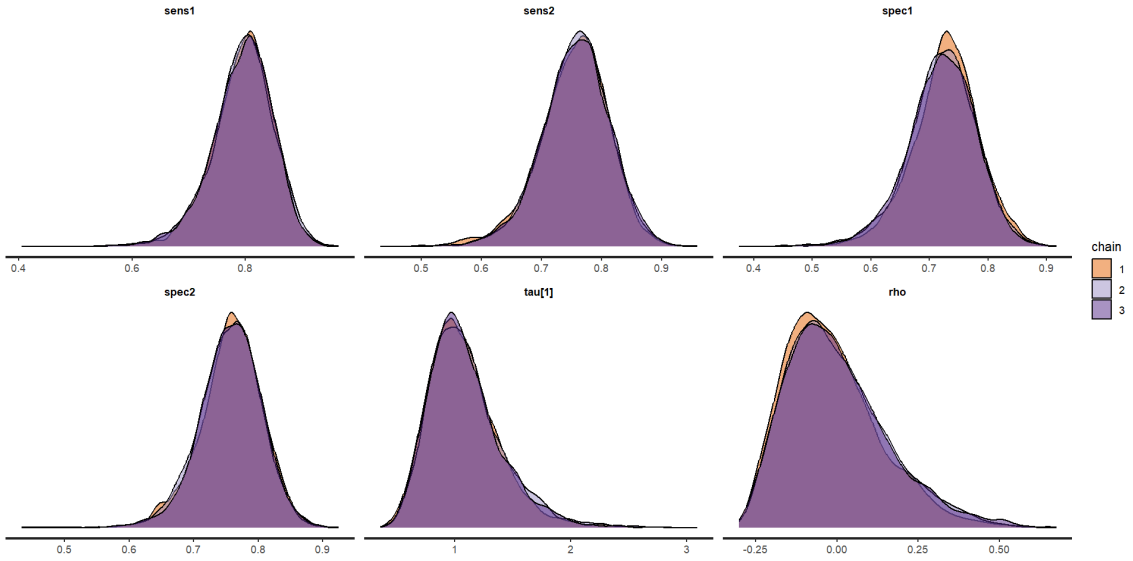
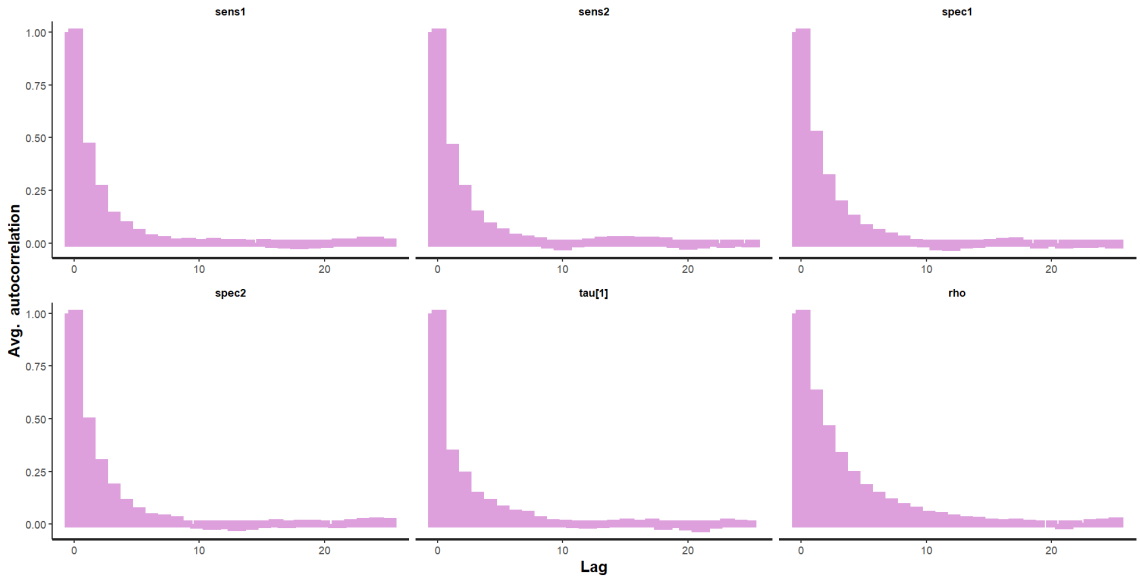


Figure C.3: Autocorrelation plots obtained from the meta-regression analysis comparing the diagnostic accuracy of amyloid-β and phosphorylated tau. Autocorrelation is averaged over the three chains.



C.4 Convergence diagnostics for the bivariate Gaussian copula meta-analysis model

Figure C.4: Trace plots obtained from the bivariate Gaussian copula meta-analysis comparing the diagnostic accuracy of amyloid-β and phosphorylated tau.

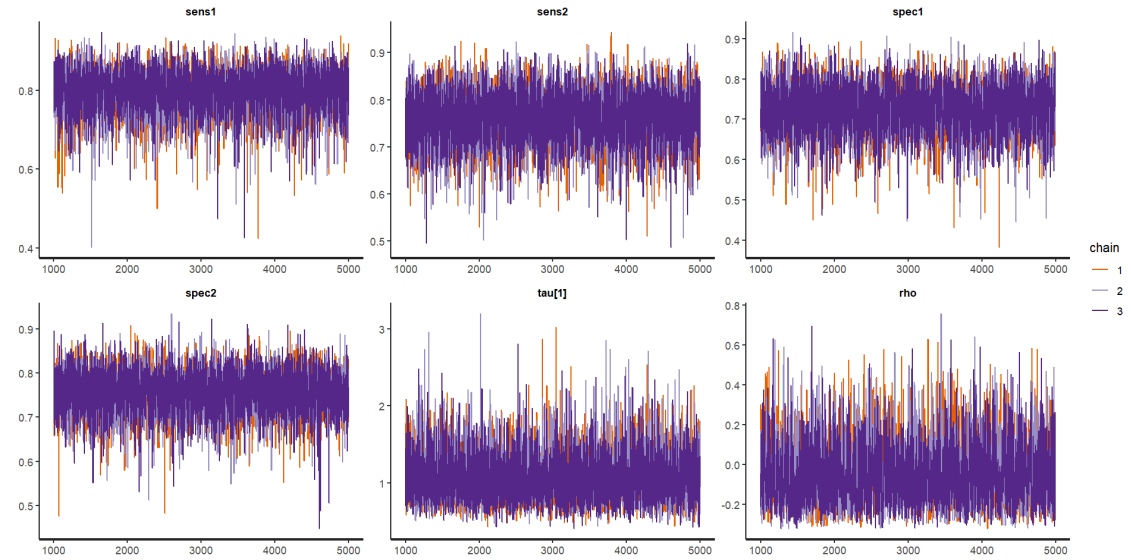


Figure C.5: Density plots obtained from the bivariate Gaussian copula meta-analysis comparing the diagnostic accuracy of amyloid-β and phosphorylated tau.

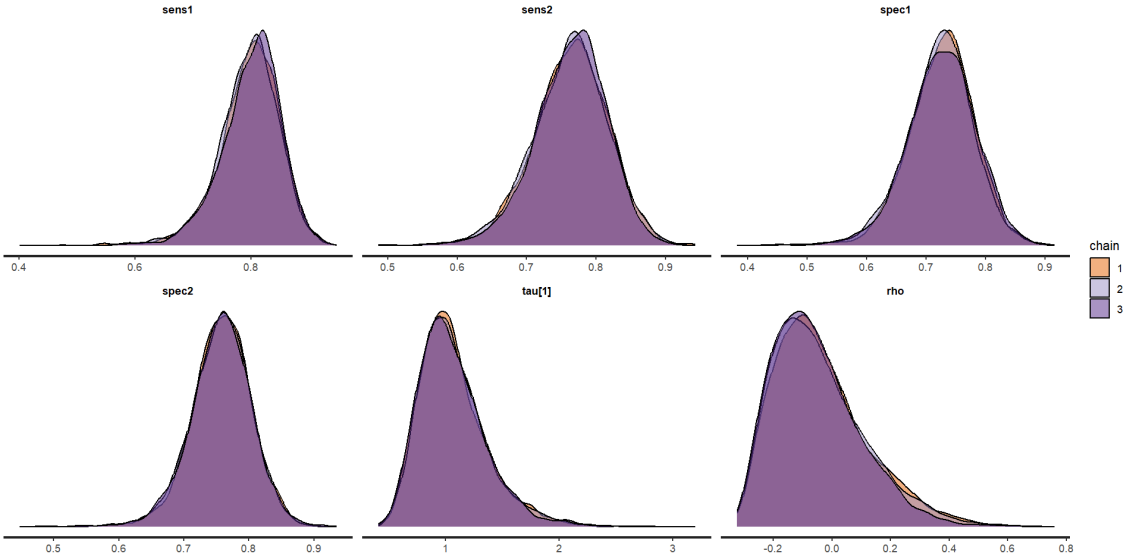
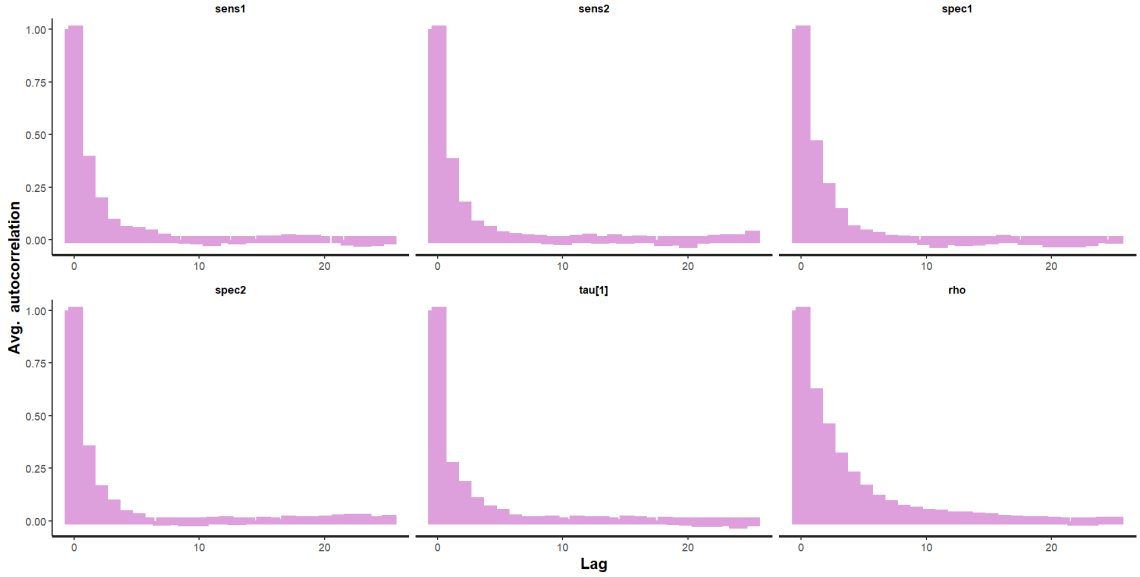


Figure C.6: Autocorrelation plots obtained from the bivariate Gaussian copula meta-analysis comparing the diagnostic accuracy of amyloid-β and phosphorylated tau. Autocorrelation is averaged over the three chains.



C.5 Convergence diagnostics for the bivariate Frank copula meta-analysis model

Figure C.7: Trace plots obtained from the bivariate Frank copula meta-analysis comparing the diagnostic accuracy of amyloid-β and phosphorylated tau.

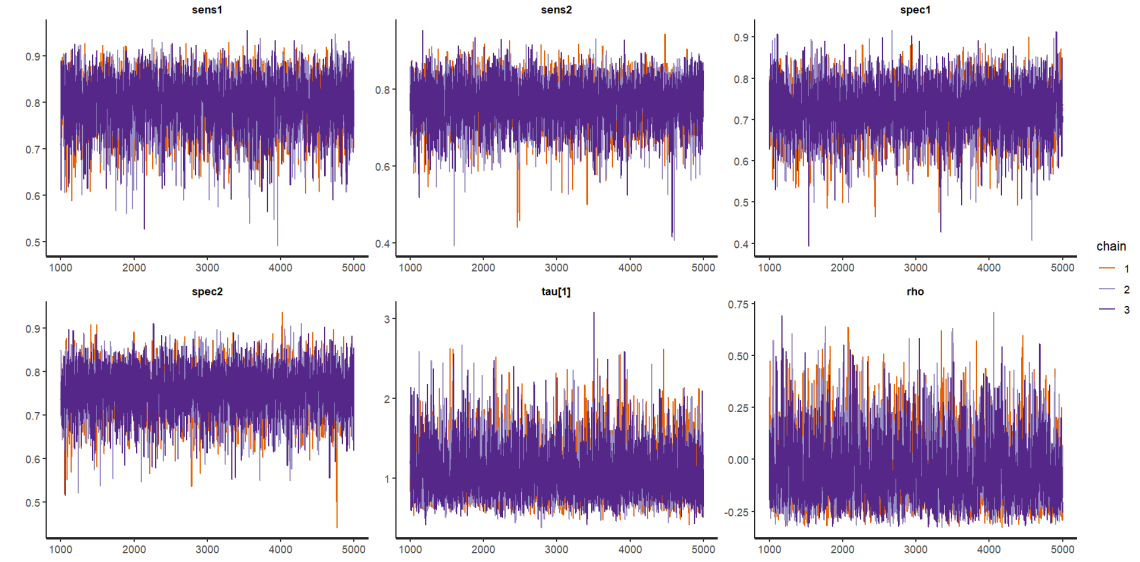


Figure C.8: Density plots obtained from the bivariate Frank copula meta-analysis comparing the diagnostic accuracy of amyloid-β and phosphorylated tau.

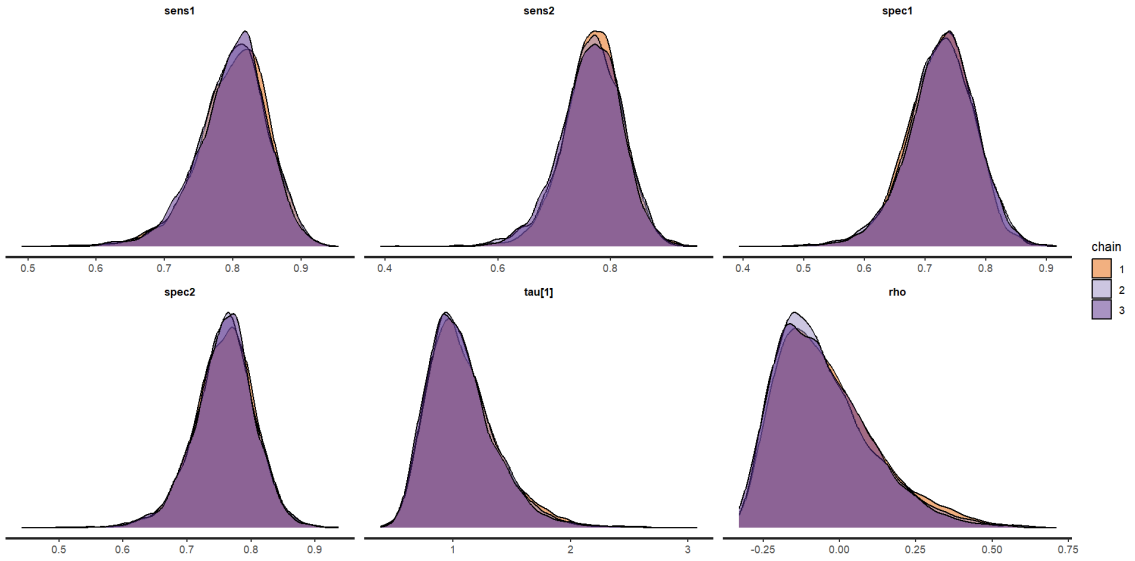
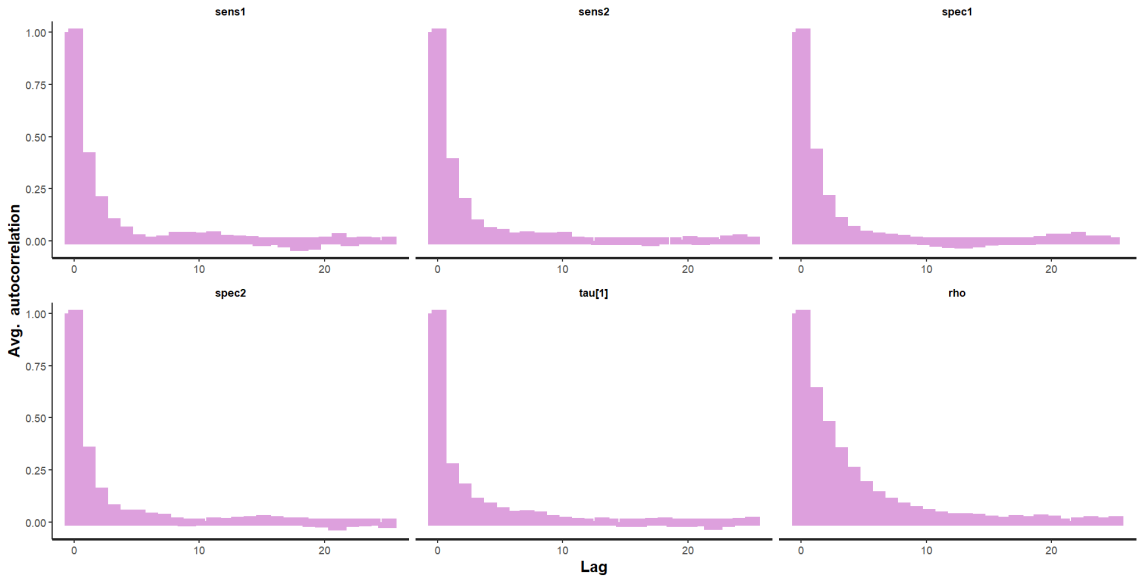


Figure C.9: Autocorrelation plots obtained from the bivariate Frank copula meta-analysis comparing the diagnostic accuracy of amyloid-β and phosphorylated tau. Autocorrelation is averaged over the three chains.



C.6 Convergence diagnostics for the bivariate Gumbel copula meta-analysis model

Figure C.10: Trace plots obtained from the bivariate Gumbel copula meta-analysis comparing the diagnostic accuracy of amyloid-β and phosphorylated tau.

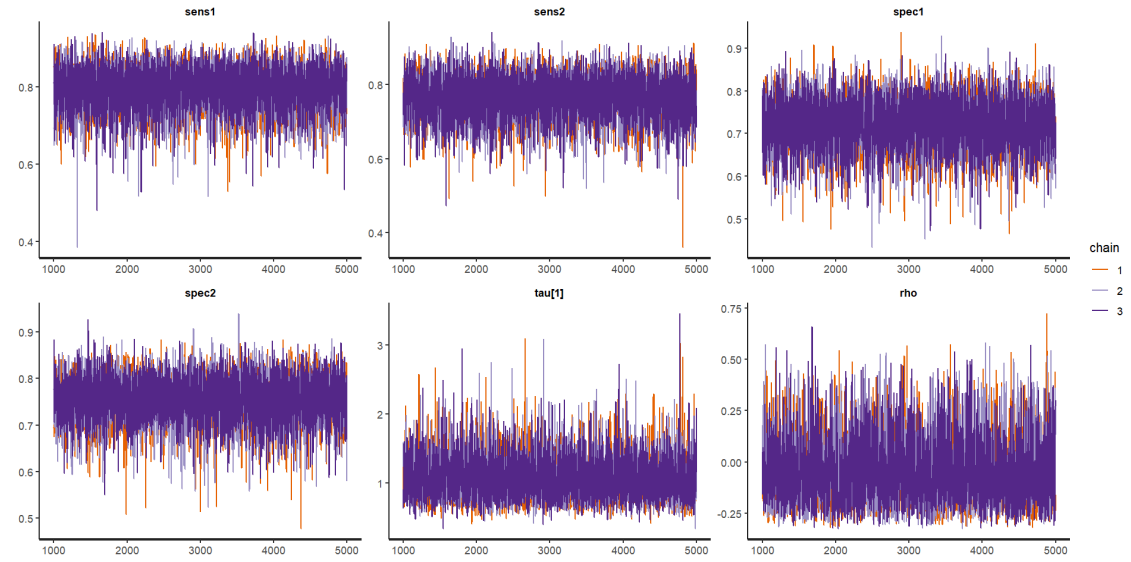


Figure C.11: Density plots obtained from the bivariate Gumbel copula meta-analysis comparing the diagnostic accuracy of amyloid-β and phosphorylated tau.

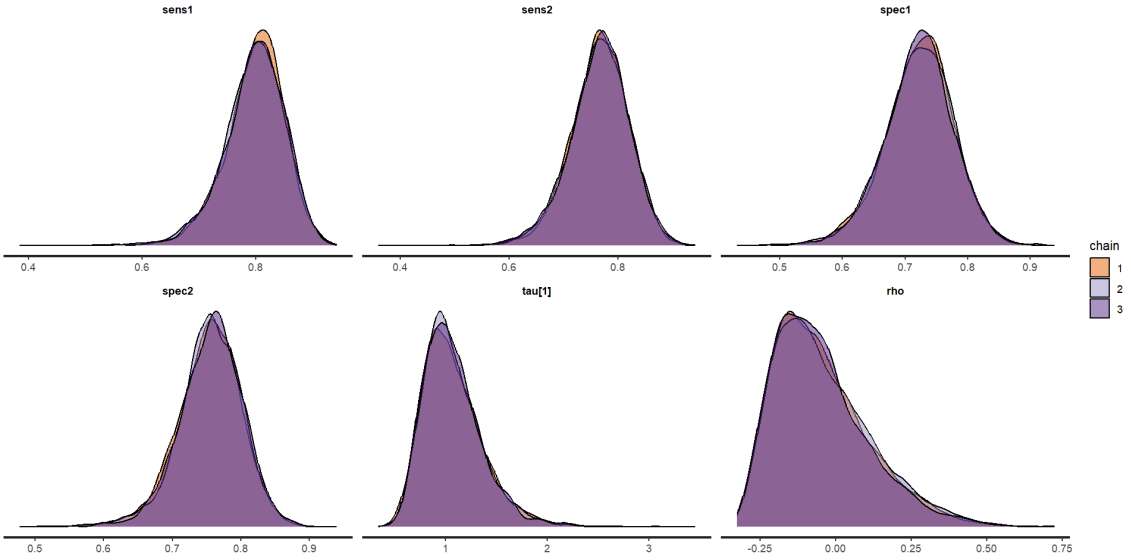
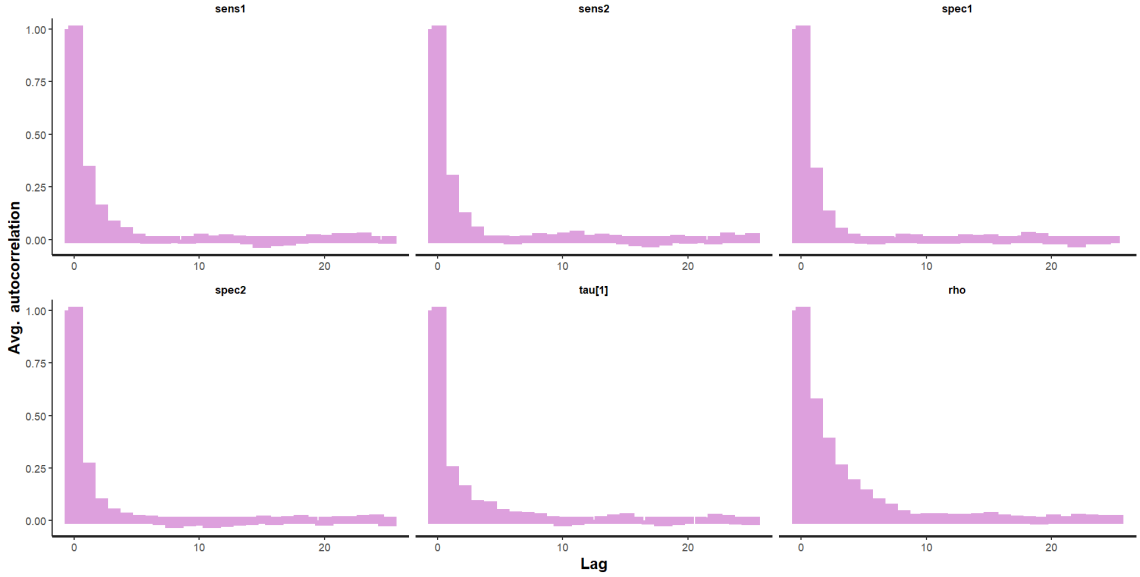


Figure C.12: Autocorrelation plots obtained from the bivariate Gumbel copula meta-analysis comparing the diagnostic accuracy of amyloid-β and phosphorylated tau. Autocorrelation is averaged over the three chains.



C.7 Convergence diagnostics for the bivariate Clayton copula meta-analysis model

Figure C.13: Trace plots obtained from the bivariate Clayton copula meta-analysis comparing the diagnostic accuracy of amyloid-β and phosphorylated tau.

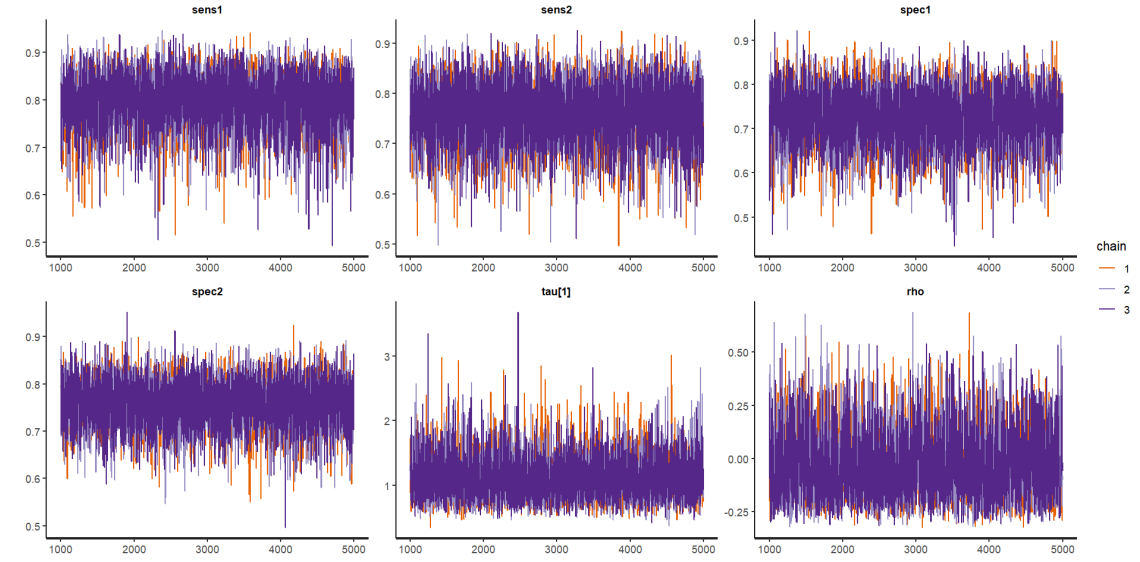


Figure C.14: Density plots obtained from the bivariate Clayton copula meta-analysis comparing the diagnostic accuracy of amyloid-β and phosphorylated tau.

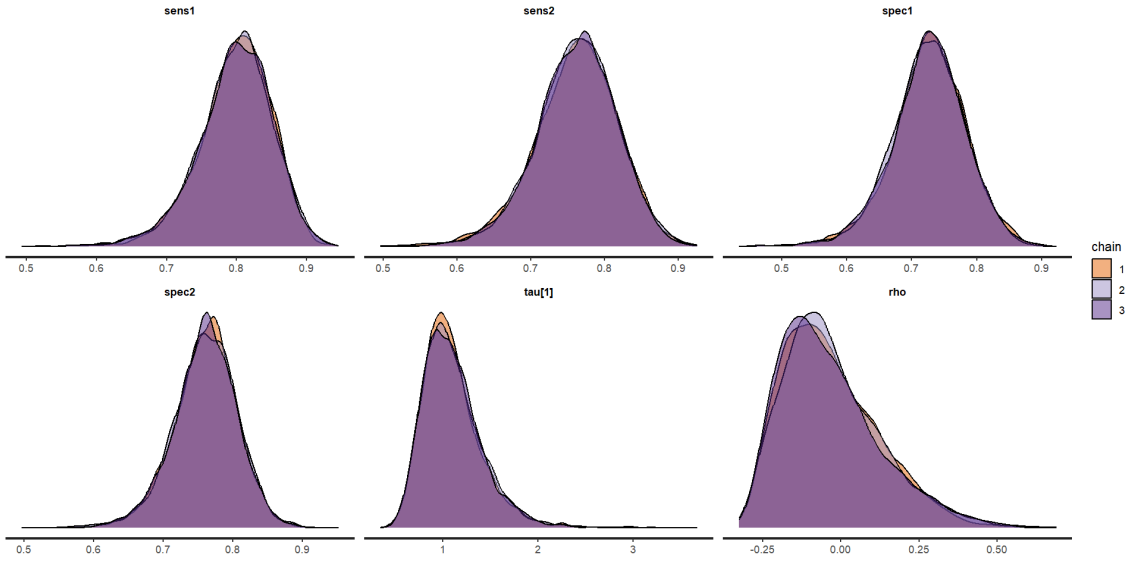
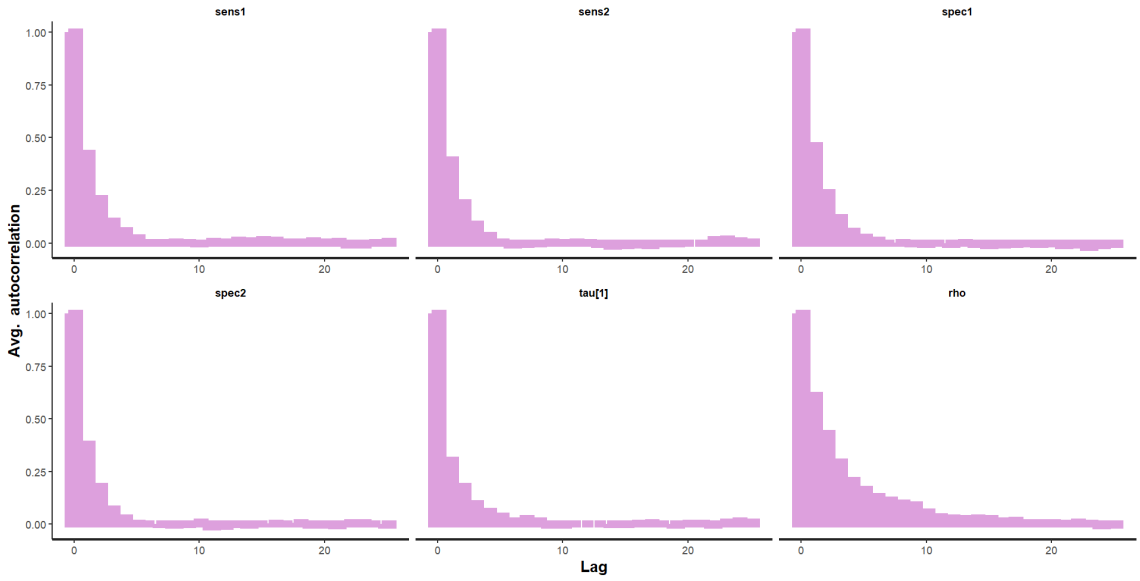


Figure C.15: Autocorrelation plots obtained from the bivariate Clayton copula meta-analysis comparing the diagnostic accuracy of amyloid-β and phosphorylated tau. Autocorrelation is averaged over the three chains.



C.8 Convergence diagnostics for the bivariate Clayton 180°copula meta-analysis model

Figure C.16: Trace plots obtained from the bivariate Clayton 180°copula meta-analysis comparing the diagnostic accuracy of amyloid-β and phosphorylated tau.

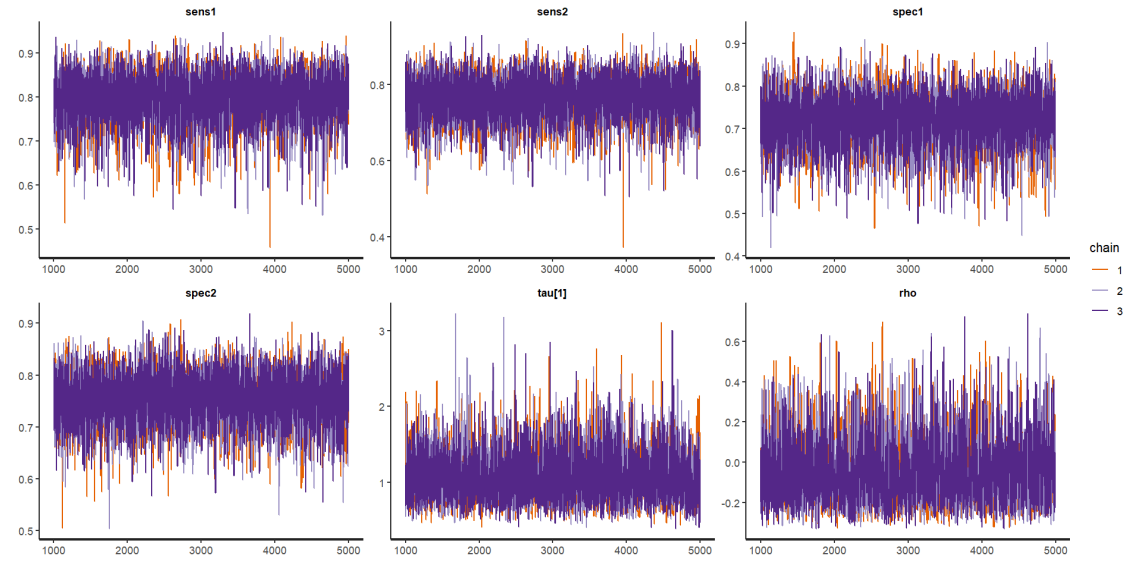


Figure C.17: Density plots obtained from the bivariate Clayton 180°copula meta-analysis comparing the diagnostic accuracy of amyloid-β and phosphorylated tau.

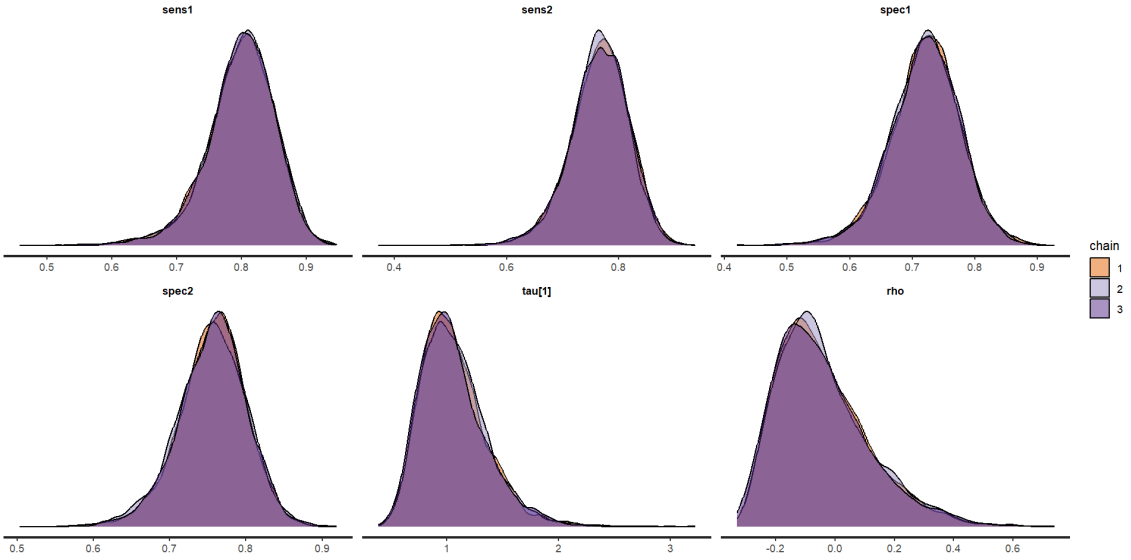
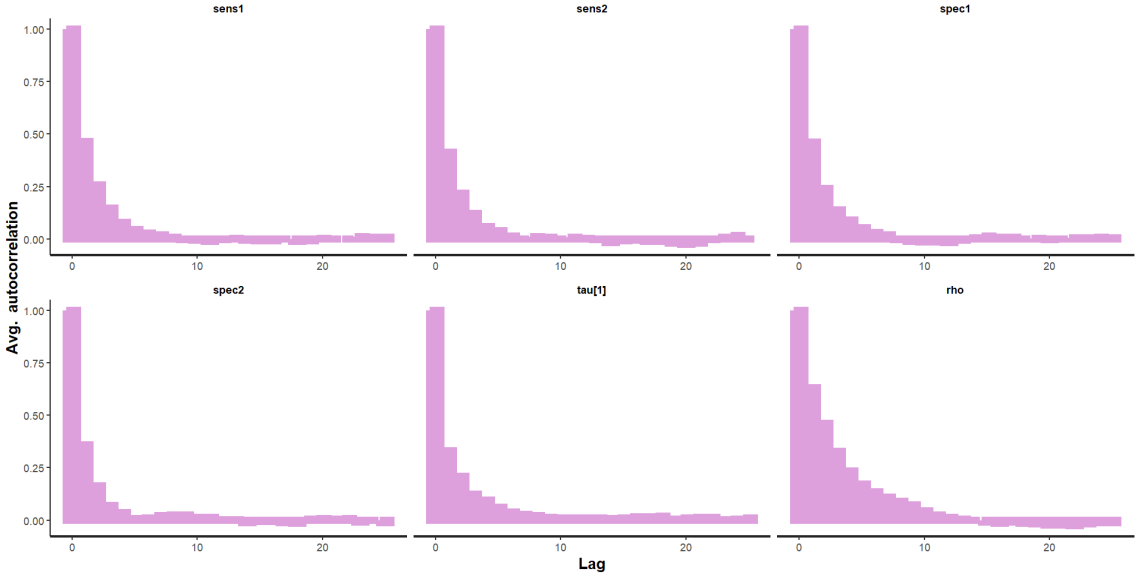


Figure C.18: Autocorrelation plots obtained from the bivariate Clayton 180°copula meta-analysis comparing the diagnostic accuracy of amyloid-β and phosphorylated tau. Autocorrelation is averaged over the three chains.



Appendix D

D.1 Bootstrapping method for trivariate copula models

R code for double bootstrap method, used to obtain estimates of the trivariate copula dependence parameters using individual participant data reconstructed from full cross-classifications from a single study. The function below should be looped over cross-classified data for each study to obtain a pair of copula dependence parameters per study. More detail on the double bootstrap method is available in Section 7.4.4.

```
pfrank <- function(u, v, w, cpar) {
  cdf <- -1 / cpar * log1p((expm1((-cpar*u)) *
         (expm1(-cpar*v)) * (expm1(-cpar*w)) / ((expm1(-cpar))^2)))
  cdf}

pgumbel <- function(u, v, w, cpar) {
  cdf <- exp(-((-log(u))^cpar + (-log(v))^cpar +
                                (-log(w))^cpar)^(cpar^-1))
  cdf}

pclayton <- function(u, v, w, cpar) {
  cdf <- (u^(-cpar) + v^(-cpar) + w^(-cpar) - 2)^(-1 / cpar);
  cdf}

cop.pmf <- function(theta, u, v, w, u1, v1, w1, copula) {
  if (copula == "frank") {
    pmf <- pfrank(u, v, w, theta) - pfrank(u1, v, w, theta) -
```

```
          pfrank(u, v1, w, theta) - pfrank(u, v, w1, theta) +
          pfrank(u1, v1, w, theta) + pfrank(u, v1, w1, theta) +
          pfrank(u1, v, w1, theta) - pfrank(u1, v1, w1, theta)
  } else if (copula == "gumbel") {
    pmf <- pgumbel(u, v, w, theta) - pgumbel(u1, v, w, theta) -
          pgumbel(u, v1, w, theta) - pgumbel(u, v, w1, theta) +
          pgumbel(u1, v1, w, theta) + pgumbel(u, v1, w1, theta) +
          pgumbel(u1, v, w1, theta) - pgumbel(u1, v1, w1, theta)
  } else if (copula == "clayton") {
    pmf <- pclayton(u, v, w, theta) - pclayton(u1, v, w, theta) -
          pclayton(u, v1, w, theta) - pclayton(u, v, w1, theta) +
          pclayton(u1, v1, w, theta) + pclayton(u, v1, w1, theta) +
          pclayton(u1, v, w1, theta) - pclayton(u1, v1, w1, theta)
  } else {
    print("Invalid choice of copula")}}

bootstrap_copula <- function (df, Nb, copula) {
  # df = dataframe containing IPD, Nb = number of bootstrap samples
  names(df) <- paste(c('n1_00', 'n1_01', 'n1_10', 'n1_11',
                       'n0_00', 'n0_01', 'n0_10', 'n0_11'))
  s <- length(df$n1_00) # number of observations in the data
  tp1 <- tp2 <- x1_11 <- tn1 <- tn2 <- x0_00 <- array(1, Nb)
  for (k in 1:Nb) {
    while ((tp1[k] == 1 && tp2[k] == 1 && x1_11[k] == 1) |
             (tn1[k] == 1 && tn2[k] == 1 && x0_00[k] == 1)) {
      sam <- sample(s, replace = TRUE)
      boot.1 <- df$n1_00[sam]
      boot.2 <- df$n1_01[sam]
      boot.3 <- df$n1_10[sam]
      boot.4 <- df$n1_11[sam]
      boot.5 <- df$n0_00[sam]
      boot.6 <- df$n0_01[sam]
      boot.7 <- df$n0_10[sam]
      boot.8 <- df$n0_11[sam]
      tp1[k] <- sum(boot.3 + boot.4)
      tp2[k] <- sum(boot.2 + boot.4)
```

```
    x1_11[k] <- sum(boot.4)
    tn1[k] <- sum(boot.5 + boot.6)
    tn2[k] <- sum(boot.5 + boot.7)
    x0_00[k] <- sum(boot.5)}}
llik1 <- function(p) -
           sum(dbinom(tp1, prob = p, size = s, log = TRUE))
llik2 <- function(p) -
           sum(dbinom(tp2, prob = p, size = s, log = TRUE))
llik3 <- function(p) -
           sum(dbinom(x1_11, prob = p, size = s, log = TRUE))
llik4 <- function(p) -
           sum(dbinom(tn1, prob = p, size = s, log = TRUE))
llik5 <- function(p) -
           sum(dbinom(tn2, prob = p, size = s, log = TRUE))
llik6 <- function(p) -
           sum(dbinom(x0_00, prob = p, size = s, log = TRUE))
p.tp1.hat <- optimize(llik1, c(0,1))$min
p.tp2.hat <- optimize(llik2, c(0,1))$min
p.x1_11.hat <- optimize(llik3, c(0,1))$min
p.tn1.hat <- optimize(llik4, c(0,1))$min
p.tn2.hat <- optimize(llik5, c(0,1))$min
p.x0_00.hat <- optimize(llik6, c(0,1))$min
u.tp   <- as.numeric(pbinom(tp1, s, p.tp1.hat))
u.tp.1 <- as.numeric(pbinom(tp1 - 1, s, p.tp1.hat))
v.tp   <- as.numeric(pbinom(tp2, s, p.tp2.hat))
v.tp.1 <- as.numeric(pbinom(tp2 - 1, s, p.tp2.hat))
w.tp   <- as.numeric(pbinom(x1_11, s, p.x1_11.hat))
w.tp.1 <- as.numeric(pbinom(x1_11 - 1, s, p.x1_11.hat))
u.tn   <- as.numeric(pbinom(tn1, s, p.tn1.hat))
u.tn.1 <- as.numeric(pbinom(tn1 - 1, s, p.tn1.hat))
v.tn   <- as.numeric(pbinom(tn2, s, p.tn2.hat))
v.tn.1 <- as.numeric(pbinom(tn2 - 1, s, p.tn2.hat))
w.tn   <- as.numeric(pbinom(x0_00, s, p.x0_00.hat))
w.tn.1 <- as.numeric(pbinom(x0_00 - 1, s, p.x0_00.hat))
if (copula == "frank") {
  fA <- function(theta1) {-sum(log(cop.pmf(theta1, u.tp, v.tp,
```

```
            w.tp, u.tp.1, v.tp.1, w.tp.1, "frank")))}
    fB <- function(theta2) {-sum(log(cop.pmf(theta2, u.tn, v.tn,
            w.tn, u.tn.1, v.tn.1, w.tn.1, "frank")))}
    optf1 <- nlminb(c(.2), fA,
               control = list(iter.max = 1000, eval.max = 1000))
    optf2 <- nlminb(c(.2), fB,
               control = list(iter.max = 1000, eval.max = 1000))
  } else if (copula == "gumbel") {
    fA <- function(theta1) {-sum(log(cop.pmf(theta1, u.tp, v.tp,
            w.tp, u.tp.1, v.tp.1, w.tp.1, "gumbel")))}
    fB <- function(theta2) {-sum(log(cop.pmf(theta2, u.tn, v.tn,
            w.tn, u.tn.1, v.tn.1, w.tn.1, "gumbel")))}
    optf1 <- nlminb(c(.2), fA, lower = 1,
               control = list(iter.max = 1000, eval.max = 1000))
    optf2 <- nlminb(c(.2), fB, lower = 1,
               control = list(iter.max = 1000, eval.max = 1000))
  } else if (copula == "clayton") {
    fA <- function(theta1) {-sum(log(cop.pmf(theta1, u.tp, v.tp,
            w.tp, u.tp.1, v.tp.1, w.tp.1, "clayton")))}
    fB <- function(theta2) {-sum(log(cop.pmf(theta2, u.tn, v.tn,
            w.tn, u.tn.1, v.tn.1, w.tn.1, "clayton")))}
    optf1 <- nlminb(c(.2), fA, lower = 0.01,
               control = list(iter.max = 1000, eval.max = 1000))
    optf2 <- nlminb(c(.2), fB, lower = 0.01,
               control = list(iter.max = 1000, eval.max = 1000))
  } else {
    print("Invalid choice of copula")
  }
return(list(thetaf1 = optf1$par, thetaf2 = optf2$par,
            diag1 = optf1$convergence, diag2 = optf2$convergence))
}
```

D.2 Trivariate copula models for joint meta-analysis of diagnostic accuracy data on two tests

Stan code for implementing the novel Bayesian meta-analysis models for synthesising diagnostic accuracy data on two tests evaluated using a paired study design, in which all patients undergo both tests plus a reference standard. The model specification is described in Section 7.4.3. Three types of trivariate copula capture within-study dependencies between two tests: Frank, Gumbel and Clayton.

```
functions {
  // Frank copula CDF
  real fcop(real theta, real u, real v, real w) {
    real a = -1 / theta * log1p((expm1(-theta * u)) *
 (expm1(-theta * v)) * (expm1(-theta * w)) /
 ((expm1(-theta))^2));
    return a;}

  // Gumbel copula CDF
  real fcop2(real theta, real u, real v, real w) {
    real a;
    real t1 = u;
    real t2 = v;
    real t3 = w;
    real neg_log_u;
    real neg_log_v;
    real neg_log_w;
    if (t1 > .9999999) {t1 = .9999999;} // boundary condition
    if (t2 > .9999999) {t2 = .9999999;} // boundary condition
    if (t3 > .9999999) {t3 = .9999999;} // boundary condition
    neg_log_u = -log(t1);
    neg_log_v = -log(t2);
    neg_log_w = -log(t3);
    a = exp(-(neg_log_u^theta + neg_log_v^theta +
              neg_log_w^theta)^(1 / theta));
    return a;}
```

```
  // Clayton copula CDF
  real fcop3(real theta, real u, real v, real w) {
    real a = (u^(-theta) + v^(-theta) +
                w^(-theta) - 2)^(-1 / theta);
    return a;}

// Trivariate pmf to model binomial aggregate data jointly
real Trivfcop_lpmf(int[] r,int n, real theta, vector mu) {
  real p1 = inv_logit(mu[1]);
  real p2 = inv_logit(mu[2]);
  real p3 = inv_logit(mu[3]);
  real f11 = binomial_cdf(r[1] - 1, n, p1);
  real f12 = binomial_cdf(r[2] - 1, n, p2);
  real f13 = binomial_cdf(r[3] - 1, n, p3);
  real f1 = f11 + exp(binomial_logit_lpmf(r[1] |n, mu[1]));
  real f2 = f12 + exp(binomial_logit_lpmf(r[2] |n, mu[2]));
  real f3 = f13 + exp(binomial_logit_lpmf(r[3] |n, mu[3]));
  real prob = fcop2(theta,f1,f2,f3) -
              fcop2(theta,f11,f2,f3) - fcop2(theta,f1,f12,f3) -
              fcop2(theta,f1,f2,f13) + fcop2(theta,f11,f12,f3) +
              fcop2(theta,f1,f12,f13) + fcop2(theta,f11,f2,f13) -
              fcop2(theta,f11,f12,f13);
  return log(prob);}}

data {
  int<lower = 0> Ns;
  int<lower = 0> x1[Ns,3]; // x1dot, xdot1, x11
  int<lower = 0> x0[Ns,3]; // x0dot, xdot0, x00
  int<lower = 0> disease[Ns];
  int<lower = 0> nodisease[Ns];
  real theta1[Ns];
  real theta2[Ns];}

parameters {
  real rr;
```

```
  vector[6] b;
  vector<lower = 0>[6] tau;
  vector[6] z[Ns];}

transformed parameters {
  matrix[6,6] Tau;
  matrix[6,6] L;
  vector[6] mu[Ns];
  real<lower = -1, upper = 1> rho;
  rho = tanh(rr);
  for (j in 1:6) {
    Tau[j,j] = tau[j]^2;
    for (k in (j+1):6) {
      Tau[j,k] = tau[j]*tau[k]*rho;
      Tau[k,j] = Tau[j,k];}}
  L = cholesky_decompose(Tau);
  //non-centred parameterisation
  for (i in 1:Ns) {
    mu[i] = b + L*z[i];}}

model {
  //priors
  rr ~ normal(0,0.8);
  b ~ normal(0,10);
  tau ~ normal(0,2.5);
  for (i in 1:Ns) {
    z[i] ~ std_normal();
    //likelihoods
    x1[i,1:3] ~ Trivfcop(disease[i], theta1[i], mu[i,1:3]);
    x0[i,1:3] ~ Trivfcop(nodisease[i], theta2[i], mu[i,4:6]);}}

generated quantities {
  real<lower = 0, upper = 1> sens1;
  real<lower = 0, upper = 1> sens2;
  real<lower = 0, upper = 1> spec1;
  real<lower = 0, upper = 1> spec2;
```

```
real<lower = 0, upper = 1> jsens;
real<lower = 0, upper = 1> jspec;
real<lower = 0, upper = 1> AND;
real<lower = 0, upper = 1> OR;
sens1 = inv_logit(b[1]);
sens2 = inv_logit(b[2]);
jsens = inv_logit(b[3]);
spec1 = inv_logit(b[4]);
spec2 = inv_logit(b[5]);
jspec = inv_logit(b[6]);
AND = jsens;
OR = sens1 + sens2 - jsens;
vector[Ns] log_lik;
for (n in 1:Ns) {
  //monitor log-likelihood to calculate WAIC post-estimation
  log_lik[n] = Trivfcop_lpmf(x1[n,1:3] | disease[n], theta1[n],
                             mu[n,1:3]);}}
```

D.3 Convergence diagnostics for the trivariate Frank copula meta-analysis model

High sensitivities and specificities with moderate within-study associations

Figure D.1: Trace plots obtained from the trivariate Frank copula meta-analysis comparing the diagnostic accuracy of two tests. Sensitivities and specificities are high and within-study associations are moderate.

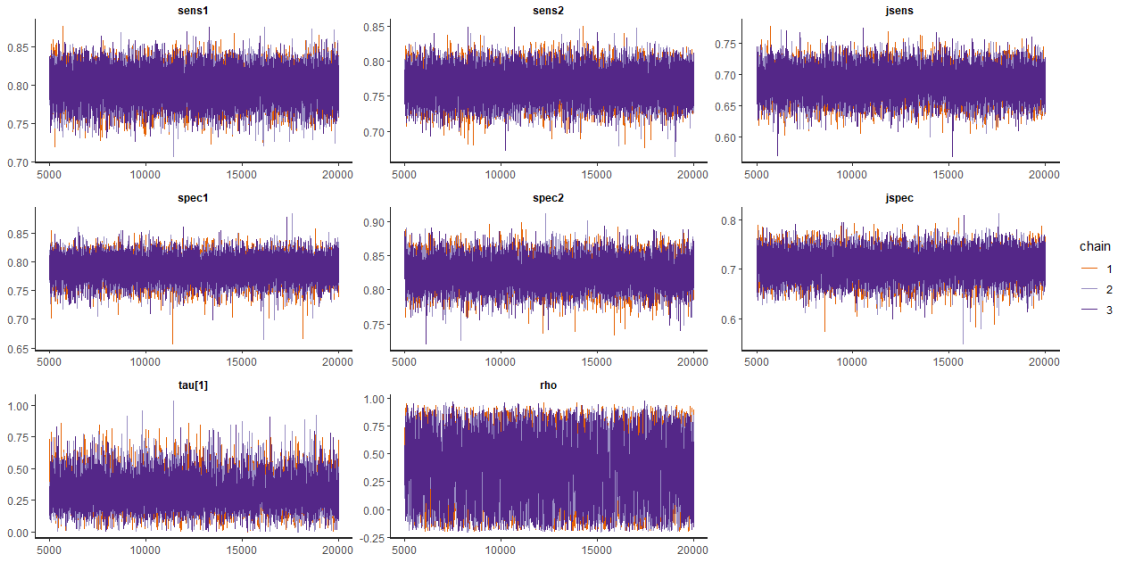


Figure D.2: Density plots obtained from the trivariate Frank copula meta-analysis comparing the diagnostic accuracy of two tests. Sensitivities and specificities are high and within-study associations are moderate.

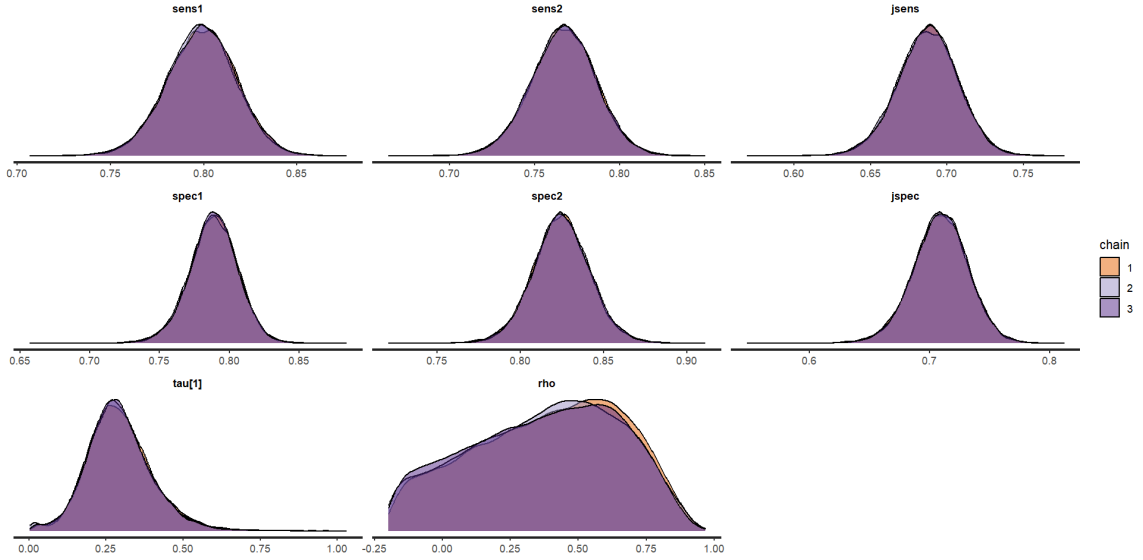
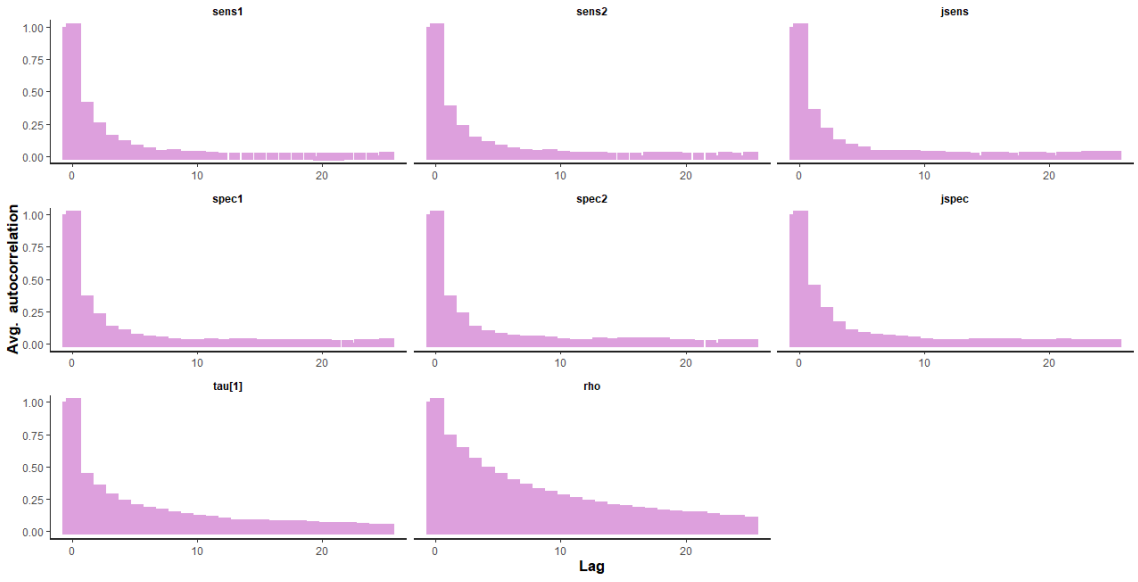


Figure D.3: Autocorrelation plots obtained from the trivariate Frank copula meta-analysis comparing the diagnostic accuracy of two tests. Sensitivities and specificities are high and within-study associations are moderate. Autocorrelation is averaged over the three chains.



Low sensitivities and specificities with moderate within-study associations

Figure D.4: Trace plots obtained from the trivariate Frank copula meta-analysis comparing the diagnostic accuracy of two tests. Sensitivities and specificities are low and within-study associations are moderate.

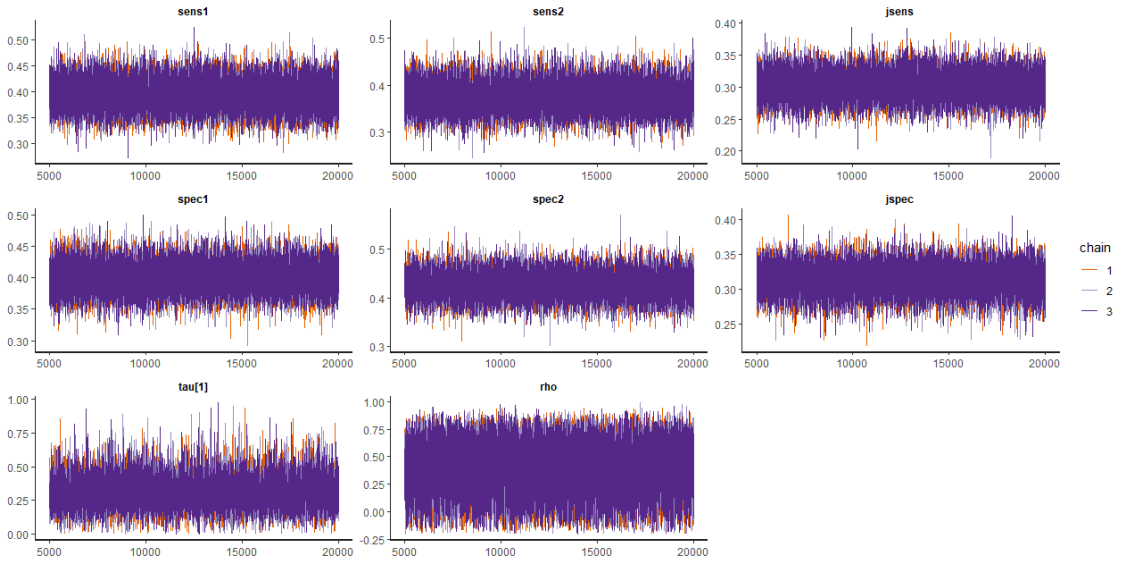


Figure D.5: Density plots obtained from the trivariate Frank copula meta-analysis comparing the diagnostic accuracy of two tests. Sensitivities and specificities are low and within-study associations are moderate.

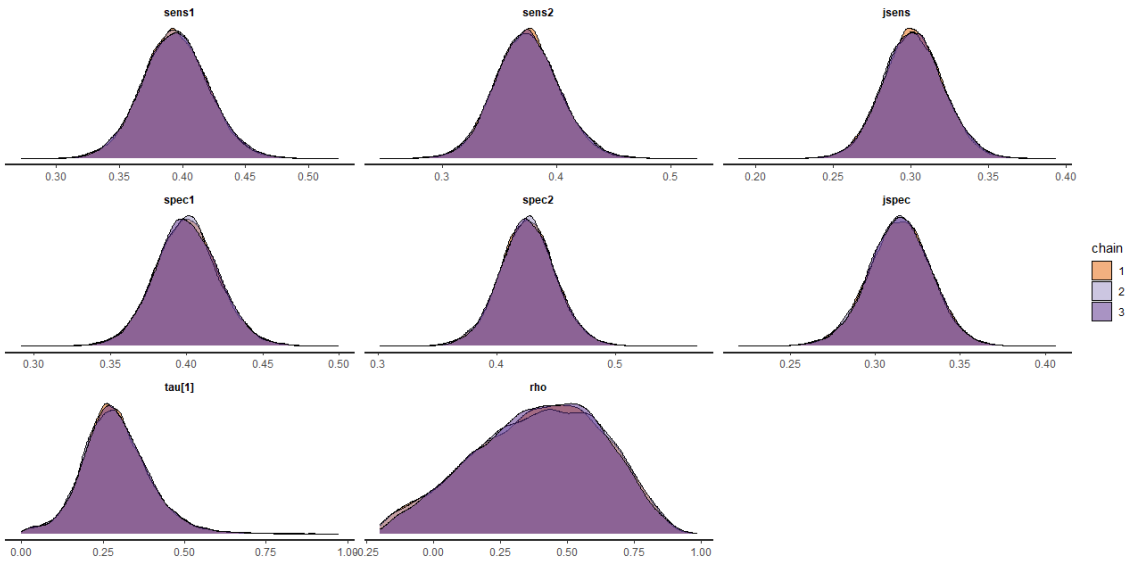
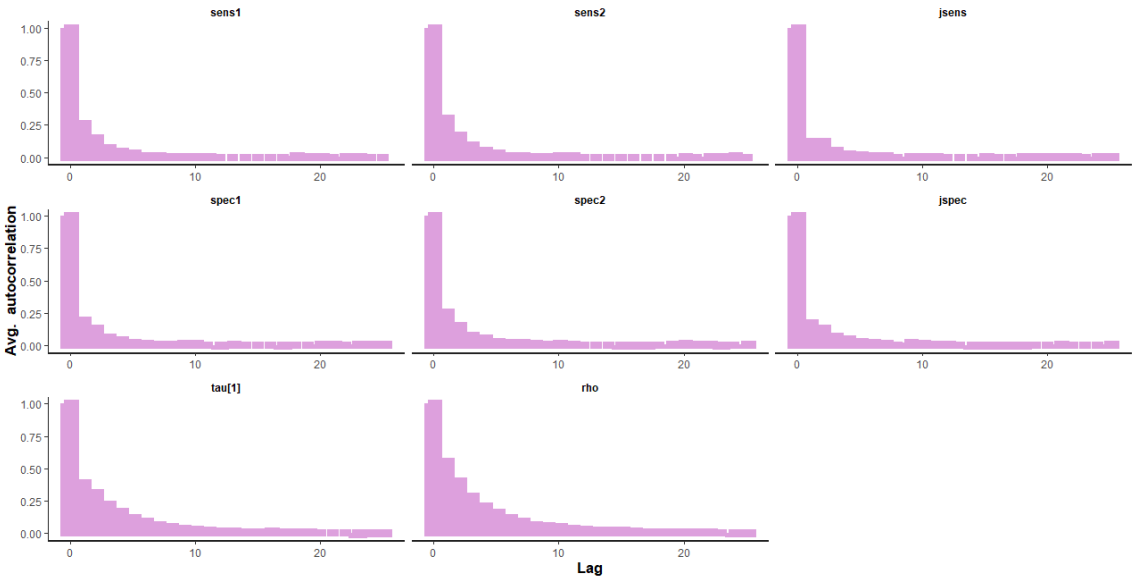


Figure D.6: Autocorrelation plots obtained from the trivariate Frank copula meta-analysis comparing the diagnostic accuracy of two tests. Sensitivities and specificities are low and within-study associations are moderate. Autocorrelation is averaged over the three chains.



D.4 Convergence diagnostics for the trivariate Gumbel copula meta-analysis model

High sensitivities and specificities with moderate within-study associations

Figure D.7: Trace plots obtained from the trivariate Gumbel copula meta-analysis comparing the diagnostic accuracy of two tests. Sensitivities and specificities are high and within-study associations are moderate.

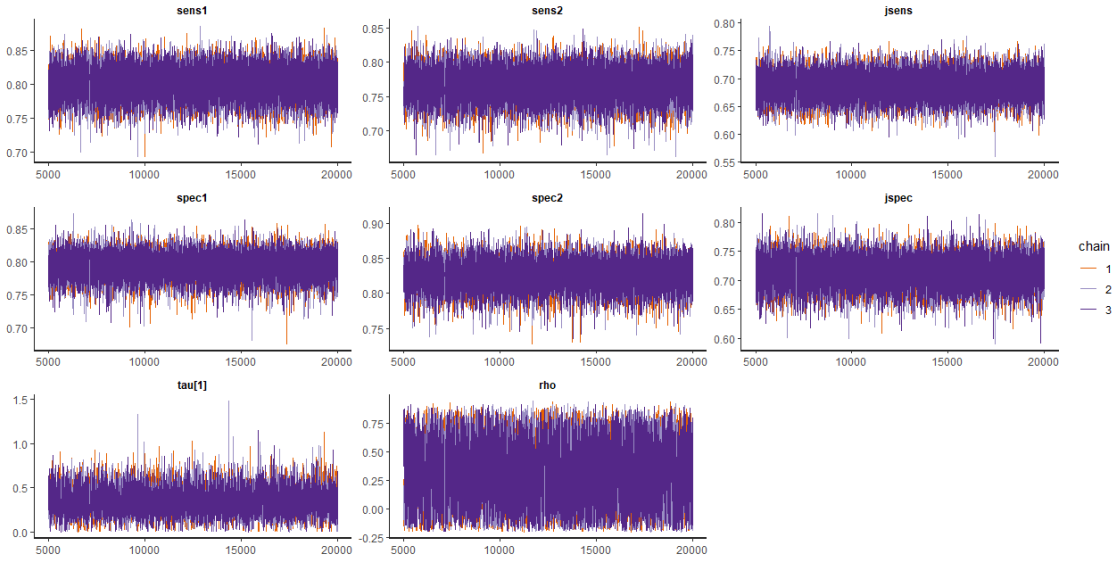


Figure D.8: Density plots obtained from the trivariate Gumbel copula meta-analysis comparing the diagnostic accuracy of two tests. Sensitivities and specificities are high and within-study associations are moderate.

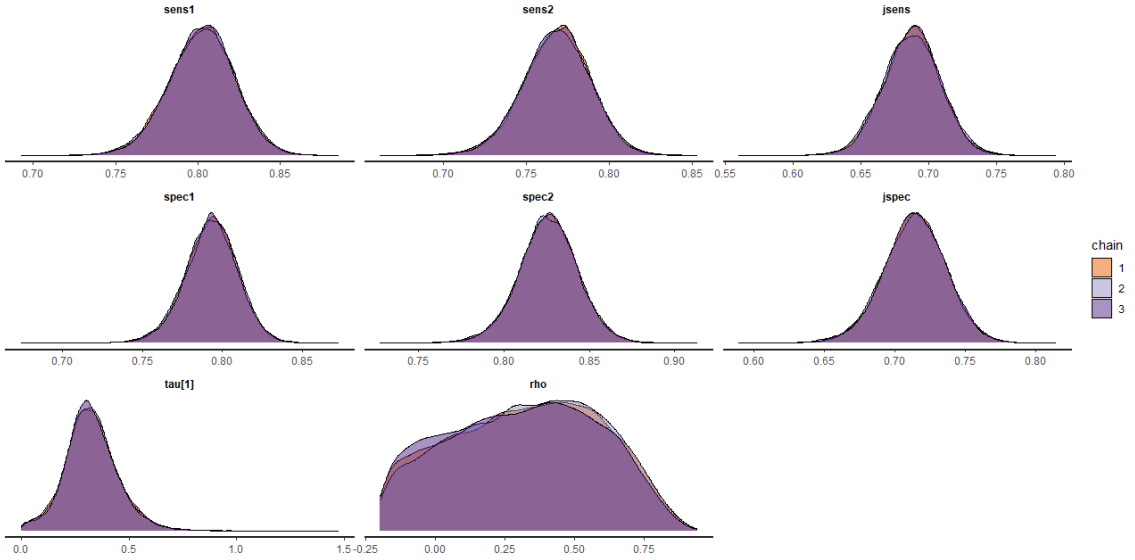
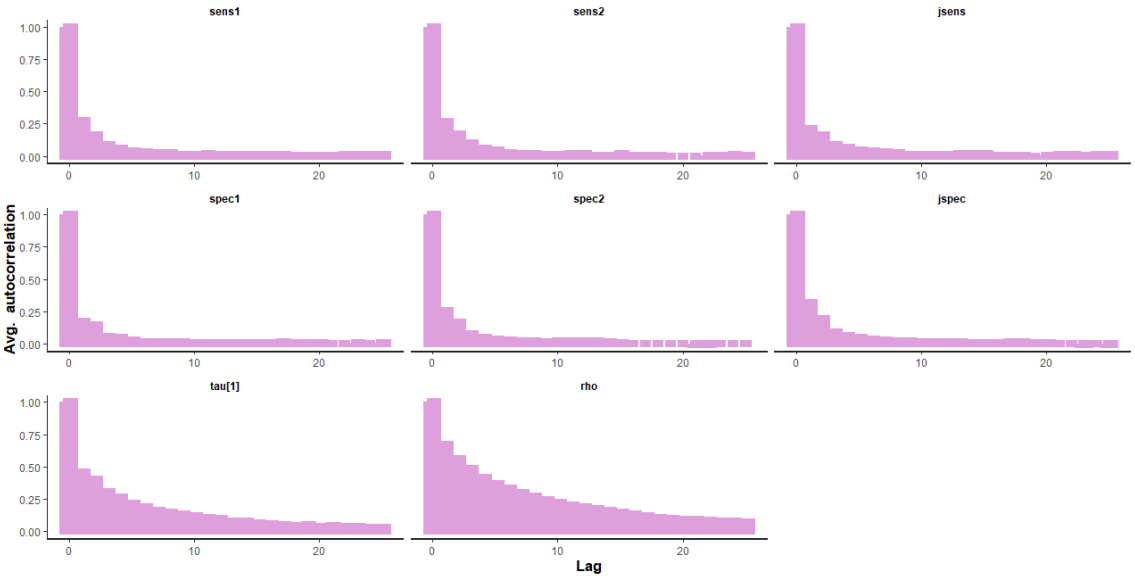


Figure D.9: Autocorrelation plots obtained from the trivariate Gumbel copula meta-analysis comparing the diagnostic accuracy of two tests. Sensitivities and specificities are high and within-study associations are moderate. Autocorrelation is averaged over the three chains.



Low sensitivities and specificities with moderate within-study associations

Figure D.10: Trace plots obtained from the trivariate Gumbel copula meta-analysis comparing the diagnostic accuracy of two tests. Sensitivities and specificities are low and within-study associations are moderate.

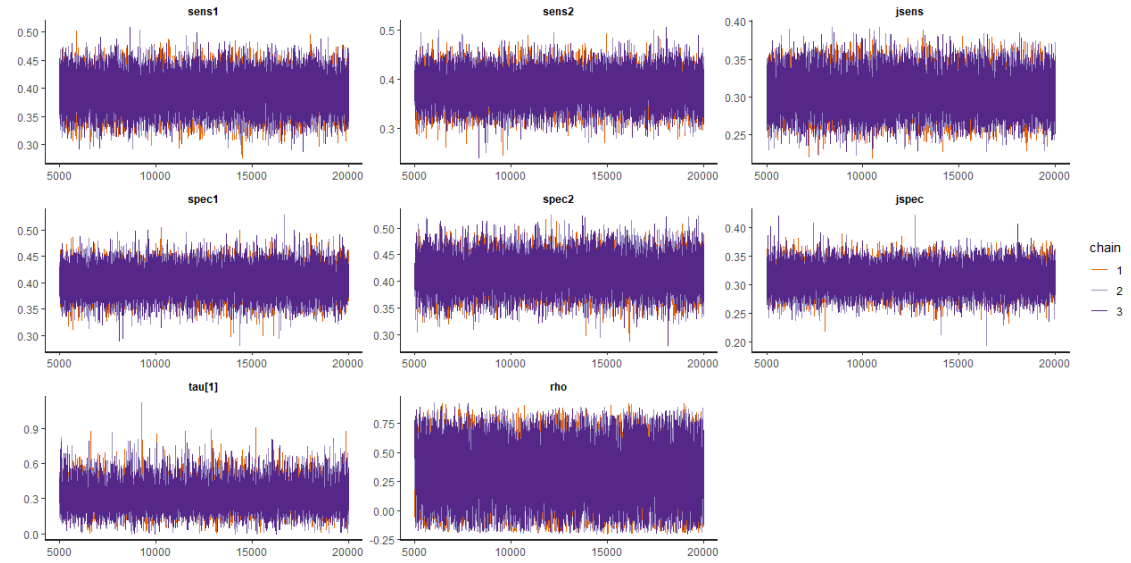


Figure D.11: Density plots obtained from the trivariate Gumbel copula meta-analysis comparing the diagnostic accuracy of two tests. Sensitivities and specificities are low and within-study associations are moderate.

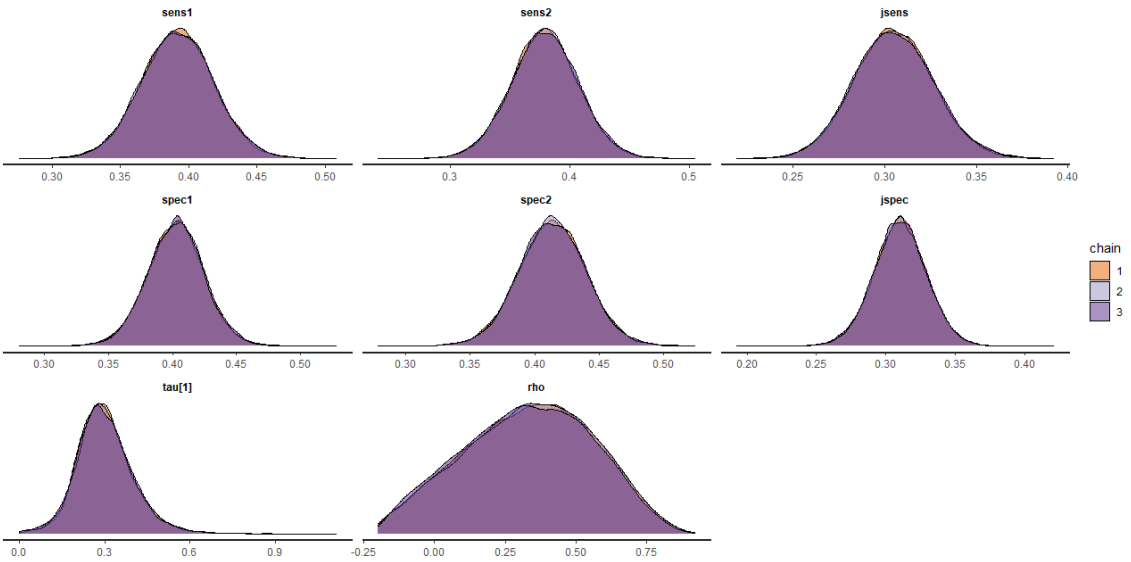
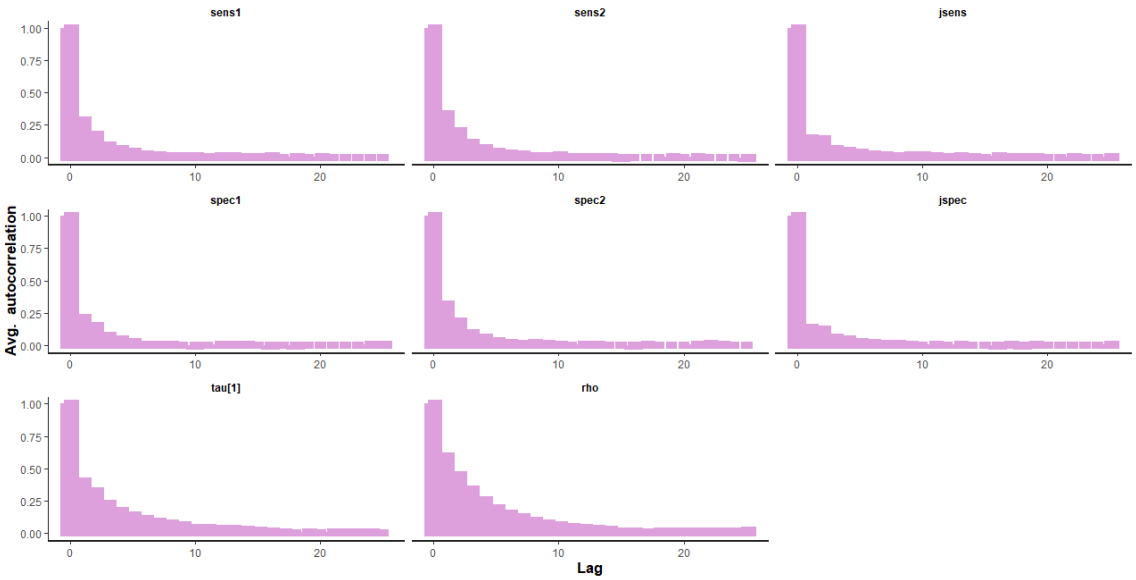


Figure D.12: Autocorrelation plots obtained from the trivariate Gumbel copula meta-analysis comparing the diagnostic accuracy of two tests. Sensitivities and specificities are low and within-study associations are moderate. Autocorrelation is averaged over the three chains.



D.5 Convergence diagnostics for the trivariate Clayton copula meta-analysis model

High sensitivities and specificities with moderate within-study associations

Figure D.13: Trace plots obtained from the trivariate Clayton copula meta-analysis comparing the diagnostic accuracy of two tests. Sensitivities and specificities are high and within-study associations are moderate.

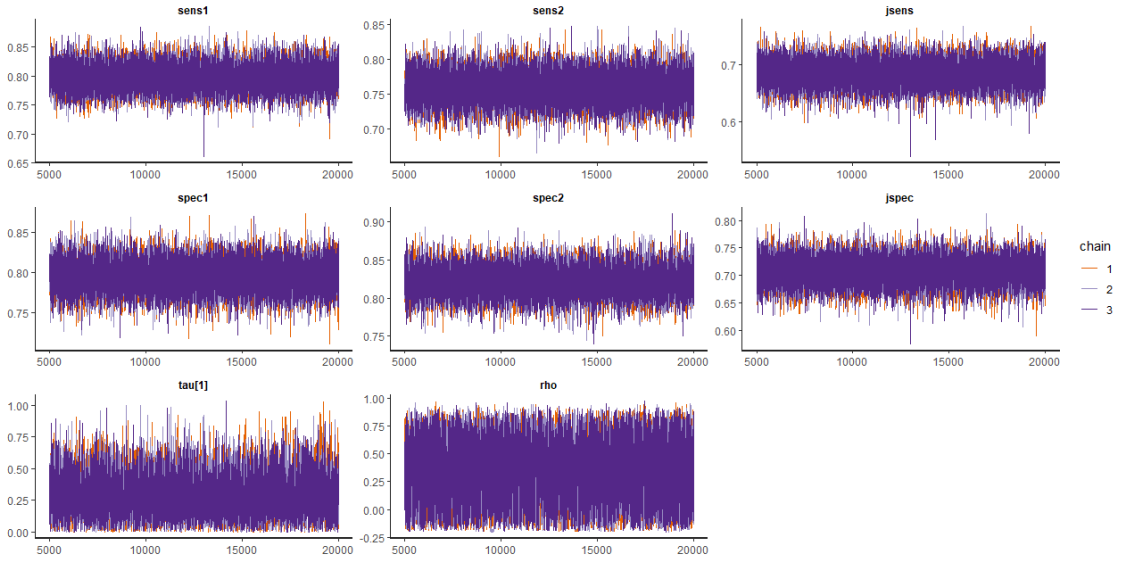


Figure D.14: Density plots obtained from the trivariate Clayton copula meta-analysis comparing the diagnostic accuracy of two tests. Sensitivities and specificities are high and within-study associations are moderate.

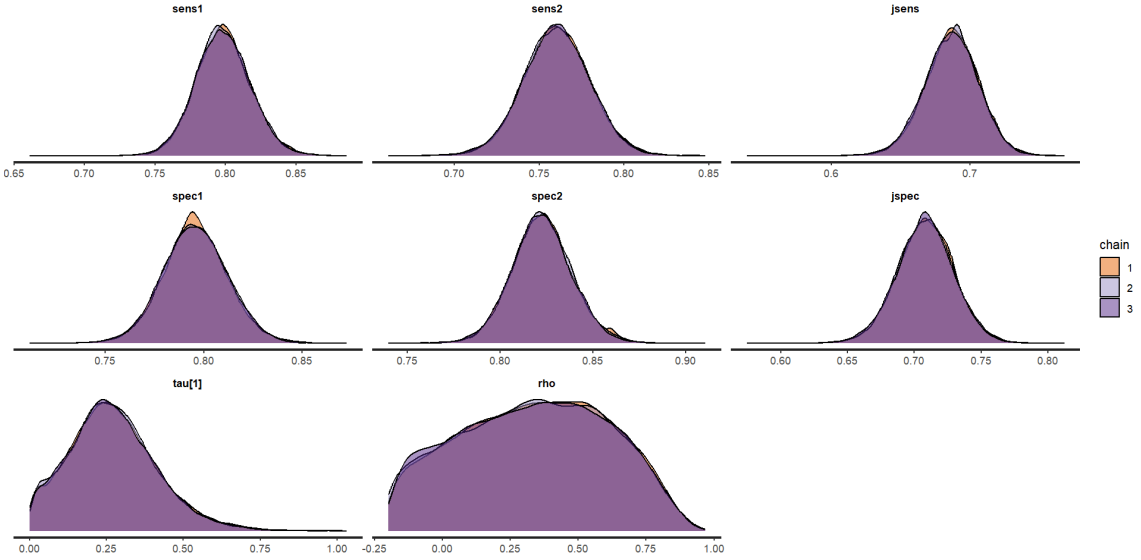
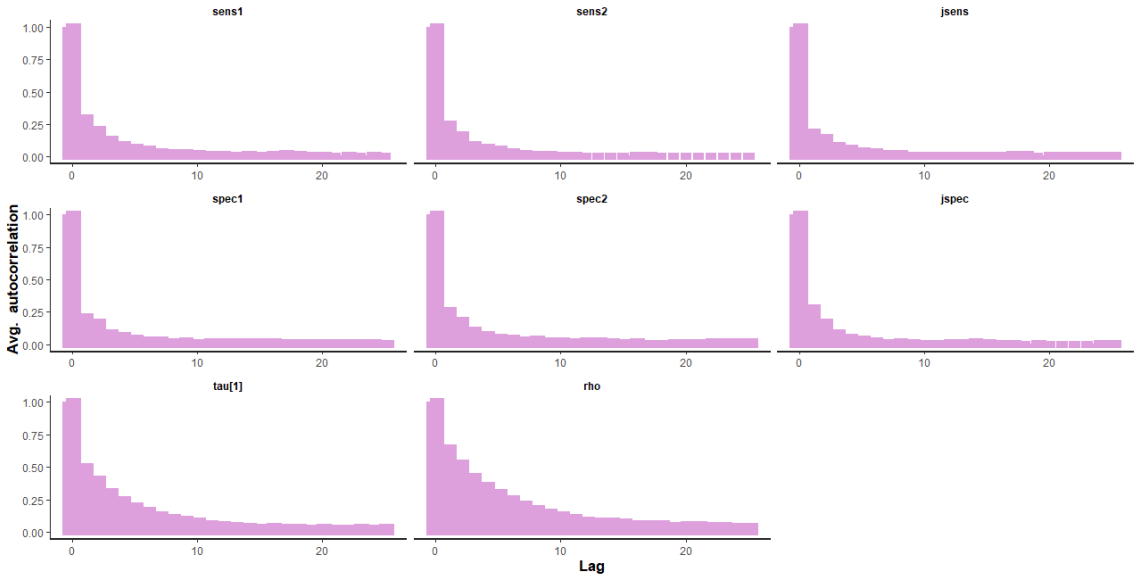


Figure D.15: Autocorrelation plots obtained from the trivariate Clayton copula meta-analysis comparing the diagnostic accuracy of two tests. Sensitivities and specificities are high and within-study associations are moderate. Autocorrelation is averaged over the three chains.



Low sensitivities and specificities with moderate within-study associations

Figure D.16: Trace plots obtained from the trivariate Clayton copula meta-analysis comparing the diagnostic accuracy of two tests. Sensitivities and specificities are low and within-study associations are moderate.

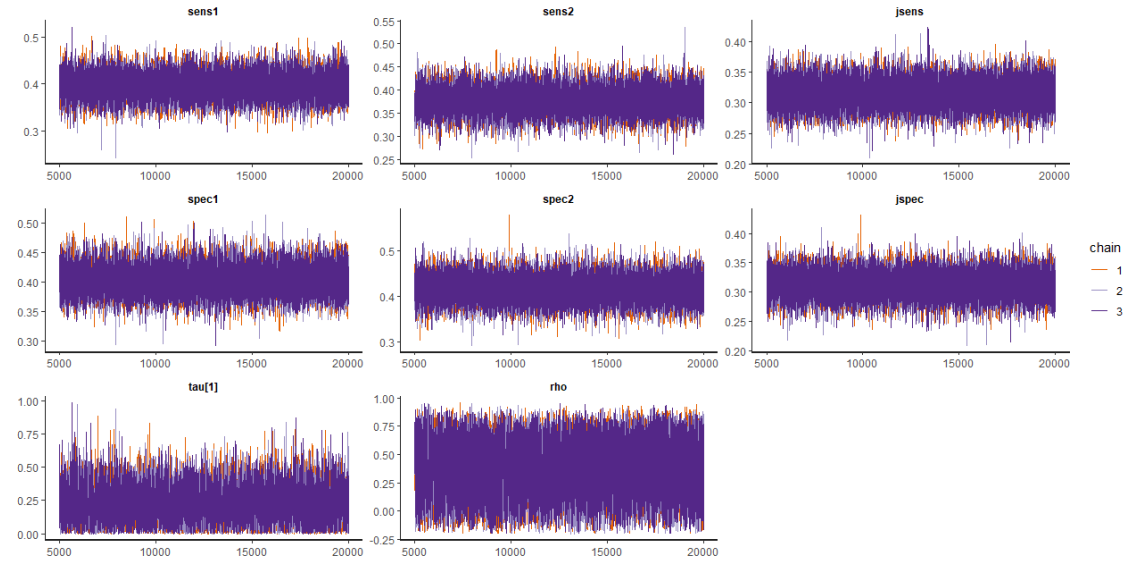


Figure D.17: Density plots obtained from the trivariate Clayton copula meta-analysis comparing the diagnostic accuracy of two tests. Sensitivities and specificities are low and within-study associations are moderate.

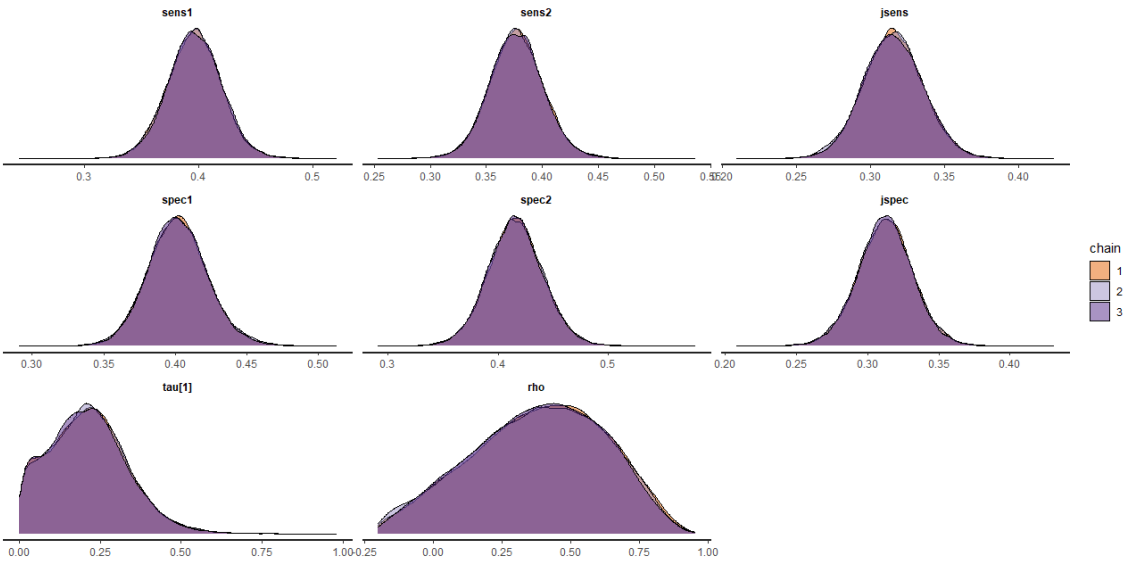
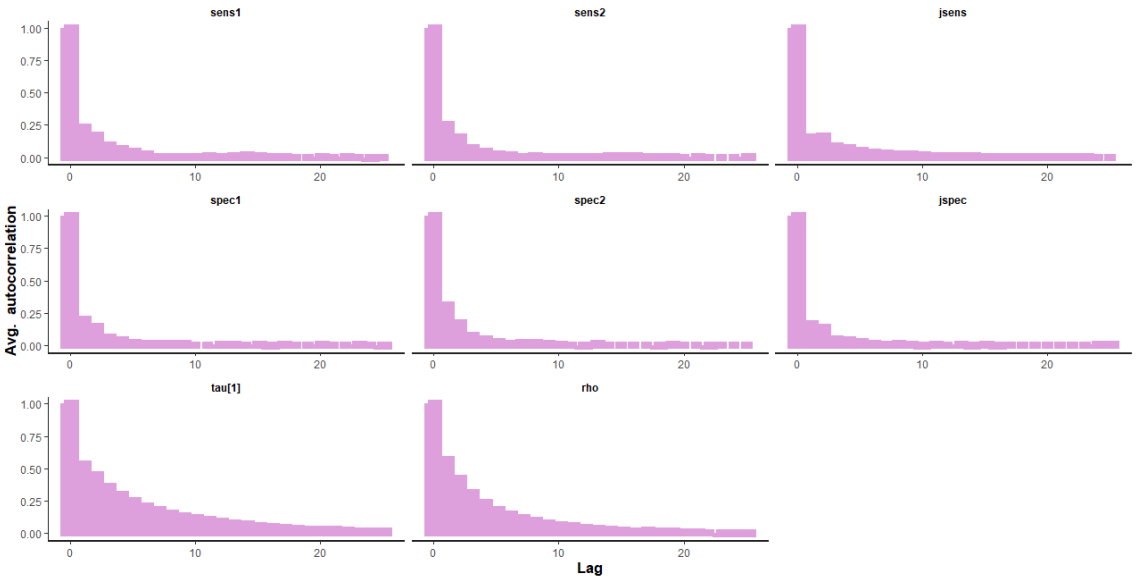


Figure D.18: Autocorrelation plots obtained from the trivariate Clayton copula meta-analysis comparing the diagnostic accuracy of two tests. Sensitivities and specificities are low and within-study associations are moderate. Autocorrelation is averaged over the three chains.



Appendix E

E.1 BMC Medical Research Methodology Paper

Manuscript currently under review at BMC Medical Research Methodology.

Title:

Joint meta-analysis of two diagnostic tests using bivariate copulas to model within-study dependencies

Author's names:

Athena L Sheppard[1,2] (ORCID: 0000-0003-1564-0740), Tasos Papanikos[3] (ORCID: 0000-0001-8971-6221), Terence J Quinn[4] (ORCID: 0000-0003-1401-0181), Keith R Abrams[5,6] (ORCID: 0000-0002-7557-1567), Sylwia Bujkiewicz[2] (ORCID: 0000-0002-3003-9403), Rhiannon K Owen[1] (ORCID: 0000-0001-5977-376X)

Author's affiliations:

1. Population Data Science, Swansea University Medical School, Swansea University, Swansea, UK
2. Biostatistics Research Group, Department of Population Health Sciences, University of Leicester, Leicester, UK
3. GlaxoSmithKline R&D Centre, GlaxoSmithKline, Stevenage, UK
4. School of Cardiovascular & Metabolic Health, University of Glasgow, Glasgow, UK
5. Department of Statistics & Warwick Medical School (WMS), University of Warwick, Coventry, UK
6. Centre for Health Economics, University of York, York, UK

Author's email addresses:

ALS: athena.sheppard@swansea.ac.uk; TP: anastasios.x.papanikos@gsk.com; TJQ: terry.quinn@glasgow.ac.uk; KRA: keith.abrams@warwick.ac.uk; SB: sylwia.bujkiewicz@leicester.ac.uk; RKO: r.k.owen@swansea.ac.uk

Corresponding author:

Athena L Sheppard

Swansea University Medical School

Swansea University

Swansea, UK

Email: athena.sheppard@swansea.ac.uk

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

All data generated or analysed during this study are included in this published article. All statistical code is included within the additional files and is available from GitHub (https://github.com/athenasheppard/bivariate-copula).

Competing interests

ALS declares that she has no competing interests.

TP is an employee of GlaxoSmithKline (GSK).

TJQ was coordinating editor of Cochrane Dementia. He has received grant funding from Chief Scientist Office, and Stroke Association for research around test accuracy in dementia.

KRA is a member of the National Institute for Health and Care Excellence (NICE) Diagnostics Advisory Committee, the NICE Decision and Technical Support Units, and is a National Institute for Health Research (NIHR) Senior Investigator *Emeritus* [NF-SI-0512-10159]. He has served as a paid consultant, providing unrelated methodological and strategic advice, to the pharmaceutical and life sciences industry generally, as well as to DHSC/NICE, and has received unrelated research funding from Association of the British Pharmaceutical Industry (ABPI), European Federation of Pharmaceutical Industries & Associations (EFPIA), Pfizer, Sanofi and Swiss Precision Diagnostics/Clearblue. He has also received course fees from ABPI and the University of Bristol, and is a Partner and Director of Visible Analytics Limited, a health technology assessment consultancy company.

SB is a member of the NICE Decision Support Unit (DSU) and the NICE Guidelines Technical Support Unit (TSU). She has served as a paid consultant, providing methodological advice, to NICE, CROs and pharmaceutical industry, has received payments for educational events from Roche and the University of Bristol, funding to attend conferences from CROs and Roche, research funding from European Federation of Pharmaceutical Industries & Associations (EEPIA) and Johnson & Johnson and research support in kind from AstraZeneca and Roche.

RKO is a member of the National Institute for Health and Care Excellence (NICE) Technology Appraisal Committee, member of the NICE Decision Support Unit (DSU), and associate member of the NICE Technical Support Unit (TSU). RKO has served as a paid consultant to the pharmaceutical industry and international reimbursement agencies, providing unrelated methodological advice generally. She reports teaching fees from the Association of British Pharmaceutical Industry (ABPI) and the University of Bristol.

Funding

This research was funded by the Medical Research Council, Methodology Research Panel (grant no. MR/T025166/1) and Health Data Research UK, an initiative funded by UK Research and Innovation, Department of Health, and Social Care (England) and the devolved administrations, and leading

medical research charities. ALS was supported by Health Data Research UK (HDRUK Studentship (NIWA1)). SB was also supported by the NIHR Leicester Biomedical Centre (BRC). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. RKO is supported by a Springboard award (SBF006\1122) funded by the Academy of Medical Sciences, Wellcome Trust, Government Department of Business, Energy and Industrial Strategy, British Heart Foundation, and Diabetes UK. ALS and RKO are supported by the Health and Care Research Wales Evidence Centre.

Author contributions

ALS: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Vizualization, Project administration. TP: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing - Review & Editing, Vizualization. TJQ: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing - Review & Editing, Vizualization, Funding acquisition. KRA: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing - Review & Editing, Vizualization, Funding acquisition. SB: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing - Review & Editing, Visualization, Supervision, Funding acquisition. RKO: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing - Review & Editing, Visualization, Supervision, Funding acquisition.

Acknowledgements

Not applicable

Abstract

Background

Several meta-analysis models have recently been proposed to synthesise data on the accuracy of two diagnostic tests. Many of these approaches do not account for correlations between multiple tests within the same patient group owing to a lack of reporting of fully cross-classified data, akin to individual participant-level data, in comparative diagnostic accuracy studies. This paper describes a novel application of copula models to capture the within-study dependencies between two diagnostic tests evaluated in the same patient group, when cross-classified data are not available for all studies.

Methods

We developed Bayesian meta-analysis models for evaluating the accuracy of two diagnostic tests in the same patients, using bivariate copulas to flexibly capture within-study dependencies between the two tests. Five bivariate copula models capturing different relationships between the joint sensitivities and specificities of the two tests were described: Gaussian, Frank, Gumbel, Clayton and Clayton rotated 180°. The models were compared to the currently recommended meta-regression approach for modelling two tests through their application to a motivating example in Alzheimer's disease dementia.

Results

The diagnostic accuracy of two cerebrospinal fluid biomarker tests, amyloid-β 42 ($A\beta_{42}$) and total tau (t-tau), for Alzheimer's disease dementia were compared. $A\beta_{42}$ and t-tau demonstrated sensitivities of 80.9% (95% credible interval: 73.4%, 87.5%) and 76.4% (69.4%, 83.1%), respectively. Summary specificity was 70.3% (61.3%, 78.4%) and 72.5% (63.7%, 81.3%), respectively. There was strong evidence that the bivariate copula models resulted in a better fit compared to the meta-regression model. The bivariate copula framework resulted in similar estimates of summary sensitivities and

specificities to the meta-regression model, but increased precision around estimates by as much as 15%.

Conclusions

The bivariate copula approach led to improved model fit compared to the current meta-regression method, which ignores associations between tests. This novel methodological development is applicable to a broad range of disease areas. It aids healthcare decision-making by allowing the comparison of two tests for the same condition while accounting for complex dependence structures arising between multiple tests. The models relax the need for cross-classified diagnostic accuracy data, making better use of the available evidence base.

Keywords

3-10 keywords: diagnostic test accuracy; meta-analysis; test comparison; health technology assessment; copula; Bayesian analysis

List of abbreviations

Aβ, amyloid-β; BRMA, bivariate random effects meta-analysis; CDF, cumulate distribution function; CrI, credible interval; CSF, cerebrospinal fluid; HTA, health technology assessment; IPD, individual participant data; NICE, National Institute for Health and Care Excellence; RCT, randomised controlled trial; STARD, Standards for Reporting of Diagnostic Accuracy Studies; TSD, Technical Support Document; t-tau, total tau; WAIC, widely applicable information criterion

1. Introduction

There is an increasing availability of diagnostic tests in contemporary healthcare, with clinicians wishing to know which of the available options is 'best'. While a diagnostic test can be comprehensively evaluated in isolation, most clinically relevant questions are comparative in nature.[1] Whether a novel test is sufficiently accurate to use in practice depends on the accuracy of similar, existing tests for the same condition.[2] Furthermore, joint analysis of two or more diagnostic tests is vital to assess the accuracy of different testing strategies and diagnostic pathways, which often comprise of multiple components.

Meta-analysis methods allow the integration of findings from multiple studies, producing quantitative summaries of often-variable study results. In the context of health technology assessment (HTA), meta-analysis is a powerful tool for informing evidence-based practice.[3-4] In a meta-analysis of multiple diagnostic tests there are two sources of association – at the between-study level and at the within-study level. Diagnostic accuracy studies most commonly report pairs of sensitivity (the proportion of participants with the target condition that are correctly identified by the test) and specificity (the proportion of participants without the target condition that are correctly identified by the test).[5] When diagnostic accuracy studies are combined in a meta-analysis, heterogeneity in sensitivity and specificity often arises from between-study variation in patient- or study-level characteristics, such as the measurement threshold to determine test-positivity. Such between-study heterogeneity for both parameters generates negative association between sensitivities and specificities, known as between-studies correlation. Test accuracy may be underestimated if such correlation is not adequately captured.[6] When inference is focussed on point estimates of sensitivity and specificity, this correlation is often modelled using a multivariate normal distribution.[7-8]

Unlike intervention studies, in which patients are frequently randomised to independent treatment groups, diagnostic accuracy studies comparing two index tests often evaluate both tests in the same patient group compared to a reference standard. Under this paired design, within-study dependencies

arise between sensitivities and specificities of the two tests. Cross-classified tables for each study – containing all possible combinations of test results compared to true disease status, equivalent to individual participant data (IPD) – are typically required to account for within-study associations between multiple tests. At present, there is a lack of clear guidance on how to best analyse data from comparative diagnostic studies.[9] Several meta-analysis models for evaluating two or more diagnostic tests have been proposed in recent years.[10-21] Whilst the existing models adequately capture between-studies association, within-study association is often ignored. Models that account for within-study dependencies often make restrictive assumptions about the distributions of the marginal variables, and result in slow convergence and long computation times.[9]

Copula theory is used to flexibly capture dependence between two (or more) random variables.[22] Copulas have previously been used to model between-studies correlation between sensitivity and specificity in a meta-analytic framework. For a single test, Kuss et al [23] and Nyaga et al [24] proposed using beta-binomial marginal distributions, linked by various coupling functions, as a natural choice to model sensitivity and specificity, which are bound between 0 and 1. Nyaga et al [13] extended their previous work to evaluate the accuracy of two tests, while Cheng [15] suggested a similar extension to Kuss et al's model. Hoyer and Kuss,[17] and later Nikoloulopoulos,[21] described four-dimensional copula models to capture between-studies dependence between the sensitivities and specificities. Within the context of surrogate endpoint meta-analysis, copulas have been used to model within-study association; either between the surrogate endpoint and the final clinical outcome in a meta-analysis of randomised controlled trials (RCTs) with time-to-event IPD by Burzykowski et al [25] or, as proposed by Papanikos et al, between the treatment effects on the surrogate endpoint and the final clinical outcome within each study in an aggregate-level meta-analysis of RCTs with binary outcomes.[26] The application of copula models offers a potentially novel solution to capture within-study dependencies arising between two tests assessed in the same patients.

Building on the model proposed by Papanikos et al in the field of surrogate endpoint evaluation and the meta-analytic models for diagnostic test accuracy,[24,26] we propose Bayesian meta-analysis models for evaluating the accuracy of two diagnostic tests, using bivariate copulas to capture within-study dependencies between the tests. Section 2 introduces the motivational data example in Alzheimer's disease dementia that the developed methodology is applied to. Section 3 describes the existing and proposed methodology for the meta-analysis of two tests, as well as providing a short overview of copula theory. The application of the methods is demonstrated in Section 4, where the results of fitting both the new models and the currently recommended meta-regression approach to the motivational example are presented and compared. Section 5 concludes the article with a discussion.

2. Motivational example

Dementia is the leading cause of disability and dependency in older adults, and is recognised as a global research priority.[27] It is characterised by progressive impairment in cognitive function beyond normal ageing, describing a range of cognitive, psychological and behavioural symptoms affecting memory, thinking and activities of daily living. Dementia is thought to progress through a series of stages, from biologically active but clinically silent, to impairments of memory and thinking that are not sufficient to interfere with daily activity (mild cognitive impairment), to overt dementia. In 2019, it was estimated that 55 million people globally were living with dementia, with the figure due to rise to 139 million by 2050. Despite this, it is thought that 75% of people with dementia have not received a diagnosis.[28] International healthcare systems are looking to redesign their diagnostic pathways for dementia, and new technologies such as biomarkers are expanding the diagnostic options available.[28]

Alzheimer's disease is the most common cause of dementia, attributable in 60-70% of cases.[27] Alzheimer's disease dementia is the result of the accumulation of abnormal protein structures in the brain – such as amyloid plaques or tau tangles – that disrupt the connection between nerve cells.[29] Alzheimer's disease dementia is diagnosed through a combination of cognitive, imaging and biomarker testing. Diagnostic rates are low and misdiagnosis is common,[28] impeding management of the condition, development of targeted treatments, and impacting quality of life for people with dementia, carers and families. Optimising the diagnostic pathway for Alzheimer's disease dementia requires comparison of the available tests using appropriate statistical methods.

Data on which we base the methodological development described in this paper were collected through a review of systematic reviews of diagnostic test accuracy studies for Alzheimer's disease dementia published in The Cochrane Library. The Cochrane Library was searched up to August 2023 for articles containing 'Alzheimer's disease' within the title, abstract or keywords, filtered by diagnostic test accuracy reviews. Diagnostic accuracy studies included within the reviews were used as a source of comparative accuracy data. Where available, 2x2 and/or 2x4 (cross-classified, see Section 3.1) data on the accuracy of two or more tests assessed in the same patient group were extracted from the primary studies. The reviews evaluated the accuracy of cognitive, imaging and cerebrospinal fluid (CSF) biomarker tests for Alzheimer's disease dementia, recruiting patients from community, primary and secondary care settings.

Seventy comparative diagnostic studies were identified within 13 relevant systematic reviews published in The Cochrane Library. In total, 251 2x2 tables were extracted, of which cross-classified data were reported in 14 (20%). From the total study pool, two tests were selected that maximised the availability of comparative data. These were amyloid-β 42 ($A\beta_{42}$) and total tau (t-tau), measured in the CSF, on which 36 2x2 tables from 18 studies were extracted. Two studies contained cross-classified data and the remaining 16 reported 2x2 data only (Table 1).

Table 1: Diagnostic accuracy data of $A\beta_{42}$ (test 1) and t-tau (test 2) for Alzheimer's disease dementia.

Study	Alzheimer's disease dementia									No Alzheimer's disease dementia								
	tp_1	tp_2	fn_1	fn_2	x_{11}^D	x_{10}^D	x_{01}^D	x_{00}^D	N^D	tn_1	tn_2	fp_1	fp_2	$x_{11}^{\bar{D}}$	$x_{10}^{\bar{D}}$	$x_{01}^{\bar{D}}$	$x_{00}^{\bar{D}}$	$N^{\bar{D}}$
Bjerke et al, 2009.[30]	18	12	2	8	-	-	-	-	20	99	130	43	12	-	-	-	-	142
Blom et al, 2009.[31]	9	7	5	7	-	-	-	-	14	5	11	9	3	-	-	-	-	14
Chiasserini et al, 2010.[32]	17	12	6	11	-	-	-	-	23	16	16	2	2	-	-	-	-	18
Frölich et al, 2017.[33]	17	24	11	4	-	-	-	-	28	65	46	22	41	-	-	-	-	87
Gaser et al, 2013.[34]	59	58	7	8	-	-	-	-	66	12	13	21	20	-	-	-	-	33
Hampel et al, 2004.[35]	24	26	5	3	-	-	-	-	29	13	11	10	12	-	-	-	-	23
Hansson et al, 2006.[36]	56	55	1	2	54	2	1	0	57	50	47	27	30	13	14	17	33	77
Herukka et al, 2008.[37]	6	6	2	2	-	-	-	-	8	11	7	2	6	-	-	-	-	13
Hertze et al, 2010.[38]	47	38	5	14	46	1	4	1	52	75	82	32	25	20	12	46	29	107
Kester et al, 2011.[39]	32	35	10	7	-	-	-	-	42	42	29	16	29	-	-	-	-	58
Monge-Argilés et al, 2011.[40]	9	8	2	3	-	-	-	-	11	16	18	10	8	-	-	-	-	26
Nesteruk et al, 2016.[41]	7	6	2	3	-	-	-	-	9	18	20	13	11	-	-	-	-	31
Palmqvist et al, 2012.[42]	47	42	5	10	-	-	-	-	52	56	58	25	23	-	-	-	-	81
Parnetti et al, 2006.[43]	4	5	7	6	-	-	-	-	11	30	32	3	1	-	-	-	-	33
Parnetti et al, 2012.[44]	18	20	14	12	-	-	-	-	32	56	51	2	7	-	-	-	-	58
Prestia et al, 2013.[45]	17	11	1	7	-	-	-	-	18	9	15	9	3	-	-	-	-	18
Rhodius-Meester et al, 2016.[46]	58	72	27	13	-	-	-	-	85	38	35	14	17	-	-	-	-	52
Vos et al, 2013.[47]	62	65	29	26	-	-	-	-	91	82	95	41	28	-	-	-	-	123

The table contains 2x2 data for each test within each study, and fully cross-classified data where available.
tp_j is the number of true positives, fn_j the number of false negatives, tn_j the number of true negatives, and fp_j the number of false positives for each of the $j = 1,2$ tests.
x_{kl}^D is the number of participants with the target condition with each combination of test results. $x_{kl}^{\bar{D}}$ is the number of participants with the target condition with each combination of test results. $k, l = 0,1$ denote the first and second test, respectively, where 0 indicates a negative test result and 1 indicates a positive test result.
N^D and $N^{\bar{D}}$ are the number of patients with and without disease, respectively.
Aβ42, amyloid-β 42; t-tau, total tau

3. Methods

3.1 Diagnostic accuracy studies of two tests compared to a common reference standard

Comparative accuracy studies evaluate the accuracy of two diagnostic tests against true disease status, inferred through a common reference standard, allowing direct comparison of their sensitivity and specificity in a single group of participants. Non-comparative studies that evaluate a single test may be included in a comparative meta-analysis to increase the study pool, but have been shown to introduce bias to summary estimates.[48]

Results of studies that utilise this paired design are most commonly reported as aggregate, 2x2 tables of the results of each test compared to the reference standard, containing the number of true positives (tp_j), false negatives (fn_j), true negatives (tn_j) and false positives (fp_j) for each of the $j = 1,2$ tests. The number of patients with and without the target condition are denoted N^D and $N^{\bar{D}}$, respectively (Table 2).

Table 2: Two-by-two diagnostic accuracy data on two tests for a single study.

	Diseased	**Non-diseased**
Test 1 +	tp_1	fp_1
Test 1 −	fn_1	tn_1
Total	N^D	$N^{\bar{D}}$

	Diseased	**Non-diseased**
Test 2 +	tp_2	fp_2
Test 2 −	fn_2	tn_2
Total	N^D	$N^{\bar{D}}$

tp_j, true positives; fn_j, false negatives; tn_j, true negatives; fp_j, false positives; $j = 1,2$ tests; N^D, total diseased; $N^{\bar{D}}$, total non-diseased

Less frequently, joint cross-classifications of test results compared to a common reference standard may be reported, producing a 2x4 table from which it is possible to reconstruct IPD (Table 3). The number of patients with each combination of test results with the target condition (x_{kl}^D) and without the target condition ($x_{kl}^{\bar{D}}$) are presented. $k, l = 0,1$ denote the first and second test, respectively, where 0 indicates a negative test result and 1 indicates a positive test result. Cross-classified data

allow estimation of within-study dependencies and evaluation of the accuracy of the tests in combination or sequence, although in practice they are often not reported.[49-50]

Table 3: Fully cross-classified diagnostic accuracy data for two tests for a single study.

	Diseased			**Non-diseased**		
	Test 2 +	**Test 2 −**	**Total**	**Test 2 +**	**Test 2 −**	**Total**
Test 1 +	x_{11}^{D}	x_{10}^{D}	tp_1	$x_{11}^{\bar{D}}$	$x_{10}^{\bar{D}}$	fp_1
Test 1 −	x_{01}^{D}	x_{00}^{D}	fn_1	$x_{01}^{\bar{D}}$	$x_{00}^{\bar{D}}$	tn_1
Total	tp_2	fn_2	N^{D}	fp_2	tn_2	$N^{\bar{D}}$

tp_j, true positives; fn_j, false negatives; tn_j, true negatives; fp_j, false positives; $j = 1,2$ tests; x_{kl}^{D}, number of participants with the target condition with each combination of test results; $x_{kl}^{\bar{D}}$, number of participants without the target condition with each combination of test results; $k, l = 0,1$, denote the first and second test, respectively, where 0 indicates a negative test result and 1 indicates a positive test result; N^{D}, total diseased; $N^{\bar{D}}$, total non-diseased

3.2 Bivariate random effects meta-analysis

Reitsma et al proposed a bivariate meta-analysis approach to summarise the accuracy of a single test at a common threshold, assuming $logit$-transformed sensitivity and specificity are normally distributed around a mean value.[7] Chu and Cole suggested that using exact binomial likelihoods are a more natural choice to model within-study variability in sensitivity and specificity,[8] a correction that is widely accepted and referred to throughout this article as the bivariate random effects meta-analysis (BRMA) model. This method forms the basis for most meta-analysis models for synthesising data on two diagnostic tests, including the novel bivariate copula models proposed in Section 3.4.

At the within-study level, the number of true positive (tp_i) and true negative (tn_i) results for study $i = 1, \ldots, I$ are assumed to follow independent binomial distributions. These counts are modelled independently as the populations of diseased and non-diseased are mutually exclusive:

$$tp_i \sim Binomial\left(se_i, N_i^{D}\right), \quad (1)$$

$$tn_i \sim Binomial\left(sp_i, N_i^{\bar{D}}\right)$$

where se_i and sp_i denote the sensitivity and specificity in the i^{th} study, respectively, and N_i^D and $N_i^{\bar{D}}$ the number of patients with and without the target condition, respectively. At the between-studies level, *logit*-transformed study-specific sensitivities ($\mu_{i,se}$) and specificities ($\mu_{i,sp}$) are jointly modelled using a bivariate normal distribution centred around *logit*-transformed summary sensitivity and specificity, μ_{se} and μ_{sp}, accounting for between-studies correlation arising between sensitivity and specificity due to differences in study characteristics. Between-studies variances are denoted σ_{se}^2 and σ_{sp}^2, respectively, while ρ_b represents the between-studies correlation parameter.

$$logit(se_i) = \mu_{i,se}, \quad logit(sp_i) = \mu_{i,sp}, \tag{2}$$

$$\begin{pmatrix}\mu_{i,se}\\ \mu_{i,sp}\end{pmatrix} \sim Normal\left(\begin{pmatrix}\mu_{se}\\ \mu_{sp}\end{pmatrix}, \begin{pmatrix}\sigma_{se}^2 & \rho_b\sigma_{se}\sigma_{sp}\\ \rho_b\sigma_{se}\sigma_{sp} & \sigma_{sp}^2\end{pmatrix}\right)$$

3.3 Meta-regression with test type as a covariate

The BRMA model, described in Eq.1 and Eq.2, can be extended to incorporate test type as a binary covariate (i.e. meta-regression), allowing the comparison of summary sensitivities and specificities of two tests.[51] This model requires study-level 2x2 data on each test compared to a common reference standard (Table 2). Consider a meta-analysis of $i = 1, \dots, I$ comparative diagnostic test accuracy studies in which all patients undergo both tests. The number of true positives ($tp_{i,j}$) and true negatives ($tn_{i,j}$) for the $j = 1,2$ tests follow independent binomial distributions. In a meta-regression with test type as a covariate, the tests are treated as independent, and, as such, within-study dependencies between tests are ignored, and can be written as:

$$\begin{aligned} tp_{i,1} &\sim Binomial\left(se_{i,1}, N_i^D\right), \quad tp_{i,2} \sim Binomial\left(se_{i,2}, N_i^D\right), \\ tn_{i,1} &\sim Binomial\left(sp_{i,1}, N_i^{\bar{D}}\right), \quad tn_{i,2} \sim Binomial\left(sp_{i,2}, N_i^{\bar{D}}\right) \end{aligned} \tag{3}$$

$se_{i,j}$ and $sp_{i,j}$ are the sensitivity and specificity of the j^{th} test in the i^{th} study, and N_i^D and $N_i^{\bar{D}}$ the number of patients with and without the target condition, respectively. To assess the accuracy of two diagnostic tests, two pairs of *logit*-transformed study-specific sensitivities ($\mu_{i,sej}$) and specificities

($\mu_{i,spj}$) are modelled. Therefore, the bivariate normal distribution at the between-studies level of the BRMA model for single tests (Eq. 2) is replaced by a four-dimensional multivariate normal distribution. A common between-studies correlation parameter is estimated across pairs of sensitivities and specificities to minimise the number of model parameters and reduce the likelihood of non-convergence:

$$logit(se_{i,1}) = \mu_{i,se1},\ \ logit(se_{i,2}) = \mu_{i,se2}, \tag{4}$$

$$logit(sp_{i,1}) = \mu_{i,sp1},\ \ logit(sp_{i,2}) = \mu_{i,sp2},$$

$$\begin{pmatrix}\mu_{i,se1}\\ \mu_{i,se2}\\ \mu_{i,sp1}\\ \mu_{i,sp2}\end{pmatrix} \sim Normal\left(\begin{pmatrix}\mu_{se1}\\ \mu_{se2}\\ \mu_{sp1}\\ \mu_{sp2}\end{pmatrix}, \begin{pmatrix}\sigma_{se1}^2 & \rho_b\sigma_{se1}\sigma_{se2} & \rho_b\sigma_{se1}\sigma_{sp1} & \rho_b\sigma_{se1}\sigma_{sp2}\\ & \sigma_{se2}^2 & \rho_b\sigma_{se2}\sigma_{sp1} & \rho_b\sigma_{se2}\sigma_{sp2}\\ & & \sigma_{sp1}^2 & \rho_b\sigma_{sp1}\sigma_{sp2}\\ & & & \sigma_{sp2}^2\end{pmatrix}\right)$$

3.4 Bivariate copula models

3.4.1 Introduction to copula modelling

Copulas enable the separation of the marginals from the dependence structure of a multivariate distribution, allowing separate specification of both the random variables and the relationship between them. A bivariate copula is a bivariate cumulative distribution function (CDF) with uniform marginal distributions on the interval [0,1].[22] Let there be two continuous, correlated random variables x and y. In accordance with Sklar's theorem,[52] any bivariate distribution, H, can be expressed in terms of univariate marginal distribution functions, F_1 and F_2, and a copula, C, which describes their relationship to one another. θ is the copula dependence parameter, which captures the association between the two random variables:

$$H(x, y, \theta) = C(F_1(x), F_2(y), \theta) \tag{5}$$

For discrete variables – such as those used to model diagnostic accuracy data, which are summarised by counts – the joint probability mass function is derived through finite differences:

$$h(tp_{i,1}, tp_{i,2} \mid se_{i,1}, se_{i,2}, N_i^D, \theta_{i,se}) = \quad (6)$$
$$C(F_1(tp_{i,1}), F_2(tp_{i,2}), \theta_{i,se}) - C(F_1(tp_{i,1}-1), F_2(tp_{i,2}), \theta_{i,se})$$
$$- C(F_1(tp_{i,1}), F_2(tp_{i,2}-1), \theta_{i,se}) + C(F_1(tp_{i,1}-1), F_2(tp_{i,2}-1), \theta_{i,se})$$

$$h(tn_{i,1}, tn_{i,2} \mid sp_{i,1}, sp_{i,2}, N_i^{\bar{D}}, \theta_{i,sp}) =$$
$$C(F_1(tn_{i,1}), F_2(tn_{i,2}), \theta_{i,sp}) - C(F_1(tn_{i,1}-1), F_2(tn_{i,2}), \theta_{i,sp})$$
$$- C(F_1(tn_{i,1}), F_2(tn_{i,2}-1), \theta_{i,sp}) + C(F_1(tn_{i,1}-1), F_2(tn_{i,2}-1), \theta_{i,sp})$$

$F_1(tp_{i,1})$, $F_2(tp_{i,2})$, $F_1(tn_{i,1})$ and $F_2(tn_{i,2})$ are the CDFs of the binomial marginal distributions on the number of true positives and true negatives, and $C(\cdot,\cdot)$ is the bivariate copula. $\theta_{i,se}$ and $\theta_{i,sp}$ are the copula dependence parameters, representing within-study dependencies between the sensitivities and specificities, respectively.

Several families of copulas with a variety of properties have been described. In this paper, we apply five types of bivariate copula to model within-study dependencies between two tests evaluated in the same patient group: Gaussian, Frank, Gumbel, Clayton and Clayton 180° (visualised in Figure 1). Elliptical copulas join univariate marginals through an elliptical distribution, most commonly the Gaussian copula - a symmetric copula with weak dependence in the tails of the distribution. Archimedean copulas are expressed as an explicit formula and depend on a single parameter that dictates the strength of the dependence. The Frank copula is a symmetric Archimedean copula with weak (positive or negative) tail dependence. The Gumbel copula, of the Archimedean family, is an asymmetric copula with positive right-hand tail dependence. The Clayton copula is another asymmetric Archimedean copula exhibiting positive dependence. When evaluating the accuracy of two diagnostic tests, it is likely that there will be stronger dependence in the right-hand tail (i.e. when

sensitivities and specificities are close to 1) than in the left (when sensitivities and specificities are low). This can be induced by rotating the copula function by 180°.

Figure 1: Simulated samples from the bivariate copulas.
Footnote: 4000 samples were simulated for each copula type, with fixed Spearman's correlation coefficient ρ_s = 0.95.

3.4.2 Model specification

We propose an extension of the BRMA model for single (Section 3.2) to jointly synthesise data on two diagnostic tests evaluated in the same patients. Bivariate copulas capture within-study dependencies between sensitivities and specificities for the two tests. For study $i = 1, \dots, I$, the number of true positive and true negative events for each $j = 1,2$ test follow bivariate distributions:

$$\begin{pmatrix} tp_{i,1} \\ tp_{i,2} \end{pmatrix} \sim h\left(se_{i,1}, se_{i,2}, N_i^D, \theta_{i,se}\right) \qquad \begin{pmatrix} tn_{i,1} \\ tn_{i,2} \end{pmatrix} \sim h\left(sp_{i,1}, sp_{i,2}, N_i^{\bar{D}}, \theta_{i,sp}\right) \tag{7}$$

with binomial marginal distributions. $se_{i,j}$ and $sp_{i,j}$ denote the true study-specific sensitivities and specificities of the two tests, respectively, N_i^D and $N_i^{\bar{D}}$ the number of patients with and without the target condition as determined by the reference standard and $\theta_{i,se}$ and $\theta_{i,sp}$ are the copula dependence parameters, representing the within-study dependencies between the sensitivities and specificities, respectively. A separate dependence parameter can be estimated for each study i, allowing within-study dependencies to vary across studies. In practice, cross-classified data required to estimate the dependence parameters may not be available for all studies. In this case, a pair of common dependence parameters, θ_{se} and θ_{sp}, can be assumed across studies (see Section 3.5). Where only 2x2 data are available, informative prior distributions for the dependence parameters can be constructed using external sources of evidence.[5,26] We apply five types of bivariate copula in this paper: Gaussian, Frank, Gumbel, Clayton and Clayton 180° (see the Supplementary Materials for full specification).

At the between-studies level, the *logit*-transformed study-specific sensitivities and specificities follow a four-dimensional multivariate normal distribution with means μ_{sej} and μ_{spj}, between-study

variances σ^2_{sej} and σ^2_{spj}, and $j = 1,2$. A common between-studies correlation parameter is estimated across pairs of sensitivities and specificities:

$$logit(se_{i,1}) = \mu_{i,se1}, \;\; logit(se_{i,2}) = \mu_{i,se2}, \tag{8}$$

$$logit(sp_{i,1}) = \mu_{i,sp1}, \;\; logit(sp_{i,2}) = \mu_{i,sp2},$$

$$\begin{pmatrix} \mu_{i,se1} \\ \mu_{i,se2} \\ \mu_{i,sp1} \\ \mu_{i,sp2} \end{pmatrix} \sim Normal\left(\begin{pmatrix} \mu_{se1} \\ \mu_{se2} \\ \mu_{sp1} \\ \mu_{sp2} \end{pmatrix}, \begin{pmatrix} \sigma^2_{se1} & \rho_b\sigma_{se1}\sigma_{se2} & \rho_b\sigma_{se1}\sigma_{sp1} & \rho_b\sigma_{se1}\sigma_{sp2} \\ & \sigma^2_{se2} & \rho_b\sigma_{se2}\sigma_{sp1} & \rho_b\sigma_{se2}\sigma_{sp2} \\ & & \sigma^2_{sp1} & \rho_b\sigma_{sp1}\sigma_{sp2} \\ & & & \sigma^2_{sp2} \end{pmatrix} \right)$$

3.5 Bootstrapping methods to obtain copula dependence parameter

The copula dependence parameters were estimated from available cross-classifications, which allow the reconstruction of IPD, using a double bootstrap method.[26,53] Double bootstrapping allows the estimation of the association between outcomes with uncertainty. Prior to bootstrapping, IPD was recreated for each study with cross-classified data by transforming the counts into a data set of zeros and ones indicating each patients' test results and disease status. Bootstrapping involves repeatedly sampling the IPD (with replacement) from a single study to create many simulated data sets. Sensitivities and specificities were estimated for each simulated data set, then the association between them estimated across the multiple bootstrap samples. Where cross-classifications are available for each of the $i = 1, \dots, I$ studies, the simulated data sets are used to estimate $\theta_{i,se}$ and $\theta_{i,sp}$ for each study using maximum likelihood estimation. In the motivating example used in this article, cross-classified data were available for two of the 18 studies.[36,38] A common pair of dependence parameters, θ_{se} and θ_{sp}, were assumed across studies by applying the double bootstrap method to each study and estimating the mean of the two for each dependence parameter. Code for the bootstrapping method is provided in the Supplementary Materials.

3.6 Estimation

All models were implemented in a Bayesian framework using Stan version 2.32.2 within R version 4.2.1 via the *rstan* version 2.32.5 package.[54-56] Stan, a Bayesian sampler for performing Hamiltonian Monte Carlo simulation, is particularly computationally efficient and stable when parameters are highly correlated in the posterior distribution – as in diagnostic meta-analysis. A non-centred parameterisation was used for all models to reduce dependencies between successive levels of the hierarchical structure, further increasing the efficiency of the sampler.[57] Copula dependence parameters were estimated using the bootstrap method described in Section 3.5 using R version 4.2.1.[55] Model convergence was assessed using trace plots, density plots and autocorrelation plots. After discarding 1,000 burn-in iterations, posterior estimates were obtained using three chains initialised at different starting values, consisting of 5,000 iterations each. 95% credible intervals (CrIs) were computed as highest posterior density intervals.

To implement the models in a Bayesian framework, prior distributions were placed on the unknown parameters. *Logit*-transformed summary sensitivities and specificities were assumed to follow a minimally informative Normal(0,10^2) prior distribution. The between-studies variance parameters were restricted to positive values through a Half-Normal(0,2.5^2) prior distribution. For the between-studies correlation parameter, the Fisher z-transformation was used, i.e. ρ_b = tanh(z), $z \sim$ Normal(0,0.8). This transformation produces an approximately normal distribution bound between [-1,1]. Model fit was compared by calculating the widely applicable information criterion (WAIC), also known as the Watanabe-Akaike information criterion, with a smaller WAIC indicating better model fit.[58] Code for the meta-regression model and the five bivariate copula models, as well as convergence diagnostic plots for a selection of models, is available in the Supplementary Materials.

4. Results

4.1 Comparison of model fit

Table 4 presents the values of the WAIC across the six fitted models. The meta-regression model was the poorest fit, corresponding to the largest WAIC of 179.7. There is strong evidence that the bivariate copula models result in a better fit compared to the meta-regression model, with reductions in WAIC ranging from 20.3-28.4. Of the five bivariate copula models, the Gumbel copula was the best fit for the data with the smallest WAIC of 151.3. The differences in WAIC between the Gaussian (WAIC = 159.6), Frank (WAIC = 155.4), Clayton (WAIC = 151.6) and Clayton 180° (WAIC = 156.9) copulas were small, however, and the evidence to support one copula over another is marginal. No issues with convergence or mixing of chains were detected for any of the models (see Supplementary Materials).

Table 4: Values of the widely applicable information criterion (WAIC) across models.

Model	WAIC	Change in WAIC compared to the meta-regression model
Meta-regression	179.7	-
Bivariate copula (Gaussian)	159.6	-20.1
Bivariate copula (Frank)	155.4	-24.3
Bivariate copula (Gumbel)	151.3	-28.4
Bivariate copula (Clayton)	151.6	-28.1
Bivariate copula (Clayton 180°)	156.9	-22.8

4.2 Summary sensitivities and specificities

The results of fitting the meta-regression and five bivariate copula models to the motivational example in Alzheimer's disease dementia are presented in Table 5. Figure 2 displays the posterior medians and 95% CrIs for key test accuracy parameters across each of the models. Based on the best fitting Gumbel copula model, CSF $A\beta_{42}$ and t-tau demonstrated 80.9% (95% CrI: 73.4, 87.5) and 76.4% (95% CrI: 69.4, 83.1) sensitivity to differentiate Alzheimer's disease dementia from mild cognitive impairment. Summary specificity was 70.3% (95% CrI: 61.3, 78.4) and 72.5% (95% CrI: 63.7, 81.3), respectively. The

clinical role of biomarkers for dementia is evolving. At present, UK guidelines recommend using CSF markers if diagnostic subtype of dementia is uncertain. Where there is already a clinical suspicion of Alzheimer's dementia, test accuracy may be high in this scenario. Overall, similar inferences about sensitivities and specificities were drawn regardless of the model used.

Posterior median summary sensitivities and specificities were similar across all the models; however, the bivariate copula models yielded narrower 95% CrIs for these parameters than the meta-regression model. The Gumbel copula produced the narrowest CrIs for the sensitivity and specificity estimates, with a 15% and 12% decrease in CrI width for sensitivity of $A\beta_{42}$ and t-tau, respectively, compared to the meta-regression model.

4.3 Between-studies standard deviations

Posterior median between-studies standard deviation estimates of logit-transformed sensitivities and specificities of $A\beta_{42}$ and t-tau from the Gumbel copula model were 0.80 (0.44, 1.23), 0.64 (0.36, 1.00), 0.73 (0.38, 1.19) and 0.83 (0.50, 1.27), respectively. Estimates were larger for the meta-regression than the bivariate copula models. 95% CrIs corresponding to the standard deviations were also wider for the meta-regression model compared to the bivariate copula models, with the Gumbel copula model resulting in a 15% and 17% reduction in CrI width for between-studies standard deviations in sensitivity of $A\beta_{42}$ and t-tau, respectively, compared to the meta-regression model.

4.4 Between-studies correlation

Posterior median between-studies correlation was estimated as -0.21 (-0.33, 0.03) using the Gumbel copula model, indicating negative association between sensitivities and specificities. Estimates were consistently lower, indicating stronger negative association between sensitivities and specificities across studies, for the bivariate copula models compared to the meta-regression model. However, the width of the 95% CrIs for the bivariate copula models exceeded the interval for the meta-regression method. The CrIs for between-studies correlation yielded by all five models were wide, and all spanned 0.

Table 5: Results of fitting the meta-regression and bivariate copula models to the motivating example.

Parameter	Meta-regression model, posterior median (95% CrI)	Bivariate copula model, posterior median (95% CrI)		
		Gaussian	Frank	Gumbel
Sensitivity (%)				
Aβ$_{42}$, $logit^{-1}(\mu_{se1})$	80.22 (72.00, 88.36)	80.82 (73.11, 87.47)	81.03 (73.23, 87.75)	80.88 (73.43, 87.52)
t-tau, $logit^{-1}(\mu_{se2})$	76.22 (67.90, 83.34)	76.36 (69.17, 82.94)	76.53 (68.86, 83.00)	76.39 (69.43, 83.13)
Specificity (%)				
Aβ$_{42}$, $logit^{-1}(\mu_{sp1})$	70.44 (61.04, 79.45)	70.33 (61.86, 78.72)	70.40 (61.36, 78.77)	70.26 (61.34, 78.36)
t-tau, $logit^{-1}(\mu_{sp2})$	72.89 (63.41, 82.50)	72.87 (63.44, 81.08)	72.97 (63.89, 81.96)	72.48 (63.66, 81.27)
Between-studies SD				
Sensitivity Aβ$_{42}$, σ_{se1}	0.914 (0.506, 1.429)	0.802 (0.453, 1.261)	0.829 (0.472, 1.279)	0.798 (0.439, 1.231)
Sensitivity t-tau, σ_{se2}	0.723 (0.397, 1.168)	0.640 (0.356, 1.013)	0.674 (0.374, 1.021)	0.642 (0.355, 1.002)
Specificity Aβ$_{42}$, σ_{sp1}	0.803 (0.442, 1.279)	0.750 (0.406, 1.212)	0.762 (0.429, 1.224)	0.733 (0.375, 1.188)
Specificity t-tau, σ_{sp2}	0.880 (0.542, 1.329)	0.846 (0.516, 1.265)	0.853 (0.528, 1.281)	0.831 (0.502, 1.268)
Between-studies correlation, ρ_b	-0.176 (-0.311, 0.034)	-0.212 (-0.333, 0.014)	-0.194 (-0.325, 0.038)	-0.205 (-0.333, 0.030)

Parameter	Bivariate copula model, posterior median (95% CrI)	
	Clayton	Clayton 180°
Sensitivity (%)		
Aβ$_{42}$, $logit^{-1}(\mu_{se1})$	81.12 (73.02, 88.12)	80.67 (73.09, 87.61)
t-tau, $logit^{-1}(\mu_{se2})$	76.60 (68.90, 83.47)	76.29 (68.92, 83.00)
Specificity (%)		
Aβ$_{42}$, $logit^{-1}(\mu_{sp1})$	70.52 (61.28, 79.19)	70.27 (61.72, 78.56)
t-tau, $logit^{-1}(\mu_{sp2})$	73.02 (63.90, 82.18)	72.50 (63.61, 81.39)
Between-studies SD		
Sensitivity Aβ$_{42}$, σ_{se1}	0.842 (0.486, 1.329)	0.803 (0.451, 1.264)
Sensitivity t-tau, σ_{se2}	0.681 (0.370, 1.051)	0.654 (0.367, 1.018)
Specificity Aβ$_{42}$, σ_{sp1}	0.772 (0.406, 1.232)	0.742 (0.383, 1.202)
Specificity t-tau, σ_{sp2}	0.861 (0.532, 1.325)	0.831 (0.505, 1.278)
Between-studies correlation, ρ_b	-0.178 (-0.328, 0.069)	-0.207 (-0.333, 0.022)

μ_{se1}, μ_{se2}, μ_{sp1} and μ_{sp2} denote *logit*-transformed summary sensitivities and specificities of test 1 and test 2, respectively.

σ_{se1}, σ_{se2}, σ_{sp1} and σ_{sp2} denote between-studies standard deviation in *logit*-transformed sensitivities and specificities of test 1 and test 2, respectively.

ρ_b denotes the between-studies correlation parameter.

Aβ$_{42}$, amyloid-β 42; CrI, credible interval; SD, standard deviation; t-tau, total tau

Figure 2: Posterior medians (solid dots) and 95% CrIs (solid bars) of test accuracy parameters.

Footnote: $A\beta_{42}$, amyloid-β 42; BC, bivariate copula; CrI, credible interval; SD, standard deviation; t-tau, total tau

5. Discussion

We have developed novel bivariate copula models for synthesising evidence on the accuracy of two diagnostic tests, accounting for associations between and within studies present between two tests evaluated in the same patients. The new models offer a robust yet flexible approach to modelling comparative test accuracy data, maximising the available evidence base by making use of both study- and individual-level data where available. The bivariate copula method resulted in improved model fit compared to the currently recommended meta-regression approach.

When applied to our motivating example, the bivariate copula models resulted in lower point estimates of the between-studies standard deviation and correlation parameters. It has been hypothesised that when within-study dependencies are not taken into account, the 'excess' of the association manifests itself as an upwardly biased estimate of the between-studies heterogeneity parameters.[26,59] The bivariate copula models yielded narrower CrIs for summary sensitivity and specificity parameters, likely due to the additional evidence on within-study associations utilised by the method. In a HTA context, sensitivity and specificity estimates are used in decision-making to determine whether a novel diagnostic test should be used in clinical practice over other technologies. Increased precision in these estimates aids the evaluation of clinical and cost-effectiveness, enabling more precise decisions on the most efficient use of health resources. Where the estimates are used to populate a health economic model, it is vital to capture the uncertainty in the parameter estimates to provide appropriate recommendations for reimbursement.[60]

There are a number of existing meta-analysis models for jointly synthesising data on multiple diagnostic tests.[10-21] Trikalinos et al [10] proposed extending the BRMA model (Section 3.2), using

multinomial likelihoods to capture within-study dependencies present between two tests evaluated in the same patients. Multivariate methods such as this make strong distributional assumptions about the marginal variables, and require cross-classified data from all studies. A scoping review of meta-analysis models for three or more diagnostic tests by Veroniki et al found that the majority of existing methods require cross-classified test results,[61] although publications of diagnostic accuracy studies rarely facilitate access to this data.[49-51] Indeed, in the review undertaken to identify a motivating example for this paper, full cross-classifications were reported in only 20% of comparative studies. Copula methodology relaxes the requirement for cross-classifications across all studies, reflecting current reporting standards and making its application in systematic reviews more generalizable.

The flexibility of copulas, of which a number of types have been defined and explored, make them a natural approach to model diagnostic accuracy data, which are often non-normal and likely to exhibit strong tail dependence when marginal sensitivities (and specificities) are high. The adaptable nature of copulas increases their suitability for a range of diagnostic data sets, or indeed any type of data in which multiple, correlated outcomes are present. An advantage of a copula approach to modelling related variables is the estimation of a single association parameter, θ, in contrast to the need to estimate both standard deviation and correlation parameters when fitting a multivariate normal distribution. If these parameters conflict with one another, for example when extreme values are sampled from the prior distribution, it can cause the Bayesian sampler to fail to converge. Indeed, the heterogeneity parameters estimated using a four-dimensional normal distribution at the between-studies level in the bivariate copula models were noted to be highly sensitive to choice of prior distribution and starting values.[62-63]

The proposed methodology is subject to a number of potential limitations. Inference focussed on sensitivity and specificity, but alternative test accuracy measures may be of greater clinical utility in practice when true disease status is unknown. Nonetheless, it is straightforward to derive estimates of predictive values, likelihood ratios and diagnostic odds ratios from posterior estimates of sensitivity

and specificity. Meta-analysis of these measures directly is subject to limitations. Predictive values are often more heterogeneous than sensitivity and specificity due to increased variation with disease prevalence, which leads to reduced goodness of fit for meta-analysis models.[64] Bivariate meta-analysis of likelihood ratios can lead to implausible corresponding values of sensitivity and specificity (i.e. <0).[65] Synthesising diagnostic odds ratios rather than paired test accuracy measures results in the loss of ability to distinguish between tests with high sensitivity and high specificity.[66] Conversely, paired test accuracy measures, including sensitivity and specificity, hinder the ranking of tests. Trade-off between sensitivity and specificity should be considered in the context of the clinical use of the test; the potential consequences of a false positive or false negative result may not be equal.[50]

It has been highlighted that further evaluation of meta-analysis models for multiple diagnostic tests is needed before their adoption into HTA.[2,9,48,51] The performance of novel meta-analysis models for two tests could be formally compared to the currently recommended meta-regression approach through a simulation study to quantify potential bias from selecting a more simplistic model. This may lead to a circular argument, however; data would need to be simulated from a copula model to capture strong tail dependencies, therefore the copula the data was simulated from would be the best fitting model. In selecting an appropriate copula model, consideration should be given to model fit, using measures such as WAIC. The Gumbel copula was the best fit for the motivational example; however, underlying dependence structures may vary by patient and test characteristics and different copula types should be compared as part of the model fitting process.

The bivariate copula framework could be extended to include direct estimation of joint accuracy measures, for example using a trivariate copula that incorporates both marginal and joint sensitivities and specificities. Alternative methods for synthesising data on two diagnostic tests incorporate other, desirable features that could be considered for future copula model development. Menten and Lesaffre [11] and Lian et al [20] described meta-analysis models that account for imperfect reference standards, the former using latent class analysis and the latter by comparing tests within a missing

data framework. Methods introduced by Owen et al [18] and Hoyer and Kuss [19] allow for multiple thresholds per test by including threshold information as a covariate, making use of a greater proportion of the available literature. Several models allow the addition of single arm studies.[12,16,20] Sensitivity and specificity are known to vary with disease prevalence.[67] Hoyer and Kuss [68] and Nikoloulopoulos [69] suggested meta-analysis models for a single diagnostic test, accounting for disease prevalence using a trivariate copula. The models could also be adapted to include a copula at the between-studies level.[13,15,17,21]

To enable the application of models to appropriately account for within-study associations between multiple tests, such as the bivariate copula models proposed here, we recommend that comparative diagnostic accuracy studies report cross-classified data wherever possible. Reporting checklists for diagnostic accuracy studies, such as the Standards for Reporting of Diagnostic Accuracy Studies (STARD),[70] have been developed to increase reporting quality and reduce research waste. The addition of an item stipulating that cross tabulation of all index test results should be reported where appropriate would encourage this practice and increase awareness of the need for comparative data to comprehensively evaluate testing strategies. Cross tabulation of index test results was identified as an item in the dementia-specific extension to STARD, STARDdem.[71] Technical Support Documents (TSDs), produced by the National Institute for Health and Care Excellence (NICE) Decision Support Unit, make recommendations on the implementation of methods for technology appraisal. TSDs developed for multivariate meta-analysis of interventional research could be extended to further aid in the translation of diagnostic meta-analysis models for multiple tests into healthcare decision-making.[72]

5.1 Conclusions

We have developed novel Bayesian meta-analysis models for synthesising data on two diagnostic tests evaluated in the same patient group (and compared to a common reference standard), using bivariate copulas to capture within-study dependencies between multiple tests. The methodology introduced in this paper is applicable to a wide range of disease areas. The models make use of a broader evidence

base by utilising both study-level and IPD, where available. This new approach aids evidence synthesis of commonly reported data items from diagnostic test accuracy studies to improve healthcare policy and decision-making.

References

[1] Takwoingi Y, Quinn TJ. Review of Diagnostic Test Accuracy (DTA) studies in older people. Age and Ageing. 2018;47(3):349-55.

[2] Leeflang MMG, Reitsma JB. Systematic reviews and meta-analyses addressing comparative test accuracy questions. Diagnostic and Prognostic Research. 2018;2:17.

[3] Spiegelhlater DJ, Abrams KR, Myles JP. Bayesian Approaches to Clinical Trials and Health-Care Evaluation. John Wiley & Sons; 2004.

[4] Welton, NJ, Sutton AJ, Cooper N, Abrams KR, Ades AE. Evidence Synthesis for Decision Making in Healthcare. John Wiley & Sons; 2012.

[5] Honest H, Khan KS. Reporting of measures of accuracy in systematic reviews of diagnostic literature. BMC Health Services Research. 2002;2:1.

[6] Deeks JJ. Systematic reviews in health care: systematic reviews of evaluations of diagnostic and screening tests. BMJ. 2001;323:157-62.

[7] Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of Clinical Epidemiology. 2005;58(10):982-90.

[8] Chu H, Cole SE. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. Journal of Clinical Epidemiology. 2006;59(12):1331-2.

[9] Welton NJ, Phillippo D, Owen R, Jones H, Dias S, Bujkiewicz S, et al. DSU Report. CHTE2020 Sources and Synthesis of Evidence; Update to Evidence Synthesis Methods. 2020.

[10] Trikalinos TA, Hoaglin DC, Small KM, Terrin N, Schmid CH. Methods for the joint meta-analysis of multiple tests. Research Synthesis Methods. 2014;5(4):294-312.

[11] Menten J, Lesaffre E. A general framework for comparative Bayesian meta-analysis of diagnostic studies. BMC Medical Research Methodology. 2015;15:70.

[12] Nyaga VN, Aerts M, Arbyn M. ANOVA model for network meta-analysis of diagnostic test accuracy data. Statistical Methods in Medical Research. 2018;27(6):1766-84.

[13] Nyaga VN, Arbyn M, Aerts M. Beta-binomial analysis of variance model for network meta-analysis of diagnostic test accuracy data. Statistical Methods in Medical Research. 2018;27(8):2554-66.

[14] Dimou NL, Adam M, Bagos PG. A multivariate method for meta-analysis and comparison of diagnostic tests. Statistics in Medicine. 2016;35:3509-23.

[15] Cheng W. Network meta-analysis of diagnostic accuracy studies [PhD thesis]. Providence (RI): Brown University; 2016. https://repository.library.brown.edu/studio/item/bdr:674079. Accessed 8th February 2024.

[16] Ma X, Lian Q, Chu H, Ibrahim JG, Chen Y. A Bayesian hierarchical model for network meta-analysis of multiple diagnostic tests. Biostatistics. 2018;19(1):87-102.

[17] Hoyer A, Kuss O. Meta-analysis for the comparison of two diagnostic tests—A new approach based on copulas. Statistics in Medicine. 2018;37:739-48.

[18] Owen RK, Cooper NJ, Quinn TJ, Lees R, Sutton AJ. Network meta-analysis of diagnostic test accuracy studies identifies and ranks the optimal diagnostic tests and thresholds for health care policy and decision-making. Journal of Clinical Epidemiology. 2018;99:64-74.

[19] Hoyer A, Kuss O. Meta-analysis for the comparison of two diagnostic tests to a common gold standard: A generalized linear mixed model approach. Statistical Methods in Medical Research. 2018;27(5):1410-21.

[20] Lian Q, Hodges JS, Chu H. A Bayesian Hierarchical Summary Receiver Operating Characteristic Model for Network Meta-Analysis of Diagnostic Tests. Journal of the American Statistical Association. 2019;114(527):949-61.

[21] Nikoloulopoulos AK. A D-vine copula mixed model for joint meta-analysis and comparison of diagnostic tests. Statistical Methods in Medical Research. 2019;28(10-11):3286-300.

[22] Joe, H. Dependence modeling with copulas. New York: Chapman and Hall/CRC Press; 2014.

[23] Kuss O, Hoyer A, Solms A. Meta-analysis for diagnostic accuracy studies: a new statistical model using beta-binomial distributions and bivariate copulas. Statistics in Medicine. 2014;33(1):17-30.

[24] Nyaga VN, Arbyn M, Aerts M. CopulaDTA: An R Package for Copula-Based Bivariate Beta-Binomial Models for Diagnostic Test Accuracy Studies in a Bayesian Framework. Journal of Statistical Software. 2017;82(1):1-27.

[25] Burzykowski T, Molenberghs G, Buyse M, Geys H, Renard D. Validation of surrogate end points in multiple randomized clinical trials with failure time end points. Journal of the Royal Statistical Society Series C: Applied Statistics. 2001;50(4):405-22.

[26] Papanikos T, Thompson JR, Abrams KR, Bujkiewicz S. Use of copula to model within-study association in bivariate meta-analysis of binomial data at the aggregate level: A Bayesian approach and application to surrogate endpoint evaluation. Statistics in Medicine. 2022;41(25):4961-81.

[27] World Health Organization. Dementia. https://www.who.int/news-room/fact-sheets/detail/dementia. Accessed 8th February 2024.

[28] Gauthier S, Rosa-Neto P, Morais JA, Webster C. World Alzheimer Report 2021: Journey through the diagnosis of dementia. Alzheimer's Disease International. 2021.

[29] McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR Jr., Kawas CH, et al. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & Dementia. 2011;7(3):263-9.

[30] Bjerke M, Andreasson U, Rolstad S, Nordlund A, Lind K, Zetterberg H, et al. Subcortical Vascular Dementia Biomarker Pattern in Mild Cognitive Impairment. Dementia and Geriatric Cognitive Disorders. 2009;28(4):348-56.

[31] Blom ES, Giedraitis V, Zetterberg H, Fukumoto H, Blennow K, Hyman BT, et al. Rapid Progression from Mild Cognitive Impairment to Alzheimer's Disease in Subjects with Elevated Levels of Tau in Cerebrospinal Fluid and the APOE ε4/ε4 Genotype. Dementia and Geriatric Cognitive Disorders. 2009;27(5):458-64.

[32] Chiasserini D, Parnetti L, Andreasson U, Zetterberg H, Giannandrea D, Calabresi P, et al. CSF Levels of Heart Fatty Acid Binding Protein are Altered During Early Phases of Alzheimer's Disease. Journal of Alzheimer's Disease. 2010;22(4):1281-8.

[33] Frölich L, Peters O, Lewczuk P, Gruber O, Teipel SJ, Gertz HJ, et al. Incremental value of biomarker combinations to predict progression of mild cognitive impairment to Alzheimer's dementia. Alzheimer's Research & Therapy. 2017;9:84.

[34] Gaser C, Franke K, Klöppel S, Koutsouleris N, Sauer H. *BrainAGE* in Mild Cognitive Impaired Patients: Predicting the Conversion to Alzheimer's Disease. PLoS ONE. 2013;8(6): e67346.

[35] Hampel H, Teipel SJ, Fuchsberger T, Andreasen N, Wiltfang J, Otto M, et al. Value of CSF b-amyloid1–42 and tau as predictors of Alzheimer's disease in patients with mild cognitive impairment. Molecular Psychiatry. 2004;9(7):705-10.

[36] Hansson O, Zetterberg H, Buchhave P, Londos E, Blennow K. Association between CSF biomarkers and incipient Alzheimer's disease in patients with mild cognitive impairment: a follow-up study. The Lancet Neurology. 2006;5(3):228-34.

[37] Herukka SK, Pennanen C, Soininen H, Pirttila T. CSF AB42, Tau and Phosphorylated Tau Correlate with Medial Temporal Lobe Atrophy. Journal of Alzheimer's Disease. 2008;14(1):51-7.

[38] Hertze J, Minthon L, Zetterberg H, Vanmechelen E, Blennow K, Hansson O. Evaluation of CSF Biomarkers as Predictors of Alzheimer's Disease: A Clinical Follow-Up Study of 4.7 Years. Journal of Alzheimer's Disease. 2010;21(4):1119-28.

[39] Kester, MI, Verwey NA, van Elk EJ, Blankenstein MA, Scheltens P, van der Flier WM. Progression from MCI to AD: Predictive value of CSF Aβ42 is modified by APOE genotype. Neurobiology of Aging. 2011;32(8):1372-8.

[40] Monge-Argilés JA, Muñoz-Ruiz C, Pampliega-Pérez A, Gómez-López MJ, Sánchez-Payá J, Rodríguez Borja E, et al. Biomarkers of Alzheimer's Disease in the Cerebrospinal Fluid of Spanish Patients With Mild Cognitive Impairment. Neurochemical Research. 2011;36(6):986-93.

[41] Nesteruk M, Nesteruk T, Styczyńska M, Mandecka M, Barczak A, Barcikowska M. Combined use of biochemical and volumetric biomarkers to assess the risk of conversion of mild cognitive impairment to Alzheimer's disease. Folia Neuropathologica. 2016;54(4):369-74.

[42] Palmqvist S, Hertze J, Minthon L, Wattmo C, Zetterberg H, Blennow K, et al. Comparison of Brief Cognitive Tests and CSF Biomarkers in Predicting Alzheimer's Disease in Mild Cognitive Impairment: Six-Year Follow-Up Study. PLoS ONE. 2012;7(6): e38639.

[43] Parnetti L, Lanari A, Silvestrelli G, Saggese E, Reboldi P. Diagnosing prodromal Alzheimer's disease: Role of CSF biochemical markers. Mechanisms of Ageing and Development. 2006;127(2):129-32.

[44] Parnetti L, Chiasserini D, Eusebi P, Giannandrea D, Bellomo G, De Carlo C, et al. Performance of $A\beta_{1-40}$, $A\beta_{1-42}$, Total Tau, and Phosphorylated Tau as Predictors of Dementia in a Cohort of Patients with Mild Cognitive Impairment. Journal of Alzheimer's Disease. 2012;29(1):229-38.

[45] Prestia A, Caroli A, Herholz K, Reiman E, Chen K, Jagust WJ, et al. Diagnostic accuracy of markers for prodromal Alzheimer's disease in independent clinical series. Alzheimer's & Dementia. 2013;9(6):677-86.

[46] Rhodius-Meester HFM, Koikkalainen J, Mattila J, Teunissen CE, Barkhof F, Lemstra A, et al. Integrating Biomarkers for Underlying Alzheimer's Disease in Mild Cognitive Impairment in

Daily Practice: Comparison of a Clinical Decision Support System with Individual Biomarkers. Journal of Alzheimer's Disease. 2016;50(1):261-70.

[47] Vos SJB, van Rossum IA, Verhey F, Knol DK, Soininen H, Wahlund L-O, et al. Prediction of Alzheimer disease in subjects with amnestic and nonamnestic MCI. Neurology. 2013;80(12):1124-32.

[48] Takwoingi Y, Leeflang MMG, Deeks JJ. Empirical Evidence of the Importance of Comparative Studies of Diagnostic Test Accuracy. Annals of Internal Medicine. 2013;158(7):544-54.

[49] Rücker, G. Network Meta-Analysis of Diagnostic Test Accuracy Studies. In: Zoccai, G. (ed.) Diagnostic Meta-Analysis. Springer; 2018. p. 183-97.

[50] Veroniki A, Tsokani S, Rücker G, Mavridis D, Takwoingi Y. Challenges in Comparative Meta-Analysis of the Accuracy of Multiple Diagnostic Tests. In: Evangelou E, Veroniki AA (eds.) Meta-Research. Humana, New York, NY; 2022. p. 299-316.

[51] Deeks JJ, Bossuyt PM, Leeflang MM, Takwoingi Y (editors). Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. Version 2.0. Cochrane. 2023. https://training.cochrane.org/handbook-diagnostic-test-accuracy/current. Accessed 8th February 2024.

[52] Sklar M. Fonctions de repartition an dimensions et leurs marges. Publications de l'Institut Statistique de l'Université de Paris. 1959;8:229-31.

[53] Bujkiewicz S, Thompson JR, Sutton AJ, Cooper NJ, Harrison MJ, Symmons DPM, et al. Multivariate meta-analysis of mixed outcomes: a Bayesian approach. Statistics in Medicine. 2013;32(22):3926-43.

[54] Stan Development Team. Stan Modeling Language Users Guide and Reference Manual, version 2.34. 2024. https://mc-stan.org. Accessed 8th February 2024.

[55] R Core Team. R: A language and environment for statistical computing, version 4.2.1. Vienna, Austria; 2022. https://www.R-project.org. Accessed 8th February 2024.

[56] Stan Development Team. RStan: the R interface to Stan, version 2.32.5. 2024. https://mc-stan.org. Accessed 8th February 2024.

[57] Papaspiliopoulos O, Roberts GO, Sköld M. A General Framework for the Parametrization of Hierarchical Models. Statistical Science. 2007;22(1):59-73.

[58] Watanabe S. Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. Journal of Machine Learning Research. 2010;11(116):3571-94.

[59] Riley RD. Multivariate meta-analysis: the effect of ignoring within-study correlation. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2009;172(4):789-811.

[60] Soares MO, Walker S, Palmer SJ, Sculpher MJ. Establishing the Value of Diagnostic and Prognostic Tests in Health Technology Assessment. Medical Decision Making. 2018;38(4):495-508.

[61] Veroniki AA, Tsokani S, Agarwal R, Pagkalidou E, Rücker G, Mavridis D, et al. Diagnostic test accuracy network meta-analysis methods: A scoping review and empirical assessment. Journal of Clinical Epidemiology. 2022;146:86-96.

[62] Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. Statistics in Medicine. 2005;24(15):2401-28.

[63] Gelman A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian Analysis. 2006;1(3):515-34.

[64] Leeflang MMG, Deeks JJ, Rutjes AWS, Reitsma JB, Bossuyt PMM. Bivariate meta-analysis of predictive values of diagnostic tests can be an alternative to bivariate meta-analysis of sensitivity and specificity. Journal of Clinical Epidemiology. 2012;65(10):1088-97.

[65] Zwinderman AH, Bossuyt PM. We should not pool diagnostic likelihood ratios in systematic reviews. Statistics in Medicine. 2008;27(5):687-97.

[66] Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PMM. The diagnostic odds ratio: a single indicator of test performance. Journal of Clinical Epidemiology. 2003;56(11):1129-35.

[67] Leeflang MMG, Rutjes AWS, Reitsma JB, Hooft L, Bossuyt PMM. Variation of a test's sensitivity and specificity with disease prevalence. Canadian Medical Association Journal. 2013;185(11):E537-E544.

[68] Hoyer A, Kuss O. Meta-analysis of diagnostic tests accounting for disease prevalence: a new model using trivariate copulas. Statistics in Medicine. 2015;34(11):1912-24.

[69] Nikoloulopoulos AK. A vine copula mixed effect model for trivariate meta-analysis of diagnostic test accuracy studies accounting for disease prevalence. Statistical Methods in Medical Research. 2017;25(5):2270-86.

[70] Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open 2016;6:e012799.

[71] Noel-Storr AH, McCleery JM, Richard E, Ritchie CW, Flicker L, Cullum SJ, et al. Reporting standards for studies of diagnostic test accuracy in dementia: The STARDdem Initiative. Neurology. 2014;83(4):364-73.

[72] Bujkiewicz S, Achana F, Papanikos T, Riley RD, Abrams KR. NICE DSU Technical Support Document 20: Multivariate meta-analysis of summary data for combining treatment effects on correlated outcomes and evaluating surrogate endpoints. 2019. https://www.sheffield.ac.uk/nice-dsu/tsds/multivariate-meta-analysis. Accessed 8th February 2024.

1

Glossary

Alzheimer's disease: A progressive, neurodegenerative disease thought to be due to the accumulation of amyloid plaques or tau tangles in the brain which disrupt the connection between nerve cells, and the most common underlying cause of dementia.

Bayesian inference: Based on Bayes theorem, direct probability statements are made about posterior parameter estimates, conditional on both the observed data and prior knowledge or beliefs.

Copula: Statistical technique used to model the dependence between two or more random variables. A copula is a multivariate cumulative distribution function with marginals that follow a standard uniform distribution.

Dementia: A progressive, clinical syndrome characterised by decline in cognitive function that impacts activities of daily living, beyond what might be expected during the normal ageing process.

Fixed effects meta-analysis: A meta-analysis model that assumes homogeneity between studies, meaning effect estimates across studies are estimating the same underlying true effect. Variation in effect estimates is due to sampling error alone.

Health technology assessment: A systematic process for evaluating the clinical and cost-effectiveness of a health technology used in the prevention, diagnosis or treatment of a condition, as well as its impact on the broader health system.

Index test: A new or existing diagnostic test of interest whose performance is being evaluated.

Meta-analysis: The statistical combination of results from two or more studies that answer the same research question.

Random effects meta-analysis: A meta-analysis model that assumes heterogeneity between studies, meaning the true effect varies between studies according to a pre-specified distribution.

Reference standard: The test against which the index test is compared, used to verify the presence or absence of the target condition. Usually the best available method for detecting the target condition. Sometimes referred to as a gold standard.

Sensitivity: The proportion of patients who have the target condition that are correctly identified by the test.

Specificity: The proportion of patients who do not have the target condition that are correctly identified by the test.

Systematic review: A summary of the evidence on a clearly formulated research question, using systematic, explicit, and reproducible methods to identify, appraise, and synthesise all relevant studies. May or may not include a meta-analytic component.

Target condition: The condition that a diagnostic test aims to detect. **Threshold:** A criteria, such as a numerical cut-off, applied to a test to define a positive result (indicating the presence of disease) or a negative result (indicating the absence of disease).

Bibliography

[1] Jack, Jr C. R., Knopman D. S., Jagust W. J., Shaw L. M., Aisen P. S., Weiner M. W., Petersen R. C., Trojanowski J. Q. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *The Lancet Neurology*, 9(1):119–128, 2010.

[2] Ritchie C. W., Terrera G. M., Quinn T. J. Dementia trials and dementia tribulations: methodological and analytical challenges in dementia research. *Alzheimer's Research & Therapy*, 7(1):31, 2015.

[3] Spiegelhlater D. J., Abrams K. R., Myles J. P. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. John Wiley & Sons, 2004.

[4] Deeks J. J., Bossuyt P. M., Leeflang M. M., Takwoingi Y. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy version 2.0. Available from: https://training.cochrane.org/handbook-diagnostic-test-accuracy [accessed 2024-01-12], 2023.

[5] Higgins J. P. T., Thomas J., Chandler J., Cumpston M., Li T., Page M. J., Welch V. A. Cochrane Handbook for Systematic Reviews of Interventions version 6.4. Available from: https://training.cochrane.org/handbook [accessed 2024-01-12], 2023.

[6] Honest H. Khan K. S. Reporting of measures of accuracy in systematic reviews of diagnostic literature. *BMC Health Services Research*, 2:4, 2002.

[7] Deeks J. J. Systematic reviews of evaluations of diagnostic and screening tests. *BMJ*, 323(7305):157–162, 2001.

[8] Takwoingi Y., Leeflang M. M. G., Deeks J. J. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. *Annals of Internal Medicine*, 158(7):544–554, 2013.

[9] Takwoingi Y. *Meta-analytic approaches for summarising and comparing the accuracy of medical tests*. Dissertation, University of Birmingham, Birmingham (UK), 2016.

[10] Rucker G. *Diagnostic Meta-Analysis*, chapter Network Meta-Analysis of Diagnostic Test Accuracy Studies, pages 183–197. Springer, 2018.

[11] Veroniki A. A., Tsokani S., Rücker G., Mavridis D., Takwoingi Y. *Meta-Research*, chapter Challenges in Comparative Meta-Analysis of the Accuracy of Multiple Diagnostic Tests, pages 299–316. Humana, New York, NY, 2022.

[12] Leeflang M. M. G. Reitsma J. B. Systematic reviews and meta-analyses addressing comparative test accuracy questions. *Diagnostic and Prognostic Research*, 2:17, 2018.

[13] Battista R. N. Hodge M. J. The evolving paradigm of health technology assessment: reflections for the millenium. *Canadian Medical Journal Association*, 160(10):1464–1467, 1999.

[14] National Institute for Health and Care Excellence. Diagnostics Assessment Programme manual. Available from: https://www.nice.org.uk/Media/Default/About/what-we-do/NICE-guidance/NICE-diagnostics-guidance/Diagnostics-assessment-programme-manual.pdf [accessed 2024-01-12], 2011.

[15] Welton N. J., Phillippo D. M., Owen R., Jones H. E., Dias S., Bujkiewicz S., Ades A. E., Abrams K. R. CHTE2020 Sources and Synthesis of Evidence; Update to Evidence Synthesis Methods. Available from: https://www.sheffield.ac.uk/nice-dsu/methods-development/chte2020-sources-and-synthesis-evidence [accessed 2024-01-12], 2020.

[16] Welton N. J., Sutton A. J., Cooper N., Abrams K. R., Ades A. E. *Evidence Synthesis for Decision Making in Healthcare*. John Wiley & Sons, 2012.

[17] Spiegelhalter D., Myles J., Jones D., Abrams K. An introduction to bayesian methods in health technology assessment. *BMJ*, 319(7208):508–512, 1999.

[18] Robinson L., Tang E., Taylor J.-P. Dementia: timely diagnosis and early intervention. *BMJ*, 350:h3029, 2015.

[19] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association, fifth edition, 2013.

[20] World Health Organisation. Dementia. Available from: https://www.who.int/news-room/fact-sheets/detail/dementia [accessed 2024-01-12], 2023.

[21] World Health Organization. Global action plan on the public health response to dementia 2017–2025. Available from: https://www.who.int/publications/i/item/global-action-plan-on-the-public-health-response-to-dementia-2017—2025 [accessed 2024-01-12], 2017.

[22] Wittenberg R., Hu B., Jagger C., Kingston A., Knapp M., Comas-Herrera A., King D., Rehill A., Banerjee S. Projections of care for older people with dementia in England: 2015 to 2040. *Age and Ageing*, 49(2):264–269, 2020.

[23] Coope B., Ballard C., Saad K., Patel A., Bentham P., Bannister C., Graham C., Wilcock G. The prevalence of depression in the carers of dementia sufferers. *International Journal of Geriatric Psychiatry*, 10(3):237–242, 1995.

[24] Cooper C., Balamurali T. B. S., Livingston G. A systematic review of the prevalence and covariates of anxiety in caregivers of people with dementia. *International Psychogeriatrics*, 19(2):175–195, 2007.

[25] Brodaty H. Donkin M. Family caregivers of people with dementia. *Dialogues in Clinical Neuroscience*, 11(2):217–228, 2009.

[26] Prince M., Knapp M., Guerchet M., McCrone P., Prina M., Comas-Herrera A., Wittenberg R., Adelaja B., Hu B., King D., Rehill A., Salimkumar D. *Dementia UK: Second Edition - Overview*. Alzheimer's Society, 2014.

[27] McKhann G. M., Knopman D. S., Chertkow H., Hyman B. T., Jack C. R. J., Kawas C. H., Klunk W. E., Koroshetz W. J., Manly J. J., Mayeux R., Mohs R. C., Morris J. C., Rossor M. N., Scheltens P., Carrillo M. C., Thies B., Weintraub S., Phelps C. H. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, 7(3):263–269, 2011.

[28] Schneider J. A., Arvanitakis Z., Bang W., Bennett D. A. Mixed brain pathologies account for most dementia cases in community-dwelling older persons. *Neurology*, 69(24):2197–2204, 2007.

[29] Fernando M. S. Ince P. G. Vascular pathologies and cognition in a population-based cohort of elderly people. *Journal of the Neurological Sciences*, 226(1-2):13–17, 2004.

[30] Alzheimer's Disease International. World Alzheimer Report 2022. Available from: https://www.alzint.org/resource/world-alzheimer-report-2022 [accessed 2024-01-12], 2022.

[31] Alzheimer's Society. Westminster Hall Debate on 'Addressing deterioration in people with long-term conditions during the Covid-19 pandemic'. Available from: https://www.alzheimers.org.uk/sites/default/files/2022-03/220310 Alzheimer's Society Briefing - Deterioration during the pandemic for people living with dementia.pdf [accessed 2024-01-12], 2022.

[32] Alzheimer's Society. Improve dementia diagnosis. Available from: https://www.alzheimers.org.uk/get-involved/our-campaigns/improve-dementia-diagnosis [accessed 2024-01-12], 2022.

[33] Alzheimer's Disease International. World Alzheimer Report 2021. Available from: https://www.alzint.org/resource/world-alzheimer-report-2021 [accessed 2024-01-12], 2021.

[34] Food and Drug Administration. FDA Grants Accelerated Approval for Alzheimer's Drug. Available from: https://www.fda.gov/news-events/press-announcements/fda-grants-accelerated-approval-alzheimers-drug [accessed 2024-01-12], 2021.

[35] Food and Drug Administration. FDA Grants Accelerated Approval for Alzheimer's Disease Treatment. Available from: https://www.fda.gov/news-events/press-announcements/fda-grants-accelerated-approval-alzheimers-disease-treatment [accessed 2024-01-12], 2023.

[36] The National Institute for Health and Care Excellence. Dementia: assessment, management and support for people living with dementia and their carers. Available from: https://www.nice.org.uk/guidance/ng97 [accessed 2024-01-12], 2018.

[37] Alzheimer's Society. 91% of people affected by dementia see clear benefits to getting a diagnosis. Available from: https://www.alzheimers.org.uk/news/2022-05-16/91-people-affected-dementia-see-clear-benefits-getting-diagnosis [accessed 2024-01-12], 2022.

[38] Hampel H., Bürger K., Teipel S. J., Bokde A. L. W., Zetterberg H., Blennow K. Core candidate neurochemical and imaging biomarkers of Alzheimer's disease. *Alzheimer's & Dementia*, 4(1):38–48, 2008.

[39] Ritchie C., Smailagic N., Noel-Storr A. H., Takwoingi Y., Flicker L., Mason S. E., McShane R. Plasma and cerebrospinal fluid amyloid beta for the diagnosis of Alzheimer's disease dementia and other dementias in people with mild cognitive impairment (MCI). *The Cochrane Database of Systematic Reviews*, 6:CD008782, 2014.

[40] Ritchie C., Smailagic N., Noel-Storr A. H., Ukoumunne O., Ladds E. C., Martin S. CSF tau and the CSF tau/ABeta ratio for the diagnosis of Alzheimer's disease dementia and other dementias in people with mild cognitive impairment (MCI). *The Cochrane Database of Systematic Reviews*, 3:CD010803, 2017.

[41] Sabbagh M. N., Hendrix S., Harrison J. E. FDA position statement "Early Alzheimer's disease: Developing drugs for treatment, Guidance for Industry". *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 5(1):13–19, 2019.

[42] Gardner M. J. Altman D. G. *Statistics with Confidence*, chapter Confidence intervals rather than P values, pages 15–27. BMJ Books, second edition, 2000.

[43] Juat N., Meredith M., Kruschke J. Highest Posterior Density Intervals. Available from: https://cran.r-project.org/web/packages/HDInterval [accessed 2024-01-12], 2022. R package version 0.2.4.

[44] Lambert P. C., Sutton A. J., Burton P. R., Abrams K. R., Jones D. R. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine*, 24(15):2401–2428, 2005.

[45] Metropolis N., Rosenbluth A. W., Rosenbluth M. N., Teller A. H., Teller E. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[46] Hastings W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[47] Gelman A., Carlin J. B., Stern H. S., Dunson D. B., Vehtari A., Rubin D. B. *Bayesian Data Analysis*. Chapman and Hall/CRC Press, third edition, 2014.

[48] Geman S. Geman D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984.

[49] Lunn D. J., Thomas A., Best N., Spiegelhalter D. WinBUGS — A Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10:325–337, 2000.

[50] Lunn D., Spiegelhalter D., Thomas A., Best N. The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, 28(25):3049–3067, 2009.

[51] Plummer M. A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, 2003.

[52] Hoffman M. D. Gelman A. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(47):1593–1623, 2014.

[53] Stan Development Team. Stan Modeling Language Users Guide and Reference Manual, version 2.34. Available from: https://mc-stan.org [accessed 2024-01-12], 2024.

[54] Stan Development Team. RStan: the R interface to Stan. Available from: http://mc-stan.org/ [accessed 2024-01-12], 2024. R package version 2.32.5.

[55] Papaspiliopoulos O., Roberts G. O., Sköld M. A General Framework for the Parametrization of Hierarchical Models. *Statistical Science*, 22(1):59–73, 2007.

[56] Betancourt M. Girolami M. *Current Trends in Bayesian Methodology with Applications*, chapter Hamiltonian Monte Carlo for Hierarchical Models, pages 79–100. CRC Press, 2015.

[57] Lunn D., Jackson C., Best N., Thomas A., Spiegelhalter D. *The BUGS Book: A Practical Introduction to Bayesian Analysis*. Chapman and Hall/CRC Press, 2012.

[58] Watanabe S. Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research*, 11(116):3571–3594, 2010.

[59] Gelman A., Hwang J., Vehtari A. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014.

[60] Sackett D. L. Haynes R. B. The architecture of diagnostic research. *BMJ*, 324(7336):539–541, 2002.

[61] Whiting P. F., Rutjes A. W. S., Westwood M. E., Mallett S. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. *Journal of Clinical Epidemiology*, 66(10):1093–1104, 2013.

[62] Lijmer J. G., Leeflang M., Bossuyt P. M. M. Proposals for a Phased Evaluation of Medical Tests. *Medical Decision Making*, 29(5):E13–E21, 2009.

[63] Horvath A. R., Lord S. J., StJohn A., Sandberg S., Cobbaert C. M., Lorenz S., Monaghan P. J., Verhagen-Kamerbeek W. D. J., Ebert C., Bossuyt P. M. M. From biomarkers to medical tests: The changing landscape of test evaluation. *Clinica Chimica Acta*, 427:49–57, 2014.

[64] Altman D. G. *Statistics with Confidence*, chapter Diagnostic tests, pages 105–119. BMJ Books, second edition, 2000.

[65] Trevethan R. Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice. *Frontiers in Public Health*, 5:307, 2017.

[66] Glas A. S., Lijmer J. G., Prins M. H., Bonsel G. J., Bossuyt P. M. M. The diagnostic odds ratio: a single indicator of test performance. *Journal of Clinical Epidemiology*, 56(11):1129–1135, 2003.

[67] Gabelica M., Bojčić R., Puljak L. Many researchers were not compliant with their published data sharing statement: a mixed-methods study. *Journal of Clinical Epidemiology*, 150:33–41, 2022.

[68] Honeyford K., Expert P., Mendelsohn E. E., Post B., Faisal A. A., Glampson B., Mayer E. K., Costelloe C. E. Challenges and recommendations for high quality research using electronic health records. *Frontiers in Digital Health*, 4:940330, 2022.

[69] Tang M. On simultaneous assessment of sensitivity and specificity when combining two diagnostic tests. *Statistics in Medicine*, 23(23):3593–3605, 2004.

[70] Sutton A. J., Abrams K. R., Jones D. R., Sheldon T. A., Song F. *Methods for Meta-Analysis in Medical Research*, chapter Random Effects Models for Combining Study Estimates, pages 73–86. Wiley, first edition, 2000.

[71] Sutton A. J., Abrams K. R., Jones D. R., Sheldon T. A., Song F. *Methods for Meta-Analysis in Medical Research*, chapter Fixed Effects Methods for Combining Study Estimates, pages 57–72. Wiley, first edition, 2000.

[72] DerSimonian R. Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188, 1986.

[73] Hamza T. H., van Houwelingen H. C., Stijnen T. The binomial distribution of meta-analysis was preferred to model within-study variability. *Journal of Clinical Epidemiology*, 61(1):41–51, 2008.

[74] Takwoingi Y., Guo B., Riley R. D., Deeks J. J. Performance of methods for meta-analysis of diagnostic test accuracy with few studies or sparse data. *Statistical Methods in Medical Research*, 26(4):1896–1911, 2017.

[75] Reitsma J. B., Glas A. S., Rutjes A. W. S., Scholten R. J. P. M., Bossuyt P. M., Zwinderman A. H. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology*, 58(10):982–990, 2005.

[76] Chu H. Cole S. R. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *Journal of Clinical Epidemiology*, 59(12):1331–1332, 2006.

[77] Rutter C. M. Gatsonis C. A. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine*, 20(19):2865–2884, 2001.

[78] Moses L. E., Shapiro D., Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: Data-analytic approaches and some additional considerations. *Statistics in Medicine*, 12(14):1293–1316, 1993.

[79] Littenberg B. Moses L. E. Estimating Diagnostic Accuracy from Multiple Conflicting Reports: A New Meta-analytic Method. *Medical Decision Making*, 13(4):313–321, 1993.

[80] Harbord R. M., Deeks J. J., Egger M., Whiting P., Sterne J. A. C. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics*, 8(2):239–251, 2007.

[81] Riley R. D., Abrams K. R., Sutton A. J., Lambert P. C., Thompson J. R. Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC Medical Research Methodology*, 7:3, 2007.

[82] Harbord R. M., Whiting P., Sterne J. A. C., Egger M., Deeks J. J., Shang A., Bachmann L. M. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. *Journal of Clinical Epidemiology*, 61(11):1095–1103, 2008.

[83] Dinnes J., Mallett S., Hopewell S., Roderick P. J., Deeks J. J. The Moses-Littenberg meta-analytical method generates systematic differences in test accuracy compared to hierarchical meta-analytical models. *Journal of Clinical Epidemiology*, 80:77–87, 2016.

[84] Takwoingi Y., Riley R. D., Deeks J. J. Meta-analysis of diagnostic accuracy studies in mental health. *Evidence Based Mental Health*, 18(4):103–109, 2015.

[85] Trikalinos T. A., Hoaglin D. C., Small K. M., Terrin N., Schmid C. H. Methods for the joint meta-analysis of multiple tests. *Research Synthesis Methods*, 5(4):294–312, 2014.

[86] Ouzzani M., Hammady H., Fedorowicz Z., Elmagarmid A. Rayyan - a web and mobile app for systematic reviews. *Systematic Reviews*, 5(210), 2016.

[87] Veroniki A. A., Tsokani S., Agarwal R., Pagkalidou E., Rücker G., Mavridis D., Takwoingi Y. Diagnostic test accuracy network meta-analysis methods: A scoping review and empirical assessment. *Journal of Clinical Epidemiology*, 146:86–96, 2022.

[88] Menten J. Lesaffre E. A general framework for comparative Bayesian meta-analysis of diagnostic studies. *BMC Medical Research Methodology*, 15:70, 2015.

[89] Nyaga V. N., Aerts M., Arbyn M. ANOVA model for network meta-analysis of diagnostic test accuracy data. *Statistical Methods in Medical Research*, 27(6):1766–1784, 2016.

[90] Nyaga V. N., Arbyn M., Aerts M. Beta-binomial analysis of variance model for network meta-analysis of diagnostic test accuracy data. *Statistical Methods in Medical Research*, 27(8):2554–2566, 2016.

[91] Dimou N. L., Adam M., Bagos P. G. A multivariate method for meta-analysis and comparison of diagnostic tests. *Statistics in Medicine*, 35(20):3509–3523, 2016.

[92] Cheng W. *Network meta-analysis of diagnostic accuracy studies.* Dissertation, Brown University, Providence (RI), 2016.

[93] Ma X., Lian Q., Chu H., Ibrahim J. G., Chen Y. A Bayesian hierarchical model for network meta-analysis of multiple diagnostic tests. *Biostatistics*, 19(1):87–102, 2018.

[94] Hoyer A. Kuss O. Meta-analysis for the comparison of two diagnostic tests - A new approach based on copulas. *Statistics in Medicine*, 37(5):739–748, 2018.

[95] Owen R. K., Cooper N. J., Quinn T. J., Lees R., Sutton A. J. Network meta-analysis of diagnostic test accuracy studies identifies and ranks the optimal diagnostic tests and thresholds for health care policy and decision-making. *Journal of Clinical Epidemiology*, 99:64–74, 2018.

[96] Hoyer A. Kuss O. Meta-analysis for the comparison of two diagnostic tests to a common gold standard: A generalized linear mixed model approach. *Statistical Methods in Medical Research*, 27(5):1410–1421, 2018.

[97] Lian Q., Hodges J. S., Chu H. A Bayesian Hierarchical Summary Receiver Operating Characteristic Model for Network Meta-Analysis of Diagnostic Tests. *Journal of the American Statistical Association*, 114(527):949–961, 2019.

[98] Nikoloulopoulos A. K. A D-vine copula mixed model for joint meta-analysis and comparison of diagnostic tests. *Statistical Methods in Medical Research*, 28(10-11):3286–3300, 2019.

[99] Kuss O., Hoyer A., Solms A. Meta-analysis for diagnostic accuracy studies: a new statistical model using beta-binomial distributions and bivariate copulas. *Statistics in Medicine*, 33(1):17–30, 2014.

[100] Hoyer A. Kuss O. Meta-analysis of diagnostic tests accounting for disease prevalence: a new model using trivariate copulas. *Statistics in Medicine*, 34(11):1912–1924, 2015.

[101] Nikoloulopoulos A. K. A vine copula mixed effect model for trivariate meta-analysis of diagnostic test accuracy studies accounting for disease prevalence. *Statistical Methods in Medical Research*, 26(5):2270–2286, 2015.

[102] NHS Digital. Recorded Dementia Diagnoses, September 2022. Available from: https://digital.nhs.uk/data-and-information/publications/statistical/recorded-dementia-diagnoses/september-2022 [accessed 2024-01-12], 2022.

[103] Arevalo-Rodriguez I., Smailagic N., Roqué-Figuls M., Ciapponi A., Sanchez-Perez E., Giannakou A., Pedraza O. L., Cosp X. B., Cullum S. Mini-Mental State Examination (MMSE) for the early detection of dementia in people with mild cognitive impairment (MCI). *Cochrane Database of Systematic Reviews*, 7:CD010783, 2021.

[104] Chan C. C. H., Fage B. A., Burton J. K., Smailagic N., Gill S. S., Herrmann N., Nikolaou V., Quinn T. J., Noel-Storr A. H., Seitz D. P. Mini-Cog for the detection of dementia within a secondary care setting. *Cochrane Database of Systematic Reviews*, 7:CD011414, 2021.

[105] Creavin S. T., Wisniewski S., Noel-Storr A. H., Trevelyan C. M., Hampton T., Rayment D., Thom V. M., Nash K. J. E., Elhamoui H., Milligan R., et al. Mini-Mental State Examination (MMSE) for the detection of dementia in clinically unevaluated people aged 65 and over in community and primary care populations. *The Cochrane Database of Systematic Reviews*, 1:CD011145, 2016.

[106] Davis D. H. J., Creavin S. T., Yip J. L. Y., Noel-Storr A. H., Brayne C., Cullum S. Montreal Cognitive Assessment for the detection of dementia. *Cochrane Database of Systematic Reviews*, 7:CD010775, 2021.

[107] Fage B. A., Chan C. C. H., Gill S. S., Noel-Storr A. H., Herrmann N., Smailagic N., Nikolaou V., Seitz D. P. Mini-Cog for the detection of dementia within a community setting. *Cochrane Database of Systematic Reviews*, 7:CD010860, 2021.

[108] Seitz D. P., Chan C. C., Newton H. T., Gill S. S., Herrmann N., Smailagic N., Nikolaou V., Fage B. A. Mini-Cog for the diagnosis of Alzheimer's disease dementia and other dementias within a primary care setting. *The Cochrane Database of Systematic Reviews*, 2:CD011415, 2021.

[109] Teunissen C. E., Verberk I. M. W., Thijssen E. H., Vermunt L., Hansson O., Zetterberg H., van der Flier W. M., Mielke M. M., del Campo M. Blood-based biomarkers for Alzheimer's disease: towards clinical implementation. *The Lancet Neurology*, 21(1):66–77, 2022.

[110] Hansson O., Lehmann S., Otto M., Zetterberg H., Lewczuk P. Advantages and disadvantages of the use of the CSF Amyloid β (AB) 42/40 ratio in the diagnosis of Alzheimer's Disease. *Alzheimer's Research & Therapy*, 11(1):34, 2019.

[111] Hansson O., Batrla R., Brix B., Carrillo M. C., Corradini V., Edelmayer R. M., Esquivel R. N., Hall C., Lawson J., Bastard N. L., Molinuevo J. L., Nisenbaum L. K., Rutz S., Salamone S. J., Teunissen C. E., Traynham C., Umek R. M., Vanderstichele H., Vandijck M., Wahl S., Weber C. J., Zetterberg H., Blennow K. The Alzheimer's Association international guidelines for handling of cerebrospinal fluid for routine clinical measurements of amyloid β and tau. *Alzheimer's & Dementia*, 17(9):1575–1582, 2021.

[112] Bloom G. S. Amyloid-β and Tau: The Trigger and Bullet in Alzheimer Disease Pathogenesis. *JAMA Neurology*, 71(4):505–508, 2014.

[113] Hansson O., Seibyl J., Stomrud E., Zetterberg H., Trojanowski J. Q., Bittner T., Lifke V., Corradini V., Eichenlaub U., Batrla R., Buck K., Zink K., Rabe C., Blennow K., Shaw L. M. CSF biomarkers of Alzheimer's disease concord with amyloid-β PET and predict clinical progression: A study of fully automated immunoassays in BioFINDER and ADNI cohorts. *Alzheimer's & Dementia*, 14(11):1470–1481, 2018.

[114] McInnes M. D. F., Moher D., Thombs B. D., McGrath T. A., Bossuyt P. M., Clifford T., Cohen J. F., Deeks J. J., Gatsonis C., Hooft L., Hunt H. A., Hyde C. J., Korevaar D. A., Leeflang M. M. G., Macaskill P., Reitsma J. B., Rodin R., Rutjes A. W. S., Salameh J.-P., Stevens A., Takwoingi Y., Tonelli M., Weeks L., Whiting P., Willis B. H. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. *JAMA*, 319(4):388–396, 2018.

[115] Archer H. A., Smailagic N., John C., Holmes R. B., Takwoingi Y., Coulthard E. J., Cullum S. Regional Cerebral Blood Flow Single Photon Emission Computed Tomography for detection of Frontotemporal dementia in people with

suspected dementia. *Cochrane Database of Systematic Reviews*, 6:CD010896, 2015.

[116] McCleery J., Morgan S., Bradley K. M., Noel-Storr A. H., Ansorge O., Hyde C. Dopamine transporter imaging for the diagnosis of dementia with Lewy bodies. *Cochrane Database of Systematic Reviews*, 1:CD010633, 2015.

[117] Creavin S. T., Noel-Storr A. H., Langdon R. J., Richard E., Creavin A. L., Cullum S., Purdy S., Ben-Shlomo Y. Clinical judgement by primary care physicians for the diagnosis of all-cause dementia or cognitive impairment in symptomatic people. *Cochrane Database of Systematic Reviews*, 6:CD012558, 2022.

[118] Kokkinou M., Beishon L. C., Smailagic N., Noel-Storr A. H., Hyde C., Ukoumunne O., Worrall R. E., Hayen A., Desai M., Ashok A. H., Paul E. J., Georgopoulou A., Casoli T., Quinn T. J., Ritchie C. W. Plasma and cerebrospinal fluid ABeta42 for the differential diagnosis of Alzheimer's disease dementia in participants diagnosed with any dementia subtype in a specialist care setting. *Cochrane Database of Systematic Reviews*, 2:CD010945, 2021.

[119] Lombardi G., Crescioli G., Cavedo E., Lucenteforte E., Casazza G., Bellatorre A.-G., Lista C., Costantino G., Frisoni G., Virgili G., Filippini G. Structural magnetic resonance imaging for the early diagnosis of dementia due to Alzheimer's disease in people with mild cognitive impairment. *Cochrane Database of Systematic Reviews*, 3:CD009628, 2020.

[120] Martínez G., Vernooij R. W., Fuentes Padilla P., Zamora J., Flicker L., Bonfill Cosp X. 18F PET with florbetaben for the early diagnosis of Alzheimer's disease dementia and other dementias in people with mild cognitive impairment (MCI). *The Cochrane Database of Systematic Reviews*, 11:CD012883, 2017.

[121] Martínez G., Vernooij R. W., Fuentes Padilla P., Zamora J., Flicker L., Bonfill Cosp X. 18F PET with flutemetamol for the early diagnosis of Alzheimer's disease dementia and other dementias in people with mild cognitive impairment (MCI). *The Cochrane Database of Systematic Reviews*, 11:CD012884, 2017.

[122] Martínez G., Vernooij R. W., Fuentes Padilla P., Zamora J., Bonfill Cosp X., Flicker L. 18F PET with florbetapir for the early diagnosis of Alzheimer's

disease dementia and other dementias in people with mild cognitive impairment (MCI). *The Cochrane Database of Systematic Reviews*, 11:CD012216, 2017.

[123] Smailagic N., Vacante M., Hyde C., Martin S., Ukoumunne O., Sachpekidis C. ^{18}F-FDG PET for the early diagnosis of Alzheimer's disease dementia and other dementias in people with mild cognitive impairment (MCI). *The Cochrane Database of Systematic Reviews*, 1:CD010632, 2015.

[124] Zhang S., Smailagic N., Hyde C., Noel-Storr A. H., Takwoingi Y., McShane R., Feng J. ^{11}C-PIB-PET for the early diagnosis of Alzheimer's disease dementia and other dementias in people with mild cognitive impairment (MCI). *The Cochrane Database of Systematic Reviews*, 7:CD010386, 2014.

[125] Petersen R. C. Mild cognitive impairment as a diagnostic entity. *Journal of Internal Medicine*, 256(3):183–194, 2004.

[126] Tschanz J. T., Welsh-Bohmer K. A., Lyketsos C. G., Corcoran C., Green R. C., Hayden K., Norton M. C., Zandi P. P., Toone L., West N. A., Breitner J. C. S. Conversion to dementia from mild cognitive disorder: the Cache County Study. *Neurology*, 67(2):229–234, 2006.

[127] Anchisi D., Borroni B., Franceschi M., Kerrouche N., Kalbe E., Beuthien-Beumann B., Cappa S., Lenz O., Ludecke S., Marcone A., Mielke R., Ortelli P., Padovani A., Pelati O., Pupi A., Scarpini E., Weisenbach S., Herholz K., Salmon E., Holthoff V., Sorbi S., Fazio F., Perani D. Heterogeneity of Brain Glucose Metabolism in Mild Cognitive Impairment and Clinical Progression to Alzheimer Disease. *Archives of Neurology*, 62(11):1728–1733, 2005.

[128] Bjerke M., Andreasson U., Rolstad S., Nordlund A., Lind K., Zetterberg H., Edman Å., Blennow K., Wallin A. Subcortical Vascular Dementia Biomarker Pattern in Mild Cognitive Impairment. *Dementia and Geriatric Cognitive Disorders*, 28(4):348–356, 2009.

[129] Blom E. S., Giedraitis V., Zetterberg H., Fukumoto H., Blennow K., Hyman B. T., Irizarry M. C., Wahlund L.-O., Lannfelt L., Ingelsson M. Rapid Progression from Mild Cognitive Impairment to Alzheimer's Disease in Subjects with Elevated Levels of Tau in Cerebrospinal Fluid and the APOE $\epsilon 4/\epsilon 4$ Genotype. *Dementia and Geriatric Cognitive Disorders*, 27(5):458–464, 2009.

[130] Borson S., Scanlan J. M., Watanabe J., Tu S.-P., Lessig M. Simplifying Detection of Cognitive Impairment: Comparison of the Mini-Cog and Mini-Mental State Examination in a Multiethnic Sample. *Journal of the American Geriatrics Society*, 53(5):871–874, 2005.

[131] Brettschneider J., Petzold A., Schöttle D., Claus A., Riepe M., Tumani H. The Neurofilament Heavy Chain (NfH^{SMI35}) in the Cerebrospinal Fluid Diagnosis of Alzheimer's Disease. *Dementia and Geriatric Cognitive Disorders*, 21(5-6):291–295, 2006.

[132] Brys M., Pirraglia E., Rich K., Rolstad S., Mosconi L., Switalski R., Glodzik-Sobanska L., Santi S. D., Zinkowski R., Mehta P., Pratico D., Louis L. A. S., Wallin A., Blennow K., de Leon M. J. Prediction and longitudinal study of CSF biomarkers in mild cognitive impairment. *Neurobiology of Aging*, 30(5):682–690, 2009.

[133] Buchhave P., Stomrud E., Warkentin S., Blennow K., Minthon L., Hansson O. Cube Copying Test in Combination with rCBF or CSF Aβ_{42} Predicts Development of Alzheimer's Disease. *Dementia and Geriatric Cognitive Disorders*, 25(6):544–552, 2008.

[134] Carmichael O. T., Kuller L. H., Lopez O. L., Thompson P. M., Dutton R. A., Lu A., Lee S. E., Lee J. Y., Aizenstein H. J., Meltzer C. C., Liu Y., Toga A. W., Becker J. T. Cerebral Ventricular Changes Associated With Transitions Between Normal Cognitive Function, Mild Cognitive Impairment, and Dementia. *Alzheimer Disease & Associated Disorders*, 21(1):14–24, 2007.

[135] Caroli A., Testa C., Geroldi C., Nobili F., Barnden L. R., Guerra U. P., Bonetti M., Frisoni G. B. Cerebral perfusion correlates of conversion to Alzheimer's disease in amnestic mild cognitive impairment. *Journal of Neurology*, 254(12):1698–1707, 2007.

[136] Chiasserini D., Parnetti L., Andreasson U., Zetterberg H., Giannandrea D., Calabresi P., Blennow K. CSF Levels of Heart Fatty Acid Binding Protein are Altered During Early Phases of Alzheimer's Disease. *Journal of Alzheimer's Disease*, 22(4):1281–1288, 2010.

[137] Clerx L., van Rossum I. A., Burns L., Knol D. L., Scheltens P., Verhey F., Aalten P., Lapuerta P., van de Pol L., van Schijndel R., de Jong R., Barkhof F., Wolz R., Rueckert D., Bocchetta M., Tsolaki M., Nobili F., Wahlund L.-O.,

Minthon L., Frölich L., Hampel H., Soininen H., Visser P. J. Measurements of medial temporal lobe atrophy for prediction of Alzheimer's disease in subjects with mild cognitive impairment. *Neurobiology of Aging*, 34(8):2003–2013, 2013.

[138] deToledo Morrell L., Stoub T. R., Bulgakova M., Wilson R. S., Bennett D. A., Leurgans S., Wuu J., Turner D. A. MRI-derived entorhinal volume is a good predictor of conversion from MCI to AD. *Neurobiology of Aging*, 25(9):1197–1203, 2004.

[139] Devanand D. P., Pradhaban G., Liu X., Khandji A., De Santi S., Segal S., Rusinek H., Pelton G. H., Honig L. S., Mayeux R., Stern Y., Tabert M. H., de Leon M. J. Hippocampal and entorhinal atrophy in mild cognitive impairment: Prediction of Alzheimer disease. *Neurology*, 68(11):828–836, 2007.

[140] Devanand D. P., Liu X., Tabert M. H., Pradhaban G., Cuasay K., Bell K., de Leon M. J., Doty R. L., Stern Y., Pelton G. H. Combining Early Markers Strongly Predicts Conversion from Mild Cognitive Impairment to Alzheimer's Disease. *Biological Psychiatry*, 64(10):871–879, 2008.

[141] Eckerström C., Olsson E., Bjerke M., Malmgren H., Edman r., Wallin A., Nordlund A. A Combination of Neuropsychological, Neuroimaging, and Cerebrospinal Fluid Markers Predicts Conversion from Mild Cognitive Impairment to Dementia. *Journal of Alzheimer's Disease*, 36(3):421–431, 2013.

[142] Erten-Lyons D., Howieson D., Moore M. M., Quinn J., Sexton G., Silbert L., Kaye J. Brain volume loss in MCI predicts dementia. *Neurology*, 66(2):233–235, 2006.

[143] Fei M., Jianghua W., Rujuan M., Wei Z., Qian W. The relationship of plasma Aβ levels to dementia in aging individuals with mild cognitive impairment. *Journal of the Neurological Sciences*, 305(1-2):92–96, 2011.

[144] Fellgiebel A., Scheurich A., Bartenstein P., Müller M. J. FDG-PET and CSF phospho-tau for prediction of cognitive decline in mild cognitive impairment. *Psychiatry Research: Neuroimaging*, 155(2):167–171, 2007.

[145] Frölich L., Peters O., Lewczuk P., Gruber O., Teipel S. J., Gertz H. J., Jahn H., Jessen F., Kurz A., Luckhaus C., Huell M., Pantel J., Reischies F. M., Schroeder J., Wagner M., Rienhoff O., Wolf S., Bauer C., Schuchhardt J.,

Heuser I., Rüther E., Henn F., Maier W., Wiltfang J., Kornhuber J. Incremental value of biomarker combinations to predict progression of mild cognitive impairment to Alzheimer's dementia. *Alzheimer's Research & Therapy*, 9:84, 2017.

[146] Galluzzi S., Geroldi C., Ghidoni R., Paghera B., Amicucci G., Bonetti M., Zanetti O., Cotelli M., Gennarelli M., Frisoni G. B. The new Alzheimer's criteria in a naturalistic series of patients with mild cognitive impairment. *Journal of Neurology*, 257(12):2004–2014, 2010.

[147] Galton C. J., Erzinçlioglu S., Sahakian B. J., Antoun N., Hodges J. R. A Comparison of the Addenbrooke's Cognitive Examination (ACE), Conventional Neuropsychological Assessment, and Simple MRI-Based Medial Temporal Lobe Evaluation in the Early Diagnosis of Alzheimer's Disease. *Cognitive and Behavioral Neurology*, 18(3):144–150, 2005.

[148] Gaser C., Franke K., Klöppel S., Koutsouleris N., Sauer H. *BrainAGE* in Mild Cognitive Impaired Patients: Predicting the Conversion to Alzheimer's Disease. *PLoS ONE*, 8(6):e67346, 2013.

[149] Hampel H., Teipel S. J., Fuchsberger T., Andreasen N., Wiltfang J., Otto M., Shen Y., Dodel R., Du Y., Farlow M., Mø"ller H.-J., Blennow K., Buerger K. Value of CSF β-amyloid$_{1–42}$ and tau as predictors of Alzheimer's disease in patients with mild cognitive impairment. *Molecular Psychiatry*, 9(7):705–710, 2004.

[150] Hansson O., Zetterberg H., Buchhave P., Londos E., Blennow K., Minthon L. Association between CSF biomarkers and incipient Alzheimer's disease in patients with mild cognitive impairment: a follow-up study. *The Lancet Neurology*, 5(3):228–234, 2006.

[151] Hansson O., Zetterberg H., Buchhave P., Andreasson U., Londos E., Minthon L., Blennow K. Prediction of Alzheimer's Disease Using the CSF Aβ42/Aβ40 Ratio in Patients with Mild Cognitive Impairment. *Dementia and Geriatric Cognitive Disorders*, 23(5):316–320, 2007.

[152] Hertze J., Minthon L., Zetterberg H., Vanmechelen E., Blennow K., Hansson O. Evaluation of CSF Biomarkers as Predictors of Alzheimer's Disease: A Clinical Follow-Up Study of 4.7 Years. *Journal of Alzheimer's Disease*, 21(4):1119–1128, 2010.

[153] Herukka S. K., Pennanen C., Soininen H., Pirttila T. CSF Aβ42, Tau and Phosphorylated Tau Correlate with Medial Temporal Lobe Atrophy. *Journal of Alzheimer's Disease*, 14(1):51–57, 2008.

[154] Jang J.-W., Park J. H., Kim S., Park Y. H., Pyun J.-M., Lim J.-S., Kim Y., Youn Y. C., Kim S. A 'Comprehensive Visual Rating Scale' for predicting progression to dementia in patients with mild cognitive impairment. *PLoS ONE*, 13(8):e0201852, 2018.

[155] Kapaki E., Paraskevas G. P., Zalonis I., Zournas C. CSF tau protein and β-amyloid (1-42) in Alzheimer's disease diagnosis: discrimination from normal ageing and other dementias in the Greek population. *European Journal of Neurology*, 10(2):119–128, 2003.

[156] Kester M. I., Verwey N. A., van Elk EJ, Blankenstein M. A., Scheltens P., van der Flier WM. Progression from MCI to AD: Predictive value of CSF Aβ42 is modified by APOE genotype. *Neurobiology of Aging*, 32(8):1372–13778, 2011.

[157] Knapskog A.-B., Braekhus A., Engedal K. The Effect of Changing the Amyloid $abeta_{42}$ Cut-off of Cerebrospinal Fluid Biomarkers on Alzheimer Disease Diagnosis in a Memory Clinic Population in Norway. *Alzheimer Disease & Associated Disorders*, 33(1):72–74, 2019.

[158] Koivunen J., Pirttilä T., Kemppainen N., Aalto S., Herukka S.-K., Jauhianen A. M., Hänninen T., Hallikainen M., Någren K., Rinne J. O., Soininen H. PET Amyloid Ligand [^{11}C]PIB Uptake and Cerebrospinal Fluid β-Amyloid in Mild Cognitive Impairment. *Dementia and Geriatric Cognitive Disorders*, 26(4):378–383, 2008.

[159] Lee J.-Y., Lee D. W., Cho S.-J., Na D. L., Jeon H. J., Kim S.-K., Lee Y. R., Youn J.-H., Kwon M., Lee J.-H., Cho M. J. Brief Screening for Mild Cognitive Impairment in Elderly Outpatient Clinic: Validation of the Korean Version of the Montreal Cognitive Assessment. *Journal of Geriatric Psychiatry and Neurology*, 21(2):104–110, 2008.

[160] Ledig C., Schuh A., Guerrero R., Heckemann R. A., Rueckert D. Structural brain imaging in Alzheimer's disease and mild cognitive impairment: biomarker analysis and shared morphometry database. *Scientific Reports*, 8:11258, 2018.

[161] Lewczuk P., Esselmann H., Otto M., Maler J. M., Henkel A. W., Henkel M. K., Eikenberg O., Antz C., Krause W.-R., Reulbach U., Kornhuber J., Wiltfang J. Neurochemical diagnosis of Alzheimer's dementia by CSF Aβ42, Aβ42/Aβ40 ratio and total tau. *Neurobiology of Aging*, 25(3):273–281, 2004.

[162] Maddalena A., Papassotiropoulos A., Müller-Tillmanns B., Jung H. H., Hegi T., Nitsch R. M., Hock C. Biochemical Diagnosis of Alzheimer Disease by Measuring the Cerebrospinal Fluid Ratio of Phosphorylated tau Protein to β-Amyloid Peptide42. *Archives of Neurology*, 60(9):1202–1206, 2003.

[163] Milian M., Leiherr A.-M., Straten G., Müller S., Leyhe T., Eschweiler G. W. The Mini-Cog versus the Mini-Mental State Examination and the Clock Drawing Test in daily clinical practice: screening value in a German Memory Clinic. *International Psychogeriatrics*, 24(5):766–774, 2012.

[164] Modrego P. J., Fayed N., Pina M. A. Conversion From Mild Cognitive Impairment to Probable Alzheimer's Disease Predicted by Brain Magnetic Resonance Spectroscopy. *American Journal of Psychiatry*, 162(4):667–675, 2005.

[165] Modrego P. J. Gazulla J. The Predictive Value of the Memory Impairment Screen in Patients With Subjective Memory Complaints. *The Primary Care Companion For CNS Disorders*, 15(1):12m01435, 2013.

[166] Monge-Argilés J. A., Muñoz-Ruiz C., Pampliega-Pérez A., Gómez-López M. J., Sánchez-Payá J., Borja E. R., Ruiz-Vegara M., Montoya-Gutiérrez F. J., Leiva-Santana C. Biomarkers of Alzheimer's Disease in the Cerebrospinal Fluid of Spanish Patients With Mild Cognitive Impairment. *Neurochemical Research*, 36(6):986–993, 2011.

[167] Montine T. J., Kaye J. A., Montine K. S., McFarland L., Morrow J. D., Quinn J. F. Cerebrospinal Fluid Aβ42, Tau, and F_2-Isoprostane Concentrations in Patients With Alzheimer Disease, Other Dementias, and in Age-Matched Controls. *Archives of Pathology & Laboratory Medicine*, 125(4):510–512, 2001.

[168] Mosconi L., Perani D., Sorbi S., Herholz K., Nacmias B., Holthoff V., Salmon E., Baron J.-C., Cristofaro M. T. R. D., Padovani A., Borroni B., Franceschi M., Bracco L., Pupi A. MCI conversion to dementia and the APOE genotype: A prediction study with FDG-PET. *Neurology*, 63(12):2332–2340, 2004.

[169] Nesteruk M., Nesteruk T., Styczyńska M., Mandecka M., Barczak A., Barcikowska M. Combined use of biochemical and volumetric biomarkers to assess the risk of conversion of mild cognitive impairment to Alzheimer's disease. *Folia Neuropathologica*, 54(4):369–374, 2016.

[170] Nobili F., Salmaso D., Morbelli S., Girtler N., Piccardo A., Brugnolo A., Dessi B., Larsson S. A., Rodriguez G., Pagani M. Principal component analysis of FDG PET in amnestic MCI. *European Journal of Nuclear Medicine and Molecular Imaging*, 35(12):2191–2202, 2008.

[171] Ong K. T., Villemagne V. L., Bahar-Fuchs A., Lamb F., Langdon N., Catafau A. M., Stephens A. W., Seibyl J., Dinkelborg L. M., Reininger C. B., Putz B., Rohde B., Masters C. L., Rowe C. C. Aβ imaging with 18F-florbetaben in prodromal Alzheimer's disease: a prospective outcome study. *Journal of Neurology, Neurosurgery & Psychiatry*, 86(4):431–436, 2015.

[172] Ossenkoppele R., Tolboom N., Foster-Dingley J. C., Adriaanse S. F., Boellaard R., Yaqub M., Windhorst A. D., Barkhof F., Lammertsma A. A., Scheltens P., van der Flier W. M., van Berckel B. N. M. Longitudinal imaging of Alzheimer pathology using [^{11}C]PIB, [^{18}F]FDDNP and [^{18}F]FDG PET. *European Journal of Nuclear Medicine and Molecular Imaging*, 39(6):990–1000, 2012.

[173] Ossenkoppele R., Prins N. D., Pijnenburg Y. A. L., Lemstra A. W., van der Flier W. M., Adriaanse S. F., Windhorst A. D., Handels R. L. H., Wolfs C. A. G., Aalten P., Verhey F. R. J., Verbeek M. M., van Buchem M. A., Hoekstra O. S., Lammertsma A. A., Scheltens P., van Berckel B. N. M. Impact of molecular imaging on the diagnostic process in a memory clinic. *Alzheimer's & Dementia*, 9(4):414–421, 2012.

[174] Palmqvist S., Hertze J., Minthon L., Wattmo C., Zetterberg H., Blennow K., Londos E., Hansson O. Comparison of Brief Cognitive Tests and CSF Biomarkers in Predicting Alzheimer's Disease in Mild Cognitive Impairment: Six-Year Follow-Up Study. *PLoS ONE*, 7(6):e38639, 2012.

[175] Parnetti L., Lanari A., Silvestrelli G., Saggese E., Reboldi P. Diagnosing prodromal Alzheimer's disease: role of CSF biochemical markers. *Mechanisms of Ageing and Development*, 127(2):129–132, 2006.

[176] Parnetti L., Chiasserini D., Eusebi P., Giannandrea D., Bellomo G., Carlo C. D., Padiglioni C., Mastrocola S., Lisetti V., Calabresi P. Performance

of A$\beta_{1\text{-}40}$, A$\beta_{1\text{-}42}$, Total Tau, and Phosphorylated Tau as Predictors of Dementia in a Cohort of Patients with Mild Cognitive Impairment. *Journal of Alzheimer's Disease*, 29(1):229–238, 2012.

[177] Perani D., Cerami C., Caminiti S. P., Santangelo R., Coppi E., Ferrari L., Pinto P., Passerini G., Falini A., Iannaccone S., Cappa S. F., Comi G., Gianolli L., Magnani G. Cross-validation of biomarkers for the early differential diagnosis and prognosis of dementia in a clinical setting. *European Journal of Nuclear Medicine and Molecular Imaging*, 43(3):499–508, 2016.

[178] Pozueta A., Rodriguez-Rodriguez E., Vazquez-Higuera J. L., Mateo I., Sanchez-Juan P., Gonzalez-Perez S., Berciano J., Combarros O. Detection of early Alzheimer's disease in MCI patients by the combination of MMSE and an episodic memory test. *BMC Neurology*, 11:78, 2011.

[179] Prestia A., Caroli A., van der Flier W. M., Ossenkoppele R., Berckel B. V., Barkhof F., Teunissen C. E., Wall A. E., Carter S. F., Scholl M., Choo I. H., Nordberg A., Scheltens P., Frisoni G. B. Prediction of dementia in MCI patients based on core diagnostic markers for Alzheimer disease. *Neurology*, 80(11):1048–1056, 2013.

[180] Prestia A., Caroli A., Herholz K., Reiman E., Chen K., Jagust W. J., Frisoni G. B. Diagnostic accuracy of markers for prodromal Alzheimer's disease in independent clinical series. *Alzheimer's & Dementia*, 9(6):677–686, 2013.

[181] Rhodius-Meester H. F. M., Koikkalainen J., Mattila J., Teunissen C. E., Barkhof F., Lemstra A. W., Scheltens P., Lötjönen J., van der Flier W. Integrating Biomarkers for Underlying Alzheimer's Disease in Mild Cognitive Impairment in Daily Practice: Comparison of a Clinical Decision Support System with Individual Biomarkers. *Journal of Alzheimer's Disease*, 50(1):261–270, 2016.

[182] Rosler N., Wichart I., Jellinger K. A. Clinical significance of neurobiochemical profiles in the lumbar cerebrospinal fluid of Alzheimer's disease patients. *Journal of Neural Transmission*, 108(2):231–246, 2001.

[183] Schreiber S., Landau S. M., Fero A., Schreiber F., Jagust W. J. Comparison of Visual and Quantitative Florbetapir F 18 Positron Emission Tomography Analysis in Predicting Mild Cognitive Impairment Outcomes. *JAMA Neurology*, 72(10):1183–1190, 2015.

[184] Smach M. A., Charfeddine B., Lammouchi T., Harrabi I., Othman L. B., Dridi H., Bennamou S., Limem K. CSF β-amyloid 1-42 and tau in Tunisian patients with Alzheimer's disease: The effect of APOE ϵ4 allele. *Neuroscience Letters*, 440(2):145–149, 2008.

[185] Spies P. E., Slats D., Sjogren J. M. C., Kremer B. P. H., Verhey F. R. J., Rikkert M. G. M. O., Verbeek M. M. The Cerebrospinal Fluid Amyloid β42/40 Ratio in the Differentiation of Alzheimers Disease from Non-Alzheimers Dementia. *Current Alzheimer Research*, 7(5):470–476, 2010.

[186] Tapiola T. Relationship between apoE genotype and CSF β-amyloid (1-42) and tau in patients with probable and definite Alzheimer's disease. *Neurobiology of Aging*, 21(5):735–740, 2000.

[187] Tariciotti L., Casadei M., Honig L. S., Teich A. F., Guy M. McKhann I. I., Tosto G., Mayeux R. Clinical Experience with Cerebrospinal Fluid Aβ_{42}, Total and Phosphorylated Tau in the Evaluation of 1,016 Individuals for Suspected Dementia. *Journal of Alzheimer's Disease*, 65(4):1417–1425, 2018.

[188] Thurfjell L., Lötjönen J., Lundqvist R., Koikkalainen J., Soininen H., Waldemar G., Brooks D. J., Vandenberghe R. Combination of Biomarkers: PET [^{18}F]Flutemetamol Imaging and Structural MRI in Dementia and Mild Cognitive Impairment. *Neurodegenerative Diseases*, 10(1-4):246–249, 2012.

[189] van der Flier W. M., van der Vlies A. E., Weverling-Rijnsburger A. W. E., de Boer N. L., Admiraal-Behloul F., Bollen E. L. E. M., Westendorp R. G. J., van Buchem M. A., Middelkoop H. A. M. MRI measures and progression of cognitive decline in nondemented elderly attending a memory clinic. *International Journal of Geriatric Psychiatry*, 20(11):1060–1066, 2005.

[190] Visser P. J., Scheltens P., Verhey F. R. J., Schmand B., Launer L. J., Jolles J., Jonker C. Medial temporal lobe atrophy and memory dysfunction as predictors for dementia in subjects with mild cognitive impairment. *Journal of Neurology*, 246(6):477–485, 1999.

[191] Visser P. J., Verhey F. R. J., Hofman P. A. M., Scheltens P., Jolles J. Medial temporal lobe atrophy predicts Alzheimer's disease in patients with minor cognitive impairment. *Journal of Neurology, Neurosurgery & Psychiatry*, 72(4):491–497, 2002.

[192] Visser P. J., Verhey F., Knol D. L., Scheltens P., Wahlund L.-O., Freund-Levi Y., Tsolaki M., Minthon L., Wallin Å. K., Hampel H., Bürger K., Pirttila T., Soininen H., Rikkert M. O., Verbeek M. M., Spiru L., Blennow K. Prevalence and prognostic value of CSF markers of Alzheimer's disease pathology in patients with subjective cognitive impairment or mild cognitive impairment in the DESCRIPA study: a prospective cohort study. *The Lancet Neurology*, 8(7):619–627, 2009.

[193] Vos S. J. B., van Rossum I. A., Verhey F., Knol D. L., Soininen H., Wahlund L.-O., Hampel H., Tsolaki M., Minthon L., Frisoni G. B., Froelich L., Nobili F., van der Flier W., Blennow K., Wolz R., Scheltens P., Visser P. J. Prediction of Alzheimer disease in subjects with amnestic and nonamnestic MCI. *Neurology*, 80(12):1124–1132, 2013.

[194] Wang L., Miller J. P., Gado M. H., McKeel D. W., Rothermich M., Miller M. I., Morris J. C., Csernansky J. G. Abnormalities of hippocampal surface structure in very mild dementia of the Alzheimer type. *NeuroImage*, 30(1):52–60, 2006.

[195] Wood R. A., Moodley K. K., Lever C., Minati L., Chan D. Allocentric Spatial Memory Testing Predicts Conversion from Mild Cognitive Impairment to Dementia: An Initial Proof-of-Concept Study. *Frontiers in Neurology*, 7:215, 2016.

[196] Xu G., Meyer J. S., Thornby J., Chowdhury M., Quach M. Screening for mild cognitive impairment (MCI) utilizing combined mini-mental-cognitive capacity examinations for identifying dementia prodromes. *International Journal of Geriatric Psychiatry*, 17(11):1027–1033, 2002.

[197] Waffenschmidt S., Knelangen M., Sieben W., Bühn S., Pieper D. Single screening versus conventional double screening for study selection in systematic reviews: a methodological systematic review. *BMC Medical Research Methodology*, 19(1), 2019.

[198] Petersen H., Poon J., Poon S. K., Loy C. Increased Workload for Systematic Review Literature Searches of Diagnostic Tests Compared With Treatments: Challenges and Opportunities. *JMIR Medical Informatics*, 2(1):e11, 2014.

[199] Korevaar D. A., van Enst W. A., Spijker R., Bossuyt P. M. M., Hooft L. Reporting quality of diagnostic accuracy studies: a systematic review and meta-

analysis of investigations on adherence to STARD. *Evidence Based Medicine*, 19(2):47–54, 2013.

[200] Macaskill P., Walter S. D., Irwig L., Franco E. L. Assessing the gain in diagnostic performance when combining two diagnostic tests. *Statistics in Medicine*, 21(17):2527–2546, 2002.

[201] Burton A., Altman D. G., Royston P., Holder R. L. The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24):4279–4292, 2006.

[202] Morris T. P., White I. R., Crowther M. J. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102, 2019.

[203] R Core Team. R: A language and environment for statistical computing, version 4.3.1. Available from: https://www.R-project.org [accessed 2024-01-12], 2023.

[204] Bachmann L. M., Puhan M. A., Riet G. t., Bossuyt P. M. Sample sizes of studies on diagnostic accuracy: literature survey. *BMJ*, 332(7550):1127–1129, 2006.

[205] Riley R. D., Thompson J. R., Abrams K. R. An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. *Biostatistics*, 9(1):172–186, 2008.

[206] Rosenberger K. J., Chu H., Lin L. Empirical comparisons of meta-analysis methods for diagnostic studies: a meta-epidemiological study. *BMJ Open*, 12(5):e055336, 2022.

[207] Papanikos T., Thompson J. R., Abrams K. R., Bujkiewicz S. Use of copula to model within-study association in bivariate meta-analysis of binomial data at the aggregate level: A Bayesian approach and application to surrogate endpoint evaluation. *Statistics in Medicine*, 41(25):4961–4981, 2022.

[208] Cohen J. F., Korevaar D. A., Altman D. G., Bruns D. E., Gatsonis C. A., Hooft L., Irwig L., Levine D., Reitsma J. B., de Vet H. C. W., Bossuyt P. M. M. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open*, 6(11):e012799, 2016.

[209] Bujkiewicz S., Achana F., Papanikos T., Riley R. D., Abrams K. R. NICE DSU Technical Support Document 20: Multivariate meta-analysis of summary data for combining treatment effects on correlated outcomes and evaluating surrogate endpoints. Available from: https://www.sheffield.ac.uk/nice-dsu/tsds/multivariate-meta-analysis [accessed 2024-01-12], 2019.

[210] Takwoingi Y. Quinn T. J. Review of Diagnostic Test Accuracy (DTA) studies in older people. *Age and Ageing*, 47(3):349–355, 2018.

[211] Joe H. *Dependence Modeling with Copulas*. Chapman and Hall/CRC Press, first edition, 2014.

[212] Nyaga V., Arbyn M., Aerts M. CopulaDTA: An R Package for Copula-Based Bivariate Beta-Binomial Models for Diagnostic Test Accuracy Studies in a Bayesian Framework. *Journal of Statistical Software*, 82(Code Snippet 1):1–27, 2017.

[213] Burzykowski T., Molenberghs G., Buyse M., Geys H., Renard D. Validation of surrogate end points in multiple randomized clinical trials with failure time end points. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 50(4):405–422, 2001.

[214] Nelson R. B. *An Introduction to Copulas*. Springer Series in Statistics. Springer, second edition, 2006.

[215] Sklar M. Fonctions de Répartition à n Dimensions et Leurs Marges. *Publications de l'Institut Statistique de l'Université de Paris*, 8:229–231, 1959.

[216] Song P. X. Multivariate Dispersion Models Generated From Gaussian Copula. *Scandinavian Journal of Statistics*, 27(2):305–320, 2000.

[217] Kojadinovic I. Yan J. Comparison of three semiparametric methods for estimating dependence parameters in copula models. *Insurance: Mathematics and Economics*, 47(1):52–63, 2010.

[218] Genest C. Frank's family of bivariate distributions. *Biometrika*, 74(3):549–555, 1987.

[219] Bujkiewicz S., Thompson J. R., Sutton A. J., Cooper N. J., Harrison M. J., Symmons D. P. M., Abrams K. R. Multivariate meta-analysis of mixed outcomes: a Bayesian approach. *Statistics in Medicine*, 32(22):3926–3943, 2013.

[220] Vehtari A., Gabry J., Magnusson M., Yao Y., Bürkner P., Paananen T., Gelman A. loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models, version 2.5.1. Available from: https://mc-stan.org/loo [accessed 2024-01-12], 2022.

[221] Riley R. D. Multivariate meta-analysis: the effect of ignoring within-study correlation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(4):789–811, 2009.

[222] Soares M. O., Walker S., Palmer S. J., Sculpher M. J. Establishing the Value of Diagnostic and Prognostic Tests in Health Technology Assessment. *Medical Decision Making*, 38(4):495–508, 2018.

[223] Mavridis D. Salanti G. A practical introduction to multivariate meta-analysis. *Statistical Methods in Medical Research*, 22(2):133–158, 2012.

[224] Riley R. D., Price M. J., Jackson D., Wardle M., Gueyffier F., Wang J., Staessen J. A., White I. R. Multivariate meta-analysis using individual participant data. *Research Synthesis Methods*, 6(2):157–174, 2014.

[225] Gelman A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534, 2006.

[226] Leeflang M. M. G., Deeks J. J., Rutjes A. W. S., Reitsma J. B., Bossuyt P. M. M. Bivariate meta-analysis of predictive values of diagnostic tests can be an alternative to bivariate meta-analysis of sensitivity and specificity. *Journal of Clinical Epidemiology*, 65(10):1088–1097, 2012.

[227] Zwinderman A. H. Bossuyt P. M. We should not pool diagnostic likelihood ratios in systematic reviews. *Statistics in Medicine*, 27(5):687–697, 2008.

[228] Leeflang M. M. G., Rutjes A. W. S., Reitsma J. B., Hooft L., Bossuyt P. M. M. Variation of a test's sensitivity and specificity with disease prevalence. *Canadian Medical Association Journal*, 185(11):E537–E544, 2013.

[229] Achana F. A., Cooper N. J., Dias S., Lu G., Rice S. J. C., Kendrick D., Sutton A. J. Extending methods for investigating the relationship between treatment effect and baseline risk from pairwise meta-analysis to network meta-analysis. *Statistics in Medicine*, 32(5):752–771, 2012.

[230] Dias S., Sutton A. J., Welton N. J., Ades A. E. Evidence Synthesis for Decision Making 3: Heterogeneity - Subgroups, Meta-Regression, Bias, and Bias-Adjustment. *Medical Decision Making*, 33(5):618–640, 2013.

[231] Zhang J., Ko C.-W., Nie L., Chen Y., Tiwari R. Bayesian hierarchical methods for meta-analysis combining randomized-controlled and single-arm studies. *Statistical Methods in Medical Research*, 28(5):1293–1310, 2018.

[232] Bowrin K., Briere J.-B., Levy P., Toumi M., Millier A. Use of real-world evidence in meta-analyses and cost-effectiveness models. *Journal of Medical Economics*, 23(10):1053–1060, 2020.

[233] Riley R. D., Dodd S. R., Craig J. V., Thompson J. R., Williamson P. R. Meta-analysis of diagnostic test studies using individual patient data and aggregate data. *Statistics in Medicine*, 27(29):6111–6136, 2008.

[234] Noel-Storr A. H., McCleery J. M., Richard E., Ritchie C. W., Flicker L., Cullum S. J., Davis D., Quinn T. J., Hyde C., Rutjes A. W. S., Smailagic N., Marcus S., Black S., Blennow K., Brayne C., Fiorivanti M., Johnson J. K., Köpke S., Schneider L. S., Simmons A., Mattsson N., Zetterberg H., Bossuyt P. M. M., Wilcock G., McShane R. Reporting standards for studies of diagnostic test accuracy in dementia: The STARDdem Initiative. *Neurology*, 83(4):364–373, 2014.

[235] Stan Development Team. shinystan: Interactive Visual and Numerical Diagnostics and Posterior Analysis for Bayesian Models, version 2.6.0. Available from: http://mc-stan.org/ [accessed 2024-01-12], 2022.

[236] Smith-Bindman R., Hosmer W., Feldstein V. A., Deeks J. J., Goldberg J. D. Second-Trimester Ultrasound to Detect Fetuses With Down Syndrome: A Meta-Analysis. *JAMA*, 285(8):1044–1055, 2001.

[237] Teunissen C. E., Chiu M.-J., Yang C.-C., Yang S.-Y., Scheltens P., Zetterberg H., Blennow K. Plasma Amyloid-β (Aβ42) Correlates with Cerebrospinal Fluid Aβ42 in Alzheimer's Disease. *Journal of Alzheimer's Disease*, 62(4):1857–1863, 2018.

[238] Bauermeister S., Orton C., Thompson S., Barker R. A., Bauermeister J. R., Ben-Shlomo Y., Brayne C., Burn D., Campbell A., Calvin C., Chandran S., Chaturvedi N., Chêne G., Chessell I. P., Corbett A., Davis D. H. J., Denis

M., Dufouil C., Elliott P., Fox N., Hill D., Hofer S. M., Hu M. T., Jindra C., Kee F., Kim C.-H., Kim C., Kivimaki M., Koychev I., Lawson R. A., Linden G. J., Lyons R. A., Mackay C., Matthews P. M., McGuiness B., Middleton L., Moody C., Moore K., Na D. L., O'Brien J. T., Ourselin S., Paranjothy S., Park K.-S., Porteous D. J., Richards M., Ritchie C. W., Rohrer J. D., Rossor M. N., Rowe J. B., Scahill R., Schnier C., Schott J. M., Seo S. W., South M., Steptoe M., Tabrizi S. J., Tales A., Tillin T., Timpson N. J., Toga A. W., Visser P.-J., Wade-Martins R., Wilkinson T., Williams J., Wong A., Gallacher J. E. J. The Dementias Platform UK (DPUK) Data Portal. *European Journal of Epidemiology*, 35(6):601–611, 2020.