Statistical methods for meta-analysis to compare two diagnostic tests to a common gold standard

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität-Düsseldorf

vorgelegt von

Annika Hoyer
aus Aschersleben

Düsseldorf, Januar 2016

aus dem Institut für Biometrie und Epidemiologie des
Deutschen Diabetes-Zentrums,
Leibniz-Zentrum für Diabetes-Forschung an
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referent: Prof. Dr. Holger Schwender

Koreferent: Prof. Dr. Oliver Kuß

Tag der mündlichen Prüfung: 7. März 2016

»Eins noch!«Smeik hielt Rumo mit einem seiner Ärmchen fest. »Hast Du die Antwort auf meine Frage gefunden? *Was wird immer kürzer, je länger es wird?*« »Oh«, antwortete Rumo, »das war leicht. Die Antwort ist natürlich *Das Leben.*« »Natürlich«, grinste Smeik.

aus Walter Moers: Rumo & Die Wunder im Dunkeln

Contents

Abstract

Meta-analysis of diagnostic studies is still a rapidly developing area of biostatistical research. In particular, there is an increasing interest in methods to compare different tests to a common gold standard. Restricting to the case of two diagnostic tests, in these meta-analyses the parameters of interest are the differences of sensitivities and specificities (with their corresponding confidence intervals) between the two diagnostic tests while accounting for the various associations within single studies, between the two tests and within patients. We propose statistical models with a quadrivariate response (where sensitivity of test 1, specificity of test 1, sensitivity of test 2, and specificity of test 2 are the four responses) as a sensible approach to this task. Using a quadrivariate generalized linear mixed model (GLMM) naturally generalizes the common standard bivariate model of meta-analysis for a single diagnostic test. An alternative is given by copula models, whereby we use a quadrivariate Gaussian copula and quadrivariate vine copula constructions which are based on bivariate Plackett copulas. All models are compared by an extensive simulation study. Finally, we illustrate our models by two examples. The first one compares two tests in the diagnosis of coronary artery disease. The second example is from diabetes research, where two screening methods for the diagnosis of type 2 diabetes mellitus are compared.

Zusammenfassung

Die Meta-Analyse von diagnostischen Studien ist ein sich stetig weiterentwickelndes Feld in der biostatistischen Forschung. Im Besonderen besteht ein wachsendes Interesse an statistischen Methoden, die es erlauben, verschiedene Tests mit einem gemeinsamen Goldstandard zu vergleichen. Bei Beschränkung auf den Fall zweier diagnostischer Tests handelt es sich bei den interessierenden Parametern um die Differenzen zwischen den Sensitivitäten und Spezifitäten (mit den zugehörigen Konfidenzintervallen) dieser Tests, während die verschiedenen Assoziationen zwischen den Studien, zwischen den Tests und innerhalb der Patienten zu beachten sind. Wir schlagen Modelle mit einer vierdimensionalen Zielgröße (wobei sich diese aus Sensitivität und Spezifität von Test 1, sowie Sensitivität und Spezifität von Test 2 zusammensetzt) als eine sinnvolle Herangehensweise vor. Die Verwendung eines vierdimensionalen generalisierten linearen gemischten Modells (GLMM) stellt hierbei eine Verallgemeinerung des bekannten bivariaten Standardmodells zur Meta-Analyse eines einzelnen diagnostischen Tests dar. Eine alternative Möglichkeit ist die Verwendung von Copula-Modellen, wobei wir eine vierdimensionale Gauß-Copula und drei vierdimensionale Vine-Copulas, die auf bivariaten Plackett-Copulas basieren, nutzen. Die Modelle werden anhand einer umfassenden Simluationsstudie miteinander verglichen. Abschließend werden die Modelle anhand von zwei Beispielen illustriert. Das erste befasst sich mit dem Vergleich von zwei Möglichkeiten zur Diagnose der koronaren Herzkrankheit. Beim zweiten Beispiel werden zwei verschiedene Tests zur Diagnose von Typ 2 Diabetes mellitus verglichen.

Acknowledgement

First of all, my deepest thanks go to Prof. Dr. Oliver Kuß for his excellent support and interesting, inspiring and motivating conversations. I also would like to thank Prof. Dr. Holger Schwender for his guidance and his interest and thoughts on that research topic. The same goes to my mentor for this thesis, Prof. Dr. Peter Kern. Many warm thanks go to my colleagues from the Institute for Biometry and Epidemiology from the German Diabetes Center in Düsseldorf for stimulating exchange as well as critical reading this thesis. A very special thank goes to Dr. Ralph Brinks who has spent much time with me sharing the same office during this thesis. Thank you for the motivating words whenever they were needed.
Finally I want to thank my friends and family for their love and support.

List of Abbreviations

ADA	American Diabetes Association
BIC	Bayesian Information Criterion
cdf	Cumulative distribution function
D^-	Non-diseased probands
D^+	Diseased probands
DOR	Diagnostic odds ratio
DPCP	Detectable preclinical phase
DSL	DerSimonian-Laird
FP	False positive
FN	False negative
FNF	False negative fraction
FPF	False positive fraction
FPG	Fasting plasma glucose
GLMM	Generalized linear mixed model
GQ	Gaussian quadrature
HbA_{1c}	Glycated hemoglobin A1c
HSROC	Hierarchical summary receiver operating characteristic
IPD	Individual patient data
JFPF	Joint false positive fraction

JSe	Joint sensitivity
OGTT	Oral glucose tolerance test
PCC	Pair-copula construction
pmf	Probability mass function
pp	Percentage points
PQL	Penalized-quasi-likelihood
Pre	Prevalence
QUADAS	Quality Assessment of Diagnostic Accuracy Studies
REML	Restricted maximum likelihood
ROC	Receiver operating characteristic
Se	Sensitivity
Sp	Specificity
SROC	Summary receiver operating characteristics
STARD	Standard for Reporting of Diagnostic Accuracy
TN	True negative
TP	True positive

1 Introduction

In 1997, Van Houwelingen pointed out that meta-analysis is one of his nightmares and he hoped that it would not haunt us (Van Houwelingen 1997). Since his publication, a few years have passed and the overall opinion has changed.

Meta-analysis, as a statistical method to summarize information from different studies, has become a growing field of biostatistical research with many recent developments. A reason is found in the increasing importance of evidence-based medicine which is a field in which meta-analysis plays an important role. 'Evidence based medicine is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients' and 'integrating individual clinical expertise with the best available external clinical evidence from systematic research' as it was pointed out by Sackett et al. (1996). Furthermore, meta-analysis can be included in medical decision making. An overview of the meta-analytic developments is given by Sutton and Higgins (2008) and by Ades and Sutton (2006).

While methods for clinical studies are well-established using standard statistical approaches like mixed models as it is presented by DerSimonian and Laird (1986), in case of diagnostic studies the situation becomes more complex and challenging. Therefore there is still an ongoing interest in new statistical methods for meta-analysis of diagnostic accuracy studies.

As a basis for such analyses, systematic reviews are used. Thereby, a number of studies according to the same topic are collected. Using the information from these single studies, a meta-analysis is conducted to summarize them. In case of diagnostic studies, every single study reports at least one fourfold table. Using that information the corresponding sensitivities (the conditional probabilities of a positive test result in presence of the disease) and specificities (the conditional probabilities of negative test result in absence of the disease) can be calculated. These effect measures should be summarized in a meta-analysis. This leads to at least bivariate approaches, because sensitivity and specificity have to be modelled simultaneously. In doing so, some points should be accounted for. Sensitivity and specificity are generally negatively correlated across studies (Harbord et al. 2008). The potential

heterogeneity, i.e. the possibility of different true study effects, should be considered, too. Such approaches are statistically very challenging. In case of meta-analysing one diagnostic test, different approaches are reported in literature. There is the possibility to use a fixed effect model which does not account for between-study heterogeneity (Deeks 2001). This model is outdated and should be avoided in many situations. The most widely used and accepted model is proposed by Reitsma et al. (2005) and Chu and Cole (2006). They recommend a bivariate logistic regression model with random effects to account for potential heterogeneity. A few alternatives and modifications of that model can be found in literature, for example Paul et al. (2010), Zapf et al. (2015) or Chen et al. (2014). A very flexible alternative is proposed by Kuss et al. (2014). They used a bivariate copula approach and extend their model to three dimensions to account additionally for disease prevalence (Hoyer and Kuss 2015). Such an approach is also given by Chu et al. (2009).

It is surprising that there is yet no further extension that allows the comparison of two diagnostic tests to a common gold standard in a meta-analytic way. The need for such models is pointed out by different authors because such studies occur more often than expected as it was shown by Takwoingi et al. (2013). They found 466 systematic reviews which compared the accuracy of two or more tests. Additionally, Leeflang et al. (2008) accented that 'policymakers and guideline developers may be particularly interested in comparative accuracy' of diagnostic tests. Tatsioni et al. (2005) wrote that 'frequently, meta-analyses assess several diagnostic tests for the same condition. In such cases, we may wish not only to report the performance of each test but also to compare performance between tests.'

In the medical research area of diabetes, we found two systematic reviews (Kodama et al. 2013, Bennett et al. 2007) that compare HbA_{1c} (Glycated hemoglobin A1c) to fasting plasma glucose (FPG) for the population-based screening of type 2 diabetes mellitus and illustrate this need in practice. Both systematic reviews report the results of the included studies only qualitatively without summary measures as differences of sensitivities and specificities.

Avowedly, different approaches for the meta-analysis for the comparison of diagnostic tests have been published in literature (Trikalinos et al. 2014, Siadaty et al. 2004, Siadaty and Shu 2004). However, they come with several disadvantages as combining the information of sensitivity and specificity using only a single measure or assuming independent tests.

In this thesis, we propose new approaches for the meta-analysis for the comparison of two diagnostic tests to a common gold standard. Thereby, we extend the mod-

els from Chu and Cole (2006) and Kuss et al. (2014) to four dimensions and use the differences of sensitivities and specificities as measures of interest which can be interpreted in an easy way without loss of information.

The thesis is organized as follows. First, we give a brief review on diagnostic tests and meta-analysis including the motivation for our new models. In the next chapter the statistical methods are explained. Afterwards, the comprehensive simulation study and the corresponding results are presented. Then the models are illustrated by two practical examples. Finally, we discuss our models with respect to their advantages and disadvantages. We conclude the work with an outlook.

2 Basic principles

In medical research, the most common known type of study is the clinical trial or intervention study. It is conducted for example to evaluate new drugs or therapies. The outcomes of interest are clinical endpoints and could be the cure of a disease. For a nice introduction to intervention studies see for example Schumacher and Schulgen (2008). Contrary to that, in other cases researchers are interested in a binary outcome according to the disease status (1: diseased (D^+), 0: non-diseased (D^-)). Considering that, diagnostic tests are in demand.

The following sections will give short introductions to diagnostic tests and meta-analyses.

2.1 Diagnostic tests

In medical practice, a diagnostic test is conducted for diagnosing a disease, i.e. to evaluate if a person is diseased or non-diseased. An example is the diagnosis of cancer via biopsy. It is possible that the actual standard diagnostic procedure, the so-called gold standard or reference test, is expensive or too invasive for patients and should be replaced by a new test. Mainly, there exist three basic roles for a new test: replacement, triage and add-on. They are well described by Leeflang et al. (2008). Roughly spoken, the aim of replacement is to replace an existing test based on a comparison with the new test applied in the same population using the same gold standard test. Triage means that patients first undergo the new test. Only these with a conspicuous test result continue with the existing reference test. The add-on role of a diagnostic test is the opposite of triage. The patients undergo the new test after the existing one to identify false-positive or false negative results (Leeflang et al. 2008). For these reasons, diagnostic studies enables us to compare a new test (the test under evaluation or index test), which is cheaper or less invasive, to the gold standard. The evaluation of these studies is one aim in medical research. The diagnostic test accuracy describes how efficient a test separates between patients with disease and those without. The aim of diagnostic studies is then to estimate

this accuracy measured in general by two values, the sensitivity (Se) and specificity (Sp) which are described more detailed in the next paragraph. Well-designed studies could be used for example as a basis in the field of medical decision making (Leeflang et al. 2008).

A diagnostic study is then conducted in the following way. Every study participant is evaluated using the gold standard method to determine the true disease status. Additionally, each of them undergoes the new procedure. This can lead to four different classifications which are shown in Table 2.1. A person could be classified as 'True Positive (TP)' if the new test and the gold standard are positive. A proband with two negative results is tested as 'True Negative (TN)'. On the other hand there are two possibilities of misclassification. A person could be classified as 'False Positive (FP)' in case the new test is positive but the standard method leads to a negative result. The last possibility is that a study participant gets a negative results when the new test is applied but is positive tested by the gold standard. Then, he is a so-called 'False Negative (FN)'. All in all that means the results of a diagnostic study can be summarized using a simple contingency table.

Table 2.1: Contingency Table

	Gold Standard	
Diagnostic Test	+	−
+	TP	FP
−	FN	TN

To evaluate the new diagnostic test in a second step, mainly two conditional probabilities are required. Initially, there is the probability to get a positive test result conditional on being diseased defined by the gold standard, the sensitivity. The second important measure is the specificity, the probability to get a negative test result conditional on being non-diseased.

$$Sensitivity = Se = P(T^+ \mid D^+) = \frac{TP}{TP + FN}$$

$$Specificity = Sp = P(T^- \mid D^-) = \frac{TN}{TN + FP}$$

Other important measures to characterize the accuracy of a diagnostic test are the predictive values. They are also conditional probabilities, but more meaningful from a patient view because they are based on the condition of the test status which is known in practice. They depend on the prevalence (Pre), the a priori probability

of being diseased, and are calculated as follows using Bayes Theorem:

$$Positive\,Predictive\,Value = P(D^+ \mid T^+) = \frac{Se \cdot Pre}{Se \cdot Pre + (1 - Sp)(1 - Pre)}$$

$$Negative\,Predictive\,Value = P(D^- \mid T^-) = \frac{Sp(1 - Pre)}{Sp(1 - Pre) + (1 - Se)Pre}.$$

In practice, the most frequently used measures to evaluate the index test are sensitivity and specificity. They are summary measures which merge all information from a single study in only two values. In some cases this could be a disadvantage because studies report different thresholds for declaring a test result positive. According to that fact, sensitivity and specificity vary with the used threshold. To investigate such variations, the concept of receiver operating characteristics (ROC) is applied.

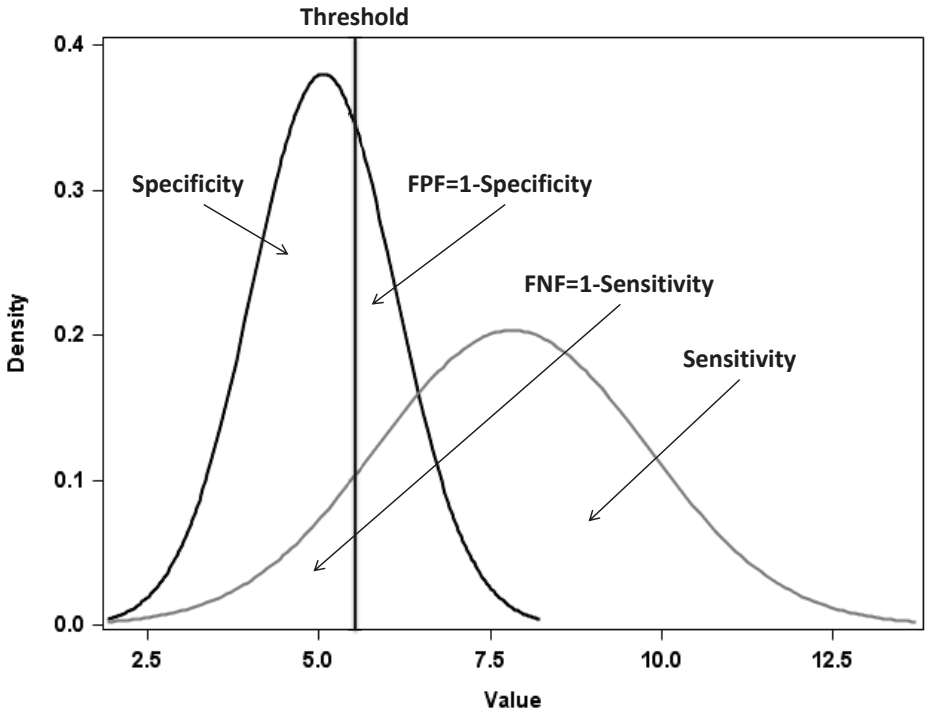

Figure 2.1: Distribution of marker values in different populations: black - non-diseased participants, grey - diseased participants.

In a first step, the density of the marker values in the population of healthy and

diseased subjects is visualised like in Figure 2.1. Using such a plot, sensitivity, specificity, the false negative fraction (FNF, 1-sensitivity) and the false positive fraction (FPF, 1-specificity) are assumed as realizations from distribution functions at selected thresholds. It is also obvious that the values vary with different thresholds. In a next step, this plot is transformed into the ROC-space what is seen in Figure 2.2. Here, estimated sensitivity is plotted against 1-specificity using different thresholds. Such ROC curves are used when multiple thresholds are investigated and the different sensitivities and specificities should be presented. After this, many studies recommend the threshold where the estimated sum of sensitivity and specificity is the largest to be used in practice. Such a selection is based on the technical principle of the Youden index which is defined as $Se + Sp - 1$ (Youden 1950).

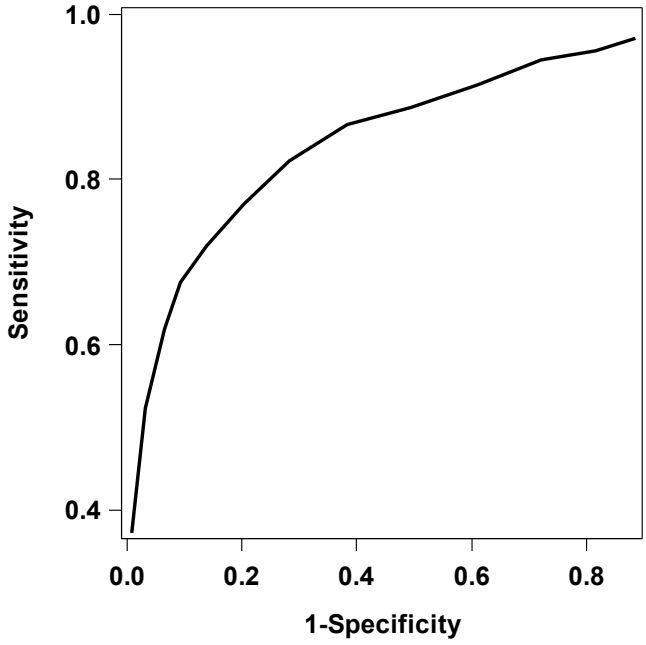

Figure 2.2: Receiver operating characteristics of a study from Choi et al. (2011).

2.2 Meta-analysis

Meta-analysis is a method to summarize and combine results from multiple single studies following a predefined question. It is known as a principle tool underlying

evidence-based medicine (Ades and Sutton 2006) but is performed in a wide field of other research areas, like economy or psychology. It is shown by Sutton and Higgins (2008) that the number of meta-analysis papers in medical research (including methods papers) increased in the last years. Figure 2.3 illustrates the number of published papers on meta-analysis from 1990 to 2015 as collected in PubMed (search term: meta-analysis). It underpins that there exists the necessity for meta-analysis because there are more and more studies according addressing the same medical problem which need to be summarized. Additionally, evidence-based medicine earns more and more a primary role in medical research. Against the background of this trend, the influence of meta-analysis has grown tremendously.

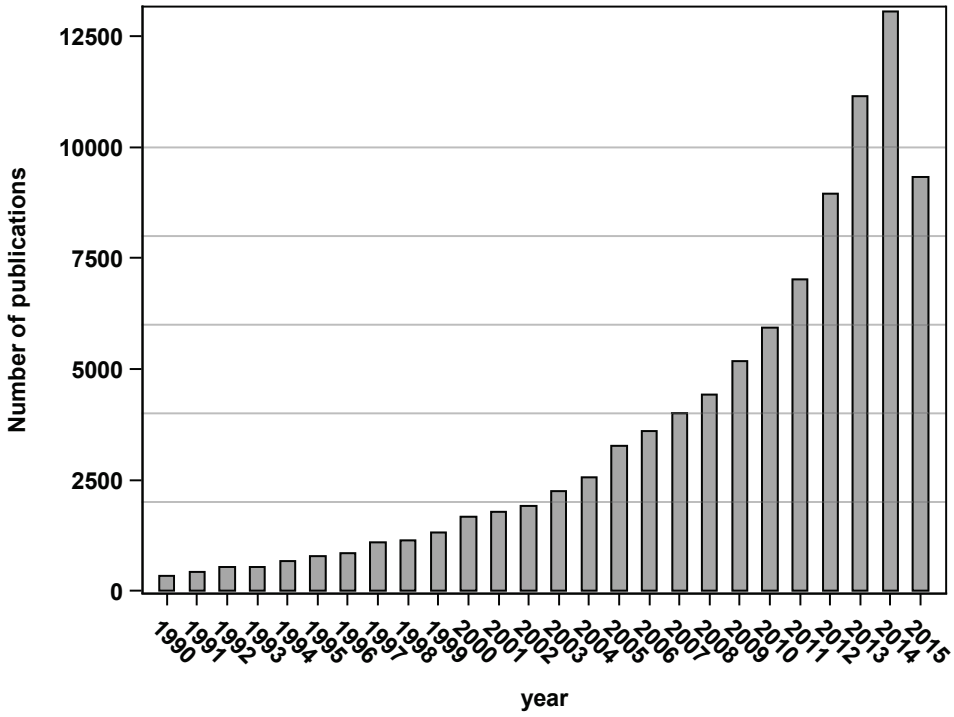

Figure 2.3: Numbers of publications concerning meta-analysis as collected in PubMed (http://www.ncbi.nlm.nih.gov/pubmed) on July 12, 2015

The basis for every meta-analysis is a systematic review, a collection of separate studies belonging to the medical context which is investigated. It includes a systematic and documented literature search, where studies are selected in consideration of different requirements like specified inclusion and exclusion criteria. After this step, a feasible meta-analysis could be performed. Such reviews are endeavoured in differ-

ent ways, for example by the Cochrane Collaboration (http://www.cochrane.org/). They offer the possibility to publish systematic reviews and meta-analyses in the Cochrane Database of Systematic Reviews. Cochrane reviews are not used to make recommendations in any particular clinical context but they should be undertaken to inform specific decisions (Sutton and Higgins 2008).

Statistically, a meta-analysis is a method to summarize the results of separate studies following a predefined question in a quantitative way (Boissel et al. 1988). The most known and widely accepted meta-analysis method is the usage of the weighted average of the point estimates from the single studies. Thereby, we have to differentiate between the assumption of equal treatment effects in the single studies (a common effect), the so-called homogeneity, and the assumption of heterogeneity which means different true treatment effects in the separate studies. In case of homogeneous studies, the fixed-effect meta-analysis is well established. Here, the inverse variances of the estimates are used as weights.

We give a short introduction to the fixed effects model based on the book by Hartung et al. (2011) which is also referred for further information. In case of a fixed effects model, we consider I independent studies where study i leads to the estimated effect size T_i. Furthermore, T_i is an estimate of the population effect size θ_i. Let $\hat{\sigma}^2(T_i)$ be the estimated variance of T_i. In general T_i, as it is based on a random sample of size n_i, is approximately normal distributed with mean θ_i and variance $\sigma^2(T_i) = \sigma^2_{\theta_i;n_i}$. In cases where the variance depends on the unknown parameter θ_i, $\hat{\sigma}^2(T_i)$ is an estimator of $\sigma^2_{\theta_i;n_i}$. Assuming homogeneity, i.e. $\theta_1 = \cdots = \theta_k = \theta$, a combined estimate of θ is obtained by a weighted combination of the T_i

$$\hat{\theta} = \frac{\sum_{i=1}^{I} w_i T_i}{\sum_{i=1}^{I} w_i},$$

where w_i is a nonnegative weight corresponding to study i. The optimal weights, i.e. weights which make $\mathrm{Var}(\hat{\theta})$ the smallest, are

$$w_i = \frac{1}{\sigma^2_{\theta_i;n_i}}.$$

These weights are often unknown, because $\sigma^2_{\theta_i;n_i}$ is usually unknown. Replacing $\sigma^2_{\theta_i;n_i}$ by $\hat{\sigma}^2(T_i)$, we get

$$\tilde{\theta} = \frac{\sum_{i=1}^{I} T_i/\hat{\sigma}^2(T_i)}{\sum_{i=1}^{I} 1/\hat{\sigma}^2(T_i)}$$

with the estimated variance

$$\hat{\sigma}^2(\theta) = \widehat{\mathrm{Var}}(\tilde{\theta}) \approx \frac{1}{\sum_{i=1}^{I} 1/\hat{\sigma}^2(T_i)}.$$

Fixed-effects methods are actually under discussion because they assume that there is the same underlying effect in each study which cannot vary. Therefore the used standard are random effects methods. Such analyses model possible heterogeneity and include the among-study variation of effects into the weights (DerSimonian and Laird 1986).

In case of a random effects model, we additionally assume that

$$\hat{\theta}_i \sim N\left(\theta, \tau^2 + \sigma_i^2(\theta_i)\right),$$

where τ^2 is the parameter for the between-study variance or the heterogeneity parameter (Hartung et al. 2011).
Using $w_i(\tau^2) = 1/[\tau^2 + \hat{\sigma}_i^2(\theta_i)]$ and

$$\hat{\theta}(\tau^2) = \frac{\sum_{i=1}^{I} w_i(\tau^2)\hat{\sigma}_i}{\sum_{i=1}^{I} w_i(\tau^2)}$$

a REML (Restricted maximum likelihood) estimate of τ^2 can be constructed numerically by iterating

$$\tau^2 = \frac{\sum_{i=1}^{I} w_i^2(\tau^2)\left\{[\hat{\theta} - \hat{\theta}(\tau^2)]^2 - \hat{\sigma}_i^2(\theta_i)\right\}}{\sum_{i=1}^{I} w_i^2(\tau^2)} + \frac{1}{\sum_{i=1}^{I} w_i(\tau^2)}$$

with an initial value of τ^2 on the right-hand side.
As an alternative, the DSL (DerSimonian-Laird) estimator as given by

$$\hat{\tau}^2_{DSL} = \frac{Q - (I-1)}{\sum_{i=1}^{I} \hat{v}_i - \sum_{i=1}^{I} \hat{v}_i^2 / \sum_{i=1}^{I} \hat{v}_i}$$

with $Q = \sum_{i=1}^{I} \hat{v}_i \left(\hat{\theta}_i - \tilde{\theta}\right)^2, \hat{v}_i = 1/\hat{\sigma}_i^2(\theta_i)$ and $\tilde{\theta} = \sum_{i=1}^{I} \hat{v}_i\hat{\theta}_i / \sum_{i=1}^{I} \hat{v}_i$. In some cases, the DSL estimator provides a negative estimates and has to be truncated to 0.

2.3 Meta-analysis of diagnostic studies

Meta-analysis of diagnostic studies is still an growing field of biostatistical research. Actually, there is an increasing interest in systematic reviews and meta-analysis (Harbord et al. 2008). This fact is visualized in Figure 2.4, where the rising number of publications concerning meta-analysis of diagnostic accuracy studies is shown. The plot is based on data from PubMed using the search terms meta-analysis and (diagnostic test or diagnostic accuracy study). This development is also pointed out by Gatsonis et al. (2006). Moreover, they explain possible reasons for that. They can be found in the increasing reference to evidence-based medicine where meta-analysis is an important tool for medical decision making. Today, diagnostic studies are in the focus of meta-analysis, because they support physicians in their decisions according to cost effectiveness (because of increasing health care costs) or evaluating of performance of tests that are less invasive. Here, meta-analysis is a powerful tool to summarize results of different, independent single studies following a common underlying question. In practice, it is done as follows.

After the first step, where separate studies are collected in a systematic review, a meta-analysis is carried out to combine the results. Such meta-analyses involves studies which meet the inclusion criteria. In the optimal case, the studies are reported using the current guidelines as the STARD (Standard for Reporting of Diagnostic Accuracy) statement (Bossuyt et al. 2003) and the QUADAS (Quality Assessment of Diagnostic Accuracy Studies) checklist (Whiting et al. 2003, Whiting et al. 2011) are included. But summarizing diagnostic studies comes with some difficulties which are also explained by Harbord et al. (2008).

In case of diagnostic studies, we analyse I different studies where each of them evaluates the same diagnostic test. The situation is illustrated in Table 2.2. Now, we are interested in a weighted estimator, which accounts for study size and heterogeneity. At this point, there is a huge difference between diagnostic studies and clinical trials because diagnostic tests are usually quantified by two measures, sensitivity and specificity (Harbord et al. 2008). Using other single measures, like the Diagnostic odds ratio

$$DOR = \frac{TP/FN}{FP/TN}, \tag{2.1}$$

leads to a loss of information (Deeks 2001). Therefore we have to model at least a bivariate outcome. Additionally, sensitivity and specificity are usually negative

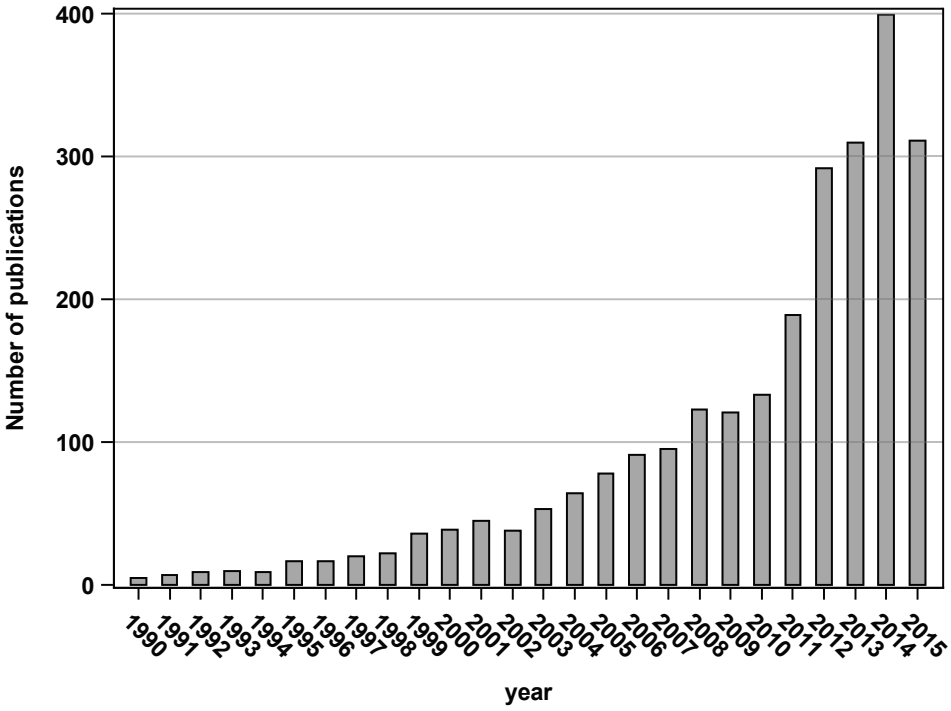

Figure 2.4: Numbers of meta-analysis publications concerning diagnostic tests as collected in PubMed (http://www.ncbi.nlm.nih.gov/pubmed) on July 12, 2015

correlated in meta-analysis. This is explained by the fact that different studies may use different thresholds for declaring a test positive. Therefore, it is not sufficient to fit two univariate models for sensitivity and specificity because the used statistical approach should model a possible correlation, too. Last but not least, the heterogeneity of the studies should be accounted for. Considering all the presented facts leads to complex models when meta-analysing diagnostic studies to a common gold standard compared to clinical trials where statistical methods are much more developed.

In the following, two different proposed approaches for the meta-analysis of a single diagnostic test are presented.

Table 2.2: Visualization of meta-analysis of diagnostic studies

Study	**Sensitivity**	**Specificity**
1	$\mathrm{TP}_1/\mathrm{D}_1^+$	$\mathrm{TN}_1/\mathrm{D}_1^-$
2	$\mathrm{TP}_2/\mathrm{D}_2^+$	$\mathrm{TN}_2/\mathrm{D}_2^-$
...	...	...
	$\hat{\theta}_{Se}$=...	$\hat{\theta}_{Sp}$=...

2.3.1 Summary ROC approach

The summary ROC (SROC) approach was developed by Moses and Littenberg (Littenberg and Moses 1993, Moses et al. 1993). The aim of that method is to construct and estimate a smooth curve through the the plotted sensitivities and specificities in the ROC space (Reitsma et al. 2005). For that reason, the sensitivity and FPF (1− specificity) scale are logit-transformed to fit a linear regression model. For that aim, we define D and S as follows:

$$D = \ln\left(\frac{Se}{1-Se}\right) - \ln\left(\frac{FPF}{1-FPF}\right) = \ln(\mathrm{DOR}) \tag{2.2}$$

$$S = \ln\left(\frac{Se}{1-Se}\right) + \ln\left(\frac{FPF}{1-FPF}\right). \tag{2.3}$$

D is the difference of the logits, i.e. the log of the DOR as it is explained in Equation (2.1), and S their sum. S corresponds to the threshold of the diagnostic test. If $S = 0$ then sensitivity equals specificity. S is positive if sensitivity is higher than specificity and negative in studies with a higher specificity.
A linear regression is used to show the relation between ln(DOR) and the implicit threshold. This leads to

$$D = \alpha + \beta S + \epsilon \tag{2.4}$$

This model can be solved using weighted or unweighted least squares linear regression (Moses et al. 1993). Furthermore, α and β are used to obtain estimated sensitivities depending on a chosen specificity or vice versa (Macaskill et al. 2010, Reitsma et al.

2005). In case of sensitivity, we obtain:

$$E(Se) = \frac{1}{1 + \frac{1}{\exp(-[\alpha+(1+\beta)\text{logit}(1-Sp)]/(1-\beta))}}$$

with $\text{logit}(x) = \log(x/(1-x))$.
It is not straightforward to interpret the intercept and slope of the regression model (2.4) as it was pointed out by Reitsma et al. (2005). In case of $\beta \approx 0$, the DOR does not depend on S and α is a summary estimate for the DOR. For $\beta \neq 0$, the DOR varies with S. However, there is no direct interpretation of the regression coefficient, but it considerably influences the shape of the SROC curve.

The main disadvantage of SROC approach is that summary estimates for sensitivity and specificity are not directly available, because the DOR is the outcome measure (Reitsma et al. 2005).
To solve that problem, a bivariate logistic regression model including random effects, is proposed.

2.3.2 Generalized linear mixed models

Before presenting the bivariate model for the meta-analysis of a single diagnostic test, we give a short introduction to generalized linear mixed models (GLMM) based on the book by Fahrmeir et al. (2009). Further information can be found in Molenberghs and Verbeke (2006).

Generalized linear mixed models allow us to model an outcome Y_{ij} which is not normally distributed while including random effects. In general, the indices denote the i-th subject with the j-th measurement. Thereby we assume that the Y_{ij} is conditionally independent of the $q-$dimensional vector b_i of the random effects. The corresponding density has to belong to an exponential family and is of the following form:

$$f(y_{ij} \mid b_i, \beta, \phi) = \exp\{\phi^{-1}[y_{ij}\theta_{ij} - \psi(\theta_{ij})] + c(y_{ij}, \phi)\},$$

where θ_{ij} is the canonical parameter, $\psi(\theta_{ij})$ is differentiable function with the derivations $\psi'(\theta_{ij})$ and $\psi''(\theta_{ij})$, and ϕ is a dispersion parameter. It can be shown that

$$\text{E}(y_{ij}) = \psi'(\theta_{ij}),\ \text{Var}(y_{ij}) = \phi\psi''(\theta_{ij}).$$

In case of GLMMs, it holds

$$g(\mu_{ij}) = x_{ij}^T\beta + z_{ij}^T b_i,$$

where μ_{ij} is the conditional expected value $\mathrm{E}(Y_{ij} \mid b_i)$, b_i are the multivariate normally distributed random effects ($b_i \sim N_q(0, D)$), g is a known link function and x_{ij} and z_{ij} are vectors including covariates.

The parameters of interest are estimated using the maximum likelihood principle. The likelihood function is given by

$$L(\beta, D, \phi) = \prod_{i=1}^{N} f(y_i \mid \beta, D, \phi) = \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f(y_{ij} \mid b_i, \beta, \phi) f(b_i \mid D) \mathrm{d}b_i. \tag{2.5}$$

Maximization of Equation (2.5) could be difficult because of a missing analytical solution. In the following we present two approaches to solve that problem. The first possibility is to approximate the data, known as penalized-quasi-likelihood method (PQL). A second alternative is to approximate the integral using Gaussian quadrature (GQ).

Penalized-quasi-likelihood method The aim of PQL estimation is to transform the Y_i so that the model can be written as

$$Y_i^* \approx X_i\beta + Z_i b_i + \epsilon_i^*.$$

Therefore the following decomposition is made:

$$Y_{ij} = h(x_{ij}^T\beta + z_{ij}^T b_i) + \epsilon_{ij},$$

where $\mathrm{Var}(Y_{ij} \mid b_i) = \phi v(\mu_{ij})$. Using the canonical link function leads to $v(\mu_{ij}) = h'(x_{ij}^T\beta + z_{ij}^T b_i)$. Based on a Taylor expansion, we obtain

$$\begin{aligned} Y_{ij} \approx\ & h(x_{ij}^T\hat{\beta} + z_{ij}^T\hat{b}_i) \\ & + h'(x_{ij}^T\hat{\beta} + z_{ij}^T\hat{b}_i)x_{ij}^T(\beta - \hat{\beta}) \\ & + h'(x_{ij}^T\hat{\beta} + z_{ij}^T\hat{b}_i)z_{ij}^T(b_i - \hat{b}_i) + \epsilon_{ij} \\ =\ & \hat{\mu_{ij}} + v(\hat{\mu_{ij}})x_{ij}^T(\beta - \hat{\beta}) + v(\hat{\mu_{ij}})z_{ij}^T(b_i - \hat{b}_i) + \epsilon_{ij}. \end{aligned}$$

Furthermore the following transformations are made:

$$Y_{ij} - \hat{\mu_{ij}} = v(\hat{\mu_{ij}})x_{ij}^T(\beta - \hat{\beta}) + v(\hat{\mu_{ij}})z_{ij}^T(b_i - \hat{b_i}) + \epsilon_{ij} \Leftrightarrow$$
$$v(\hat{\mu_{ij}})^{-1}(Y_{ij} - \hat{\mu_{ij}}) + x_{ij}^T\hat{\beta} + z_{ij}^T\hat{b_i} = x_{ij}^T + z_{ij}^T + v(\hat{\mu_{ij}})^{-1}\epsilon_{ij}.$$

Defining

$$Y_{ij}^* = v(\hat{\mu_{ij}})^{-1}(Y_{ij} - \hat{\mu_{ij}}) + x_{ij}^T\hat{\beta} + z_{ij}^T\hat{b_i}$$

and

$$\epsilon_{ij}^* = v(\hat{\mu_{ij}})^{-1}\epsilon_{ij},$$

we finally obtain

$$Y_i^* \approx X_i\beta + Z_ib_i + \epsilon_i^*.$$

To summarize, four steps are made to get PQL estimates of a GLMM (Fahrmeir et al. 2009):

(i) Empirical Bayes estimation of b_i with starting values of β, D, ϕ,

(ii) Generating pseudo-data Y_i^*,

(iii) Estimation of the approximate linear mixed model $\rightarrow$ update of β, D, ϕ,

(iv) Repeat step 1 to 3 up to convergence.

Gaussian quadrature As an alternative to the PQL method, it is possible to approximate the integral in Equation (2.5) via Gaussian quadrature. The idea is to use the following approximation:

$$\int f(z)\phi(z)\mathrm{d}z \approx \sum_{q=1}^{Q} w_q f(z_q),$$

where $\phi(z)$ is the density function of multivariate standard normal distribution, q is the approximation degree and w_q are appropriate weights.

In case of a GLMM, we standardize the random effects using $\delta_i = D^{-\frac{1}{2}}b_i$. Then

the integral can be rewritten as

$$f(y_i \mid \beta, D, \phi) = \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f(y_{ij} \mid b_i, \beta, \phi) f(b_i \mid D) \mathrm{d}b_i$$
$$= \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f(y_{ij} \mid \delta_i, \beta, \phi, D) f(\delta_i) \mathrm{d}\delta_i$$

which can be approximated using the relationship $\int f(z)\phi(z)\mathrm{d}z \approx \sum_{q=1}^{Q} w_q f(z_q)$. Gaussian quadrature is presented in a more comprehensive way in Deuflhard and Hohmann (2008).

2.3.3 Bivariate logistic regression model with random effects

The current bivariate standard approach is used when only one test should be analyzed in a meta-analytic sense. It is based on bivariate logistic regression models with random effects and is introduced for example by Reitsma et al. (2005), Chu et al. (2006) or Paul et al. (2010). In detail, in a first stage the TP and the TN of the i-th study are assumed to be binomially distributed given sensitivity and specificity.

$$TP_i \mid Se_i \sim Binomial(TP_i + FN_i, Se_i),$$

$$TN_i \mid Sp_i \sim Binomial(TN_i + FP_i, Sp_i).$$

To model potential between-study correlation and heterogeneity of sensitivity and specificity, a generalized linear mixed model is assumed. Therefore, in a second stage, a logit transformation of sensitivity and specificity is carried out. We get:

$$\mathrm{logit}(Se_i) = \mu + \phi_i,\ \mathrm{logit}(Sp_i) = \nu + \psi_i,$$

where μ and ν are intercepts for $\mathrm{logit}(Se_i), \mathrm{logit}(Sp_i)$ and ϕ_i and ψ_i are random effects.

Assuming a bivariate normal distribution with an expected value of 0 for the random effects, it is possible to account for between-study correlation and heterogeneity

using the parameter ρ of the random effects covariance matrix

$$\begin{pmatrix}\phi_i\\ \psi_i\end{pmatrix} \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_\phi^2 & \rho\sigma_\phi\sigma_\psi\\ \rho\sigma_\phi\sigma_\psi & \sigma_\psi^2\end{pmatrix}\right].$$

For this GLMM the likelihood function can be expressed as

$$L(Se, Sp, \sigma_\phi^2, \sigma_\psi^2, \rho) = \prod_{i=1}^{I} \iint \prod_{j=1}^{2} g\left(y_{ij}; n_{ij}, l^{-1}(x_j)\right) \boldsymbol{\phi_2}(x_1, x_2; \boldsymbol{\mu_2}, \boldsymbol{\Sigma_2}) \mathrm{d}x_1 \mathrm{d}x_2,$$

where g is the binomial probability mass function (pmf) and y_{ij} and n_{ij} are appropriate entries from the fourfold tables of the single studies, e.g. y_{11} and n_{11} represent the true positives and the number of diseased of study one, respectively. Additionally, l represents the chosen link function and $\boldsymbol{\phi_2}$ is the density of the bivariate normal distribution with mean vector $\boldsymbol{\mu_2} = (l(Se), l(Sp))^T$ and random effects matrix $\boldsymbol{\Sigma_2}$. Mostly and as in our case, the logit link as the canonical one is used but other link functions like the complementary log-log or probit are possible, too (Chu et al. 2010).

An also well-established model for the meta-analysis of one diagnostic test is the HSROC (Hierarchical summary receiver operating characteristic) approach proposed by Rutter and Gatsonis (2001). This model is equivalent to the standard bivariate approach without including covariates (Harbord et al. 2007).
Many other authors also developed alternative approaches, e.g. Kuss et al. (2014), Paul et al. (2010) or Zapf et al. (2015), which can be used in practice.

2.4 Meta-analysis for the comparison of diagnostic tests

The development of more complex meta-analytic models for diagnostic accuracy studies is still a growing field of actual research. Especially, there is an increasing interest in methods to compare diagnostic tests. This was pointed out by different authors. Leeflang et al. (2008) state that 'policymakers and guideline developers may be particularly interested in comparative accuracy' of diagnostic tests. This is supported by Tatsioni et al. (2005) who wrote that 'frequently, meta-analyses assess several diagnostic tests for the same condition. In such cases, we may wish not only to report the performance of each test but also to compare performance between

tests.' Such models are very challenging, because they are based upon sophisticated statistical knowledge, e.g. multivariate models. Actually, there exist studies which evaluate more than one test as it is shown by Takwoingi et al. (2013). Therefore there is an increasing interest in meta-analysis to compare diagnostic tests, taking health care costs and the accuracy of different tests into consideration. There is also an interest in a single measure which allows to describe which test works better. Because of this, we are reporting differences between sensitivities and specificities.

In case of comparison of two tests, every patient underwent two tests in every single study. The combination of all possible results are shown in Table 2.3.

Table 2.3: Probabilities of each combination of results in case of two diagnostic tests for study i

Test 1	**Test 2**	**Diseased**	**Non-diseased**
$-$	$-$	$p_{i,00}^{D^+}$	$p_{i,00}^{D^-}$
$-$	$+$	$p_{i,01}^{D^+}$	$p_{i,01}^{D^-}$
$+$	$-$	$p_{i,10}^{D^+}$	$p_{i,10}^{D^-}$
$+$	$+$	$p_{i,11}^{D^+}$	$p_{i,11}^{D^-}$

Some studies may report the data of every participant, i.e. the test results for each individual are known. These are the cases where the maximum of information is reported and we have individual patient data (IPD) where it is possible to model the additional correlation within patients. If this information is available, we can calculate the probabilities in Table 2.3 and we can also count the number of diseased and non-diseased per combination of test results. Assuming additionally independent tests we can calculate sensitivity and specificity for study i with respect to the respective test in the following way:

$$Sensitivity_{i,1} = p_{i,1\bullet}^{D^+} = p_{i,10}^{D^+} + p_{i,11}^{D^+}, \tag{2.6}$$

$$Sensitivity_{i,2} = p_{i,\bullet 1}^{D^+} = p_{i,01}^{D^+} + p_{i,11}^{D^+}, \tag{2.7}$$

$$Specificity_{i,1} = p_{i,0\bullet}^{D^-} = p_{i,01}^{D^-} + p_{i,00}^{D^-}, \tag{2.8}$$

$$Specificity_{i,2} = p_{i,\bullet 0}^{D^-} = p_{i,10}^{D^-} + p_{i,00}^{D^-}. \tag{2.9}$$

However, such studies are rarely reported. Mostly, aggregated data, i.e. the 2x2 contingency tables, are reported. That means each study reports two values for sensitivity and two values for specificity. This situation is illustrated in Table 2.4. It is also necessary to keep in mind that the tests could be dependent. Restricting to the case of two diagnostic tests, in these meta-analyses the parameters of interest are the differences of sensitivities and specificities (with their corresponding confidence intervals) between the two diagnostic tests while accounting for the various associations within single studies, between the two tests and within patients.

Table 2.4: Comparison of diagnostic tests

	Test 1		**Test 2**	
Study	**Sensitivity 1**	**Specificity 1**	**Sensitivity 2**	**Specificity 2**
1	$\mathrm{TP}_{11}/\mathrm{D}_{11}^{+}$	$\mathrm{TN}_{11}/\mathrm{D}_{11}^{-}$	$\mathrm{TP}_{12}/\mathrm{D}_{12}^{+}$	$\mathrm{TN}_{12}/\mathrm{D}_{12}^{-}$
2	$\mathrm{TP}_{21}/\mathrm{D}_{21}^{+}$	$\mathrm{TN}_{21}/\mathrm{D}_{21}^{-}$	$\mathrm{TP}_{22}/\mathrm{D}_{22}^{+}$	$\mathrm{TN}_{22}/\mathrm{D}_{22}^{-}$
...	...	...	...	...
	$\hat{\theta}_{Se_1} = \ldots$	$\hat{\theta}_{Sp_1} = \ldots$	$\hat{\theta}_{Se_2} = \ldots$	$\hat{\theta}_{Sp_2} = \ldots$
	$\hat{\theta}_{Se_1} - \hat{\theta}_{Se_2} = \ldots$		$\hat{\theta}_{Sp_1} - \hat{\theta}_{Sp_2} = \ldots$	

However, different authors have proposed models for the meta-analysis of two diagnostic tests. Three of them are presented in the following.

2.4.1 Approach proposed by the Cochrane Collaboration

The Cochrane Collaboration propose an approach for the meta-analytic comparison of diagnostic tests in their handbook for systematic reviews of diagnostic test accuracy (Macaskill et al. 2010). This approach is based on the bivariate model including an additional binary covariate to identify which fourfold table corresponds to each test. Therefore, they expand the model presented in Section 2.3.3 in the

following way:

$$\text{logit}(Se_i) = \mu + X_i\alpha + \phi_i,\ \text{logit}(Sp_i) = \nu + X_i\beta + \psi_i.$$

μ and ν are intercepts for $\text{logit}(Se_i), \text{logit}(Sp_i)$ and X_i is a binary covariate connecting the diagnostic test with the corresponding fourfold table. As mentioned above, ϕ_i and ψ_i are random effects following a normal distribution:

$$\begin{pmatrix}\phi_i \\ \psi_i\end{pmatrix} \sim N\left[\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}\sigma_\phi^2 & \rho\sigma_\phi\sigma_\psi \\ \rho\sigma_\phi\sigma_\psi & \sigma_\psi^2\end{pmatrix}\right].$$

The main disadvantage of this approach is that the tests are assumed to independent and that there is no possibility to model a correlation between them.

2.4.2 Approach proposed by Siadaty et al.

Siadaty et al. (2004, 2004) present another approach based on modelling the DOR. They define two indicator variables: 'Disease' for the results of the gold standard and 'Result' corresponding to the results of the index test. Both indicator variables are available if the information from all fourfold tables is extracted. Then, the results of the gold standard and index test is dichotomized. Afterwards, Siadaty et al. (2004, 2004) propose the following regression model in case of a single diagnostic test:

$$\text{logit}(R_{ij}) = \beta_0 + \beta_1 D_{ij}, \tag{2.10}$$

where i and j indicate the study and the test, respectively. R represents the test result (positive or negative) and D is the true disease status. In Equation (2.10), β_1 is the log DOR (LDOR). This can also be expressed as the log ratio of TN/FP over FN/TP. To allow for a comparison of different tests, they expand their model including more covariates and interactions between them:

$$\text{logit}(R_{ij}) = \beta_0 + \beta_1 D_{ij} + \beta_2 T_{ij} + \beta_3 D_{ij} T_{ij},$$

where β_3 is the log ratio of the two DORs from the different tests. T represents the diagnostic test. It constitutes the performance of one test vs. the other. In case of a ratio that is statistically different from 0, there is a real difference between both tests.

Siadaty et al. (2004, 2004) also show how to measure heterogeneity and the equiv-

alence to the model from Littenberg and Moses (1993). However, one main disadvantage of the model proposed by Siadaty et al. (2004, 2004) is, that the measure of interest is the DOR. Using such a measure lead to a loss of information (Deeks 2001). Another disadvantage is that it is not possible to get concrete estimators for the sensitivities and specificities of each test.

2.4.3 Approach proposed by Trikalinos et al.

Trikalinos et al (2014) suggest an alternative model. Thereby, the approach is based on a mixture of IPD, as presented in Table 2.3, and aggregated data, as shown in Table 2.4. In order to model the within patient correlation, they use Equations (2.6) - (2.9) to define the sensitivities and FPFs (1−specificities) of each test. Additionally, the 'joint sensitivity' (JSe) and the 'joint false positive fraction' (JFPF) are defined as

$$JSe_i = p_{i,11}^{D^+}$$

and

$$JFPF_i = p_{i,11}^{D^-}.$$

Afterwards, for study i and test j, a logit transformation is carried out:

$$\eta_{ij} = \text{logit}(Se_{i,j})$$

$$\xi_{ij} = \text{logit}(FPF_{i,j}).$$

This is also done for the joint measures:

$$\eta_{i^*} = \text{logit}(JSe_i)$$

$$\xi_{i^*} = \text{logit}(JFPF_i).$$

To account for potential across study correlation, random effects are included that follow a six-dimensional normal distribution:

$$\begin{pmatrix} \eta_{i1} \\ \eta_{i2} \\ \eta_{k^*} \\ \xi_{i1} \\ \xi_{i2} \\ \xi_{i^*} \end{pmatrix} \sim N\left[\begin{pmatrix} \text{H} \\ \Xi \end{pmatrix}, \text{T} \right].$$

Thereby, H and Ξ are vectors with the overall sensitivities and FPFs for both tests and T is the corresponding covariance matrix which can be chosen in different ways as it is presented by Trikalinos et al. (2014).
To sum up, Trikalinos et al. (2014) propose a six-dimensional generalized linear mixed model including random effects, to estimate the differences of sensitivities and FPFs. This model based on individual patient data which is the main disadvantage because such data is rarely reported in practice.

3 Statistical methods and model ideas

The situation of comparison of two diagnostic tests described in Chapter 2 is statistically challenging. Different authors have proposed models but they all come with some problems. Siadaty et al. (2004, 2004) propose approaches using diagnostic odds ratios and proportional odds ratio models. Using only diagnostic odds ratios lead to a loss of information (Deeks 2001). Additionally, diagnostic odds ratios are measures that are rarely used by practitioners who prefer sensitivity and specificity to evaluate diagnostic tests. Therefore it is difficult to establish such a method, too. Another approach is presented by Trikalinos et al. (2014). Here, two main disadvantages can be found. First, they assume independent tests. Second, they use a mixture of individual (IPD) and aggregated data. As it is pointed out in Chapter 2, it is really difficult to get individual patient data because such datasets are rarely reported. These two points argue against this approach. The actual used 'standard' approach for the direct comparison of diagnostic tests is given by the Cochrane Collaboration. They propose a bivariate model as Chu and Cole (2006) including a binary covariate for the test type to identify which fourfold tables corresponds to each test (Macaskill et al. 2010). Thereby, the tests are assumed to be independent and no correlation between them can be modelled.

In the following two new models are presented which compensate the disadvantages of the alternative approaches. Differences between sensitivities and specificities will be estimated which can be easily interpreted by physicians and the assumption is made that both tests are dependent. Moreover, aggregated data are used which is the most reported one in current literature.

First, a quadrivariate logistic regression model including random effects is proposed that is a natural extension of the current bivariate standard approach used in case of meta-analysis of one diagnostic test. The second presented model extends the copula approach developed by Kuss et al. (2014) to four dimensions.

3.1 Quadrivariate logistic regression model with random effects

We propose a new model with a quadrivariate response. This is a natural extension of the common bivariate random effects model for sensitivity and specificity proposed by Reitsma et al. (2005) or Chu et al. (2006). The new model jointly accounts for the four random variables sensitivity$_1$, specificity$_1$, sensitivity$_2$ and specificity$_2$ using a quadrivariate generalized linear mixed model (GLMM). This approach is also presented in Hoyer and Kuss (2016).
Restricting to the case of two diagnostic tests, every single study reports four values:

- the true positive (TP$_1$) and the true negative (TN$_1$) of test 1
- the true positive (TP$_2$) and the true negative (TN$_2$) of test 2.

Analogously to the bivariate approach, we assume that the true positive and the true negative of the i-th study ($i = 1, ..., I$) and the j-th test ($j = 1, 2$) are binomially distributed, given the sensitivities and the specificities of test j.

$$TP_{ij} \mid Se_{ij} \sim Binomial(TP_{ij} + FN_{ij}, Se_{ij}),$$

$$TN_{ij} \mid Sp_{ij} \sim Binomial(TN_{ij} + FP_{ij}, Sp_{ij}),$$

To account for potential between-study correlation and heterogeneity of the two sensitivities and the two specificities, a random effects model is suggested. First, a logit-transformation is made as proposed in the following:

$$\text{logit}(Se_{ij}) = \mu_j + \phi_{ij},\ \text{logit}(Sp_{ij}) = \nu_j + \psi_{ij}$$

μ_j and ν_j are intercepts for $\text{logit}(Se_i), \text{logit}(Sp_i)$ and ϕ_{ij} and ψ_{ij} are random effects. To specify the random effects $(\phi_{i1}, \psi_{i1}, \phi_{i2}, \psi_{i2})^T$ a quadrivariate normal distribution is assumed:

$$\begin{pmatrix} \phi_{i1} \\ \psi_{i1} \\ \phi_{i2} \\ \psi_{i2} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{\phi_1} & \rho_{\phi_1\psi_1}\sigma_{\phi_1}\sigma_{\psi_1} & \rho_{\phi_1\phi_2}\sigma_{\phi_1}\sigma_{\phi_2} & \rho_{\phi_1\psi_2}\sigma_{\phi_1}\sigma_{\psi_2} \\ & \sigma^2_{\psi_1} & \rho_{\psi_1\phi_2}\sigma_{\psi_1}\sigma_{\phi_2} & \rho_{\psi_1\psi_2}\sigma_{\psi_1}\sigma_{\psi_2} \\ & & \sigma^2_{\phi_2} & \rho_{\phi_2\psi_2}\sigma_{\phi_2}\sigma_{\psi_2} \\ & & & \sigma^2_{\psi_2} \end{pmatrix} \right].$$

14 parameters have to be estimated. The four variance parameters $\sigma^2_{\phi_1}, \sigma^2_{\psi_1}, \sigma^2_{\phi_2}, \sigma^2_{\psi_2}$ are used to describe possible between-study heterogeneity of sensitivity$_1$, specificity$_1$,

sensitivity$_2$ and specificity$_2$. The parameters $\rho_{\phi_1\psi_1}, \rho_{\phi_1\phi_2}, \rho_{\phi_1\psi_2}, \rho_{\psi_1\phi_2}, \rho_{\psi_1\psi_2}, \rho_{\phi_2\psi_2}$ capture the corresponding correlation among the random effects. Assuming the four correlation parameters $\rho_{\phi_1\phi_2}, \rho_{\phi_1\psi_2}, \rho_{\psi_1\phi_2}$ and $\rho_{\psi_1\psi_2}$ to be zero, is equivalent to fit two independent bivariate models for both diagnostic tests.
Because our four-dimensional model is a natural extension of the bivariate one, we get the following likelihood function

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{I} \int\int \int\int \prod_{j=1}^{4} g\left(y_{ij}; n_{ij}, l^{-1}(x_j)\right) \boldsymbol{\phi_4}(x_1, x_2, x_3, x_4; \boldsymbol{\mu_4}, \boldsymbol{\Sigma_4}) \mathrm{d}x_1 \mathrm{d}x_2 \mathrm{d}x_3 \mathrm{d}x_4, \tag{3.1}$$

with $\boldsymbol{\theta} = (Se_1, Sp_1, Se_2, Sp_2, \sigma^2_{\phi_1}, \sigma^2_{\psi_1}, \sigma^2_{\phi_2}, \sigma^2_{\psi_2}, \rho_{\phi_1\psi_1}, \rho_{\phi_1\phi_2}, \rho_{\phi_1\psi_2}, \rho_{\psi_1\phi_2}, \rho_{\psi_1\psi_2}, \rho_{\phi_2\psi_2})$ and the same nomenclature as in Section 2.3.3 otherwise.
$\boldsymbol{\phi_4}$ represents the density of the four-dimensional normal distribution with mean vector $\boldsymbol{\mu_4} = (l(Se_1), l(Sp_1), l(Se_2), l(Sp_2))^T$ and the random effects matrix $\boldsymbol{\Sigma_4}$.
To estimate the parameters of a generalized linear mixed model different numerical approaches are available. We compare the penalized-quasi-likelihood (PQL) method to Gaussian quadrature (GQ). We also use two different link functions, namely the identity link which allows to estimate the differences of sensitivities and specificities directly and the canonical logit link where the confidence intervals of the parameters of interest have to be calculated by hand using the delta method.

3.2 Copula models

Before introducing the quadrivariate copula model for the meta-analytic comparison of two diagnostic tests, we give a short overview on the concept of copulas.

3.2.1 Copulas

Definition 3.1 (Copulas (Nelsen 2006))
A function $C : [0,1]^d \to [0,1]$ is a d-dimensional copula, if

- $C(u_1, ..., u_{i-1}, 0, u_{i+1}, ..., u_d) = 0$ for all $u_k \in [0,1]$
- $C(1, ..., 1, u, 1, ..., 1) = u$ for all $u \in [0,1]$
- C is d-increasing, i.e. for each hyperrectangle $B = \prod_{i=1}^{d}[x_i, y_i] \subseteq [0,1]^d$ the C-volume of B is non-negative:

$$\int_B \mathrm{d}C(u) = \sum_{z \in \times_{i=1}^{d} \{x_i, y_i\}} (-1)^{N(z)} C(z) \geq 0,$$

where $N(z) = \#\{k : z_k = x_k\}$.

Copulas are a valuable tool to construct the joint distribution function of multiple random variables. To get the corresponding cumulative distribution functions, Sklar's Theorem is necessary.

Theorem 3.2 (Sklar's Theorem (Nelsen 2006) or (Sklar 1959))
Let H be a joint distribution function with margins $F_i(x)$. Then there exists a copula C such that for all $x_i \in \mathbb{R} \cup \{-\infty, \infty\}, i = 1, ..., d$:

$$H(x_1, ..., x_d) = C(F_1(x_1), ..., F_d(x_d)).$$

If the margins are continuous, then C is unique. Otherwise, C is uniquely determined on Range(F_1) $\times ... \times$ Range(F_d). The converse is also true: if C is a copula and $F_i, i = 1, ..., d$, are distribution functions, then the function $H(x_1, ..., x_d) = C(F_1(x_1), ..., F_d(x_d))$ is a joint distribution function with margins $F_i(x)$.

The proof is given by various authors, see for example Schweizer and Sklar (1983) or Rüschendorf (2009).

It exists also a conditional version of Sklar's Theorem which is of the following form:

Theorem 3.3 (Conditional version of Sklar's Theorem, (Patton 2002, Patton 2006, Cherubini et al. 2012))
Let $F_{X|Z}(. \mid z)$ be the conditional distribution function of $X \mid Z = z$ and $G_{Y|Z}(. \mid z)$ be the conditional distribution function of $Y \mid Z = z$. Let $H_{(X,Y)|Z}(., . \mid z)$ be the joint conditional distribution function of $(X, Y) \mid Z = z$. Additionally, let Z be the support of Z, $F_{X|Z}$ and $G_{Y|Z}$ be continuous in x and y. Then

(i) there exists a unique conditional copula function $C(., . \mid z)$ with

$$H_{(X,Y)|Z}(x, y \mid z) = C(F_{X|Z}(x \mid z), G_{Y|Z}(y \mid z) \mid z),$$

for all $(x, y) \in \mathbb{R}^2$ and all $z \in Z$. This is equivalent to

$$H_{(X,Y)|Z}(F^{-1}_{X|Z}(u \mid z), G^{-1}_{Y|Z}(v \mid z) \mid z) = C(u, v \mid z),$$

for all $(u, v) \in [0, 1]^2$ *and all* $z \in Z$.

(ii) for any conditional copula C*, all* $(x, y) \in \mathbb{R}^2$ *and all* $z \in Z$

$$H_C(x, y \mid z) := C(F_{X|Z}(x \mid z), G_{Y|Z}(y \mid z) \mid z)$$

is a joint two-dimensional conditional distribution function of $(X, Y) \mid Z = z$ *with conditional margins* $F_{X|Z}(. \mid z)$ *and* $G_{Y|Z}(. \mid z)$.

The proof is given by the cited authors.

3.2.2 Modelling dependence

The concept of copulas enables us to model the dependence between the marginal distributions through the copula parameters. These association parameters can be transformed into measures of correlation, like the Spearman's ($\rho_S(X_1, X_2)$) or Kendall's ($\tau(X_1, X_2)$) correlation coefficient. It was shown by Schweizer and Wolff (1981) that both measures could be used for description of dependence. In the following, the two correlation coefficients are explained in more details with respect to their relation to copulas.

Kendall's correlation coefficient

Definition 3.4 (Kendall's correlation coefficient)
Let (X_1, Y_1) and (X_2, Y_2) be independently and identically distributed random variables. Then the correlation coefficient called Kendall's τ is given by

$$\tau = \tau_{X,Y} = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0].$$

In context of copulas, Kendall's τ is defined in the following way (Nelsen 2006):

Theorem 3.5
Let X *and* Y *be continuous random variables with copula* C*. Then Kendall's* τ *is given by*

$$\tau_{X,Y} = \tau_C = Q(C, C) = 4 \iint_{I^2} C(u, v) \mathrm{d}C(u, v) - 1,$$

where $Q = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0]$.

Proof. Let F and G be the common marginal distributions of X_1 and X_2 and Y_1 and Y_2, respectively. C_1 and C_2 are the copulas of (X_1, Y_1) and (X_2, Y_2). Therefore

it follows that $H_1(x,y) = C_1(F(x), G(y))$ and $H_2(x,y) = C_2F(x), G(y))$. Q denotes the difference between the probabilities of concordance and discordance at which (X_i, Y_i) and (X_j, Y_j) are concordant if $(X_i - X_j)(Y_i - Y_j) > 0$ and discordant if $(X_i - X_j)(Y_i - Y_j) < 0$ (Nelsen 2006). For continuous random variables applies $P[(X_1 - X_2)(Y_1 - Y_2) < 0] = 1 - P[(X_1 - X_2)(Y_1 - Y_2) > 0]$ and therefore

$$Q = 2P[(X_1 - X_2)(Y_1 - Y_2) > 0] - 1. \tag{3.2}$$

Furthermore, $P[(X_1 - X_2)(Y_1 - Y_2) > 0] = P[X_1 > X_2, Y_1 > Y_2] + P[X_1 < X_2, Y_1 < Y_2]$, which can be achieved by integrating over the distribution of one of the vectors (X_1, Y_1) or (X_2, Y_2). Using (X_1, Y_1) we have

$$\begin{aligned} P[X_1 > X_2, Y_1 > Y_2] &= P[X_2 < X_1, Y_2 < Y_1] \\ &= \iint_{\mathbb{R}} P[X_2 \leq x, Y_2 \leq y] \mathrm{d}C_1(F(x), G(y)) \\ &= \iint_{\mathbb{R}} C_2(F(x), G(y)) \mathrm{d}C_1(F(x), G(y)). \end{aligned}$$

Let $u = F(x)$ and $v = G(y)$. Than we get

$$P[X_1 > X_2, Y_1 > Y_2] = \iint_{I^2} C_2(u,v) \mathrm{d}C_1(u,v).$$

Moreover, we obtain

$$\begin{aligned} P[X_1 < X_2, Y_1 < Y_2] &= \iint_{\mathbb{R}^2} P[X_2 > x, Y_2 > y] \mathrm{d}C_1(F(x), G(y)) \\ &= \iint_{\mathbb{R}^2} [1 - F(x) - G(y) + C_2(F(x), G(y))] \mathrm{d}C_1(F(x), G(y)) \\ &= \iint_{I^2} [1 - u - v + C_2(u,v)] \mathrm{d}C_1(u,v). \end{aligned}$$

Furthermore, $E(U) = E(V) = 1/2$ because C_1 is the joint distribution function of a pair (U, V) of uniformly distributed random variables. Hence,

$$P[X_1 < X_2, Y_1 < Y_2] = 1 - \frac{1}{2} - \frac{1}{2} + \iint_{I^2} C_2(u,v) \mathrm{d}C_1(u,v) = \iint_{I^2} C_2(u,v) \mathrm{d}C_1(u,v).$$

Therefore

$$P[(X_1 - X_2)(Y_1 - Y_2) > 0] = 2 \iint_{I^2} C_2(u,v) \mathrm{d}C_1(u,v).$$

It follows using Equation (3.2)

$$Q = 2P[(X_1 - X_2)(Y_1 - Y_2) > 0] - 1 = 4 \iint_{I^2} C_2(u, v) \mathrm{d}C_1(u, v) \tag{3.3}$$

Therefore we obtain for Kendall's τ:

$$\tau_{X,Y} = \tau_C = Q = Q(C, C) = 4 \iint_{I^2} C(u, v) \mathrm{d}C(u, v) - 1$$

□

Spearman's correlation coefficient

Definition 3.6 (Spearman's correlation coefficient)
Let $(X_1, Y_1), (X_2, Y_2)$ and (X_3, Y_3) be independently and identically distributed random variables. Then Spearman's correlation coefficient is defined as

$$\rho_{X,Y} = \rho_S = 3(P[(X_1 - X_2)(Y_1 - Y_3) > 0] - P[(X_1 - X_2)(Y_1 - Y_3) < 0]).$$

In context of copulas we have (Nelsen 2006):

Theorem 3.7
Let X and Y be continuous random variables with corresponding copula C. Then Spearman's ρ is given by

$$\rho_{X,Y} = \rho_C = 3Q(C, \Pi) = 12 \iint_{I^2} uv \mathrm{d}C(u, v) - 3 = 12 \iint_{I^2} C(u, v) \mathrm{d}u\mathrm{d}v - 3,$$

where $\Pi = \Pi(u, v) = uv$.

Proof. The proof is simple using Equation (3.3). Then we get

$$\begin{aligned} \rho_{X,Y} = \rho_C &= 3Q(C, \Pi) \\ &= 12 \iint_{I^2} uv \mathrm{d}C(u, v) - 3 \\ &= 12 \iint_{I^2} C(u, v) \mathrm{d}u\mathrm{d}v - 3. \end{aligned}$$

□

3.2.3 Quadrivariate copula model

As an alternative to the quadrivariate logistic regression model, it is also possible to use copula models with a four-dimensional response. It is a natural extension of the bivariate copula model from Kuss et al. (2014).

Analogously to the quadrivariate logistic regression model with random effects, we assume the TP and TN of the i-th study and the j-th test to be binomially distributed, given the two sensitivities and specificities. In a next step, we assume the sensitivities and specificities as beta distributed with parameters a_* and b_*. Contrary to the standard model, we now have a distribution on the original scale of sensitivity and specificity with

$$Se_1 \sim Beta(a_{Se_1}, b_{Se_1}),\ Sp_1 \sim Beta(a_{Sp_1}, b_{Sp_1})$$

and

$$Se_2 \sim Beta(a_{Se_2}, b_{Se_2}),\ Sp_2 \sim Beta(a_{Sp_2}, b_{Sp_2}).$$

The corresponding density function is defined by

$$f(p; a_{Se_1}, b_{Se_1}) = \begin{cases} \frac{\Gamma(a_{Se_1}+b_{Se_1})}{\Gamma(a_{Se_1})\Gamma(b_{Se_1})} p^{a_{Se_1}-1}(1-p)^{b_{Se_1}-1} & ,0<p<1 \\ 0 & ,\text{else} \end{cases}$$

for sensitivity$_1$ (analogously defined for sensitivity$_2$, specificity$_1$ and specificity$_2$) where $\Gamma(x)$ is the Gamma function with $\Gamma(n+1) = n!$.
The expected values of the three beta distributions are defined as $(a_*/(a_*+b_*))$ and describe the meta-analytic parameters of interest (Se_1, Sp_1, Se_2, Sp_2). Additionally, the estimated variances $(a_*b_*/(a_*+b_*+1)(a_*+b_*)^2)$ explain their variation which corresponds to heterogeneity.
The beta distribution is used because of its flexibility and the fact that it is conjugated to the binomial distribution. The advantage is that we get a posteriori beta-binomial distributions which are commonly known.

Theorem 3.8
The beta distribution is conjugated to the binomial distribution.

Proof. (using the example of sensitivity)
Let $(TP_{ij} = d \mid Se_j = u) \sim Binomial(D_{ij}, u)\,, u \in [0,1], d \in \mathbb{N}$ and Se_j priori beta distributed.

The Theorem of Bayes leads to:

$$f(Se_j = u \mid TP_{ij} = d) = \frac{P(TP_{ij} = d \mid Se_j = u)f(u)}{\int_0^1 P(TP_{ij} = d \mid Se_j = u)f(u)du},$$

where f is the density of the beta distribution.
It follows:

$$\begin{aligned} g_{Se_j}(d) = P(TP_{ij} = d) &= \int_0^1 P(TP_{ij} = d \mid Se_j = u)f(u; \alpha_{Se_j}, \beta_{Se_j})\mathrm{d}u \\ &= \binom{D_{ij}}{d} \frac{1}{B(\alpha_{Se_j}, \beta_{Se_j})} \int_0^1 u^{d+\alpha_{Se_j}-1}(1-u)^{D_{ij}-d+\beta_{Se_j}-1}\mathrm{d}u \\ &= \binom{D_{ij}}{d} \frac{B(\alpha_{Se_j} + d, D_{ij} - d + \beta_{Se_j})}{B(\alpha_{Se_j}, \beta_{Se_j})}, \end{aligned}$$

where B is Euler's beta function. Using the Theorem of Bayes:

$$\begin{aligned} f(Se_j = u \mid TP_{ij} = d) &= \frac{\binom{D_{ij}}{d} u^d (1-u)^{D_{ij}-d} u^{\alpha_{Se_j}-1}(1-u)^{\beta_{Se_j}-1}/B(\alpha_{Se_j}, \beta_{Se_j})}{\binom{D_{ij}}{d} B(\alpha_{Se_j} + u, \beta_{Se_j} + D_{ij} - u)/B(\alpha_{Se_j}, \beta_{Se_j})} \\ &= \begin{cases} \frac{u^{d+\alpha_{Se_j}-1}(1-u)^{D_{ij}-d+\beta_{Se_j}-1}}{B(\alpha_{Se_j}+d, \beta_{Se_j}+D_{ij}-d)} & , 0 < u < 1 \\ 0 & , \text{else} \end{cases} \end{aligned}$$

Finally, we obtain a posteriori a beta distribution for $Se_j = u \mid TP_{ij} = d$ with

$$(Se_j = u \mid TP_{ij} = d) \sim Be(\alpha_{Se_j} + u; \beta_{Se_j} + D_{ij} - u).$$

Therefore, the posteriori distribution belongs to the same family as the priori distribution. □

Finally we achieve beta-binomial distributions (Held 2008) for TP_{ij} and TN_{ij}:

$$\begin{aligned} P(TP_{ij} = d) &= \int_0^1 P(TP_{ij} = d \mid Se_j = u)f(u; a_{Se_j}, b_{Se_j})\mathrm{d}u \\ &= \binom{D_{ij}}{d} \frac{1}{B(a_{Se_j}, b_{Se_j})} \int_0^1 u^{d+a_{Se_j}-1}(1-u)^{D_{ij}-d+b_{Se_j}-1}\mathrm{d}u \\ &= \binom{D_{ij}}{d} \frac{B(a_{Se_j} + d, D_{ij} - d + b_{Se_j})}{B(a_{Se_j}, b_{Se_j})}, \end{aligned}$$

(analogously defined for TN_{ij}).
The presented model is a true random effects model. That means every subject

(here: every study) has a single parameter (here: Se_j, Sp_j). These parameters follow a common distribution. Consequently, we still have a closed-form marginal distribution (Molenberghs and Verbeke 2006), which is a contrast to the most other random effect models like the four-dimensional GLMM.

To model the potential correlation between TP_{ij} and TN_{ij}, we apply the concept of copulas.

In our case, we have to construct a four-dimensional cumulative distribution function (cdf). Therefore we get using Sklar's Theorem (Sklar 1959):

$$H(x_1, x_2, x_3, x_4) = C(F_{Se_1}(x_1), F_{Sp_1}(x_2), F_{Se_2}(x_3), F_{Sp_2}(x_4)).$$

Now we determine the probability density function (pdf) simply by differentiating the joint distribution function. We obtain (considering the Radon-Nikodym derivative):

$$h(x_1, x_2, x_3, x_4) = f_{Se_1} f_{Sp_1} f_{Se_2} f_{Sp_2} c_{1234}(F_{Se_1}(x_1), F_{Sp_1}(x_2), F_{Se_2}(x_3), F_{Sp_2}(x_4)),$$

where c_{1234} denotes a 4-variate copula density.
The density is treated as the corresponding likelihood function. Note that in the copula case the likelihood is analytically determined and has a closed form where no random effects are necessary. Therefore standard maximum likelihood methods are used for parameter estimation. Similarly to the GLMM, 14 parameters have to be estimated: 2 for each beta-binomial distribution and 6 copula parameters which are used to model the associations between the random variables.

There is one technical issue that needs to be noticed. The used beta-binomially marginal distributions are discrete. In that case, the copula density have to be determined by differentiating with respect to the counting measure. According to Sklar's Theorem uniqueness of the copula is not guaranteed. Nevertheless, in the following we approximate the discrete margins by assuming continuous forms. The problem of discrete margins will be discussed in Section 3.2.7.

It is an advantage of the copula model that in principle a large number of copulas is available allowing for different association structures. This is a contrast to the standard model where a quadrivariate normal distribution is used as a single correlation structure. The class of potential copulas is only limited by the range of association. This range should be the whole one from perfectly negative (-1) to

perfectly positive (+1) because it is commonly known and described in Chapter 2 that sensitivity and specificity are mostly negatively correlated. Considering that and based on the previous work of Kuss et al. (2014), a quadrivariate Gaussian copula and quadrivariate vine copulas based on bivariate Plackett copulas are used.

3.2.4 Gaussian copula

The quadrivariate Gaussian copula is given by

$$C_G(u_1, u_2, u_3, u_4) = \Phi_4(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \Phi^{-1}(u_3), \Phi^{-1}(u_3), \Phi^{-1}(u_4) \mid \Gamma),$$

where $\Phi_4(. \mid \Gamma)$ is the distribution function of the quadrivariate normal distribution with corresponding correlation matrix Γ. Φ^{-1} refers to the inverse distribution function of the univariate standard normal distribution. The corresponding density of the Gaussian copula is defined in the following way (Song 2002):

$$c_G(u_1, u_2, u_3) = \mid \Gamma \mid^{-1/2} \exp\left\{\frac{1}{2} q^T (I_4 - \Gamma^{-1}) q\right\},$$

where $q = (q_1, q_2, q_3, q_4)^T$ with normal scores $q_j = \Phi^{-1}(u_j)$, $j = 1, 2, 3, 4$, and the four-dimensional identity matrix I_4.
The Gaussian copula is a member of the class of elliptical copulas which can be generalized to higher dimensions in a simple way. The association parameter should be interpreted as a Pearson correlation of the normal scores. That is numerically close to a Spearman correlation of the original margins (Song 2002).

3.2.5 Vine copulas

The most common known copulas are bivariate ones. The construction of higher dimensional copulas could be difficult. Essentially, there are three possibilities:

(i) Use of elliptical copulas, like the Gaussian copula. This class can easily extended to higher dimensions.

(ii) Use of Archimedean copulas. This is another class of copulas which can be extended to more than two dimensions. In our case they are improper based on two reasons. First, the range of the copula parameter is often not the whole one from -1 to $+1$. Second, the Clayton copula (Nelsen 2006) which could be

used, leads to unsatisfactory simulation results in the bivariate case as shown by Kuss et al. (2014).

(iii) Build higher dimensional copulas on basis of bivariate copulas. The concept is called pair-copula constructions (PCC) or vine copulas.

In the following the concept of vine copulas is used and therefore it is necessary to introduce it in a more detailed way, especially conditional copulas. The statistical background is also well described by Patton (2002).
Let $X = (X_1, ..., X_n)$ be a vector of random variables with joint density function $f(x_1, ..., x_n)$ which can be factorised as follows (Aas et al. 2009):

$$f(x_1, ..., x_n) = f(x_n)f(x_{n-1} \mid x_n)f(x_{n-2} \mid x_{n-1}, x_n)...f(x_1 \mid x_2, ..., x_n), \tag{3.4}$$

where the decomposition is unique up to exchanging of the variables. The corresponding derivation is given in the following remark.

Remark 3.9. Let $X = (X_1, ..., X_n)$ be a vector of random variables with joint density function $f(x_1, ..., x_n)$. Then, this density can be factorized as follows:

$$\begin{aligned} f(x_1, ..., x_n) &= f(x_1, ..., x_n) \prod_{i=0}^{n-2} \overbrace{\frac{f(x_{n-i}, ..., x_n)}{f(x_{n-i}, ..., x_n)}}^{=1} \\ &= f(x_n) \prod_{i=0}^{n-2} \frac{f(x_{n-i-1}, ..., x_n)}{f(x_{n-i}, ..., x_n)} \\ &= f(x_n) \prod_{i=0}^{n-2} f(x_{n-i-1} \mid x_{n-i}, ..., x_n). \end{aligned}$$

Using the unconditional version of Sklar's Theorem (Sklar 1959) and assuming an absolutely continuous distribution function F and absolutely continuous margins $F_1, ..., F_n$ we obtain (Aas et al. 2009):

$$\begin{aligned} f_{X_1,...,X_n}(x_1, ..., x_n) &= \frac{\partial^n}{\partial x_1 ... \partial x_n} F_{X_1,...,X_n}(x_1, ..., x_n) \\ &= \frac{\partial^n}{\partial u_1 ... \partial u_n} C(u_1, ..., u_n)|_{(u_1,...,u_n)=(F_{X_1}(x_1),...,F_{X_n}(x_n))} \\ &\cdot \frac{\partial}{\partial x_1} F_{X_1}(x_1) ... \frac{\partial}{\partial x_n} F_{X_n}(x_n) \\ &= c_{12...n}(F_{X_1}(x_1), ..., F_{X_n}(x_n)) f_{X_1}(x_1) ... f_{X_n}(x_n), \end{aligned}$$

where $c_{12...n}$ is the n-variate copula density.
This result has to be expanded to the conditional case, where a conditional version of Sklar's Theorem is necessary.

Using Theorem 3.3 the decomposition of the joint density in the conditional case is given by:

$$\begin{aligned} f_{X_1,...,X_n|Z}(x_1,...,x_n \mid z) =& c(F_{X_1|Z}(x_1 \mid z),...,F_{X_n|Z}(x_n \mid z) \mid z) \\ & f_{X_1|Z}(x_1 \mid z)...f_{X_n|Z}(x_n \mid z). \end{aligned}$$

For simplicity, the notation F_{12} is used instead of F_{X_1,X_2}, respectively for the other functions.
In the unconditional bivariate case we get

$$f_{12}(x_1,x_2) = c_{12}(F_1(x_1),F_2(x_2))f_1(x_1)f_2(x_2),$$

where $c_{12}(.,.)$ is the pair-copula density for the pair of transformed variables $F_1(x_1)$ and $F_2(x_2)$. Based on this relationship, in the conditional case we have

$$f_{1|2}(x_1 \mid x_2) = \frac{f_{12}(x_1,x_2)}{f_2(x_2)} = c_{12}(F_1(x_1),F_2(x_2))f_1(x_1).$$

In case of three random variables this leads to

$$f_{1|23}(x_1 \mid x_2,x_3) = c_{12|3}(F_{1|3}(x_1 \mid x_3),F_{2|3}(x_2 \mid x_3))f_{1|3}(x_1 \mid x_3),$$

for the appropriate pair-copula $c_{12|3}$. An alternative is:

$$f_{1|23}(x_1 \mid x_2,x_3) = c_{13|2}(F_{1|2}(x_1 \mid x_2),F_{3|2}(x_3 \mid x_2))f_{1|2}(x_1 \mid x_2), \tag{3.5}$$

in which the pair-copulas $c_{12|3}$ and $c_{13|2}$ are different.
A further decomposition of (3.5) leads to:

$$f_{1|23}(x_1 \mid x_2,x_3) = c_{13|2}(F_{1|2}(x_1 \mid x_2),F_{3|2}(x_3 \mid x_2))c_{12}(F_1(x_1),F_2(x_2))f_1(x_1),$$

where two pair-copulas are used.
It is shown by Aas et al. (2009) that each term in (3.4) can be decomposed using

the general formula

$$f(x \mid v) = c_{xv_j|v_{-j}}(F(x \mid v_{-j}), F(v_j \mid v_{-j}))f(x \mid v_{-j}),$$

for a d-dimensional vector v, where v_{-j} denotes the vector excluding the j-th component. Based on this relationship it is obvious that a multivariate density as shown in (3.4) can be expressed as a product of pair-copulas. Therefore it is necessary to determine the marginal conditional distribution $F(x \mid v)$. It is shown by Joe (1996) that this distribution is given by

$$F(x \mid v) = \frac{\partial C_{x,v_j|v_{-j}}(F(x \mid v_{-j}), F(v_j \mid v_{-j}))}{\partial F(v_j \mid v_{-j})},$$

where $C_{ij|k}$ is a bivariate copula distribution function. In case of a univariate distributed v we get

$$F(x \mid v) = \frac{\partial C_{xv}(F_x(x), F_v(v))}{\partial F_v(v)}.$$

Assuming that x and v are uniformly distributed, the function $h(x, v, \Theta)$ is used to express the conditional cdf. This implies

$$h(x, v, \Theta) = F(x \mid v) = \frac{\partial C_{x,v}(x, v, \theta)}{\partial v}, \tag{3.6}$$

where Θ denotes the set of copula parameters.

Vine copulas are a very flexible tool and there is a huge amount of possible pair-copula constructions. Bedford and Cooke (2001, 2002) give a graphical model, called 'regular vines', with the aim to organize them. The so defined class is still very general. Therefore we restrict in the following on two special cases of regular vines: the canonical vine (C-vine) and the drawable vine (D-vine) (Kurowicka and Cooke 2004). Each of them illustrates a specific decomposition of the density using nested trees.

D-vine copulas Figure 3.1 visualizes the concept of a four-dimensional D-vine copula. It involves three trees (T_j, $j = 1, 2, 3$) where tree T_j consists of $6 - j$ nodes and $5 - j$ edges. The edges illustrate the pair-copula densities where the labels correspond to the respective copula density, e.g. $14 \mid 23$ corresponds to $c_{14|23}$. In the whole decomposition $n(n-1)/2$ edges and marginal densities are used. No node in a tree is connected with more than two edges. The corresponding joint density can

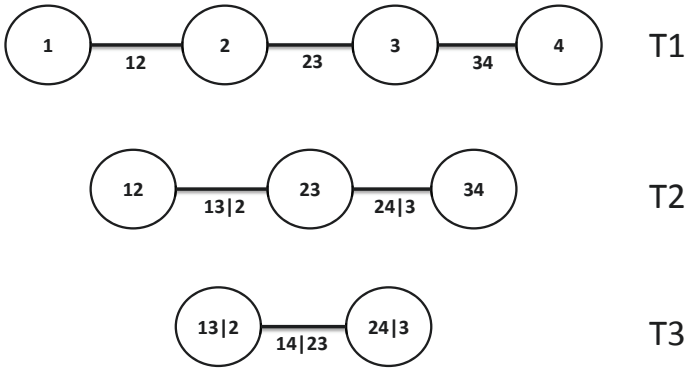

Figure 3.1: Visualization of a four-dimensional D-vine copula

be obtained trough (Aas et al. 2009)

$$\prod_{k=1}^{n} f(x_k) \prod_{j=1}^{n-1} \prod_{i=1}^{n-j} c_{i,i+j|i+1,\ldots,i+j-1}(F(x_i \mid x_{i+1}, ..., x_{i+j-1}), F(x_{i+j} \mid x_{i+1}, ..., x_{i+j-1})),$$

where j stands for the tree and i runs over the corresponding edges.

C-vine copulas A possible C-vine copula is illustrated in Figure 3.2. In such vine copulas each tree has got a unique node which is connected to $n - j$ edges. The corresponding density is written as (Aas et al. 2009)

$$\prod_{k=1}^{n} f(x_k) \prod_{j=1}^{n-1} \prod_{i=1}^{n-j} c_{j,j+i|1,\ldots,j-1}(F(x_j \mid x_1, ..., x_{j-1}), F(x_{j+i} \mid x_1, ..., x_{j-1})).$$

In our current case it is necessary to model quadrivariate joint densities using vine

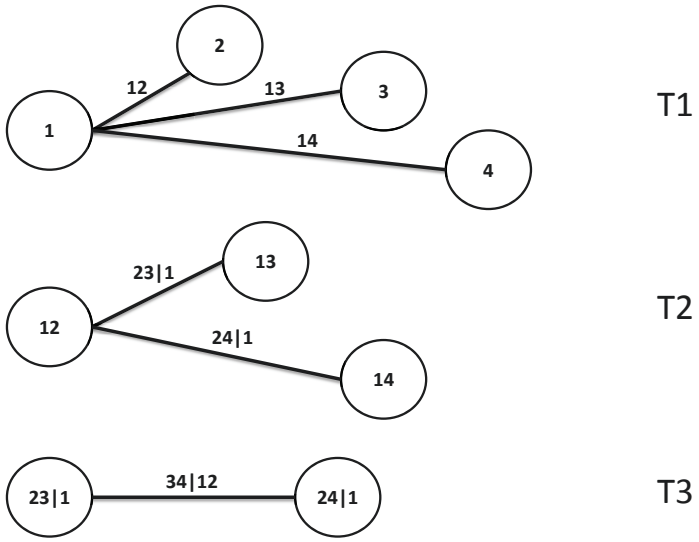

Figure 3.2: Visualization of a four-dimensional C-vine copula

copulas. Therefore we get (Aas et al. 2009)

$$\begin{aligned} f(x_1, x_2, x_3, x_4) =& f(x_1)f(x_2)f(x_3)f(x_4) \\ & c_{12}(F(x_1), F(x_2))c_{13}(F(x_1), F(x_3))c_{14}(F(x_1), F(x_4)) \\ & c_{23|1}(F(x_2 \mid x_1), F(x_3 \mid x_1))c_{24|1}(F(x_2 \mid x_1), F(x_4 \mid x_1)) \\ & c_{34|12}(F(x_3 \mid x_1, x_2), F(x_4 \mid x_1, x_2)), \end{aligned}$$

as a possible C-vine copula.
A possible D-vine structure is given by

$$\begin{aligned} f(x_1, x_2, x_3, x_4) =& f(x_1)f(x_2)f(x_3)f(x_4) \\ & c_{12}(F(x_1), F(x_2))c_{23}(F(x_2), F(x_3))c_{34}(F(x_3), F(x_4)) \\ & c_{13|2}(F(x_1 \mid x_2), F(x_3 \mid x_2))c_{24|3}(F(x_2 \mid x_3), F(x_4 \mid x_3)) \\ & c_{14|23}(F(x_1 \mid x_2, x_3), F(x_4 \mid x_2, x_3)). \end{aligned}$$

In the following, the derivation of a four-dimensional C-vine copula is given. We assume that a four-dimensional density $F_{1234}(x_1, x_2, x_3, x_4)$ is decomposed as

$$f_{1234}(x_1, x_2, x_3, x_4) = f_1(x_1)f_{2|1}(x_2 \mid x_1)f_{3|12}(x_3 \mid x_1, x_2)f_{4|132}(x_4 \mid x_1, x_3, x_2). \quad (3.7)$$

It yields

$$f_{2|1}(x_2 \mid x_1) = c_{12}(F_1(x_1), F_2(x_2))f_2(x_2),$$

and

$$\begin{aligned}
f_{3|12}(x_3 \mid x_1, x_2) &= \frac{f_{23|1}(x_2, x_3 \mid x_1)}{f_{2|1}(x_2 \mid x_1)} \\
&= \frac{c_{23|1}(F_{2|1}(x_2 \mid x_1), F_{3|1}(x_3 \mid x_1))f_{3|1}(x_3 \mid x_1)f_{2|1}(x_2 \mid x_1)}{f_{2|1}(x_2 \mid x_1)} \\
&= c_{23|1}(F_{2|1}(x_2 \mid x_1), F_{3|1}(x_3 \mid x_1))f_{3|1}(x_3 \mid x_1) \\
&= c_{23|1}(F_{2|1}((x_2 \mid x_1), F_{3|1}(x_3 \mid x_1))c_{13}(F_1(x_1), F_3(x_3))f_3(x_3).
\end{aligned}$$

Furthermore we get

$$\begin{aligned}
f_{4|132}(x_4 \mid x_1, x_3, x_2) &= \frac{f_{34|12}(x_3, x_4 \mid x_1, x_2)}{f_{3|12}(x_3 \mid x_1, x_2)} \\
&= \frac{c_{34|12}(F_{3|12}(x_3 \mid x_1, x_2), F_{4|12}(x_4 \mid x_1, x_2))}{f_{3|12}(x_3 \mid x_1, x_2)} \\
&\quad \frac{f_{3|12}(x_3 \mid x_1, x_2)f_{4|12}(x_4 \mid x_1, x_2)}{f_{3|12}(x_3 \mid x_1, x_2)} \\
&= c_{34|12}(F_{3|12}(x_3 \mid x_1, x_2), F_{4|12}(x_4 \mid x_1, x_2))f_{4|12}(x_4 \mid x_1, x_2) \\
&= c_{34|12}(F_{3|12}(x_3 \mid x_1, x_2), F_{4|12}(x_4 \mid x_1, x_2))\frac{f_{24|1}(x_2, x_4 \mid x_1)}{f_{2|1}(x_2 \mid x_1)} \\
&= c_{34|12}(F_{3|12}(x_3 \mid x_1, x_2), F_{4|12}(x_4 \mid x_1, x_2)) \\
&\quad \frac{c_{24|1}(F_{2|1}(x_2 \mid x_1), F_{4|1}(x_4 \mid x_1))f_{2|1}(x_2 \mid x_1)f_{4|1}(x_4 \mid x_1)}{f_{2|1}(x_2 \mid x_1)} \\
&= c_{34|12}(F_{3|12}(x_3 \mid x_1, x_2), F_{4|12}(x_4 \mid x_1, x_2)) \\
&\quad c_{24|1}(F_{2|1}(x_2 \mid x_1), F_{4|1}(x_4 \mid x_1))f_{4|1}(x_4 \mid x_1) \\
&= c_{34|12}(F_{3|12}(x_3 \mid x_1, x_2), F_{4|12}(x_4 \mid x_1, x_2)) \\
&\quad c_{24|1}(F_{2|1}(x_2 \mid x_1), F_{4|1}(x_4 \mid x_1)c_{14}(F_1(x_1), F_4(x_4))f_4(x_4).
\end{aligned}$$

Combining with (3.7) gives

$$\begin{aligned}
f_{1234}(x_1, x_2, x_3, x_4) = f_1(x_1)f_2(x_2)f_3(x_3)f_4(x_4) \\
c_{12}(F_1(x_1), F_2(x_2))c_{13}(F_1(x_1), F_3(x_3))_{14}(F_1(x_1), F_4(x_4)) \\
c_{23|1}(F_{2|1}(x_2 \mid x_1), F_{3|1}(x_3 \mid x_1))c_{24|1}(F_{2|1}(x_2 \mid x_1), F_{4|1}(X_4 \mid x_1)) \\
c_{34|12}(F_{3|12}(x_3 \mid x_1, x_2), F_{4|12}(x_4 \mid x_1, x_2))
\end{aligned}$$

as a possible canonical (C-) vine decomposition.
Starting from a different decomposition as used in Equation (3.7), we could obtain a possible drawable (D-) vine copula.

To sum up, there exist twelve different D-vine decompositions and twelve different C-vine decompositions where none of the D-vine ones is equal to any of the C-vine ones. Therefore in total we have 24 different decompositions.

3.2.6 Plackett copula

It was shown in the bivariate case by Kuss et al. (2014) that the Plackett copula is appropriate in case of a meta-analysis of diagnostic accuracy studies. Based on their simulation results, the decision is made that the four-dimensional vine copulas are constructed on the basis of bivariate Plackett copulas.
The bivariate Plackett copula is defined as:

$$C_P(u,v) = \frac{(1+(\theta-1)(u+v)) - \sqrt{(1+(\theta-1)(u+v))^2 - 4uv\theta(\theta-1)}}{2(\theta-1)}\,, \theta > 0,\ \theta \neq 1,$$

with the corresponding density

$$c_P(u,v) = ((1+(\theta-1)(u+v))^2 - 4\theta(\theta-1)uv)^{-\frac{3}{2}}\theta(1+(\theta-1)(u+v-2uv)).$$

In case of the Plackett copula, Spearman's correlation coefficient is given by

$$\rho_S = \frac{\theta+1}{\theta-1} - \frac{2\theta}{(\theta-1)^2}\ln(\theta)$$

with $\rho_S \in (-1,1)$.
Constructing vine copulas, determination of conditional distribution functions is necessary. Using Equation (3.6) we get

$$F(x_i \mid x_j) = \frac{1}{2}\left[1 - \frac{1+(\theta_{ij}-1)(x_i+x_j) - 2x_i\theta_{ij}}{(1+(\theta_{ij}-1)(x_i+x_j)^2 - 4x_ix_j\theta_{ij}(\theta_{ij}-1)^{1/2}}\right]$$

where θ_{ij} denotes the association between the i-th and the j-th variable. The more complex conditional cdfs, like $F(x_3 \mid x_1, x_2)$, are constructed on basis of $F(x_i \mid x_j)$.

3.2.7 Copulas with discrete margins

The above presented copula approach has one main disadvantage. We use beta-binomial distributions as marginal distributions which are discrete. Based on Sklar's Theorem, the joint distribution associated with a copula is guaranteed to be unique only in case of continuous copulas. Therefore it is necessary to think about possibilities to account for the problem of discrete marginal distributions. In the recent literature there are different approaches addressing this task, for example Genest and Nešlehová (2007) or Smith and Khaled (2012). Song et al. (2002, 2009) propose a method to construct a discrete version of the Gaussian copula using multivariate normal distributions. This approach is also used in practice (He et al. 2012). In the work of Kuss et al. (2014), the authors simulate such a version of a bivariate Gaussian copula and show that it is numerically more unstable than treating the discrete margins as continuous. The approach does not lead to better results as they support by their simulation study. Additionally, it is also difficult to implement the Song approach using SAS, the software that we use, because multivariate distribution functions have to evaluated at different points. Therefore this method is not presented in details in the following. Instead, we use a recent method to construct discrete D-vine copulas as given by Panagiotelis et al. (2012). This approach is in the following illustrated using four dimensions because that is the case we need.
For $m = 4$ dimensions the probability mass function (pmf) of the joint density can be decomposed in the following way:

$$\begin{aligned} P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3, Y_4 = y_4) = {} & P(Y_1 = y_1 \mid Y_2 = y_2, Y_3 = y_3, Y_4 = y_4) \\ & \times P(Y_4 = y_4 \mid Y_2 = y_2, Y_3 = y_3) \\ & \times P(Y_3 = y_3 \mid Y_2 = y_2) \\ & \times P(Y_2 = y_2). \end{aligned} \tag{3.8}$$

Now, we rewrite the first term on the right-hand side using the concept presented by Panagiotelis et al. (2012). This leads to

$$\begin{aligned}
&P(Y_1 = y_1 \mid Y_2 = y_2, Y_3 = y_3, Y_4 = y_4) \\
&\quad = \frac{\sum_{i_1 \in \{0,1\}} \sum_{i_4 \in \{0,1\}} (-1)^{i_1+i_4} C_{14|23}(F_{1|23}(y_1 - i_1 \mid y_2, y_3), F_{4|23}(y_4 - i_4 \mid y_2, y_3))}{P(Y_4 = y_4 \mid Y_2 = y_2, Y_3 = y_3)} \\
&\quad = C_{14|23}(F_{1|23}(y_1 \mid y_2, y_3), F_{4|23}(y_4 \mid y_2, y_3)) \\
&\qquad - C_{14|23}(F_{1|23}(y_1 \mid y_2, y_3), F_{4|23}(y_4 - 1 \mid y_2, y_3)) \\
&\qquad - C_{14|23}(F_{1|23}(y_1 - 1 \mid y_2, y_3), F_{4|23}(y_4 \mid y_2, y_3)) \\
&\qquad + C_{14|23}(F_{1|23}(y_1 - 1 \mid y_2, y_3), F_{4|23}(y_4 - 1 \mid y_2, y_3)),
\end{aligned}$$

because the denominator cancels out with the second term. Using another result from the Panagiotelis publication, $F_{1|23}, F_{1|3}$ and $F_{2|3}$ can be rewritten as

$$\begin{aligned}
f_{1|23}(y_1 - i_1 \mid y_2, y_3) = &\frac{1}{P(Y_2 = y_2 \mid Y_3 = y_3)} \{ \\
&C_{12|3}(F_{1|3}(y_1 - i_1 \mid y_3), F_{2|3}(y_2 \mid y_3)) \\
&- C_{12|3}(F_{1|3}(y_1 - i_1 \mid y_3), F_{2|3}(y_2 - 1 \mid y_3))\}
\end{aligned}$$

with

$$\begin{aligned}
&P(Y_2 = y_2 \mid Y_3 = y_3) = \\
&\qquad \frac{\sum_{i_2 \in \{0,2\}} \sum_{i_3 \in \{0,1\}} (-1)^{i_2+i_3} C_{23}(F_2(y_2 - i_2), F_3(y_3 - i_3))}{P(Y_3 = y_3)} \\
&\qquad = \frac{1}{P(Y_3 = y_3)} \{C_{23}(F_2(y_2), F_3(y_3)) - C_{23}(F_2(y_2), F_3(y_3 - 1)) \\
&\qquad - C_{23}(F_2(y_2 - 1), F_3(y_3)) + C_{23}(F_2(y_2 - 1), F_3(y_3 - 1))\}
\end{aligned}$$

and

$$F_{1|3}(y_1 - i_1 \mid y_3) = \frac{C_{13}(F_1(y_1 - i_1), F_3(y_3)) - C_{13}(F_1(y_1 - i_1), F_3(y_3 - 1))}{P(Y_3 = y_3)}$$

and

$$F_{2|3}(y_2 \mid y_3) = \frac{C_{23}(F_2(y_2), F_3(y_3)) - C_{23}(F_2(y_2), F_3(y_3 - 1))}{P(Y_3 = y_3)}$$

with

$$P(Y_3 = y_3) = F_3(y_3) - F_3(y_3 - 1).$$

The term $F_{4|23}$ can be rewritten analogously. After this, the third term of decomposition (3.8) can be formulated as

$$\begin{aligned} P(Y_3 = y_3 \mid Y_2 = y_2) &= \frac{\sum_{i_3 \in \{0,1\}} \sum_{i_2 \in \{0,1\}} (-1)^{i_3 + i_2} C_{32}(F_3(y_3 - i_3), F_2(y_2 - i_2))}{P(Y_2 = y_2)} \\ &= C_{32}(F_3(y_3), F_2(y_2)) - C_{32}(F_3(y_3), F_2(y_2 - 1)) \\ &\quad - C_{32}(F_3(y_3 - 1), F_2(y_2)) + C_{32}(F_3(y_3 - 1), F_2(y_2 - 1)), \end{aligned}$$

where the term $P(Y_2 = y_2)$ cancels out with the last term in (3.8).
Finally, the complete expression for the pmf of a discrete D-vine copula is given and can be implemented for simulation using the Plackett copula as basis.

4 Simulation studies

To compare the quadrivariate generalized linear mixed model, denoted by the term 'GLMM', to the copula models, a simulation study was conducted. We also included the approach proposed by the Cochrane Collaboration (Macaskill et al. 2010) to investigate which model performs better. The simulation was divided into different sections to present it in a clear way. First, the performance of the GLMM using different link functions and integral/ data approximations was investigated and compared to the Cochrane approach. This part is also described by Hoyer and Kuss (2016). Second, the GLMM using the identity and logit link in combination with PQL was compared to the copula models. An overview of the compared models is given in Figure 4.1. The simulation program was written in SAS 9.3 (SAS Institute Inc., Cary, NC, USA).

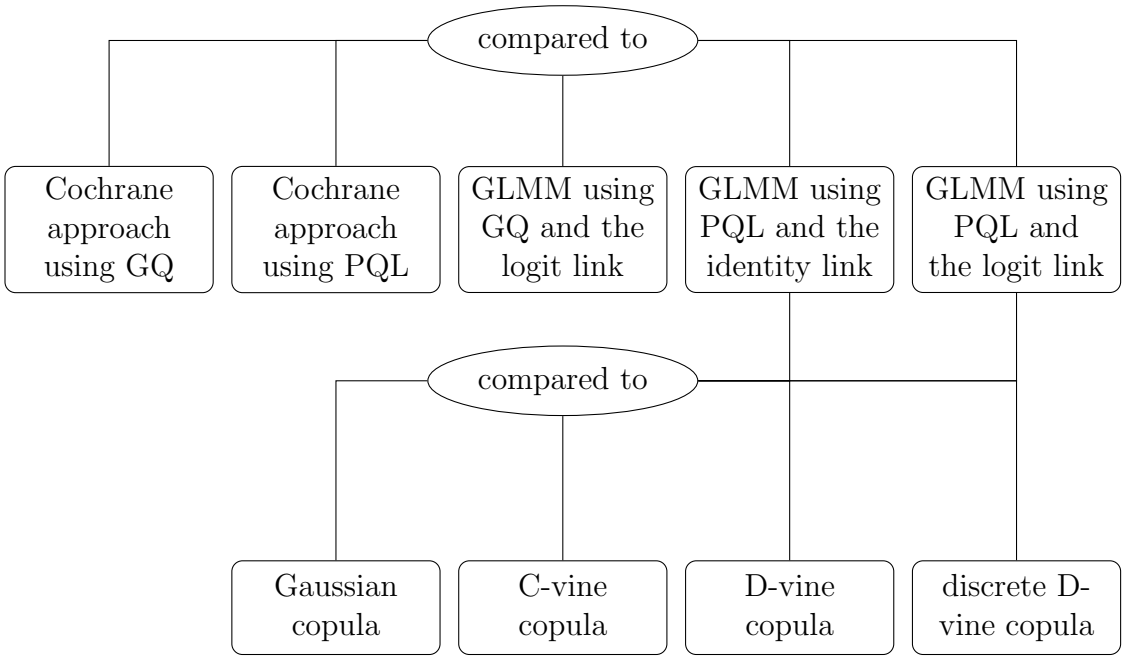

Figure 4.1: Overview of the compared models. Abbreviations: GQ: Gaussian quadrature, PQL: penalized quasi likelihood

4.1 Simulation setting

After analyzing a convenience sample from meta-analysis data (Kodama et al. 2013, Picano et al. 2000) that give a widespread impression of possible estimates of sensitivity and specificity, the following variables were varied:

- the true sensitivity$_1$ (70%), specificity$_1$ (80%), sensitivity$_2$ (65%, 70%, 80%), specificity$_2$ (75%, 80%, 90%) and as a conclusion the true difference of sensitivities (-10 percentage points (pp), 0 pp, 5 pp) and the true difference of specificities (-10 pp, 0 pp, 5 pp)

- the true association between sensitivity$_1$, specificity$_1$, sensitivity$_2$ and specificity$_2$. The three assumed correlation matrices are:

$$\Gamma_{none} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \; \Gamma_{neg} = \begin{pmatrix} 1.0 & -0.3 & -0.2 & -0.3 \\ -0.3 & 1.0 & -0.3 & -0.2 \\ -0.2 & -0.3 & 1.0 & -0.3 \\ -0.3 & -0.2 & -0.3 & 1.0 \end{pmatrix},$$

$$\Gamma_{mix} = \begin{pmatrix} 1.0 & -0.3 & 0.2 & -0.3 \\ -0.3 & 1.0 & -0.3 & 0.2 \\ 0.2 & -0.3 & 1.0 & -0.3 \\ -0.3 & 0.2 & -0.3 & 1.0 \end{pmatrix}.$$

Γ_{none} assumes that the sensitivities and specificities across the studies and even the two tests are completely independent. For the other two matrices, we have chosen a negative correlation of -0.3 between sensitivity and specificity of each test because negative correlations between these two variables are most likely to appear in reality. In case of a negative correlation structure, the sensitivities and specificities of both tests are assumed to be negatively correlated, too. Using a mixed correlation structure, we model a positive correlation of 0.2 between sensitivity$_1$ and sensitivity$_2$ and specificity$_1$ and specificity$_2$. The entries of the correlation matrices are inspired by the data set from Kodama et al. (2013).
We assume 0.27 as the true variance of sensitivity$_1$, specificity$_1$, sensitivity$_2$ and specificity$_2$ on the logit scale which corresponds to a variance of 0.02 on the original [0,1]-scale.

4.2 Data generation

During data generation, we were confronted with one problem. It is difficult to generate random samples out of vine copulas. An algorithm according to that problem is presented by Aas et al. (Aas et al. 2009), but it is not possible to use it in case when the PCC is based on bivariate Plackett copulas because of the inversion of some functions. This problem is also remarked by Hoyer and Kuss (2015). Therefore, in the following only the standard model and the Gaussian copula are used to generate data for the simulation.
After combining the design parameters, we get 54 different simulation scenarios. For each, 1000 meta-analyses were generated. The simulated number of studies varies between 10 and 30 and their sizes between 30 and 200, respectively, assuming uniform distributions. Based on the study size, the number of diseased persons were sampled from a uniform distribution. Figure 4.2 shows the process of data generation. This choice is motivated by different meta-analyses reported in practice, for example by Menke (2010) or Kodama et al. (2013).

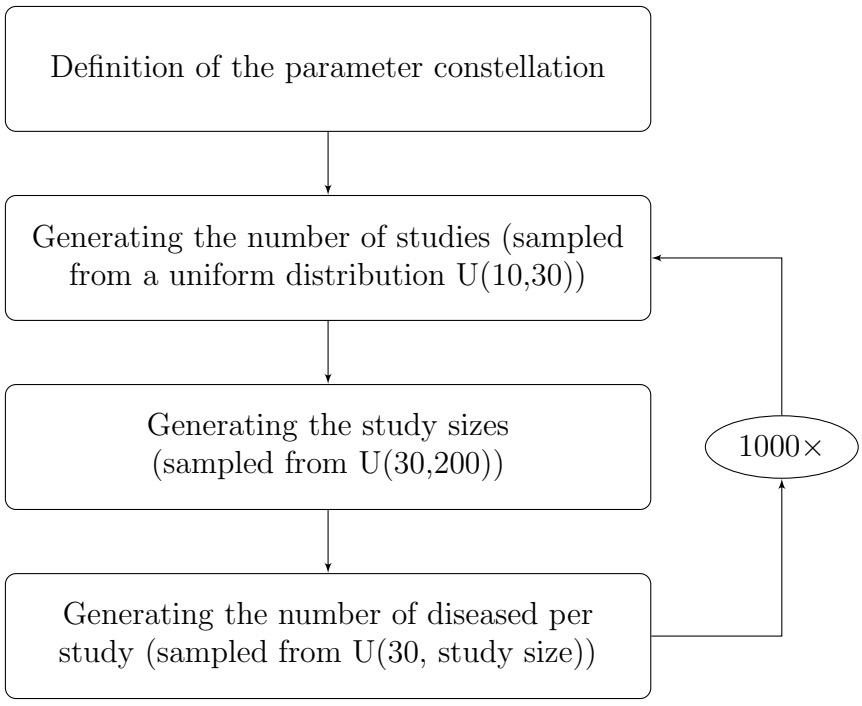

Figure 4.2: Generating the data sets for simulation

4.2.1 Generalized linear mixed model

To generate the observed numbers of true positives and true negatives of the single studies, the VNORMAL call in SAS/IML is used. This procedure enables us to create quadrivariate normally distributed random vectors where the corresponding covariance matrix is chosen to reproduce the true values of correlation:

$$u_{ik} \sim N_4(0, \Gamma_x),\ k = 1, \ldots, 4,$$

where x represents the chosen correlation structure. Then, these random numbers are added to the true values of the two sensitivities and specificities and used to calculate logit-transformed values for them:

$$\text{logit}(Se_{ij}^*) = \log\left(\frac{Se_{j,true}}{1 - Se_{j,true}}\right) + u_{ik},\ k = 1, 3,$$

$$\text{logit}(Sp_{ij}^*) = \log\left(\frac{Sp_{j,true}}{1 - Sp_{j,true}}\right) + u_{ik},\ k = 2, 4.$$

After this, an expit-transformation leads to the values for sensitivity$_1$, specificity$_1$, sensitivity$_2$ and specificity$_2$:

$$Se_{ij}^* = \frac{\exp(\text{logit}(Se_j^*))}{1 + \exp(\text{logit}(Se_j^*))},$$

$$Sp_{ij}^* = \frac{\exp(\text{logit}(Sp_j^*))}{1 + \exp(\text{logit}(Sp_j^*))}.$$

Finally, these are multiplied by the number of diseased and healthy and rounded to the nearest integer to get the numbers of true positives and true negatives of each study for both tests:

$$TP_{ij} \approx Se_{ij}^* N_{D_i^+},$$

$$TN_{ij} \approx Sp_{ij}^* (1 - N_{D_i^+}),$$

where $N_{D_i^+}$ is the number of diseased per study.

4.2.2 Gaussian copula

Generating random numbers from the Gaussian copula is done by the SOLVE statement from the SAS MODEL procedure. To achieve beta-distributed values, the relationship between the true values of the two sensitivities and specificities, their

corresponding variances, the expected random numbers and the distribution parameters is used. These parameters are the basis for the number of true positives and true negatives of each study for both tests.

4.3 Estimation methods

For each of the simulated meta-analyses, 14 parameters have to be estimated. In case of the standard model, these are: the two sensitivities, the two specificities, their variances and the correlations between them. The same number of parameters have to estimated when copula models are used. Here, these are two parameters for each beta-binomial distribution (in total 8) and the corresponding copula parameters to model the association (in total 6).
While we are using the default options in SAS, the estimated value and the corresponding 95% t-confidence interval are computed.

We compare 3 implementations of the generalized linear mixed model:

- penalized-quasi-likelihood and the logit link (implemented using PROC GLIMMIX)
- penalized-quasi-likelihood and the identity link (implemented using PROC GLIMMIX)
- Gaussian quadrature and the logit link (implemented using PROC NLMIXED).

This is done in order to check which approximation works less biased and numerically more stable. Therefore we compare a numerical integral approximation (Gaussian quadrature) to a data approximation method for generalized linear mixed models (PQL). For these statistical features we refer to the established literature, e.g. Molenberghs and Verbeke 2006 or Fahrmeir et al. 2009 and Section 2.3.2. We also compare two different link functions. In case of PQL estimation and the logit link, we have the problem that the confidence intervals have to be calculated by hand using the delta method. These models are compared to two different implementations of the Cochrane approach. Here, we also use the PQL method compared to Gaussian quadrature.
In case of copulas, we used four different ones:

- Gaussian copula,
- C-vine copula based on bivariate Plackett copulas,

- D-vine copula based on bivariate Plackett copulas, and
- discrete version of a D-vine copula based on bivariate Plackett copulas

For each copula, we construct the corresponding density as it was pointed out in Chapter 3 and use the maximum likelihood concept to estimate the parameters of interest.
Figure 4.3 shows the flow diagram of the program execution.

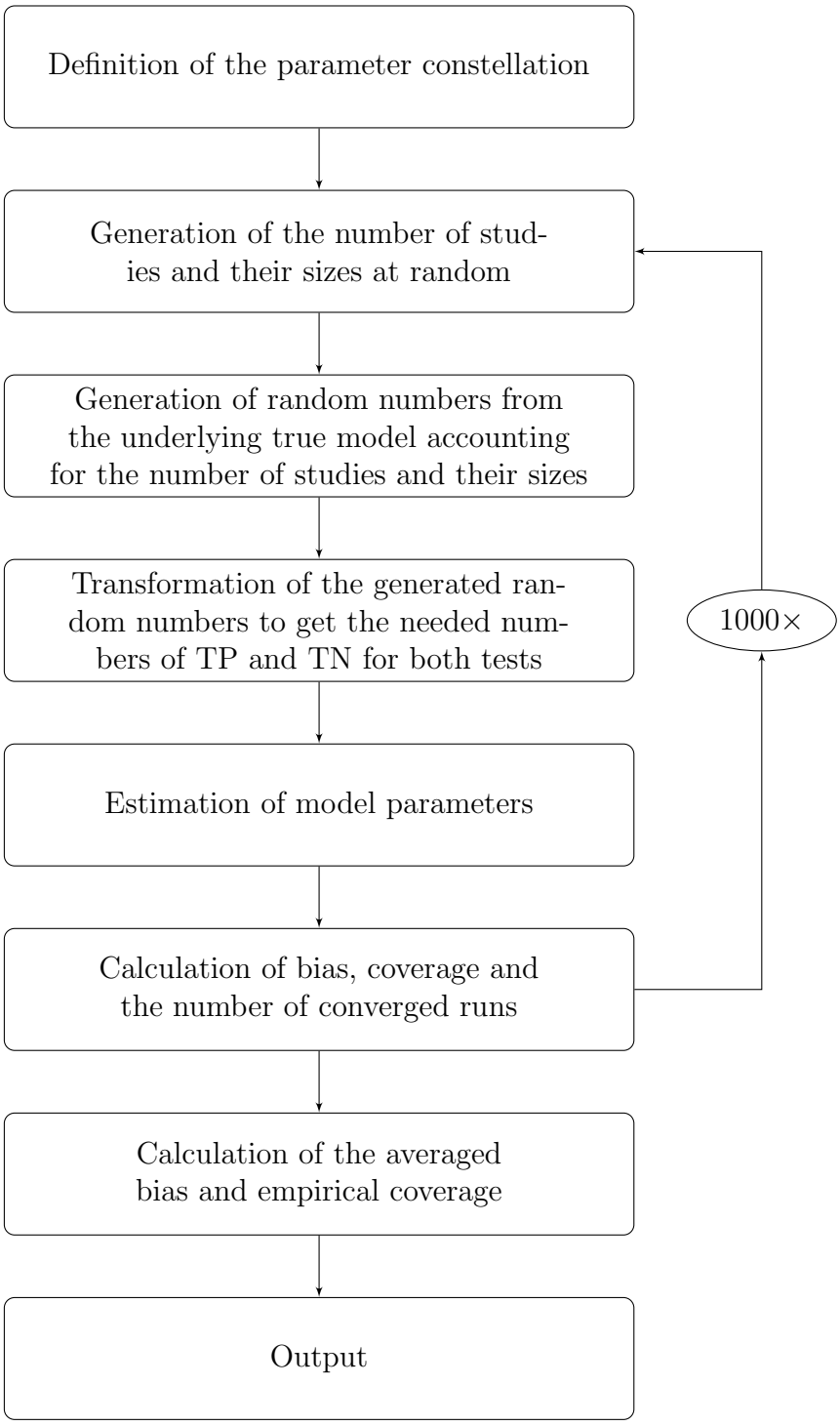

Figure 4.3: Program execution

4.4 Results

We present the results in different sections. First, the results of the different implementations of the generalized linear mixed model and the Cochrane approach are compared. Second, the copula models are compared to the GLMM.

Our parameters of interest are the differences between the sensitivities and specificities of the two tests. Therefore, in reporting of our results we restrict to them. For these parameters, we show bias, coverage and the number of converged simulation runs. Thereby, coverage means the proportion that an estimated confidence interval contains the true value. To address numerical robustness, we count the number of converged simulation runs that can reach a maximum number of 1000. Results for the estimated two sensitivities and specificities are also calculated but are not shown to avoid too much data. Nevertheless, these results are in the same order as the presented ones.

To refer to the different simulation settings, the abbreviations shown in Table 4.1 are used. The results are visualized in different figures. The numerical values are

Table 4.1: Simulated settings

Setting	Sensitivity_1	Specificity_1	Sensitivity_2	Specificity_2	Difference of sensitivities	Difference of specificities
A	0.7	0.8	0.80	0.90	-0.10	-0.10
B	0.7	0.8	0.80	0.75	-0.10	0.05
C	0.7	0.8	0.80	0.80	-0.10	0.00
D	0.7	0.8	0.65	0.90	0.05	-0.10
E	0.7	0.8	0.65	0.75	0.05	0.05
F	0.7	0.8	0.65	0.80	0.05	0.00
G	0.7	0.8	0.70	0.90	0.00	-0.10
H	0.7	0.8	0.70	0.75	0.00	0.05
I	0.7	0.8	0.70	0.80	0.00	0.00

presented in the tables in the Appendix.

4.4.1 Generalized linear mixed model

Bias In terms of bias all models performed nearly similar, except in a few situations. In cases where the first tests specificity was better and the corresponding

sensitivity was worse (scenario B), the model using penalized-quasi-likelihood and the logit link did not seem to depend on the correlation structure because the bias was nearly the same for every correlation matrix. On the contrary, the identity link performed better in situations where no correlation is present. Using that model, situations where a negative correlation structure is assumed, led to the worst results. Here, the differences of sensitivities and specificities were overestimated. Using Gaussian quadrature the worst results were reached in case of a negative correlation structure. Analogous results were achieved in cases where the sensitivity of the first test was better and the specificity was lower, respectively. In case of differences of specificities, the models seemed to be less sensitive in situations without correlation and a mixed structure. It was striking that the model using the identity link performs worst in case of negative correlations in scenario D. A similar performance of all models could be found in scenario H with negative correlations. The bias of the different implementations of the Cochrane approach had the same magnitude as the most quadrivariate models and especially a bit higher than the PQL model with logit link. To sum up, the model using the logit link seemed to perform best compared to the others. Figures 4.4 and 4.5 support these observations.

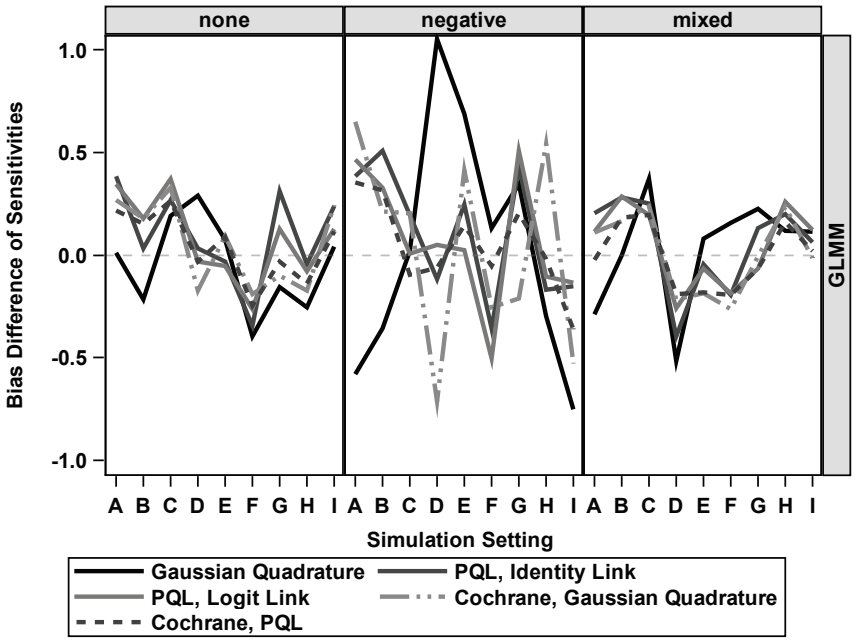

Figure 4.4: Bias difference of sensitivities: averaged values over 1000 meta-analyses. True model=GLMM with none, negative and mixed correlation structures

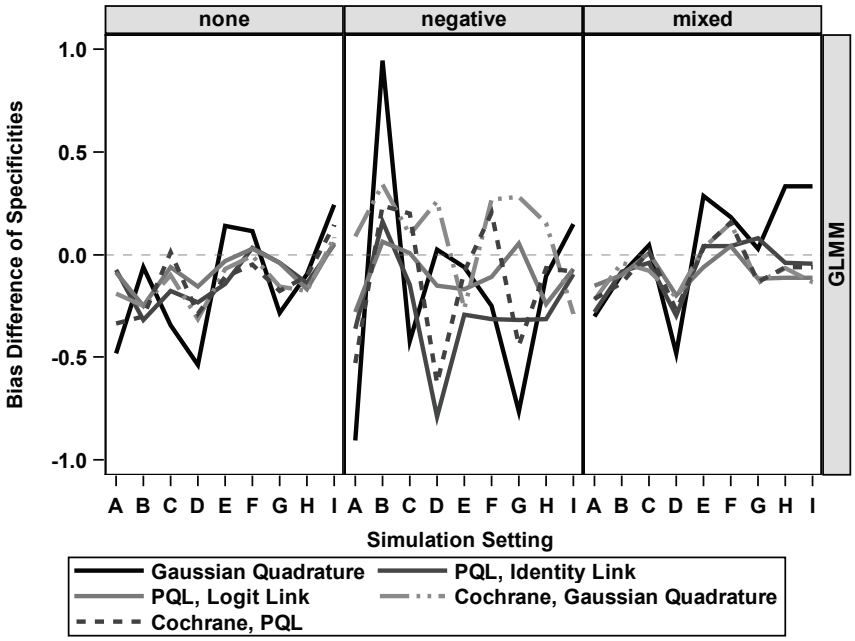

Figure 4.5: Bias difference of specificities: averaged values over 1000 meta-analyses. True model=GLMM with none, negative and mixed correlation structures

Coverage With respect to coverage, all GLMMs yielded results close to the expected 95%. The results were satisfactory, because we propose multivariate models with a quadrivariate response which are numerically extremely difficult and susceptible. In case of estimated differences of sensitivities, the coverage of all models was the best when no correlation is present. Here, Gaussian quadrature seemed to have an advantage over the others. In case of existing correlations, both models using PQL performed similar and better than the model using Gaussian quadrature. This model behaved worse in scenarios B and D, especially when negative correlations were simulated. With a view on the estimated differences of specificities, the performance was nearly similar compared to the differences of sensitivities. The results were bit worse in case of none correlation, but more consistent when there is some underlying correlation. The model using Gaussian quadrature led to some results over the expected 95% in case of a negative correlation structure (scenarios E, H). Especially in case of the Cochrane approach, the coverage was nearly in every case smaller than the expected 95%. This was not satisfactory and we suggest a worse estimation of confidence intervals using the Cochrane approach. The complete results

are shown in Figures 4.6 and 4.7.

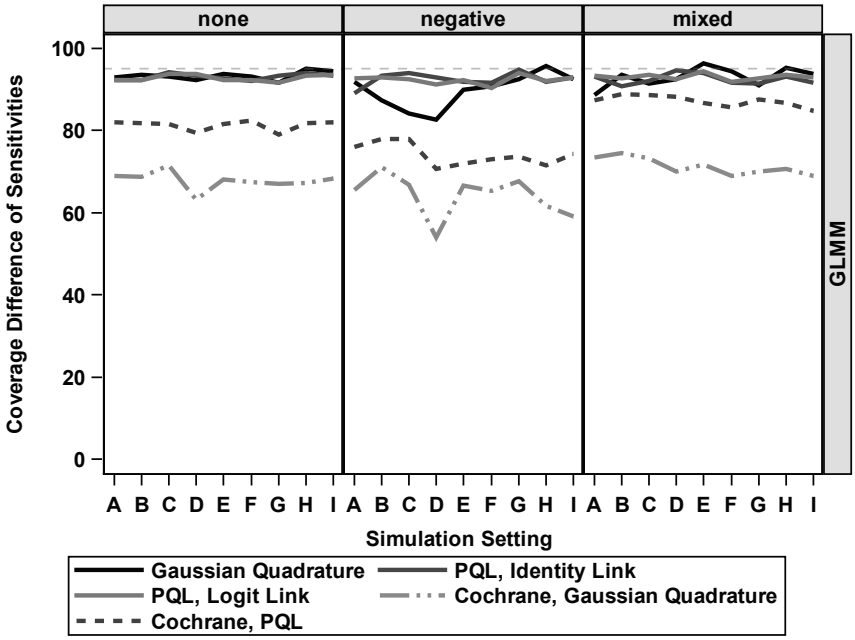

Figure 4.6: Coverage difference of sensitivities. True model=GLMM with none, negative and mixed correlation structures

Convergence In terms of convergence it was obvious that none of the models reached the maximum number of 1000 converged runs. The worst results were got in cases with negative underlying correlations. The model using the logit link was always the best and the model with Gaussian quadrature was always the worst. The performance of the model using the identity link depended on the underlying simulation setting. It seemed to be fragile in the scenarios A, D and G. Because of the bivariate character of the Cochrane model, the numerical robustness was quite good as expected in terms of convergence. The results can be found in Figure 4.8.

4.4.2 Copula models compared to the GLMM

Bias At first, the results are interpreted when the true underlying model was the GLMM. In that case, the Gaussian copula behaved similar to the generalized linear mixed model. The estimation of differences of specificities seemed to be more vulnerable. No clear dependence on the underlying correlation structure was visible,

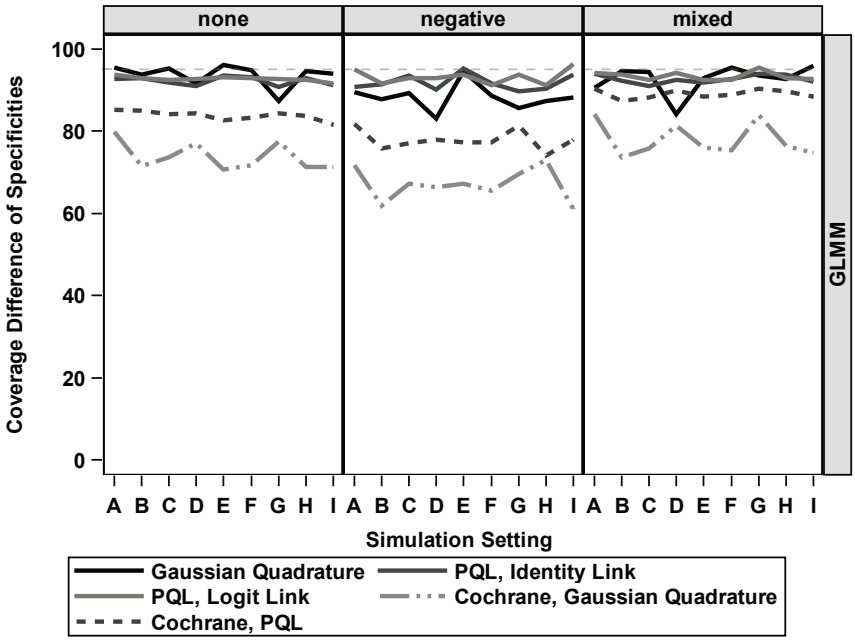

Figure 4.7: Coverage difference of specificities. True model=GLMM with none, negative and mixed correlation structures

because there exist situations where any correlation was absent and the estimation was more biased than in cases with correlation. This phenomenon occurred for example in scenario D. Using vine copulas for parameter estimation, there was sometimes a larger bias compared to the other models. Especially the D-vine copula led to the biggest deviation of all models, for example about 1.4 pp in setting A. Here, the results seemed to depend on the underlying correlation structure. That means, the vine copula models led to the largest bias in case of a negative correlation. Interestingly, the discrete copula yielded the most accurate results of the vine models. Generally, all models seemed to overestimate the difference of sensitivity, visible by a positive bias in case of a higher sensitivity of the first test and a lower one of the second test. Especially, larger differences of about 10 pp were overestimated. On that basis, mostly scenario A led to the greatest bias. Except for the D-vine copula, the models overestimated positive differences of specificity. This was shown by a negative bias in case of a lower specificity of diagnostic test 1 and a higher of test 2. To sum up, the models seemed to overestimate the true effects.

In cases where the Gaussian copula was the true underlying model, the estimates from all models are more biased. As expected, the Gaussian copula behaved in

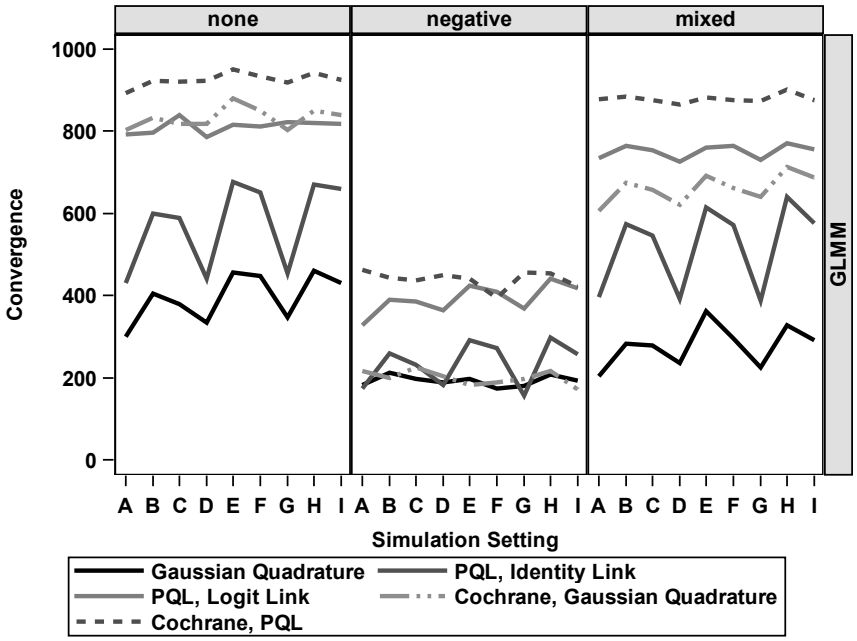

Figure 4.8: Convergence. True model=GLMM with none, negative and mixed correlation structures

cases of none and negative correlation best because itself was the underlying true model. Remarkable were worse results of that copula in case of a mixed correlation structure. Here, we observed a bias up to 6 pp. The GLMM using the logit link yielded generally to better results than the model with the identity link. Mainly setting A was the most biased one for differences of sensitivities. The C-vine and the discrete copula worked better when no or a negative correlation underlies. Contrary to that, the C- and D-vine copula behaved similar in case of a mixed correlation structure where the discrete copula worked clearly worse. The estimation of the differences of specificities was more unstable using the Gaussian copula and the GLMM. Here, the Gaussian copula led to best results in case of negative correlation. When a mixed correlation structure underlies, the standard model as well as the vine copula approaches worked better than the Gaussian copula. The estimation of the differences of specificities was more stable using vine copulas. The discrete copula behaved similar to the C-vine, except in cases of negative correlation where the C-vine led to worse results. The bias from the D-vine copula was clearly larger compared with both other in cases with none or negative correlation structures.

As a conclusion, the estimation of the effects seemed to depend on the underlying

correlation structure, especially using the vine copula models, and the estimates were mainly a bit overestimated and therefore too optimistic. Another main conclusion was that the results are worse and more unstable when the random variables are generated using the Gaussian copula.
The complete results are depicted in Figures 4.9, 4.10, 4.11 and 4.12.

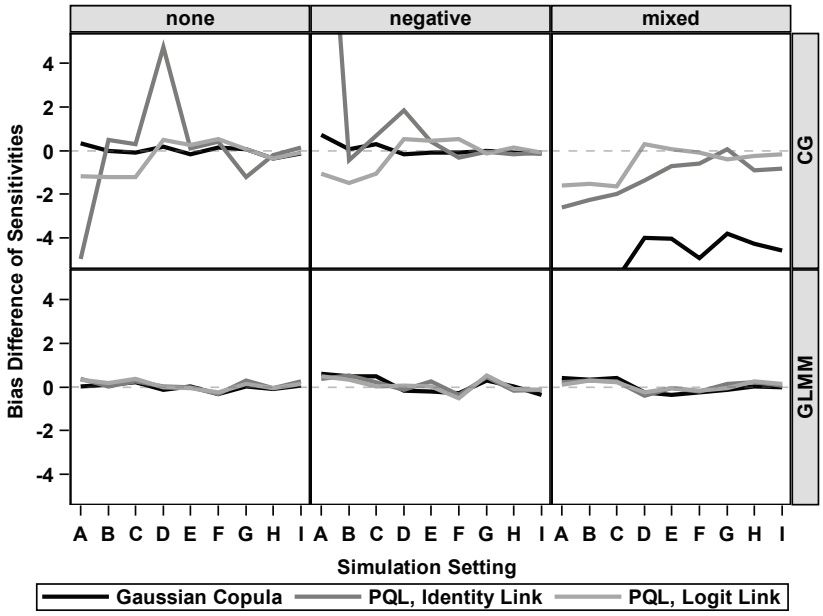

Figure 4.9: Bias difference of sensitivities: averaged values over 1000 meta-analyses. True model=GLMM or Gaussian copula (CG) with none, negative and mixed correlation structures

Coverage Generally, it was obvious that the expected value of 95% for the empirical coverage was in less cases observed. The problem estimating the confidence intervals was also reported in simpler models in the meta-analysis of only one diagnostic test, e.g. by Paul et el. (2010), Kuss et al. (2014) or Hoyer and Kuss (2015).

In terms of coverage in our conducted simulation study, we observed that both, the GLMM and the Gaussian copula model, behaved similar when the true model is the GLMM. The results were satisfactory and close to 95% in every case independent of the correlation structure. The C- and D-vine copulas worked at least as good as the Gaussian copula and the GLMM and in some cases even better. That means in some cases a coverage over 95% was reached, e.g. in scenario G and H with a mixed

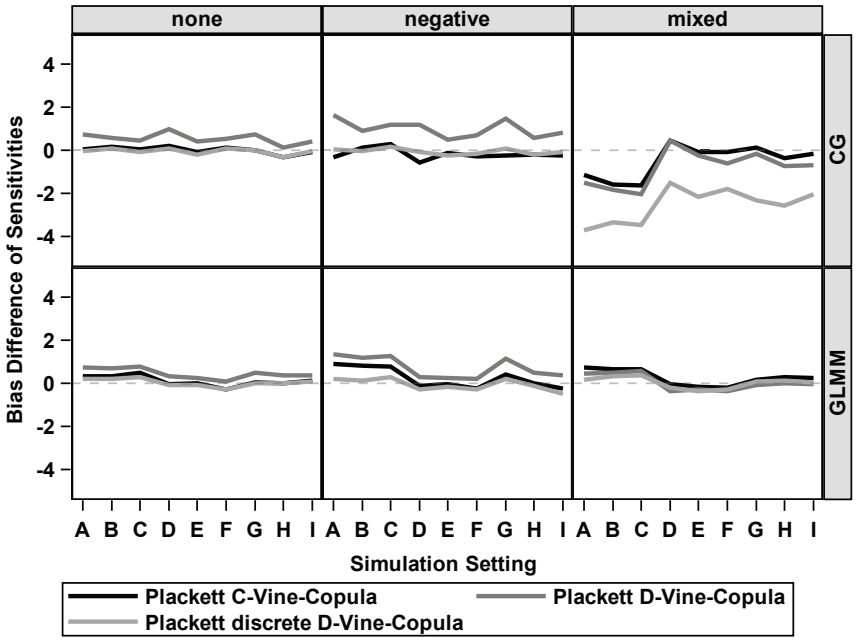

Figure 4.10: Bias difference of sensitivities: averaged values over 1000 meta-analyses. True model=GLMM or Gaussian copula (CG) with none, negative and mixed correlation structures

correlation structure. The discrete copula performed worst, especially in the mixed correlation case.

When the Gaussian copula was the true underlying model, the reached coverage of differences of sensitivities was in case of none and negative correlation satisfactory. It was noticeable that the usage of the identity link sometimes led to a coverage over 95%, e.g. in setting D and G. On the other hand, the model in setting A combined with negative correlation performed bad. The performance of the Gaussian copula was worse in case of mixed correlation which was also observed for the estimated bias. In case of none correlation, the C- and D-vine worked best. The discrete copula was always a bit worse. Under a mixed correlation, the vine copula led to the worst results. The GLMM using the logit link performed constant satisfactory for the differences of specificities, too. The usage of the identity link yielded to worse results in setting D and G under none and negative correlation. Under mixed correlation structures, the Gaussian copula performed worst even for the differences of specificities, as observed before. There was no considerable difference in the goodness of the vine copula models between the estimation of the difference of

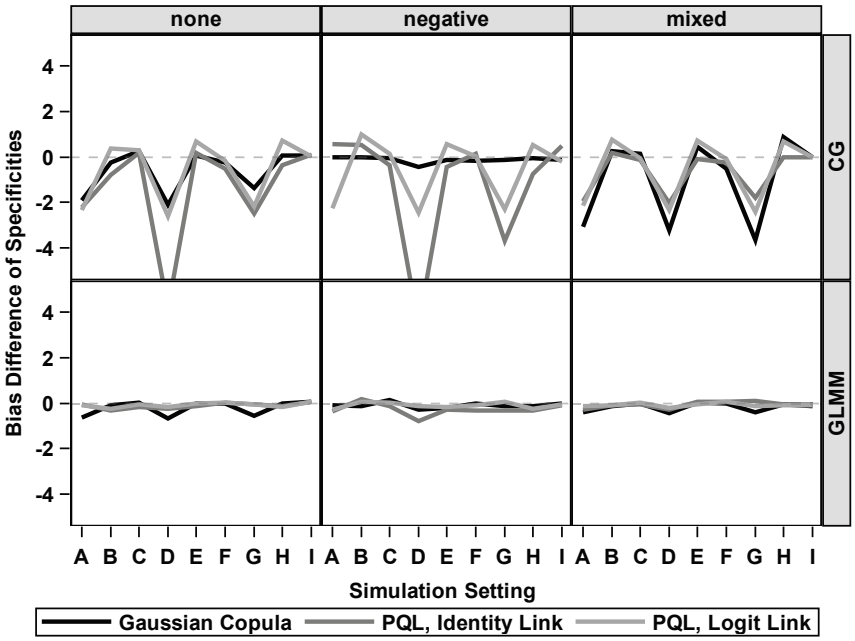

Figure 4.11: Bias difference of specificities: averaged values over 1000 meta-analyses. True model=GLMM or Gaussian copula (CG) with none, negative and mixed correlation structures

sensitivities and specificities, respectively.

As a conclusion, the results seemed to depend on the underlying correlation structure when the true model is the Gaussian copula. Especially the vine copula models performed worse with increasing complexity, e.g. mixed correlation.

The results are shown in Figures 4.13, 4.14, 4.15 and 4.16.

Convergence In terms of convergence and numerical robustness, there were clear differences between the different models.

When the GLMM was the true underlying model, then all of the copula models performed always better than the GLMM independent from the correlation structure. The GLMM led to worse results in case of negative correlation where the usage of the identity link was the worst. Scenarios A, D and G were the most problematic ones. The C- and D-vine approaches performed nearly similar and better than the Gaussian copula and the GLMM. The discrete copula converged less than the other vine copula models, especially using a mixed correlation structure.

When the true model was the Gaussian copula, the model himself converged in cases of none and negative correlation nearly in every setting. Just under a mixed

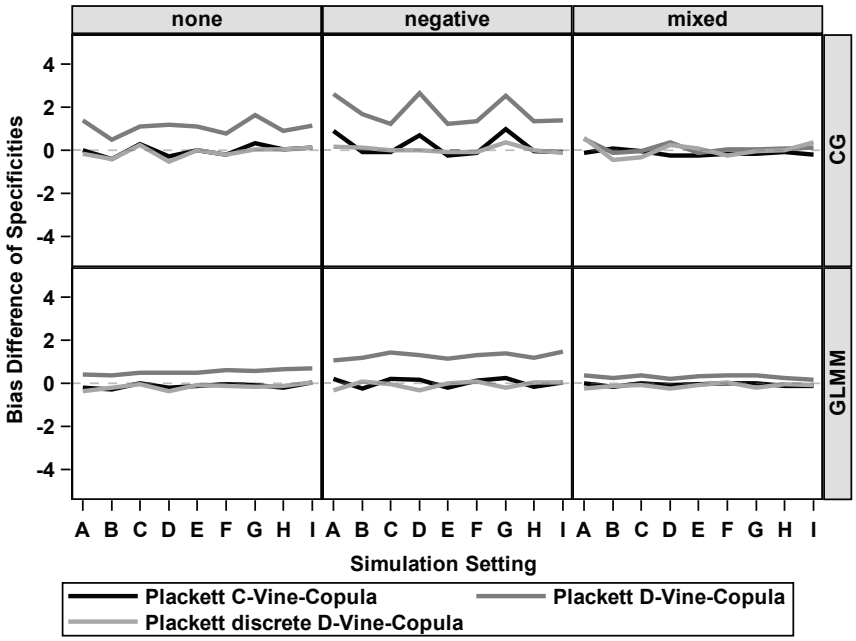

Figure 4.12: Bias difference of specificities: averaged values over 1000 meta-analyses. True model=GLMM or Gaussian copula (CG) with none, negative and mixed correlation structures

correlation, the Gaussian copula performed worse than the GLMM using a logit link. This was expected regarding the results of bias and coverage. The vine copula converged under none and negative correlation nearly ever but the results for mixed correlation structures are even worst (between 300 and 500 times). In that cases the vines are sometimes worse than the GLMM using the logit link. The GLMM with the identity link was clearly the worst in setting A.

As a conclusion, the situations with underlying mixed correlation seemed to be the hardest with some numerical problems in parameter estimation. The GLMM with the identity link converged in less cases. Especially the copula models seemed to perform numerical robust compared to the generalized linear models with random effects.

The results are visualized in Figures 4.17 and 4.18.

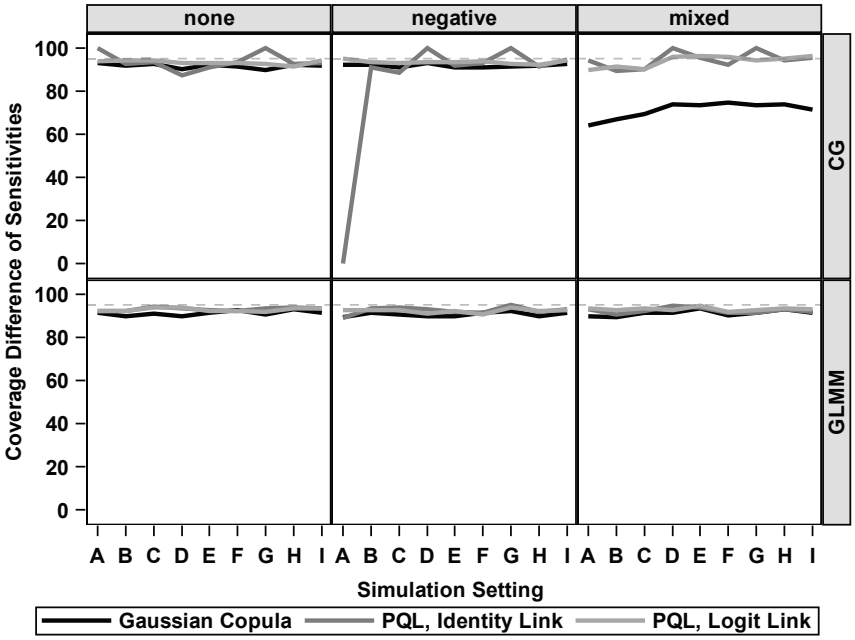

Figure 4.13: Coverage difference of sensitivities. True model=GLMM or Gaussian copula (CG) with none, negative and mixed correlation structures

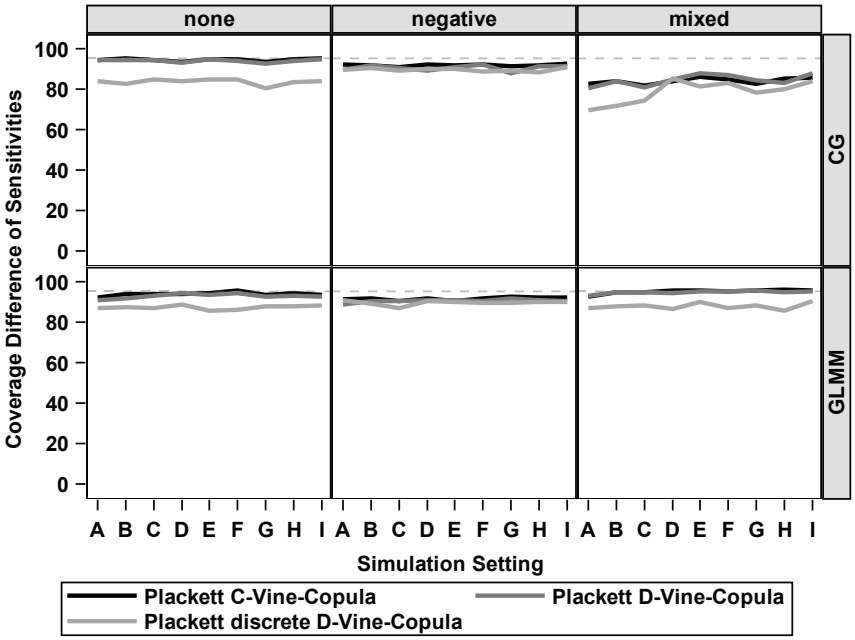

Figure 4.14: Coverage difference of sensitivities. True model=GLMM or Gaussian copula (CG) with none, negative and mixed correlation structures

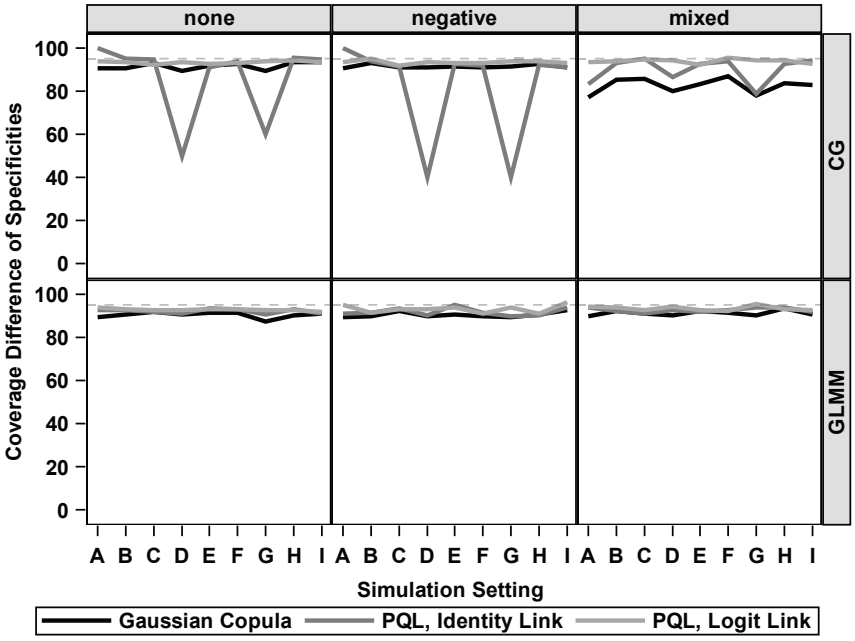

Figure 4.15: Coverage difference of specificities. True model=GLMM or Gaussian copula (CG) with none, negative and mixed correlation structures

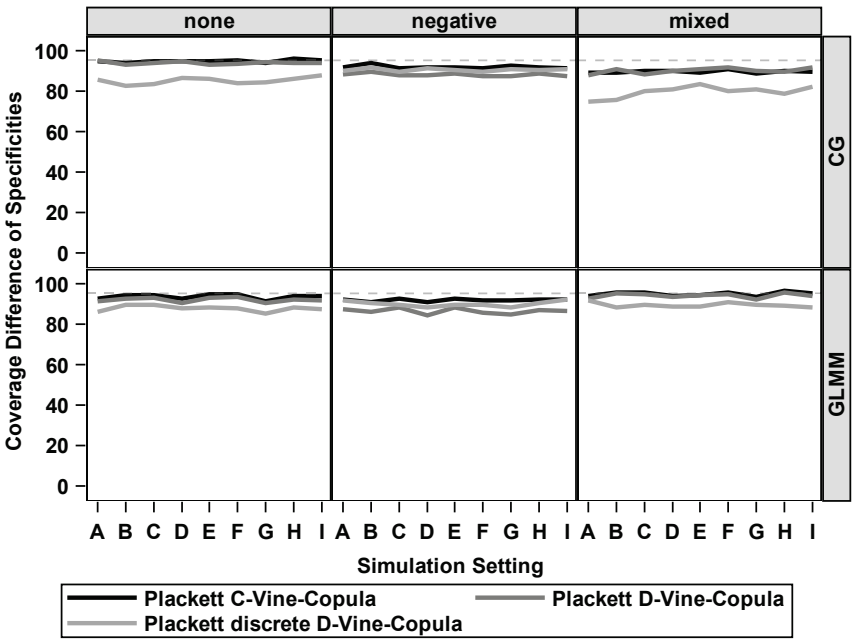

Figure 4.16: Coverage difference of specificities. True model=GLMM or Gaussian copula (CG) with none, negative and mixed correlation structures

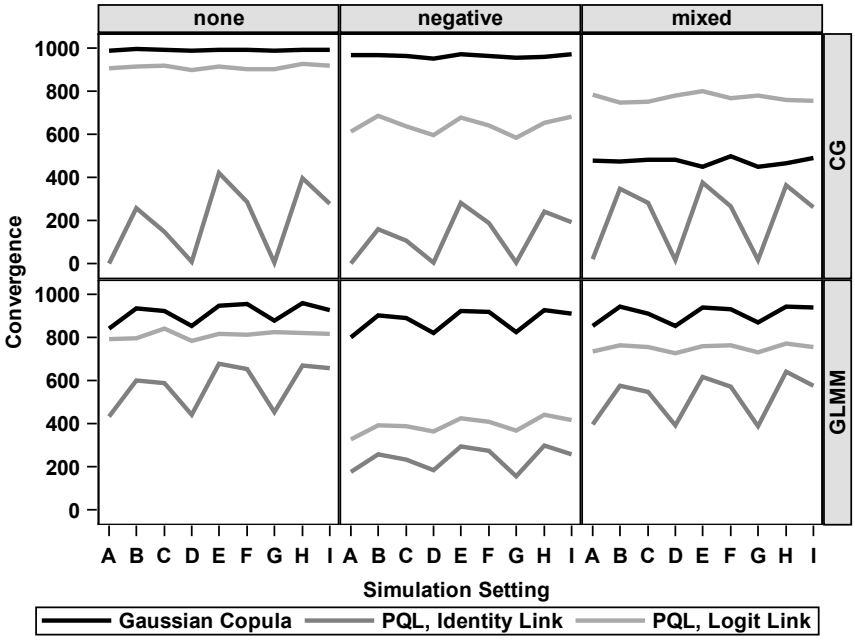

Figure 4.17: Convergence. True model=GLMM or Gaussian copula (CG) with none, negative and mixed correlation structures

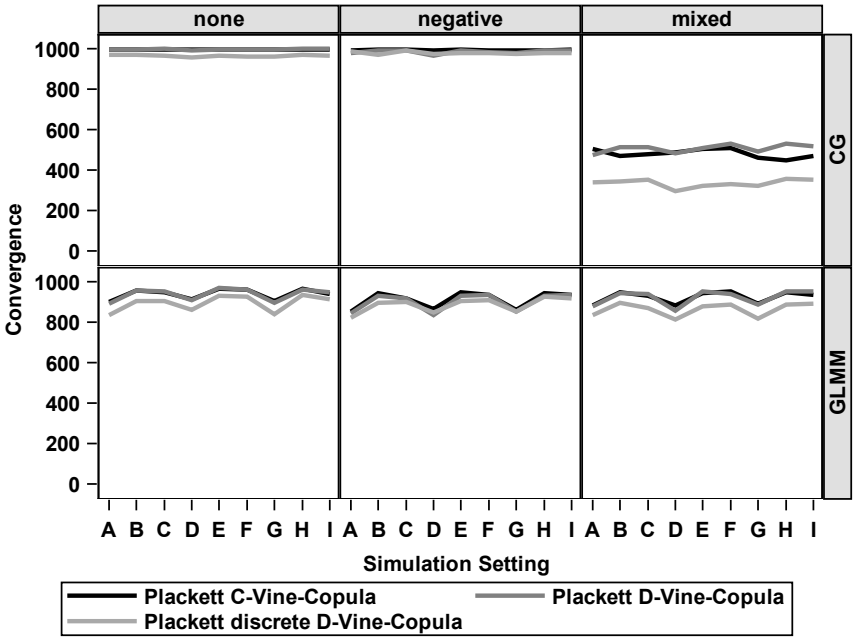

Figure 4.18: Convergence. True model=GLMM or Gaussian copula (CG) with none, negative and mixed correlation structures

5 Exemplary applications

The following section illustrates the application of the developed models using two different data sets from research practice. Both examples stem from different medical backgrounds. The first example investigates the performance of two tests in the diagnosis of coronary artery disease, the second compares two tests in the screening of type 2 diabetes mellitus. The authors of the publications are not able to give summary measures like the difference of sensitivities and specificities to address the diagnostic accuracy. Thus, it can be seen that there is a need for the development of the considered methods.

5.1 Picano data set

The aim of the meta-analysis from Picano et al. (2000) was to investigate the diagnostic accuracies of dobutamine-stress and dipyridamole-stress echocardiographies. These methods are widespread in the diagnosis of coronary artery disease. Dipyridamole is mostly used in Europe whereas dobutamine is more widespread in the USA. Several medical reasons for that are pointed out by Picano et al. (2000). As a conclusion they wanted to compare both tests according to their performance in terms of sensitivity and specificity. Finally, they collected twelve studies where each proband underwent both tests. The data are shown in Table 5.1.

The meta-analytic estimates can be found in Tables 5.2 and 5.3 as well as in Figures 5.1 and 5.2. Depending on the chosen model, the estimated sensitivity of dipyridamole ranges between 66% and 69%. This is slightly lower compared to the sensitivity of dobutamine which is about 75%. Contrary to that, the estimated specificity of dipyridamole that ranges from 92% to 95%, is higher than the specificity of dobutamine which is in between 75% and 76%. Every model estimates a clearly negative difference of sensitivities favoring dobutamine. The differences ranging from -8.2 pp using the Gaussian copula to -5.8 pp using the GLMM with the logit link and Gaussian quadrature. Both of the other GLMMs using PQL estimation lead to analogous results from about 6.8 to 7.0 pp as the Plackett copula models. The

Table 5.1: Echocardiography dataset from Picano et al. (2000), first test: dipyridamole, second test: dobutamine

Study	TP1	FN1	FP1	TN1	TP2	FN2	FP2	TN2
Martin et al.	14	11	4	5	19	6	6	3
Salustri et al.	18	10	2	16	16	12	4	14
Previtali et al.	34	23	1	22	45	12	4	19
Beleslin et al.	88	31	1	16	98	21	4	13
Gruber et al.	12	5	2	22	12	5	2	22
Dagianti et al.	13	12	1	34	18	7	1	34
Sochowski et al.	16	8	3	19	17	7	4	18
Pingitore et al.	75	17	1	17	77	15	3	15
San Roman et al.	49	14	1	38	49	14	2	37
Minardi et al.	32	12	1	2	33	11	1	2
Santoro et al.	18	15	1	26	20	13	1	26
Batlle et al.	34	7	0	15	33	8	1	14

estimated differences of specificities are always positive. That means that each model leads to a higher specificity of dipyridamole. The results range from 7.0 pp using the GLMM with the PQL method and the logit link to 9.6 pp using the Gaussian copula. The other copula models reach a difference from about 8.0 pp. It is obvious that the Gaussian copula model leads to the most conspicuous results in cases of sensitivity and specificity, respectively. Picano et al. performed a single meta-analysis per test and reached similar results but without accounting for the various association between the two tests. They got a difference of sensitivities from about -9 pp and a difference of specificities from about 10 pp. Our new models account for the potential correlations and we yield differences which are a bit smaller. Taking into account that the simulation study showed that the new models overestimate the parameters a bit, the results from Picano et al. seem to be too high and the differences between the two tests are not as large as they appear.

Table 5.2: Estimated sensitivities and specificities (in %) for the echocardiography data set using the different models

Model	Sensitivity Dipyridamole [95% CI]	Specificity Dipyridamole [95% CI]	Sensitivity Dobutamine [95% CI]	Specificity Dobutamine [95% CI]
GLMM GQ (logit link)	68.6 [64.6; 72.7]	95.4 [92.8; 98.0]	74.5 [71.0; 78.0]	86.1 [81.8; 90.3]
GLMM PQL (identity link)	68.6 [62.3; 74.8]	92.2 [87.7; 96.8]	75.3 [70.5; 80.1]	85.2 [77.4; 92.9]
GLMM PQL (logit link)	68.6 [61.9; 74.6]	92.5 [85.6; 96.3]	75.3 [70.1; 80.0]	85.6 [75.0; 92.1]
Gaussian Copula	66.3 [59.7; 72.9]	90.3 [-; -]	74.5 [68.9; 80.1]	80.6 [73.7; 87.6]
Plackett C-Vine Copula	69.1 [63.1; 75.1]	92.1 [87.6; 96.6]	76.2 [71.4; 81.0]	84.3 [76.4; 92.2]
Plackett D-Vine Copula	69.1 [63.1; 75.1]	92.1 [87.6; 96.6]	76.2 [71.4; 81.0]	84.3 [76.4; 92.2]
Plackett discrete D-Vine Copula	69.1 [63.1; 75.1]	92.1 [87.6; 96.6]	75.9 [71.2; 80.7]	84.1 [76.1; 92.0]

Table 5.3: Estimated differences of sensitivities and specificities (in percentage points) for the echocardiography data set using the different models

Model	Difference of sensitivities [95% CI]	Difference of specificities [95% CI]
GLMM GQ (logit link)	-5.8 [-9.2; -2.4]	9.3 [4.4; 14.2]
GLMM PQL (identity link)	-6.8 [-12.3; -1.3]	7.1 [1.6; 12.5]
GLMM PQL (logit link)	-6.8 [-12.6; -0.9]	7.0 [0.7; 13.3]
Gaussian Copula	-8.2 [-13.8; -2.6]	9.6 [2.7; 16.6]
Plackett C-Vine Copula	-7.0 [-14.7; 0.6]	7.8 [-1.3; 17.0]
Plackett D-Vine Copula	-7.0 [-14.7; 0.6]	7.8 [-1.3; 17.0]
Plackett discrete D-Vine Copula	-6.8 [-14.4; 0.8]	8.0 [-1.1; 17.2]

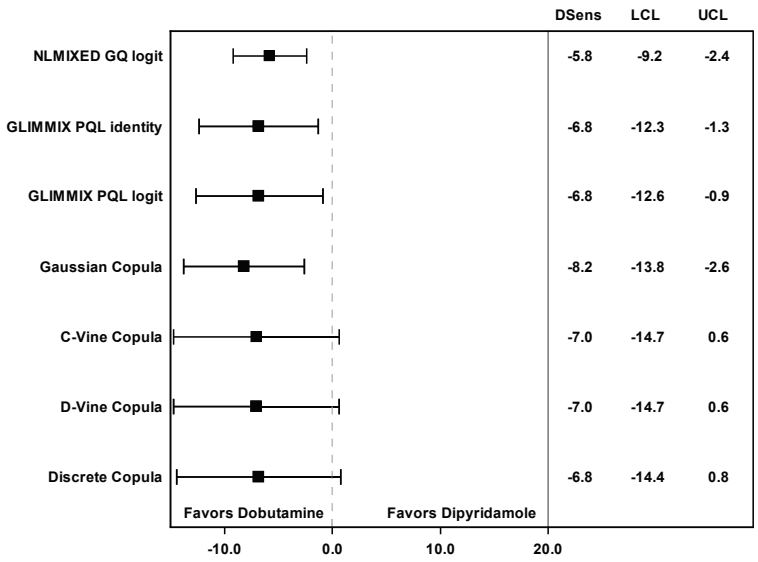

Figure 5.1: Estimates for the Picano data set: difference of sensitivities, Abbreviations: DSens - difference of sensitivities, LCL - lower confidence limit, UCL - upper confidence limit

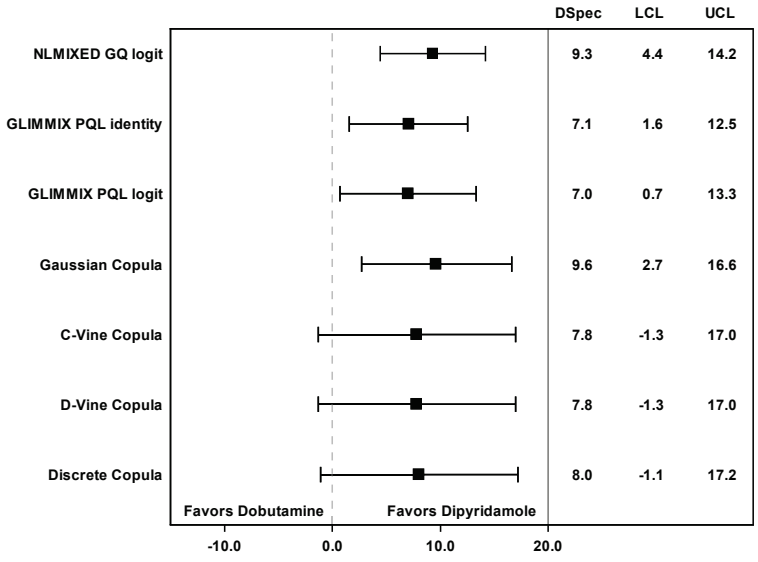

Figure 5.2: Estimates for the Picano data set: difference of specificities, Abbreviations: DSpec - Difference of specificities, LCL - lower confidence limit, UCL - upper confidence limit

5.2 Type 2 diabetes data set

With a second example, we will illustrate our new models using two existing systematic reviews (Kodama et al. 2013, Bennett et al. 2007) on the population-based screening of type 2 diabetes mellitus. This disease is due to a progressive insulin secretory defect on the background of insulin resistance (American Diabetes Association 2015). In principle, three methods are available to diagnose type 2 diabetes: the 2-h plasma glucose value after a 75-g oral glucose tolerance test (OGTT), measurement of glycated heamoglobin A1c (HbA_{1c}) and measurement of fasting plasma glucose (FPG). In the two systematic reviews, the single studies use mainly the OGTT as gold standard and compare HbA_{1c} to FPG. Unfortunately, the situation is a bit more complicated because it occurs that the study-specific reference standards also includes information on HbA_{1c} or FPG. Potentially, this may lead to favouring one of the two tests. However, we ignore these subtleties for the apllication of our models. The same is done by Kodama et al. (2013) where the different reference standards were also ignored. HbA_{1c}, as well as FPG, is more rapid and less unpleasant than the OGTT. HbA_{1c} has additional advantages to FPG and OGTT like greater convenience (because fasting is not required meaning that patients are not requested to refrain from eating and drinking any liquids other than water before the testing procedure), more preanalytical stability and less day-to-day variability during stress and illness. This is pointed out in a comprehensively by the American Diabetes Association (American Diabetes Association) (2015) and is important in a possible screening setting. In both systematic reviews no summary measures for sensitivity and specificity in a meta-analytic sense are given but results were reported only narratively. To compensate for this disadvantage, we use this example to show how our models work in practice. The used data are given in Table 5.4. We included in total 38 different studies. For the references of the single studies, see Kodama et al. (2013) and Bennett et al. (2007) and the citations therein.

The estimated sensitivities, specificities and their corresponding differences are shown in Tables 5.5 and 5.6 and in Figures 5.3 and 5.4. Our estimated differences of sensitivities and specificities show that there is a difference between HbA_{1c} and FPG but unfortunately two models do not converge. These are the models using PQL with the identity link and the discrete copula model. This example shows that both mentioned approaches can be very unstable in practice where the number and size of studies could be small. Therefore the non-canonical link function and the very complex discrete copula model should be used carefully. The other approaches lead

to a sensitivity between 71% and 74% for HbA_{1c} and between 70% and 73% for FPG, respectively. Based on this, all models except of the GLMM using PQL and the logit link, estimate a difference ranging from 1 pp (GLMM with Gaussian quadrature and Gaussian copula) to 1.7 pp (vine copulas) in terms of sensitivity favoring HbA_{1c}. All models show that FPG seems to have a higher specificity than HbA_{1c} with a range from 2 pp to 4.4 pp. These effects should be interpreted carefully because of wider confidence intervals. Based on our conducted simulation study we know that the differences are mostly a bit overestimated. That means the difference in the performance of the two tests could be a bit smaller. Taking into account that the copula models and the GLMM using the logit link perform best, even in situation as in the example, and that Gaussian quadrature leads to reasonable results when the model converges, FPG seems to have a small benefit compared to HbA_{1c}.

Table 5.4: Type 2 diabetes dataset from Kodama et al. (Kodama et al. 2013), first test: HbA_{1c}, second test: fasting plasma glucose

Study	TP1	FN1	FP1	TN1	TP2	FN2	FP2	TN2
Badings et al.	574	262	682	1389	633	203	465	1606
Choi et al.	489	146	1774	6966	445	190	524	8216
Li et al.	36	13	95	998	33	16	120	973
Schöttker et al.	338	29	2376	4060	266	101	1389	5047
Tahrani et al.	16	25	10	147	21	20	25	132
Wang et al.	424	192	121	2112	612	4	1281	952
Hu et al.	644	151	286	1217	648	147	293	1210
Zhang et al.	50	14	4	40	57	7	6	38
Zhou et al.	176	102	768	1286	206	72	823	1231
Kim et al.	72	16	46	258	75	13	35	269
Nakagami et al.	89	26	302	1382	74	41	79	1605
Salmasi et al.	23	7	5	109	16	14	21	93
Glümer et al.	181	71	1988	3877	198	54	721	5144
Anand et al., South Asia	25	2	45	243	24	3	60	228
Anand et al., China	12	2	25	268	12	2	59	234
Anand et al., Europe	13	6	35	260	9	10	40	255
Jesudason et al.	43	11	62	389	40	14	24	427
Tavintharan et al.	17	4	11	79	10	11	2	88
Ko et al.	575	52	1270	980	554	73	469	1781
Papoz et al.	100	12	108	381	77	35	103	386
Choi et al.	610	285	1692	3358	555	340	1667	3383
Heianza et al.	184	154	638	5265	262	76	1418	4485
Law et al.	58	23	129	204	22	59	25	308
Mukai et al.	195	100	718	969	199	96	580	1107
Soulimane et al., Denmark	74	40	1156	3660	80	34	771	4045
Soulimane et al., Australia	145	41	1107	4719	121	65	641	5185
Soulimane et al., France	61	31	742	2950	69	23	876	2816
Cederberg et al.	21	43	36	284	14	50	24	296
Nakagami et al.	42	15	318	814	35	22	198	934
Sato et al.	392	267	1130	5015	541	118	2116	4029
Inoue et al.	187	181	1112	8562	328	40	2411	7263
Inoue et al.	9	8	37	395	15	2	71	361
Norberg et al.	88	76	39	265	82	82	33	271
Takahashi et al.	52	13	37	79	39	26	29	87
Ko et al.	22	22	35	129	19	25	20	144
Mannucci et al.	79	1	689	223	75	5	686	226
Wiener et al.	114	64	20	203	139	39	27	196
Tanaka et al.	135	43	96	592	93	85	0	688

Table 5.5: Estimated sensitivities and specificities (in %) for the type 2 diabetes data set using the different models

Model	Sensitivity HbA_{1c} [95% CI]	Specificity HbA_{1c} [95% CI]	Sensitivity FPG [95% CI]	Specificity FPG [95% CI]
GLMM GQ (logit link)	74.1 [72.9; 75.3]	81.4 [80.8; 81.9]	73.0 [71.8; 74.2]	85.8 [85.2; 86.3]
GLMM PQL (identity link)	- [-; -]	- [-; -]	- [-; -]	- [-; -]
GLMM PQL (logit link)	72.1 [66.7; 76.9]	80.8 [76.3; 84.7]	73.1 [66.0; 79.1]	84.0 [79.0; 88.0]
Gaussian Copula	71.0 [66.3; 75.6]	78.0 [73.9; 82.2]	70.1 [64.7; 75.4]	80.1 [75.9; 84.3]
Plackett C-Vine Copula	71.6 [66.8; 76.3]	77.5 [73.1; 82.0]	69.9 [64.1; 75.8]	79.6 [75.0; 84.2]
Plackett D-Vine Copula	71.6 [66.8; 76.3]	77.5 [73.1; 82.0]	69.9 [64.0; 75.7]	79.6 75.0; 84.2]
Plackett discrete D-Vine Copula	- [-; -]	- [-; -]	- [-; -]	- [-; -]

Table 5.6: Estimated differences of sensitivities and specificities (in percentage points) for the type 2 diabetes data set using the different models

Model	Difference of sensitivities [95% CI]	Difference of specificities [95% CI]
GLMM GQ (logit link)	1.1 [-0.6; 2.8]	-4.4 [-5.1; -3.6]
GLMM PQL (identity link)	- [-; -]	- [-; -]
GLMM PQL (logit link)	-1.0 [-7.8; 5.8]	-3.1 [-8.2; 2.0]
Gaussian Copula	0.9 [-4.9; 6.6]	-2.1 [-7.1; 2.9]
Plackett C-Vine Copula	1.7 [-5.9; 9.2]	-2.1 [-8.4; 4.3]
Plackett D-Vine Copula	1.7 [-5.9; 9.2]	-2.0 [-8.4; 4.3]
Plackett discrete D-Vine Copula	- [-; -]	- [-; -]

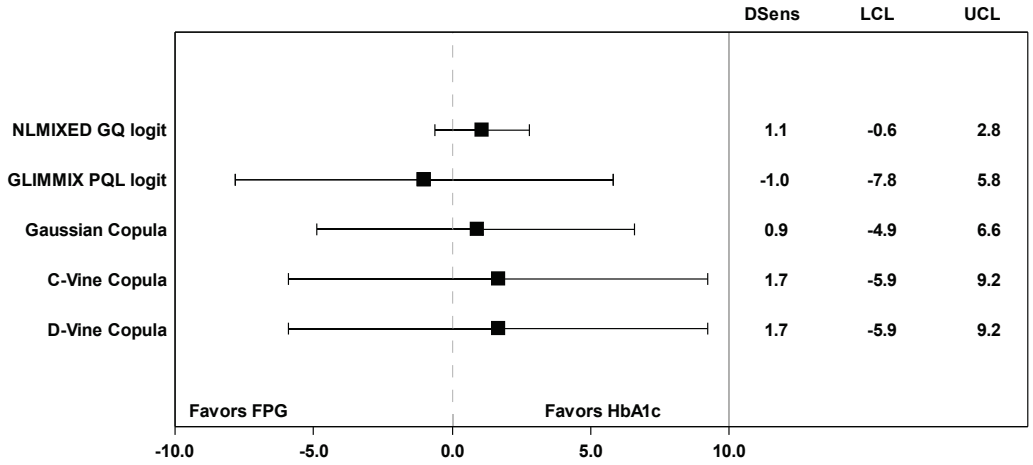

Figure 5.3: Estimates for the Kodama data set: Difference of sensitivities, Abbreviations: DSens - difference of sensitivities, LCL - lower confidence limit, UCL - upper confidence limit

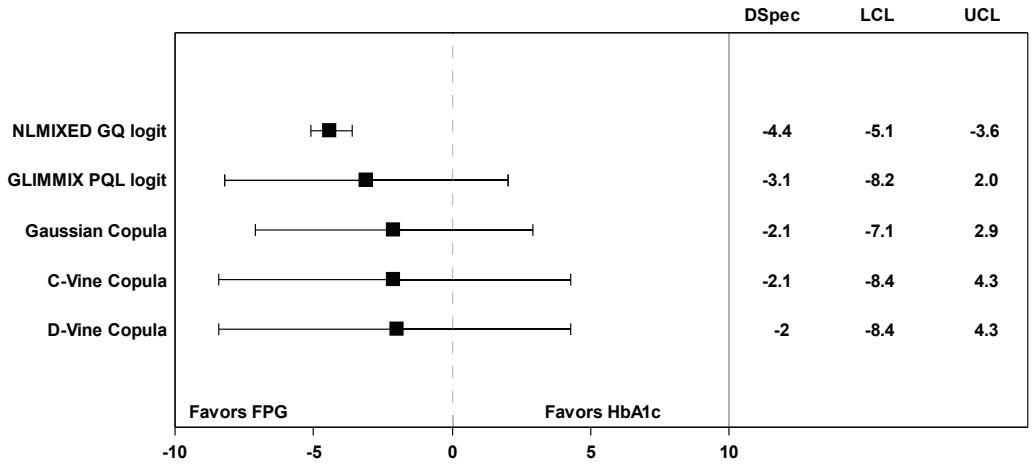

Figure 5.4: Estimates for the Kodama data set: Difference of specificities, Abbreviations: DSpec - Difference of specificities, LCL - lower confidence limit, UCL - upper confidence limit

6 Discussion

In this thesis, new models that address the topic meta-analysis to compare two diagnostic tests to a common gold standard are developed. This issue is of special concern in the research field of meta-analysis of diagnostic accuracy studies. While methods for meta-analysis of clinical trials are well-established, this is a still growing area of biostatistical research with recent developments. Generally, in meta-analyses the results of different single studies dealing with the same underlying medical problem are summarized. As a basis, a systematic review is used. This yields results which cannot be obtained from the single studies alone as it is pointed out by Boissel et al. (1988). Actually, there is still an increasing need and interest in systematic reviews and meta-analysis of diagnostic studies (Harbord et al. 2008). However, this type of studies bring along difficulties that justify the current research efforts. In case of diagnostic studies, we have at least a bivariate outcome, because each single study reports two measures, the sensitivity (conditional probability of which a diseased person is tested as diseased) and the specificity (conditional probability of which a non-diseased proband is classified as non-diseased). In a meta-analysis, we are interested in a weighted estimator for sensitivity and specificity that accounts for different study sizes and potential heterogeneity, i.e. different accuracies of the single studies. As an additional challenge, sensitivity and specificity are generally negatively correlated across studies (Harbord et al. 2008). Different approaches to this task are published (for example Reitsma et al. 2005, Chu and Cole 2006 or Kuss et al. 2014), but the most well-known and frequently used approach is the generalized linear mixed model from Chu and Cole (2006). These models are also extended to the trivariate case (Chu et al. 2009, Hoyer and Kuss 2015) including the prevalence as a third parameter of interest.

But there is still a need for approaches that allow a meta-analytic comparison of diagnostic tests as it is recommended by Leeflang et al. (2008) and Tatsioni et al. (2005). Such approaches should model a quadrivariate outcome including the sensitivities and the specificities of both tests. This is quite challenging, because of many correlations that should be accounted for. There is a correlation within

the probands and within studies, because every participant underwent both tests, and a correlation between studies. Different authors have addressed this topic but their approaches involve some disadvantages. Siadaty et al. (2004) summarized the information in only one measure, the diagnostic odds ratio, which is more difficult to interpret. Trikalinos et al. (2014) assume independent tests and need a mixture of aggregated and individual data. The Cochrane Collaboration recommend a bivariate model as in Chu and Cole (2006) including a binary covariate for the test type to identify which fourfold tables corresponds to each test (Macaskill et al. 2010). Using that approach it is impossible to account for all potential correlations.

To compensate these disadvantages we propose new models with a four-dimensional outcome and the difference of sensitivities and specificities as measures of interest. As a natural extension of the bivariate approach from Chu and Cole (2006), a quadrivariate generalized linear model with random effects is recommended where two different link functions, the identity and the logit link, are used. Another possible model is given by the usage of four-dimensional copulas, which is an extension of the bivariate copula model proposed by Kuss et al. (2014). As a basis, marginal beta-binomial distributions are used that are linked by a copula. This has the advantage that no random effects are necessary.

To compare our models, an extensive simulation study was conducted. Thereby, we use three different ways to implement the GLMM. The logit link and Gaussian quadrature, the logit link and the PQL method and the identity link with PQL estimation are used. On the other hand, four copula models were implemented where the Gaussian copula, a C- and D-vine copula based on bivariate Plackett copulas and a discrete D-vine copula are used. We have shown that in most practically relevant cases the models perform well in terms of bias and coverage. Especially the copula models work at least as good as the GLMM and in some situations frequently better. With a view to the convergence of the models, it is obvious that there is a difference. Taking the numerical robustness and the other performance measures into account, the GLMM using Gaussian quadrature works slightly worse and should be carefully used in practice. The same can be said about the GLMM using the identity link. Summing up, the copula models and the GLMM with logit link and PQL estimation seem to be valid models to compare diagnostic tests. Especially the copula models are numerically robust because no random effects are needed. The vine copulas have to be implemented in a comprehensive way. This leads to some convergence problems especially for the discrete version. This problem occurs also in practice, shown by the diabetes example. All in all, the GLMM with logit link

and the copula models lead to valid estimates and could be applied in practice.

Of course, there are some disadvantages of our models. First, in case of the mixed models four random effects are used which could cause numerical problems in meta-analyses where only small data sets are available. The same problems are observed using vine copulas with complex likelihood functions that have to be optimized. We also do not have model selection criteria to compare copula models to mixed models. A possible alternative could be cross validation. Only in case of the GLMMs it is possible to select an optimal model based on well-known criteria. This is done while performing tests on the random effects covariance matrix. By this it is tested how many parameters should be included in the matrix. Based on the BIC (Bayesian Information Criterion) and -2 loglikelihood a model selection can be made. This option is available using PROC NLMIXED. We propose this idea in our submitted paper (Hoyer and Kuss 2016). In previous publications (Kuss et al. 2014, Hoyer and Kuss 2015), our copula models were criticized, because we consider our discrete marginal distributions as continuous which allows us to determine the derivative. Our simulations in the bivariate case have shown that the there is no huge bias treating the distributions as continuous compared to discrete implementations. In the quadrivariate case, we simulated a discrete version of a vine copula and prove that there is no benefit. We conclude the same as before, that using the 'pseudo-continuous' approximation does not lead to a loss of information in the complex case of four-dimensional responses.

To model the described within-patient correlation, individual data are necessary. In practice, these are rarely reported. It was also shown by different authors that there is not a huge loss of efficiency including only summary measures in meta-analysis (Liu et al. 2015, Lin and Zeng 2010). For the rare cases where we have these data, we implemented a GLMM with an additional random effect. Such models are numerically very instable and need an extensive computation time.

Many expansions of our proposed models are possible. In practice, there are meta-analyses where more than two diagnostic tests are compared (Siadaty et al. 2004). On this basis we want to extent our approaches to more than four dimensions to cope with such complex situations. The development of model selection criteria to compare the copula models with the GLMMs is a potential future issue. Similarly, meta-analyses of ROC curves are actually requested, especially in the field of diabetes. We observed this during screening the studies where mostly more than one threshold is reported. While methods for estimating a summary ROC curve for only one test are actually in scope of interest and development (e.g. Littenberg

and Moses 1993, Dukic and Gatsonis 2003), such approaches are missing for meta-analytic comparisons of diagnostic tests. Therefore we would like to improve our model to estimate full summary ROC curves for the comparison of two tests. In a first approach we included the threshold as a covariate.

To sum up, the proposed models offer different possibilities for future work and could be used by medical researchers as well to make conclusions about the effectiveness of two diagnostic tests which are directly compared.

References

Aas, K., C. Czado, A. Frigessi, and H. Bakken (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics 44*(2), 182–198.

Ades, A. E. and A. J. Sutton (2006). Multiparameter evidence synthesis in epidemiology and medical decision-making: current approaches. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 169*(1), 5–35.

American Diabetes Association (2015). Classification and Diagnosis of Diabetes. *Diabetes Care 38*(Suppl. 1), S8–S16.

Bedford, T. and R. M. Cooke (2001). Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence 32*, 245–268.

Bedford, T. and R. M. Cooke (2002). Vines - a new graphical model for dependent random variables. *Annals of Statistics 30*(4), 1031–1068.

Bennett, C. M., M. Guo, and S. C. Dharmage (2007). HbA(1c) as a screening tool for detection of Type 2 diabetes: a systematic review. *Diabetic Medicine 24*(4), 333–343.

Boissel, J. P., H. S. Sacks, A. Leizorovicz, J. Blanchard, E. Panak, and J. C. Peyrieux (1988). Meta analysis of clinical trials: summary of an international conference. *European Journal of Clinical Pharmacology 34*(6), 535–538.

Bossuyt, P. M., J. B. Reitsma, D. E. Bruns, C. A. Gatsonis, P. P. Glasziou, L. M. Irwig, J. G. Lijmer, D. Moher, D. Rennie, and H. C. de Vet (2003). Standards for Reporting of Diagnostic Accuracy. Towards complete and accurate reporting of studies of diagnostic accuray: The STARD Initiative. *Annals of Internal Medicine 138*(1), 40–44.

Chen, Y., Y. Liu, J. Ning, L. Nie, H. Zhu, and H. Chu (2014). A composite likelihood method for bivariate meta-analysis in diagnostic systematic reviews. *Statistical Methods in Medical Research*. Epub ahead of print.

Cherubini, U., F. Gobbi, S. Mulinacci, and S. Romagnoli (2012). *Dynamic copula methods in finance*. 1. publ. The Wiley finance series. Chichester: Wiley.

Choi, S. H., T. H. Kim, S. Lim, K. S. Park, H. C. Jang, and N. H. Cho (2011). Hemoglobin A1c as a diagnostic tool for diabetes screening and new-onset diabetes prediction: a 6-year community-based prospective study. *Diabetes Care 32*(4), 944–949.

Chu, H. and S. R. Cole (2006). Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *Journal of Clinical Epidemiology 59*(12), 1331–1332.

Chu, H., H. Guo, and Y. Zhou (2010). Bivariate Random Effects Meta-analysis of Diagnostic Studies Using Generalized Linear Mixed Models. *Medical Decision Making 30*, 499–508.

Chu, H., L. Nie, S. R. Cole, and C. Poole (2009). Meta-analysis of diagnostic accuracy studies accounting for disease prevalence: Alternative parameterizations and model selection. *Statistics in Medicine 28*(18), 2384–2399.

Deeks, J. J. (2001). Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. *British Medical Journal 323*(7305), 157–162.

DerSimonian, R. and N. Laird (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials 7*(3), 177–188.

Deuflhard, P. and A. Hohmann (2008). *Numerische Mathematik 1: Eine algorithmisch orientierte Einführung*. Walter de Gruyter & Co.

Dukic, V. and C. Gatsonis (2003). Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds. *Biometrics 59*, 936–946.

Fahrmeir, L., T. Kneib, and S. Lang (2009). *Regression, Modelle, Methoden und Anwendungen*. Berlin Heidelberg: Springer.

Gatsonis, C. and P. Paliwal (2006). Meta-Analysis of Diagnostic and Screening Test Accuracy Evaluations: Methodologic Primer. *American Journal of Roentgenology 187*(2), 271–281.

Genest, C. and J. Nešlehová (2007). A Primer on Copulas for Count Data. *Astin Bulletin 37*(2), 475–515.

Harbord, R. M., J. J. Deeks, M. Egger, P. Whiting, and J. A. C. Sterne (2007).

A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics 8*(2), 239–251.

Harbord, R. M., P. Whiting, J. A. C. Sterne, M. Egger, J. J. Deeks, A. Shang, and L. M. Bachmann (2008). An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. *Journal of Clinical Epidemiology 61*(11), 1095–1103.

Hartung, J., G. Knapp, and B. K. Sinha (2011). *Statistical Meta-Analysis with Applications*, Volume 738. John Wiley & Sons.

He, J., H. Li, A. C. Edmondson, D. J. Rader, and M. Li (2012). A Gaussian copula approach for the analysis of secondary phenotypes in case-control genetic association studies. *Biostatistics 13*(3), 497–508.

Held, L. (2008). *Methoden der statistischen Inferenz. Likelihood und Bayes*. Heidelberg: Spektrum Akad. Verl.

Hoyer, A. and O. Kuss (2015). Meta-analysis of diagnostic tests accounting for disease prevalence: A new model using trivariate copulas. *Statistics in Medicine 34*(11), 1912–1924.

Hoyer, A. and O. Kuss (2016). Meta-analysis for the comparison of two diagnostic tests to common gold standard: A generalized linear mixed model approach. *Statistical Methods in Medical Research*. submitted.

Joe, H. (1996). Families of m-variate distributions with given margins and m(m-1)/2 bivariate dependence parameters. *Distributions with Fixed Marginals and Related Topics*. In L. Rüschendorf and B. Schweizer and M.D. Taylor (Ed.).

Kodama, S., C. Horikawa, K. Fujihara, R. Hirasawa, Y. Yachi, S. Yoshizawa, S. Tanaka, Y. Sone, H. Shimano, K. T. Iida, K. Saito, and H. Sone (2013). Use of high-normal levels of haemoglobin A1C and fasting plasma glucose for diabetes screening and for prediction: a meta-analysis. *Diabetes/ Metabolism Research and Reviews 29*(8), 680–692.

Kurowicka, D. and R. M. Cooke (2004). Distribution - free continuous bayesian belief nets. *Fourth International Conference on Mathematical Methods in Reliability Methodology and Practice, Santa Fe, New Mexico*.

Kuss, O., A. Hoyer, and A. Solms (2014). Meta-analysis for diagnostic accuracy studies: A new statistical model using beta-binomial distributions and bivariate copulas. *Statistics in Medicine 33*(1), 17–30.

Leeflang, M. M. G., J. J. Deeks, C. Gatsonis, and P. M. M. Bossuyt (2008). Systematic Reviews of Diagnostic Test Accuracy. *Annals of Internal Medicine 149*(12), 889–897.

Lin, D. Y. and D. Zeng (2010). On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika 97*(2), 321?332.

Littenberg, B. and L. E. Moses (1993). Estimating Diagnostic Accuracy from Multiple Conflicting Reports: A New Meta-analytic Method. *Medical Decision Making 13*(4), 313–321.

Liu, D., R. Y. Liu, and M. Xie (2015). Multivariate meta-analysis of heterogeneous studies using only summary statistics: efficiency and robustness. *Journal of the American Statistical Association 110:509*, 326–340.

Macaskill, P., C. Gatsonis, J. J. Deeks, R. M. Harbord, and Y. Takwoingi (2010). *Chapter 10: Analysing and Presenting Results. In: Deeks JJ, Bossuyt PM, Gatsonis C (editors) Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0*. The Cochrane Collaboration. Available from: http://srdta.cochrane.org/.

Menke, J. (2010). Bivariate random-effects meta-analysis of sensitivity and specificity with SAS PROC GLIMMIX. *Methods of Information in Medicine 49*(1), 54–62, 62–64.

Molenberghs, G. and G. Verbeke (2006). *Models for Discrete Longitudinal Data*. Springer: New York.

Moses, L. E., D. Shapiro, and B. Littenberg (1993). 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.

Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer: New York.

Panagiotelis, A., C. Czado, and H. Joe (2012). Pair Copula Constructions for Multivariate Discrete Data. *Journal of the American Statistical Association 107*(499), 1063–1072.

Patton, A. J. (2002). Applications of copula theory in financial econometrics. *Unpublished Ph.D. dissertation. University of California, San Diego*. URL: http://www.amstat.org/sections/bus_econ/papers/patton_dissertation.pdf.

Patton, A. J. (2006). Modelling asymmetric exchange rate dependence. *International Economic Review 47*(2), 527–556.

Paul, M., A. Riebler, L. M. Bachmann, H. Rue, and L. Held (2010). Bayesian bivariate meta-analysis of diagnostic test studies using integrated nested Laplace approximations. *Statistics in Medicine 29*(12), 1325–1339.

Picano, E., G. Bedetti, A. Varga, and E. Cseh (2000). The comparable diagnostic accuracies of dobutamine-stress and dipyridamole-stress echocardiographies: a meta-analysis. *Coronary Artery Disease 11*(2), 151–159.

Reitsma, J. B., A. S. Glas, A. W. Rutjes, R. J. Scholten, P. M. Bossuyt, and A. H. Zwinderman (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology 58*(19), 982–990.

Rüschendorf, L. (2009). On the distributional transform, Sklar's theorem, and the empirical copula process. *Journal of Statistical Planning and Inference 139*(11), 3921–3927.

Rutter, C. M. and C. A. Gatsonis (2001). A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine 20*(19), 2865–2884.

Sackett, D. L., W. M. Rosenberg, J. A. Gray, R. B. Haynes, and W. S. Richardson (1996). Evidence based medicine: what it is and what it isn't. *British Medical Journal 312*(7023), 71.

Schumacher, M. and G. Schulgen (2008). *Methodik klinischer Studien: Methodische Grundlagen der Planung, Durchführung und Auswertung*. Springer-Verlag.

Schweizer, B. and A. Sklar (1983). *Probabilistic metric spaces*. North-Holland Series in Probabilistic and Applied Mathematics, North-Holland Publishing Co., New York.

Schweizer, B. and E. F. Wolff (1981). On nonparametric measures of dependence for random variables. *Annals of Statistics 9*, 879–885.

Siadaty, M. S., J. T. Philbrick, S. W. Heim, and J. M. Schectman (2004). Repeated-measures modeling improved comparison of diagnostic tests in meta-analysis of dependent studies. *Journal of Clinical Epidemiology 57*(7), 698–711.

Siadaty, M. S. and J. Shu (2004). Proportional odds ratio model for comparison of diagnostic tests in meta-analysis. *BMC Medical Research Methodology 4:27*.

Sklar, A. (1959). Fonctions de répartition á n dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Université de Paris 8*, 229–231.

Smith, M. S. and M. A. Khaled (2012). Estimation of Copula Models With Discrete Margins via Bayesian Data Augmentation. *Journal of the American Statistical Association 107*(497), 290–303.

Song, P. X. K. (2002). Multivariate Dispersion Models Generated From Gaussian Copula. *Scandinavian Journal of Statistics 27*(2), 305–320.

Song, P. X. K., M. Li, and Y. Yuan (2009). Joint Regression Analysis of Correlated Data Using Gaussian Copulas. *Biometrics 65*(1), 60–68.

Sutton, A. J. and J. P. T. Higgins (2008). Recent developments in meta-analysis. *Statistics in Medicine 27*(5), 625–650.

Takwoingi, Y., M. M. G. Leeflang, and J. J. Deeks (2013). Empirical Evidence of the Importance of Comparative Studies of Diagnostic Test Accuracy. *Annals of internal medicine 158*(7), 544–554.

Tatsioni, A., D. A. Zarin, N. Aronson, D. J. Samson, C. R. Flamm, C. Schmid, and J. Lau (2005). Challenges in Systematic Reviews of Diagnostic Technologies. *Annals of Internal Medicine 142*(12), 1048–1055.

Trikalinos, T. A., D. C. Hoaglin, K. M. Small, N. Terrin, and C. H. Schmid (2014). Methods for the joint meta-analysis of multiple tests. *Research Synthesis Methods 5*(4), 294–312.

Van Houwelingen, H. C. (1997). The future of biostatistics: expecting the unexpected. *Statistics in Medicine 16*(24), 2773–2784.

Whiting, P. F., A. W. Rutjes, J. B. Reitsma, P. M. Bossuyt, and J. Kleijnen (2003). The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in sytematic reviews. *BMC Medical Research Methodology 3:25*.

Whiting, P. F., A. W. Rutjes, M. E. Westwood, S. Mallett, J. J. Deeks, J. B. Reitsma, M. M. Leeflang, J. A. Sterne, and P. M. Bossuyt (2011). QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine 155*(8), 529–536.

Youden, W. (1950). Index for rating diagnostic tests. *Cancer 3*(1), 32–35.

Zapf, A., A. Hoyer, K. Kramer, and O. Kuss (2015). Nonparametric meta-analysis for diagnostic accuracy studies. *Statistics in Medicine 34*(29), 3831–3841.

List of Figures

List of Tables

Appendix

Simulation Results

Table 6.1: Bias (multiplied by 100) for the differences of sensitivity and specificity on the [0, 1]- scale. Abbreviations: ΔSe=Difference of sensitivities, ΔSp=Difference of specificities, corr=correlation between Se_1, Sp_1, Se_2 and Sp_2, SN=GLMM using GQ, SI=GLMM using PQL and the identity link, SL=GLMM using PQL and the logit link, CM=Cochrane model using GQ and the logit link, CA=Cochrane model using the PQL and the logit link, True model=GLMM

True ΔSe and ΔSp	True corr	Estimated model									
		SN		SI		SL		CM		CA	
		ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp
0%/0%	non	0.0	0.2	0.2	0.0	0.1	0.1	0.2	0.1	0.1	0.1
	negative	-0.8	0.1	-0.2	-0.1	-0.1	-0.1	-0.5	-0.3	-0.4	-0.1
	mixed	0.1	0.3	0.1	-0.0	0.1	-0.1	-0.0	-0.1	0.0	-0.1
0%/-10%	non	-0.2	-0.3	0.3	-0.0	0.1	-0.0	-0.1	-0.2	-0.0	-0.2
	negative	0.3	-0.8	0.4	-0.3	0.5	0.1	-0.2	0.3	0.2	-0.4
	mixed	0.2	0.0	0.1	0.1	-0.1	-0.1	-0.0	-0.1	-0.1	-0.1
0%/5%	non	-0.3	-0.1	-0.0	-0.1	-0.1	-0.2	-0.2	-0.2	-0.1	-0.1
	negative	-0.3	-0.1	-0.2	-0.3	-0.1	-0.2	0.5	0.2	-0.0	-0.1
	mixed	0.1	0.3	0.2	-0.0	0.3	-0.1	0.2	-0.1	0.2	-0.1
-10%/0%	non	0.2	-0.3	0.3	-0.2	0.4	-0.1	0.3	-0.1	0.3	0.0
	negative	0.0	-0.4	0.2	-0.1	0.0	0.0	0.2	0.1	-0.1	0.2
	mixed	0.4	0.0	0.3	-0.0	0.2	0.0	0.2	-0.1	0.2	0.0
-10%/-10%	non	0.0	-0.5	0.4	-0.1	0.3	-0.1	0.3	-0.2	0.2	-0.3
	negative	-0.6	-0.9	0.4	-0.4	0.5	-0.3	0.7	0.1	0.4	-0.5
	mixed	-0.3	-0.3	0.2	-0.3	0.1	-0.2	0.1	-0.2	-0.0	-0.2
-10%/5%	non	-0.2	-0.1	0.0	-0.3	0.2	-0.2	0.2	-0.3	0.2	-0.3
	negative	-0.4	0.9	0.5	0.2	0.3	0.1	0.2	0.3	0.3	0.2
	mixed	-0.0	-0.1	0.3	-0.1	0.3	-0.1	0.2	-0.0	0.2	-0.1
5%/0%	non	-0.4	0.1	-0.3	0.0	-0.3	0.0	-0.2	-0.0	-0.2	-0.0
	negative	0.1	-0.2	-0.4	-0.3	-0.5	-0.1	-0.3	0.3	-0.1	0.2
	mixed	0.2	0.2	-0.2	0.0	-0.2	0.0	-0.3	0.2	-0.2	0.2
5%/-10%	non	0.3	-0.5	0.0	-0.2	-0.0	-0.2	-0.2	-0.3	-0.0	-0.3
	negative	1.1	0.0	-0.1	-0.8	0.0	-0.2	-0.7	0.3	-0.1	-0.6
	mixed	-0.5	-0.5	-0.4	-0.3	-0.3	-0.2	-0.2	-0.2	-0.2	-0.3
5%/5%	non	0.1	0.1	-0.0	-0.1	-0.1	-0.0	0.1	-0.1	0.1	-0.1
	negative	0.7	-0.1	0.2	-0.3	0.0	-0.2	0.4	-0.3	0.1	-0.1
	mixed	0.1	0.3	-0.0	0.0	-0.1	-0.1	-0.2	0.0	-0.2	0.0

Table 6.2: Empirical coverage (in %) for the 95% confidence intervals for the differences of sensitivity and specificity on the [0, 1]- scale. Abbreviations: ΔSe=Difference of sensitivities, ΔSp=Difference of specificities, corr=correlation between Se_1, Sp_1, Se_2 and Sp_2, SN=GLMM using GQ, SI=GLMM using PQL and the identity link, SL=GLMM using PQL and the logit link, CM=Cochrane model using GQ and the logit link, CA=Cochrane model using the PQL and the logit link, True model=GLMM

True ΔSe and ΔSp	True corr	Estimated model									
		SN		SI		SL		CM		CA	
		ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp
0%/0%	non	94.3	93.9	93.3	91.2	93.5	91.7	68.3	71.3	82.0	81.5
	negative	92.4	88.2	93.0	93.8	92.8	96.4	59.0	61.0	74.3	77.9
	mixed	93.7	96.0	91.7	92.0	93.0	92.7	69.0	74.7	84.8	88.4
0%/-10%	non	91.6	87.3	93.4	90.8	91.7	92.7	67.0	77.5	79.0	84.3
	negative	92.4	85.7	94.9	89.8	94.0	93.8	67.6	69.7	73.7	81.4
	mixed	91.0	93.7	91.5	94.1	92.6	95.5	70.0	84.0	87.5	90.3
0%/5%	non	95.2	94.6	94.0	93.0	93.3	92.6	67.2	71.3	81.7	83.8
	negative	95.7	87.3	91.9	90.3	92.1	91.2	61.7	73.1	71.6	74.0
	mixed	95.3	92.8	93.1	93.8	93.5	93.0	70.7	76.4	86.7	89.7
-10%/0%	non	93.1	95.3	94.2	91.9	93.7	92.5	71.6	73.6	81.7	84.3
	negative	84.1	89.3	93.9	93.5	92.5	93.0	66.9	67.1	78.0	77.1
	mixed	91.5	94.4	92.1	91.0	93.5	92.6	73.2	75.9	88.6	88.2
-10%/-10%	non	92.9	95.5	92.3	92.8	92.3	93.7	68.9	79.8	82.1	85.2
	negative	91.8	89.6	89.1	90.8	92.7	95.1	65.5	71.6	76.0	81.9
	mixed	88.5	90.6	93.2	93.9	93.3	94.1	73.4	84.1	87.4	90.3
-10%/5%	non	93.6	93.8	92.3	92.8	92.3	93.0	68.7	71.5	81.7	84.9
	negative	87.3	87.8	93.4	91.5	92.8	91.5	71.1	62.0	77.9	75.8
	mixed	93.5	94.7	90.8	92.3	92.7	93.7	74.4	73.8	88.9	87.3
5%/0%	non	93.2	94.8	92.2	93.1	92.2	92.8	67.6	71.7	82.3	83.2
	negative	90.7	88.7	91.6	91.6	90.4	91.2	65.2	65.6	72.9	77.2
	mixed	94.5	95.5	91.6	92.7	91.9	92.4	68.9	75.3	85.7	88.8
5%/-10%	non	92.4	91.7	93.4	90.9	93.8	92.7	63.3	77.1	79.4	84.3
	negative	82.6	83.1	92.9	90.2	91.2	92.8	54.0	66.4	70.7	78.0
	mixed	92.4	84.1	94.6	92.6	92.6	94.2	70.1	81.4	88.2	90.0
5%/5%	non	93.9	96.2	92.8	93.5	92.3	93.1	68.2	70.6	81.6	82.6
	negative	90.0	94.6	91.8	95.2	92.2	93.9	66.7	67.3	71.9	77.4
	mixed	96.3	93.0	94.0	91.9	94.5	92.5	71.7	75.9	86.6	88.4

Table 6.3: Number of converged runs from 1000 simulation runs. Abbreviations: ΔSe=Difference of sensitivities, ΔSp=Difference of specificities, corr=correlation between Se_1, Sp_1, Se_2 and Sp_2, SN=GLMM using GQ, SI=GLMM using PQL and the identity link, SL=GLMM using PQL and the logit link, CM=Cochrane model using GQ and the logit link, CA=Cochrane model using the PQL and the logit link, True model=GLMM

		Estimated model									
True ΔSe and ΔSp	True corr	SN		SI		SL		CM		CA	
		ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp
0%/0%	non	431	431	659	659	817	817	840	840	926	926
	negative	192	192	256	256	418	418	172	172	421	421
	mixed	292	292	575	575	756	756	687	687	876	876
0%/-10%	non	347	347	455	455	823	823	802	802	919	919
	negative	180	180	157	157	369	369	197	197	457	457
	mixed	224	224	387	387	731	731	640	640	874	874
0%/5%	non	460	460	670	670	820	820	849	849	942	942
	negative	207	207	298	298	441	441	216	216	454	454
	mixed	327	327	641	641	770	770	712	712	901	901
-10%/0%	non	378	378	589	589	839	839	818	818	921	921
	negative	198	198	231	231	386	386	224	224	437	437
	mixed	279	279	545	545	753	753	657	657	875	875
-10%/-10%	non	299	299	431	431	793	793	803	803	893	893
	negative	182	182	174	174	327	327	217	217	463	463
	mixed	203	203	396	396	735	735	606	606	878	878
-10%/5%	non	405	405	599	599	797	797	832	832	923	923
	negative	212	212	259	259	390	390	199	199	443	443
	mixed	283	283	574	574	764	764	675	675	885	885
5%/0%	non	447	447	652	652	812	812	849	849	934	934
	negative	173	173	273	273	408	408	189	189	395	395
	mixed	295	295	572	572	765	765	662	662	876	876
5%/-10%	non	335	335	441	441	785	785	819	819	923	923
	negative	188	188	183	183	363	363	204	204	450	450
	mixed	236	236	392	392	726	726	622	622	866	866
5%/5%	non	456	456	676	676	816	816	881	881	951	951
	negative	196	196	292	292	424	424	182	182	442	442
	mixed	361	361	615	615	760	760	692	692	882	882

Table 6.4: Bias (multiplied by 100) for the differences of sensitivity and specificity on the [0, 1]- scale. Abbreviations: ΔSe=Difference of sensitivities, ΔSp=Difference of specificities, corr=correlation between Se_1, Sp_1, Se_2 and Sp_2, SI=GLMM using PQL and the identity link, SL=GLMM using PQL and the logit link, CG = Gaussian Copula, CP = C-Vine Copula, CD = D-Vine Copula, DC = Discrete D-Vine Copula, True model=Gaussian copula

		Estimated model											
True ΔSe and ΔSp	True corr	SI		SL		CG		CP		CD		DC	
		ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp
0%/0%	non	0.1	0.1	-0.1	0.0	-0.1	0.1	-0.1	0.1	0.4	1.1	-0.1	0.1
	negative	-0.1	0.5	-0.1	-0.2	-0.1	-0.1	-0.2	-0.1	0.8	1.4	-0.1	-0.1
	mixed	-0.8	-0.0	-0.2	-0.0	-4.6	-0.0	-0.1	-0.2	-0.7	0.1	-2.0	0.4
0%/-10%	non	-1.2	-2.5	0.0	-2.2	0.1	-1.4	0.0	0.3	0.7	1.6	-0.0	0.1
	negative	-0.1	-3.7	-0.1	-2.3	-0.0	-0.1	-0.3	1.0	1.5	2.5	0.1	0.4
	mixed	0.0	-1.8	-0.4	-2.4	-3.8	-3.6	0.1	-0.1	-0.2	0.0	-2.3	-0.0
0%/5%	non	-0.2	-0.4	-0.4	0.7	-0.4	0.1	-0.3	0.0	0.1	0.9	-0.3	0.0
	negative	-0.2	-0.7	0.1	0.5	-0.1	-0.1	-0.2	-0.1	0.6	1.4	-0.2	0.0
	mixed	-0.9	-0.0	-0.2	0.7	-4.3	0.9	-0.4	-0.1	-0.7	0.1	-2.6	0.0
-10%/0%	non	0.3	0.2	-1.2	0.3	-0.1	0.3	0.0	0.3	0.4	1.1	-0.1	0.2
	negative	0.7	-0.4	-1.1	0.1	0.3	-0.0	0.3	-0.1	1.2	1.2	0.2	-0.0
	mixed	-2.0	-0.1	-1.6	-0.1	-5.9	0.1	-1.6	-0.0	-2.0	-0.0	-3.5	-0.3
-10%/-10%	non	-5.0	-2.2	-1.2	-2.3	0.3	-1.9	0.1	0.0	0.7	1.4	-0.1	-0.2
	negative	17.4	0.6	-1.1	-2.3	0.7	-0.0	-0.3	0.9	1.6	2.6	0.0	0.1
	mixed	-2.6	-2.0	-1.6	-2.1	-5.8	-3.1	-1.2	-0.1	-1.5	0.5	-3.7	0.6
-10%/5%	non	0.5	-0.8	-1.2	0.4	-0.0	-0.3	0.2	-0.4	0.6	0.5	0.1	-0.4
	negative	-0.5	0.5	-1.5	1.0	0.1	-0.0	0.1	-0.1	0.9	1.7	-0.0	0.1
	mixed	-2.3	0.2	-1.5	0.8	-6.1	0.3	-1.6	0.1	-1.9	-0.1	-3.4	-0.5
5%/0%	non	0.4	-0.5	0.5	-0.2	0.1	-0.2	0.1	-0.2	0.5	0.8	0.1	-0.2
	negative	-0.3	0.1	0.5	0.0	-0.1	-0.2	-0.3	-0.1	0.7	1.3	-0.1	-0.1
	mixed	-0.6	-0.2	-0.1	-0.1	-4.9	-0.5	-0.1	-0.1	-0.6	0.0	-1.8	-0.2
5%/-10%	non	4.7	-6.7	0.5	-2.6	0.2	-2.1	0.2	-0.3	1.0	1.2	0.1	-0.5
	negative	1.8	-7.4	0.5	-2.5	-0.2	-0.5	-0.6	0.7	1.2	2.7	-0.1	-0.0
	mixed	-1.4	-2.0	0.3	-2.3	-4.0	-3.2	0.5	-0.2	0.5	0.4	-1.5	0.2
5%/5%	non	0.1	0.2	0.3	0.7	-0.2	0.1	-0.1	-0.0	0.4	1.1	-0.2	0.0
	negative	0.4	-0.4	0.4	0.6	-0.1	-0.1	-0.1	-0.2	0.5	1.2	-0.2	-0.1
	mixed	-0.7	-0.1	0.1	0.7	-4.0	0.4	-0.1	-0.2	-0.3	-0.1	-2.2	0.1

Table 6.5: Bias (multiplied by 100) for the differences of sensitivity and specificity on the [0, 1]- scale. Abbreviations: ΔSe=Difference of sensitivities, ΔSp=Difference of specificities, corr=correlation between Se_1, Sp_1, Se_2 and Sp_2, SI=GLMM using PQL and the identity link, SL=GLMM using PQL and the logit link, CG = Gaussian Copula, CP = C-Vine Copula, CD = D-Vine Copula, DC = Discrete D-Vine Copula, True model=GLMM

		Estimated model											
True ΔSe and ΔSp	True corr	SI		SL		CG		CP		CD		DC	
		ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp
0%/0%	non	0.2	0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.4	0.7	0.1	0.1
	negative	-0.2	-0.1	-0.1	-0.1	-0.4	-0.0	-0.2	0.0	0.4	1.5	-0.5	0.0
	mixed	0.1	-0.0	0.1	-0.1	-0.0	-0.2	0.2	-0.1	-0.0	0.2	0.0	-0.1
0%/-10%	non	0.3	-0.0	0.1	-0.0	0.0	-0.6	0.1	-0.1	0.5	0.6	0.0	-0.2
	negative	0.4	-0.3	0.5	0.1	0.3	-0.1	0.4	0.2	1.1	1.4	0.2	-0.2
	mixed	0.1	0.1	-0.1	-0.1	-0.1	-0.4	0.2	-0.0	-0.1	0.4	0.1	-0.2
0%/5%	non	-0.0	-0.1	-0.1	-0.2	-0.1	-0.0	0.0	-0.2	0.4	0.7	-0.0	-0.1
	negative	-0.2	-0.3	-0.1	-0.2	0.0	-0.1	-0.0	-0.2	0.5	1.2	-0.1	0.0
	mixed	0.2	-0.0	0.3	-0.1	0.0	-0.1	0.3	-0.1	-0.0	0.3	0.1	-0.0
-10%/0%	non	0.3	-0.2	0.4	-0.1	0.2	0.0	0.5	-0.0	0.8	0.5	0.3	-0.0
	negative	0.2	-0.1	0.0	0.0	0.5	0.1	0.8	0.2	1.3	1.4	0.3	-0.0
	mixed	0.3	-0.0	0.2	0.0	0.4	-0.0	0.7	0.0	0.6	0.4	0.4	-0.1
-10%/-10%	non	0.4	-0.1	0.3	-0.1	0.0	-0.6	0.3	-0.2	0.7	0.4	0.2	-0.4
	negative	0.4	-0.4	0.5	-0.3	0.6	-0.1	0.9	0.2	1.4	1.0	0.2	-0.3
	mixed	0.2	-0.3	0.1	-0.2	0.4	-0.4	0.7	0.0	0.4	0.4	0.2	-0.3
-10%/5%	non	0.0	-0.3	0.2	-0.2	0.1	-0.1	0.3	-0.3	0.7	0.4	0.2	-0.2
	negative	0.5	0.2	0.3	0.1	0.5	-0.1	0.8	-0.3	1.2	1.2	0.1	0.1
	mixed	0.3	-0.1	0.3	-0.1	0.3	-0.1	0.7	-0.2	0.5	0.2	0.3	-0.1
5%/0%	non	-0.3	0.0	-0.3	0.0	-0.3	-0.0	-0.3	-0.0	0.1	0.6	-0.3	-0.1
	negative	-0.4	-0.3	-0.5	-0.1	-0.3	-0.0	-0.3	0.1	0.2	1.3	-0.3	0.1
	mixed	-0.2	0.0	-0.2	0.0	-0.2	-0.0	-0.2	-0.0	-0.4	0.4	-0.3	0.0
5%/-10%	non	0.0	-0.2	-0.0	-0.2	-0.1	-0.7	-0.0	-0.2	0.3	0.5	-0.1	-0.4
	negative	-0.1	-0.8	0.0	-0.2	-0.2	-0.3	-0.1	0.1	0.3	1.3	-0.3	-0.3
	mixed	-0.4	-0.3	-0.3	-0.2	-0.2	-0.5	-0.0	-0.1	-0.4	0.2	-0.2	-0.3
5%/5%	non	-0.0	-0.1	-0.1	-0.0	0.0	-0.0	-0.0	-0.1	0.2	0.5	-0.1	-0.1
	negative	0.2	-0.3	0.0	-0.2	-0.2	-0.2	-0.0	-0.2	0.2	1.2	-0.1	-0.0
	mixed	-0.0	0.0	-0.1	-0.1	-0.4	0.0	-0.2	-0.1	-0.3	0.3	-0.4	-0.1

Table 6.6: Empirical coverage (in %) for the 95% confidence intervals for the differences of sensitivity and specificity on the [0, 1]- scale. Abbreviations: ΔSe=Difference of sensitivities, ΔSp=Difference of specificities, corr=correlation between Se_1, Sp_1, Se_2 and Sp_2, SI=GLMM using PQL and the identity link, SL=GLMM using PQL and the logit link, CG = Gaussian Copula, CP = C-Vine Copula, CD = D-Vine Copula, DC = Discrete D-Vine Copula, True model=Gaussian copula

		Estimated model											
True ΔSe and ΔSp	True corr	SI		SL		CG		CP		CD		DC	
		ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp
0%/0%	non	93.2	94.6	94.1	93.6	91.9	93.3	95.4	95.3	94.6	93.8	84.0	87.6
	negative	94.8	91.1	94.4	92.9	92.6	91.0	92.5	91.5	91.1	87.5	90.9	90.6
	mixed	95.3	94.1	96.3	92.6	71.5	83.0	85.6	89.7	87.8	91.7	83.9	82.0
0%/-10%	non	100	60.0	92.6	93.9	89.9	89.5	93.5	94.1	92.5	94.4	80.6	84.3
	negative	100	40.0	92.5	93.7	91.2	91.3	91.1	92.8	87.6	87.5	89.0	90.9
	mixed	100	78.6	94.3	94.2	73.3	78.0	82.4	88.9	84.4	89.9	78.4	81.1
0%/5%	non	92.7	95.5	91.6	94.4	92.2	93.2	94.8	96.0	94.1	94.1	83.3	86.2
	negative	91.3	92.1	92.4	93.9	92.0	92.8	91.8	91.8	91.2	88.9	88.1	90.5
	mixed	94.2	92.5	95.0	94.3	73.7	83.7	85.3	90.1	83.2	89.6	79.9	78.8
-10%/0%	non	93.3	94.6	94.1	92.2	92.5	93.0	94.4	94.8	94.2	93.8	84.7	83.6
	negative	88.7	91.5	93.1	92.0	91.1	91.0	90.9	91.5	90.6	87.7	88.9	89.4
	mixed	90.4	95.0	90.1	94.5	69.4	85.9	81.5	89.9	81.0	88.3	74.4	79.9
-10%/-10%	non	100	100	94.0	93.8	93.1	90.4	94.3	94.8	94.2	95.1	84.0	85.8
	negative	0	100	95.0	93.5	92.0	90.8	92.3	91.6	89.7	88.4	89.6	89.9
	mixed	94.4	83.3	89.9	93.6	64.1	77.0	82.7	89.3	80.6	87.9	69.7	74.7
-10%/5%	non	92.6	94.9	94.1	93.4	91.9	90.8	95.3	94.0	94.3	93.1	82.7	82.6
	negative	91.2	93.7	93.6	94.9	92.3	93.0	91.9	94.1	91.7	89.5	90.5	91.7
	mixed	89.5	93.0	91.6	94.0	66.7	85.2	83.7	89.2	84.0	90.7	71.6	75.6
5%/0%	non	93.3	93.7	92.9	92.9	91.3	92.8	95.0	95.1	94.1	93.3	84.7	84.1
	negative	93.0	91.9	93.7	93.1	91.0	90.9	92.0	91.3	92.3	87.5	88.6	89.6
	mixed	92.3	93.9	96.0	95.6	74.5	87.1	84.8	90.7	86.7	91.6	83.0	79.8
5%/-10%	non	87.5	50.0	93.1	93.5	90.3	89.5	93.3	94.7	93.1	94.7	83.9	86.4
	negative	100	40.0	93.6	93.3	92.9	91.0	92.1	91.8	89.0	87.9	89.9	91.3
	mixed	100	86.7	95.8	94.2	74.0	80.0	83.7	90.1	84.9	90.1	85.0	80.8
5%/5%	non	91.0	91.2	93.0	92.7	92.1	91.7	94.9	95.0	94.7	93.0	84.6	86.2
	negative	91.8	92.5	93.5	92.9	90.9	91.6	91.7	91.6	90.7	88.7	90.0	90.3
	mixed	95.7	92.5	96.1	92.4	73.5	83.4	85.9	89.2	87.7	90.9	81.4	83.5

Table 6.7: Empirical coverage (in %) for the 95% confidence intervals for the differences of sensitivity and specificity on the [0, 1]- scale. Abbreviations: ΔSe=Difference of sensitivities, ΔSp=Difference of specificities, corr=correlation between Se_1, Sp_1, Se_2 and Sp_2, SI=GLMM using PQL and the identity link, SL=GLMM using PQL and the logit link, CG = Gaussian Copula, CP = C-Vine Copula, CD = D-Vine Copula, DC = Discrete D-Vine Copula, True model=GLMM

		Estimated model											
True ΔSe and ΔSp	True corr	SI		SL		CG		CP		CD		DC	
		ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp
0%/0%	non	93.3	91.2	93.5	91.7	91.5	91.0	93.4	94.0	92.4	91.7	88.1	87.2
	negative	93.0	93.8	92.8	96.4	91.2	92.7	92.0	92.1	90.4	86.3	90.2	92.1
	mixed	91.7	92.0	93.0	92.7	91.6	90.7	95.7	95.1	95.1	93.8	90.4	88.1
0%/-10%	non	93.4	90.8	91.7	92.7	90.5	87.3	93.3	91.4	92.6	90.5	87.6	85.3
	negative	94.9	89.8	94.0	93.8	92.3	89.2	92.5	91.8	91.5	84.7	89.7	88.4
	mixed	91.5	94.1	92.6	95.5	91.3	90.1	95.7	93.7	95.4	92.0	88.2	89.7
0%/5%	non	94.0	93.0	93.3	92.6	92.9	90.1	94.2	93.7	93.1	92.3	87.9	88.1
	negative	91.9	90.3	92.1	91.2	89.8	90.7	92.3	92.1	91.0	86.7	90.0	90.4
	mixed	93.1	93.8	93.5	93.0	93.0	93.6	96.1	96.3	94.9	95.6	85.6	89.1
-10%/0%	non	94.2	91.9	93.7	92.5	90.9	91.7	93.9	94.2	93.0	92.9	87.1	89.4
	negative	93.9	93.5	92.5	93.0	90.6	92.1	90.6	92.5	90.2	88.4	87.1	89.4
	mixed	92.1	91.0	93.5	92.6	91.2	91.1	94.9	95.7	94.9	95.0	88.3	89.4
-10%/-10%	non	92.3	92.8	92.3	93.7	91.3	89.3	92.2	92.6	90.7	91.2	87.1	86.1
	negative	89.1	90.8	92.7	95.1	89.5	89.2	91.4	92.4	88.6	87.4	91.0	91.6
	mixed	93.2	93.9	93.3	94.1	89.8	89.9	92.6	93.9	93.2	92.6	87.0	91.6
-10%/5%	non	92.3	92.8	92.3	93.0	89.7	90.6	93.9	94.2	91.8	92.5	87.4	89.7
	negative	93.4	91.5	92.8	91.5	91.6	89.7	91.9	90.9	90.5	85.9	89.3	90.2
	mixed	90.8	92.3	92.7	93.7	89.6	92.0	94.8	95.6	94.7	95.3	87.8	88.5
5%/0%	non	92.2	93.1	92.2	92.8	92.6	91.5	95.6	94.8	94.3	93.4	86.0	87.8
	negative	91.6	91.6	90.4	91.2	91.4	89.7	91.6	91.8	90.6	85.6	89.5	89.7
	mixed	91.6	92.7	91.9	92.4	90.3	91.3	95.1	95.6	95.4	94.8	87.0	90.7
5%/-10%	non	93.4	90.9	93.8	92.7	89.8	90.7	93.9	92.8	94.3	90.5	88.6	88.0
	negative	92.9	90.2	91.2	92.8	89.9	90.0	91.9	90.7	90.7	84.5	90.3	88.1
	mixed	94.6	92.6	92.6	94.2	91.5	90.3	95.7	93.9	94.5	93.6	86.5	88.9
5%/5%	non	92.8	93.5	92.3	93.1	91.5	91.2	94.4	94.9	93.5	92.8	85.7	88.2
	negative	91.8	95.2	92.2	93.9	89.8	90.6	90.2	92.7	91.0	88.3	89.8	89.7
	mixed	94.0	91.9	94.5	92.5	93.4	92.1	95.7	94.3	95.2	94.2	90.1	88.6

Table 6.8: Number of converged runs from 1000 simulation runs. Abbreviations: ΔSe=Difference of sensitivities, ΔSp=Difference of specificities, corr=correlation between Se_1, Sp_1, Se_2 and Sp_2, SI=GLMM using PQL and the identity link, SL=GLMM using PQL and the logit link, CG = Gaussian Copula, CP = C-Vine Copula, CD = D-Vine Copula, DC = Discrete D-Vine Copula, True model=Gaussian copula

		Estimated model											
True ΔSe and ΔSp	True corr	SI		SL		CG		CP		CD		DC	
		ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp
0%/0%	non	279	279	920	920	992	992	997	997	999	999	964	964
	negative	192	192	680	680	972	972	994	994	992	992	980	980
	mixed	261	261	754	754	490	490	469	469	517	517	352	352
0%/-10%	non	6	6	900	900	986	986	996	996	995	995	962	962
	negative	5	5	583	583	956	956	989	989	978	978	974	974
	mixed	17	17	778	778	450	450	461	461	493	493	320	320
0%/5%	non	396	396	925	925	992	992	997	997	999	999	970	970
	negative	241	241	654	654	958	958	993	993	991	991	976	976
	mixed	362	362	758	758	466	466	449	449	529	529	355	355
-10%/0%	non	149	149	919	919	990	990	997	997	1000	1000	964	964
	negative	107	107	637	637	962	962	996	996	992	992	989	989
	mixed	283	283	750	750	483	483	479	479	514	514	352	352
-10%/-10%	non	1	1	907	907	989	989	997	997	994	994	969	969
	negative	1	1	614	614	966	966	992	992	978	978	987	987
	mixed	19	19	783	783	478	478	504	504	476	476	341	341
-10%/5%	non	256	256	914	914	995	995	996	996	997	997	970	970
	negative	159	159	687	687	967	967	994	994	991	991	970	970
	mixed	347	347	748	748	472	472	469	469	512	512	342	342
5%/0%	non	284	284	903	903	990	990	996	996	997	997	961	961
	negative	188	188	639	639	965	965	989	989	984	984	976	976
	mixed	265	265	768	768	496	496	509	509	531	531	331	331
5%/-10%	non	8	8	898	898	987	987	996	996	992	992	957	957
	negative	5	5	594	594	949	949	991	991	967	967	974	974
	mixed	16	16	781	781	481	481	489	489	482	482	297	297
5%/5%	non	421	421	916	916	991	991	995	995	996	996	964	964
	negative	281	281	678	678	973	973	997	997	993	993	980	980
	mixed	375	375	799	799	451	451	505	505	509	509	324	324

Table 6.9: Number of converged runs from 1000 simulation runs. Abbreviations: ΔSe=Difference of sensitivities, ΔSp=Difference of specificities, corr=correlation between Se_1, Sp_1, Se_2 and Sp_2, SI=GLMM using PQL and the identity link, SL=GLMM using PQL and the logit link, CG = Gaussian Copula, CP = C-Vine Copula, CD = D-Vine Copula, DC = Discrete D-Vine Copula, True model=GLMM

		Estimated model											
True ΔSe and ΔSp	True corr	SI		SL		CG		CP		CD		DC	
		ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp
0%/0%	non	659	659	817	817	925	925	938	938	948	948	912	912
	negative	256	256	418	418	911	911	934	934	933	933	915	915
	mixed	575	575	756	756	938	938	935	935	952	952	889	889
0%/-10%	non	455	455	823	823	877	877	905	905	895	895	837	837
	negative	157	157	369	369	824	824	861	861	853	853	850	850
	mixed	387	387	731	731	868	868	890	890	887	887	816	816
0%/5%	non	670	670	820	820	957	957	963	963	962	962	934	934
	negative	298	298	441	441	927	927	944	944	931	931	926	926
	mixed	641	641	770	770	941	941	948	948	953	953	886	886
-10%/0%	non	589	589	839	839	924	924	948	948	954	954	902	902
	negative	231	231	386	386	890	890	915	915	917	917	898	898
	mixed	545	545	753	753	910	910	930	930	938	938	870	870
-10%/-10%	non	431	431	793	793	842	842	899	899	891	891	833	833
	negative	174	174	327	327	799	799	850	850	844	844	822	822
	mixed	396	396	735	735	852	852	883	883	876	876	836	836
-10%/5%	non	599	599	797	797	935	935	957	957	955	955	906	906
	negative	259	259	390	390	901	901	942	942	932	932	894	894
	mixed	574	574	764	764	941	941	946	946	945	945	897	897
5%/0%	non	652	652	812	812	954	954	961	961	959	959	925	925
	negative	273	273	408	408	917	917	935	935	933	933	907	907
	mixed	572	572	765	765	931	931	950	950	941	941	885	885
5%/-10%	non	441	441	785	785	854	854	912	912	907	907	860	860
	negative	183	183	363	363	822	822	863	863	836	836	848	848
	mixed	392	392	726	726	852	852	882	882	858	858	811	811
5%/5%	non	676	676	816	816	947	947	965	965	970	970	930	930
	negative	292	292	424	424	924	924	949	949	930	930	903	903
	mixed	615	615	760	760	939	939	943	943	950	950	879	879

Example SAS Code

In appendix C the SAS code to fit the generalized linear mixed model using PQL and the logit link for the diabetes data set is given.

```
* HbA1c and FPG data set;
* First test: HbA1c, second test: FPG;
DATA diabetes;
INPUT study tp1 fn1 fp1 tn1
            tp2 fn2 fp2 tn2;
      s1=tp1+fn1;
      h1=tn1+fp1;
      s2=tp2+fn2;
      h2=tn2+fp2;
DATALINES;
1   574 262  682  1389   633 203 465  1606
2   489 146  1774 6966   445 190 524  8216
3   36  13   95   998    33  16  120  973
4   338 29   2376 4060   266 101 1389 5047
5   16  25   10   147    21  20  25   132
6   424 192  121  2112   612 4   1281 952
7   644 151  286  1217   648 147 293  1210
8   50  14   4    40     57  7   6    38
9   176 102  768  1286   206 72  823  1231
10  72  16   46   258    75  13  35   269
11  89  26   302  1382   74  41  79   1605
12  23  7    5    109    16  14  21   93
13  181 71   1988 3877   198 54  721  5144
14  25  2    45   243    24  3   60   228
15  12  2    25   268    12  2   59   234
16  13  6    35   260    9   10  40   255
17  43  11   62   389    40  14  24   427
18  17  4    11   79     10  11   2   88
19  575 52   1270 980    554 73  469  1781
20  100 12   108  381    77  35  103  386
21  610 285  1692 3358   555 340 1667 3383
22  184 154  638  5265   262 76  1418 4485
23  58  23   129  204    22  59  25   308
24  195 100  718  969    199 96  580  1107
25  74  40   1156 3660   80  34  771  4045
26  145 41   1107 4719   121 65  641  5185
27  61  31   742  2950   69  23  876  2816
28  21  43   36   284    14  50  24   296
29  42  15   318  814    35  22  198  934
30  392 267  1130 5015   541 118 2116 4029
31  187 181  1112 8562   328 40  2411 7263
32  9   8    37   395    15  2   71   361
33  88  76   39   265    82  82  33   271
34  52  13   37   79     39  26  29   87
35  22  22   35   129    19  25  20   144
36  79  1    689  223    75  5   686  226
37  114 64   20   203    139 39  27   196
38  135 43   96   592    93  85  0    688
;RUN;
```

```
* Quadruplicate the data set;
DATA glimmix1;
    SET diabetes;
    DO temp1=1 TO 4; OUTPUT;END;
RUN;

* Assign the corresponding outcome;
DATA glimmix2;
    SET glimmix1;
    IF temp1=1 THEN DO; test=1; outcome="Sens"; outcomenum=0; outcomenum0=1;
                                outcomenum1=0; outcomenum2=0; outcomenum3=0; num=tp1; den=s1; END;
    IF temp1=2 THEN DO; test=1; outcome="Spec"; outcomenum=1; outcomenum0=0;
                                outcomenum1=1; outcomenum2=0; outcomenum3=0; num=tn1; den=h1; END;
    IF temp1=3 THEN DO; test=2; outcome="Sens"; outcomenum=2; outcomenum0=0;
                                outcomenum1=0; outcomenum2=1; outcomenum3=0; num=tp2; den=s2; END;
    IF temp1=4 THEN DO; test=2; outcome="Spec"; outcomenum=3; outcomenum0=0;
                                outcomenum1=0; outcomenum2=0; outcomenum3=1; num=tn2; den=h2; END;
RUN;

PROC GLIMMIX DATA=glimmix2 METHOD=rspl MAXOPT=2000;
    CLASS study outcomenum outcome;
    MODEL num/den=outcomenum / NOINT DIST=binomial LINK=logit SOLUTION;
    RANDOM outcomenum / SUBJECT=study TYPE=un;

    ESTIMATE "Sensitivity, HbA1c" outcomenum 1 0 0 0/  ILINK CL DF=10000;
    ESTIMATE "Specificity, HbA1c" outcomenum 0 1 0 0/  ILINK CL DF=10000;
    ESTIMATE "Sensitivity, FPG"   outcomenum 0 0 1 0/  ILINK CL DF=10000;
    ESTIMATE "Specificity, FPG"   outcomenum 0 0 0 1/  ILINK CL DF=10000;
    ESTIMATE "Difference of Sensitivities" outcomenum 1 0 -1  0/ ILINK CL;
    ESTIMATE "Difference of Specificities" outcomenum 0 1  0 -1/ ILINK CL;

    ODS OUTPUT Estimates=MuEstimates(keep=LABEL Mu);
    ODS OUTPUT Estimates=StdErrMuEstimates(keep=LABEL StdErrMu);
    ODS OUTPUT Estimates=PDiff(keep=LABEL Probt);
    ODS OUTPUT Estimates=glimmixestimates(drop=Estimate Statement DF tValue Probt StdErr Alpha Lower Upper
                        rename=(Mu=Estimate LowerMu=KI95Lower UpperMu=KI95Upper StdErrMu=SE));

    NLOPTIONS TECH=newrap MAXITER=1000;
RUN;

* Calculate 95% confidence intervals for the estimated differences on the original
  [0,1]-scale;
PROC TRANSPOSE DATA=MuEstimates(where=(Label in ("Sensitivity, HbA1c", "Specificity, HbA1c",
                                                 "Sensitivity, FPG", "Specificity, FPG")))
               OUT=TransMuEstimates(rename=(COL1=Sens1 COL2=Spec1 COL3=Sens2 COL4=Spec2)
                   drop=_NAME_ _LABEL_);
RUN;
PROC TRANSPOSE DATA=StdErrMuEstimates(where=(Label in ("Sensitivity, HbA1c", "Specificity, HbA1c",
                                                       "Sensitivity, FPG", "Specificity, FPG")))
               OUT=TransStdErrMuEstimates(rename=(COL1=SE_Sens1 COL2=SE_Spec1 COL3=SE_Sens2 COL4=SE_Spec2)
                   drop=_NAME_ _LABEL_);
RUN;
PROC TRANSPOSE DATA=PDiff(where=(Label in ("Difference of Sensitivities", "Difference of Specificities")))
```

```
              OUT=TransPDiff(rename=(COL1=PValue_DiffSens COL2=PValue_DiffSpec) drop=_NAME_ _LABEL_);
RUN;

DATA GLIMMIXresults;
    MERGE TransMuEstimates TransStdErrMuEstimates TransPDiff;
    diffsens=Sens1-Sens2;
    diffspec=Spec1-Spec2;

    * Calculate the standard error for the differences;
    Quantile_DiffSens=probit(1 - PValue_DiffSens/2);
    Quantile_DiffSpec=probit(1 - PValue_DiffSpec/2);

    StdErr_DiffSens=abs(diffsens)/Quantile_DiffSens;
    StdErr_DiffSpec=abs(diffspec)/Quantile_DiffSpec;

    CI95L_diffsens=diffsens - probit(0.975)*StdErr_DiffSens;
    CI95U_diffsens=diffsens + probit(0.975)*StdErr_DiffSens;
    CI95L_diffspec=diffspec - probit(0.975)*StdErr_DiffSpec;
    CI95U_diffspec=diffspec + probit(0.975)*StdErr_DiffSpec;
RUN;
PROC PRINT DATA=GLIMMIXresults NOOBS LABEL;
    VAR diffsens CI95L_diffsens CI95U_diffsens
    diffspec CI95L_diffspec CI95U_diffspec;
    LABEL diffsens="Difference of Sensitivities";
    LABEL diffspec="Difference of Specificities";
    LABEL CI95L_diffsens="Lower limit 95%-CI"; LABEL CI95U_diffsens="Upper limit 95%-CI";
    LABEL CI95L_diffspec="Lower limit 95%-CI"; LABEL CI95U_diffspec="Upper limit 95%-CI";
    TITLE "GLMM, Logit-Link, Differences of Sensitivities and Specificities with 95%-CI";
RUN;
```

Erklärung

Ich versichere an Eides Statt, dass die Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der "Grundsätze zur Sicherheit guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf" erstellt worden ist. Die Dissertation wurde in der vorgelegten oder in ähnlicher Form noch bei keiner anderen Institution eingereicht. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Annika Hoyer Düsseldorf, den 27.01.2016