Meta-analysis for the comparison of two diagnostic tests to a common gold standard: A generalized linear mixed model approach

Journal:	*Statistical Methods in Medical Research*
Manuscript ID	SMM-16-0016
Manuscript Type:	Original Article
Keywords:	Meta-analysis, Sensitivity, Specificity, Comparison, Generalized linear mixed models
Abstract:	Meta-analysis of diagnostic studies is still a rapidly developing area of biostatistical research. Especially, there is an increasing interest in methods to compare different diagnostic 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. If information on several thresholds of the tests are available, the quadrivariate model can be further generalized to yield a comparison of full ROC curves. We illustrate our model by an example where two screening methods for the diagnosis of type 2 diabetes are compared.

Journal Title
XX(X):2–??

DOI: 10.1177/ToBeAssigned
www.sagepub.com/

Meta-analysis for the comparison of two diagnostic tests to a common gold standard: A generalized linear mixed model approach

Annika Hoyer[1], Oliver Kuss[1]

Abstract

Meta-analysis of diagnostic studies is still a rapidly developing area of biostatistical research. Especially, there is an increasing interest in methods to compare different diagnostic 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. If information on several thresholds of the tests are available, the quadrivariate model can be further generalized to yield a comparison of full ROC curves. We illustrate our model by an example where two screening methods for the diagnosis of type 2 diabetes are compared.

Keywords
Meta-analysis; Sensitivity; Specificity; Comparison; Generalized linear mixed models

Introduction

Statistical methods for the meta-analysis of diagnostic studies has been a vivid research area in recent years. Although it is meanwhile accepted that the bivariate logistic regression model with random effects [1;2] should be regarded as the standard approach for such analyses, this model has been extended in several directions. In response to numerical problems when using maximum likelihood methods for estimation, more robust methods have been proposed that are guaranteed to give always estimates with confidence intervals [3;4]. We proposed to use a model with beta-binomial marginal distributions that are linked by a copula [5], which results in a closed likelihood function, thus better convergence, and offers additional flexibility for modelling the correlation between sensitivity and specificity. Moreover, it has been argued to additionally account for the disease prevalence to arrive at summary estimates for sensitivity and specificity by using trivariate models [6;7].

It is surprising that there is yet no extension that allows meta-analysis for the comparison of two diagnostic tests to a common gold standard. These studies occur more often than expected as it was shown by Takwoingi et al. [8] which found more than 450 systematic reviews which compared the accuracy of two or more tests. In line with this, other medical researchers have called for meta-analytic methods to this task. For example, Tatsioni et al. [9] wrote as early as in 2005, 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.' Leeflang et al. [10] emphasized that 'policymakers and guideline developers may be particularly interested in comparative accuracy' of diagnostic tests. In our research area of diabetes there are two systematic reviews [11;12] that compare HbA_{1c} and fasting plasma glucose for the population-based screening of type 2 diabetes mellitus. Both reviews

[1] German Diabetes Center, Leibniz Institute for Diabetes Research at Heinrich Heine University Düsseldorf, Institute for Biometry and Epidemiology

Corresponding author:
Annika Hoyer, Deutsches Diabetes-Zentrum, Institut für Biometrie und Epidemiologie, Auf'm Hennekamp 65, 40225 Düsseldorf, Germany
Email: annika.hoyer@ddz.uni-duesseldorf.de

include more than 30 studies, but report results only qualitatively. Especially, they do not report differences of sensitivities or specificities which are probably the parameters of highest interest when comparing two diagnostic tests.

Admittedly, methods have been proposed earlier for meta-analysis to compare two diagnostic tests. However, the method of Siadaty et al.[13;14] confounds the information for sensitivity and specificity and their differences by combining them in a diagnostic odds ratio, a measure which is rarely used by practitioners. Another approach was given by Trikalinos et al.[15]. This approach assumes the two tests to be independent, which would rarely be given in most clinical applications. Moreover, the method can only be applied if both, individual and aggregated proband data, are available. Finally, the Diagnostic Test Accuracy Working Group of the Cochrane Collaboration proposed an approach which allows comparisons of diagnostic tests[16]. They suggest a meta-regression extension of the bivariate model including a binary covariate for the tests to compare. However, this model does not account for potential correlations between the two tests probably compromising the statistical properties of the method. Only recently, this model has been extended to the network meta-analysis situation, allowing comparison of more than two tests by Menten and Lesaffre[17].

In the following, a new model is presented which compensates for the disadvantages of earlier approaches. It computes differences of sensitivities and specificities while fully accounting for all correlations within and heterogeneities between studies. The model is a natural quadrivariate extension of the standard bivariate model for meta-analyzing one diagnostic test. As such it inherits all the well-known and appreciated properties from this model. Additionally, it is possible to use information from multiple test thresholds if these are given in the single studies. In Section 2, we introduce the data set that motivated our research. Section 3 introduces our model and in Section 4 we report the results of a small simulation study that validates the proposed model in realistic situations. In Section 5, we come back to our data set. Finally, in Section 6, we summarize and discuss our findings, and point to advantages and drawbacks of the model.

Data set

We illustrate our method by two systematic reviews[11;12] on population-based screening of type 2 diabetes mellitus. In principle, three methods are available to diagnose diabetes: the oral glucose tolerance test (OGTT), measurement of HbA_{1c} and measurement of fasting plasma glucose (FPG). HbA_{1c} and FPG are less invasive than the OGTT, where

HbA_{1c} has the additional advantage that patients are not requested to refrain from eating and drinking any liquids other than water before the testing procedure, which is especially important in a screening setting.

In the two reviews, the single studies use mainly the OGTT as reference standard and compare HbA_{1c} to FPG. Admittedly, the actual situation is a bit more complicated, and the study-specific reference standards sometimes also includes information from HbA_{1c} or FPG, potentially favouring one of the two tests over the other. However, we ignore these subtleties here for the sake of the presentation of our method. Just aside, differences between reference standards were also ignored in the original Kodama paper[12]. Moreover, in both reviews no quantitative estimates were reported but results were given only narratively.

For a first analysis we use data from Bennett et al.[11] and Kodama et al.[12] as given in Table 1.

PLACE TABLE 1 APPROXIMATELY HERE

In a second analysis we use the same two systematic reviews, but additionally include all information on the reported thresholds of HbA_{1c} and FPG from the single studies. This was done because we noticed that a number of studies reported this additional information and we did not want it to be wasted. To this task, we re-run the search algorithm from Kodama et al.[12], but found no additional studies. One of us (AH) then read all single studies in full text and reconstructed the four-fold tables for each reported threshold. As a result, we found that in 38 studies 135 pairs of sensitivity and specificity were given which used 26 different thresholds for HbA_{1c} (ranging from 3.9 to 7.6) and 27 for FPG (ranging from 3.0 to 7.8). That is, a standard analysis that uses only a single pair of sensitivity and specificity from each study, would use only 28% of the available observations. The full data set can be found in the Supporting Web Materials.

Statistical methods

Bivariate logistic regression model

As our model is a straightforward extension of the bivariate standard model, we shortly reiterate this model. To this task, we use the following notation. We assume that each individual study (indexed by $i = 1, \ldots, I$) in the meta-analysis reports a four-fold table with the number of true positives (TP_i), true negatives (TN_i), false positives (FP_i), and false negatives (FN_i). The sensitivity of the i-th study is defined as $Se_i = TP_i/(TP_i +$

FN_i) and the specificity as $Sp_i = TN_i/(TN_i + FP_i)$. The numbers of true positives and true negatives are assumed to be binomially distributed:

$$TP_i \mid Se_i \sim Binomial(TP_i + FN_i, Se_i), \tag{1}$$

$$TN_i \mid Sp_i \sim Binomial(TN_i + FP_i, Sp_i). \tag{2}$$

To model potential across study correlation and heterogeneity of sensitivity and specificity, a generalized linear mixed model is used:

$$logit(Se_i) = \mu + \phi_i,\ logit(Sp_i) = \nu + \psi_i \tag{3}$$

with $logit(p) = log(p/(1-p))$ and random effects ϕ_i and ψ_i. The random effects are assumed to follow a bivariate normal distribution

$$\begin{pmatrix}\phi_i \\ \psi_i\end{pmatrix} \sim N\left[\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}\sigma_\phi^2 & \rho\sigma_\phi\sigma_\psi \\ \rho\sigma_\phi\sigma_\psi & \sigma_\psi^2\end{pmatrix}\right]. \tag{4}$$

That is, σ_ϕ^2 and σ_ψ^2 model the heterogeneity (on the logit scale) in sensitivities and specificities across studies, and ρ the across study correlation.

As noted in the introduction, the 'Cochrane' model[16] extends the bivariate model by a single binary covariate for the tests under comparison, resulting in

$$logit(Se_i) = \mu + \alpha + \phi_i,\ logit(Sp_i) = \nu + \beta + \psi_i, \tag{5}$$

but again, this model assumes that the two diagnostic tests are independent.

Quadrivariate logistic regression model

As written before, the quadrivariate model for comparing two tests is an extension of the bivariate model. We now assume that each study reports two four-fold tables with the number of true positives (TP_{ij}), true negatives (TN_{ij}), false positives (FP_{ij}), and false negatives (FN_{ij}) for the i-th study and the j-th diagnostic test ($j = 1, 2$). Note that we assume that the gold standard is the same for both tests, so that each individual contributes three binary pieces of information: its result for test 1, its result for test 2 and its true disease status.

Analogous to the bivariate approach, we assume that the true positives and the true negatives of the i-th study and the j-th test are binomially distributed, given the

sensitivities and the specificities of test j and study i.

$$TP_{ij} \mid Se_{ij} \sim Binomial(TP_{ij} + FN_{ij}, Se_{ij}), \tag{6}$$

$$TN_{ij} \mid Sp_{ij} \sim Binomial(TN_{ij} + FP_{ij}, Sp_{ij}), \tag{7}$$

The corresponding logit transformations are

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

where $logit(p) = log(p/(1-p))$. The random effects $(\phi_{i1}, \psi_{i1}, \phi_{i2}, \psi_{i2})^T$ are now assumed to follow a quadrivariate normal distribution

$$\begin{pmatrix} \phi_{i1} \\ \psi_{i1} \\ \phi_{i2} \\ \psi_{i2} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{\phi_1} & \rho_{\phi_1\psi_1}\sigma_{\phi_1}\sigma_{\psi_1} & \rho_{\phi_1\phi_2}\sigma_{\phi_1}\sigma_{\phi_2} & \rho_{\phi_1\psi_2}\sigma_{\phi_1}\sigma_{\psi_2} \\ & \sigma^2_{\psi_1} & \rho_{\psi_1\phi_2}\sigma_{\psi_1}\sigma_{\phi_2} & \rho_{\psi_1\psi_2}\sigma_{\psi_1}\sigma_{\psi_2} \\ & & \sigma^2_{\phi_2} & \rho_{\phi_2\psi_2}\sigma_{\phi_2}\sigma_{\psi_2} \\ & & & \sigma^2_{\psi_2} \end{pmatrix} \right]. \tag{8}$$

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 sensitivities (Se_1, Se_2) and specificities (Sp_1, Sp_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 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 fitting two independent bivariate models for both diagnostic tests separately.

Finally, the differences of sensitivities and specificities as our main parameters of interest are estimated as follows:

$$\Delta Se = \frac{\exp(\hat{\mu}_1)}{1+\exp(\hat{\mu}_1)} - \frac{\exp(\hat{\mu}_2)}{1+\exp(\hat{\mu}_2)} \tag{9}$$

for the difference of sensitivities, and analogously for ΔSp, the difference of specificities, through replacing the $\hat{\mu}_j$ by $\hat{\nu}_j$.

Accounting for multiple thresholds

Results from diagnostic tests frequently originate from dichotomizing a continuous marker at certain thresholds. The single studies in a meta-analysis thus might report several four-fold tables, one for each threshold. These additional information is frequently ignored in meta-analyses and we saw this waste of information also in

our example meta-analysis. However, it is straightforward to include the threshold information as a covariate in our model (and of course, also in the bivariate standard model) by using

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

where μ_j and ν_j are intercepts for $logit(Se_{ij})$ and $logit(Sp_{ij})$ and X_{ij} is a vector containing the threshold values from each study and each test. The threshold values themselves and also the number of them can differ for every study. To model the random effects $(\phi_{i1}, \psi_{i1}, \phi_{i2}, \psi_{i2})^T$, a quadrivariate normal distribution is assumed as before. It should be noted that accounting for thresholds simply corresponds to a meta-analysis of full ROC curves from the single studies. As such we propose here also a method for the meta-analysis of differences of ROC curves.

Simulation

To assess the statistical properties of our model in realistic situations, a simulation study was conducted. The simulation program was written in SAS 9.3 (SAS Institute Inc., Cary, NC, USA).

Setting

Being inspired by our two example meta-analyses [11;12] and another data set from cardiology [18], the following parameters were varied:

- True sensitivities and specificities:
  The true sensitivity and specificity of test 1 was held constant with 70% and 80%, respectively. The true sensitivity and specificity of test 2 were varied between (65%, 70%, 80%) and (75%, 80%, 90%), respectively. Following this, we achieved true differences in sensitivities of -10 percentage points (pp), 0 pp, and 5 pp and true differences in specificities of -10 pp, 0 pp, and 5 pp.
- The true association between sensitivities and specificities of both tests:

To this task, we assumed the following three random effect matrices (as in (8)), here given as their corresponding correlation matrices:

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

$$\Gamma_{mix} = \begin{pmatrix} 1 & -0.3 & 0.2 & -0.3 \\ -0.3 & 1 & -0.3 & 0.2 \\ 0.2 & -0.3 & 1 & -0.3 \\ -0.3 & 0.2 & -0.3 & 1 \end{pmatrix}.$$

Γ_{none} assumes that sensitivities and specificities across studies and even between two tests, that is, within the same patient, are completely independent. For Γ_{neg} we chose a negative correlation of -0.3 between sensitivity and specificity of each test because negative correlations are frequently observed and actually expected in reality. The correlation between the two sensitivities and the two specificities is assumed to be -0.2. The matrix Γ_{mix} denotes a mixed correlation structure. Based on Γ_{neg} we now assume a positive correlation of 0.2 between sensitivities and specificities, because such positive values for the correlations were seen in our diabetes data set.

We did not vary the true random effect variances $\sigma^2_{\phi_1}, \sigma^2_{\psi_1}, \sigma^2_{\phi_2}, \sigma^2_{\psi_2}$, but kept them constant at the value 0.27 on the logit scale. This value was inspired by our previous work and corresponds to a variance of sensitivity (and specificity) of 0.02 on the [0,1]-scale.

Data generation

After combining the design parameters we got 27 different simulation scenarios. For each of them, 1,000 meta-analyses were generated. The simulated number of studies within each meta-analysis was uniformly distributed and varied between 10 and 30. The study sizes were also generated from a uniform distribution and varied between 30 and 200. Finally, the number of diseased persons in each study and for a given study size was also sampled from a uniform distribution that varied between 0 and the sampled size. These choices were based on different meta-analyses reported in practice, for example

by Menke et al.[19] or Kodama et al.[12]. To generate the observed numbers of true positives and true negatives in the single studies, the VNORMAL call in SAS/IML was used to create quadrivariate normally distributed random vectors following the specifications for the respective Γ_*. These random numbers were used to calculate logit-transformed values for the two sensitivities and specificities with respect to their true values. After this, an expit-transformation led to the values for Se_{*1}, Sp_{*1}, Se_{*2}, and Sp_{*2}. These were multiplied by the number of diseased and non-diseased probands and rounded to the nearest integer to get the final numbers of true positives and true negatives.

Estimation methods

For each of the simulated meta-analyses, 14 parameters have to be estimated for the quadrivariate model. These are the two sensitivities, the two specificities, and the 10 values in the random effects covariance matrix. Parameter estimation via the maximum likelihood principle in generalized linear mixed models is complicated by the fact that integrals which can not be solved analytically, appear in the likelihood function. Well-established methods that address this problem and yield exact maximum likelihood estimates are Gaussian quadrature or Markov Chain Monte Carlo (MCMC). Approximate methods like penalized quasi-likelihood (PQL) are also available. We restrict here to Gaussian quadrature and PQL estimation because both methods can be conveniently coded in SAS procedures NLMIXED and GLIMMIX. The GLIMMIX code is given in the Supporting Web Materials. Actually, with respect to estimation methods, we compared 3 implementations:

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

We included a model with an identity link, because in this model the raw difference in sensitivities $\hat{\mu}_1 - \hat{\mu}_2$ and specificities $\hat{\nu}_1 - \hat{\nu}_2$ originate directly from the natural parameters. Opposed to this and as seen in equation (9), for the standard logit link, differences in sensitivities and specificities are linear combinations of model parameters and their confidence intervals have to be computed, with some extra effort, by the multivariate delta method. All procedures were run with their default options to ensure a fair comparison between models. Starting values for the GLIMMIX procedures are automatically generated within the procedure. In case of NLMIXED, where starting values should be given, we computed them as raw proportions of sensitivities and

specificities. The starting values for the variances and correlations were also generated using the corresponding raw values and appropriate transformations of them.

As a reference method we also included the model of the Diagnostic Test Accuracy Working Group of the Cochrane Collaboration, henceforth denoted as the 'Cochrane' method. Parameters from this model were estimated by Gaussian quadrature via the SAS NLMIXED code from Macaskill et al.[16] Additionally, we implemented SAS GLIMMIX code yielding PQL estimates as in the quadrivariate case.

For comparison of the estimation methods, mean bias and empirical coverage (to the 95% level) were calculated. Confidence intervals were calculated assuming t-distributions where we used the default numbers of degrees of freedom from the respective SAS procedure. To address the problem of numerical robustness, we report the number of converged runs, too.

Results

Our parameters of interest are the differences between sensitivities and specificities of the two tests. Therefore, in reporting our results we restrict to them.

Tables 2, 3 and 4 give the simulation results for the situation that is most similar to our diabetes example where the test with the higher sensitivity has a lower specificity as compared to the other test. However, the description of our outcomes is based on the complete simulation results which can be found in the Supporting Web Materials.

Bias In terms of bias all estimation methods performed nearly similar, except in a few situations. Averaging the different correlation structures, the overall bias from the model using Gaussian quadrature was slightly higher compared to the other estimation methods. Referring explicitly to the underlying correlation matrices, the differences of sensitivities and specificities were overestimated in case of mixed as compared to the none correlation structures. Comparing the different estimation methods, the most difficult situations, resulting in a higher bias, were these with negative underlying correlations. This occurred especially for the model using Gaussian quadrature and the model with the identity link. The quadrivariate model using PQL and the logit link was the most robust in terms of bias without huge outliers and only small deviations from the true values. Both implementations of the Cochrane approach led to biased estimates in the same range. The magnitude was a bit higher as compared to the quadrivariate model using PQL and the logit link. That means our new proposed model seems to perform better than the proposed Cochrane model in terms of bias.

PLACE TABLE 2 APPROXIMATELY HERE

Coverage In terms of coverage it is important to note that due to random error, values between 93.6% and 96.4% (95%-Wald confidence interval for a binomial proportion of 950 successes out of 1,000 trials) are still compatible with the hypothesis of a correct coverage.

In case of our quadrivariate models, all estimation methods obtained results near the expected 95%. The best results were obtained in cases where no correlation is present. Thereby, Gaussian quadrature had a small advantage over the other quadrivariate models. In case of non-zero correlations, the models using PQL performed similar and better than the model using Gaussian quadrature. The simulation results showed obviously that the Cochrane models led to worse results compared to all implemented estimation methods of the quadrivariate model. That is, our proposed model performs frequently better than the Cochrane approach in terms of coverage.

PLACE TABLE 3 APPROXIMATELY HERE

Convergence In terms of convergence none of the models reaches 1,000 converged runs and worst results were observed in cases with the negative underlying correlation structure. The methods using the the logit link were always superior and the methods using Gaussian quadrature were always inferior. The performance of the model using the identity link depended on the underlying simulation setting. It seemed to be fragile in the scenarios where the specificity of the first test is lower than the specificity of the second test. With respect to convergence, the Cochrane approach led in most cases to better results than the quadrivariate models. This was expected, as the Cochrane model is a simpler model including only two random effects.

PLACE TABLE 4 APPROXIMATELY HERE

Examples

In this section we return to our example on population-based screening of type 2 diabetes mellitus. As noted previously, we report two analyses, the first one using the original data from the two systematic reviews, the second one using the full information from all reported thresholds of HbA_{1c} and FPG.

First analysis using a single threshold per study

The estimated sensitivities, specificities and their corresponding differences are shown in Table 5. Using Gaussian quadrature we found a difference of about 1 pp between the sensitivities of the two tests, favouring HbA_{1c}. The model using PQL and the logit link finds that FPG has a higher sensitivity than HbA_{1c}, but with a high uncertainty as can be seen from the wide confidence interval. Both models judge FPG to have a higher specificity than HbA_{1c}, but again, confidence intervals are wide. Although we have seen in our simulation that the non-canonical identity link is not necessarily inferior in terms of convergence, the model with the identity link did not converge for the example data set.

PLACE TABLE 5 APPROXIMATELY HERE

Second analysis using multiple thresholds per study

In the second analysis, we proceed to use the full information on all possible thresholds for comparing HbA_{1c} and FPG. This is equivalent to perform a meta-analysis on the differences of ROC curves. This comes with the technical difficulty that HbA_{1c} and FPG, in order to compare them, must be measured on the same scale. To this task, we fit a simple linear regression model with the observed HbA_{1c} threshold values as dependent and the observed FPG threshold values as independent variable from our data set. The resulting equation $\mathrm{HbA}_{1c} = 1.348 + 0.767\mathrm{FPG}$ is used to model the relationship between the thresholds. The results are illustrated in Figures 1 and 2 where observed and estimated differences between sensitivities and specificities are given.

PLACE FIGURES 1 AND 2 APPROXIMATELY HERE

We only use the GLMM with PQL estimation and the logit link, because the simulation has shown that this model performs best if one threshold is available. The biggest difference between HbA_{1c} and FPG can be seen in the range between 6.5 and 7.0, where (in case of sensitivity) differences up to 7 pp, favouring FPG, can be found. On the other hand, the largest confidence intervals correspond to these differences. The estimated differences of sensitivities show that FPG is judged to perform better in higher ranges of thresholds. Only when lower thresholds were used, HbA_{1c} is preferred. At the threshold of 6.5 which is recommended by the American Diabetes Association[20] as well as by the WHO[21] for diagnosing diabetes, FPG performs better in terms of sensitivity while the estimated specificity of both tests is nearly identical at this value.

The example shows that it is highly beneficial from a clinical viewpoint to explicitly model the information from different thresholds: Only then sensitivities and specificities can be compared at specific thresholds. In a standard meta-analysis using only two pairs of sensitivity and specificity from each study and test, only one pair of overall differences between sensitivities and specificity would have been available, ignoring all information from the different thresholds.

Discussion

In this paper, we propose a new model for the meta-analysis of diagnostic studies that compare two diagnostic tests to a common gold standard, situations which are not that rare in medical research. Up to now it was not possible to summarize the results in a meta-analytic way, at least if one was interested in reporting differences in sensitivity and specificity between the two tests while accounting for all potential correlations between tests and populations. Our model constitutes a quadrivariate generalized linear mixed model and is thus just a straightforward extension of the current bivariate standard model as proposed by Reitsma et al.[1] and Chu and Cole[2]. As such, all the well established statistical theory and software implementations for generalized linear mixed model with a multivariate outcome can be used. In a small simulation study we showed that the standard logit link and the PQL principle for parameter estimation worked well and better than the 'Cochrane' approach in a variety of realistic scenarios. By simply adding a covariate to the linear predictor we were able to meta-analyse studies with multiple thresholds corresponding to the meta-analysis of differences of ROC curves. This is a straightforward generalization of previous methods that proposed estimation of summary ROC curves while using information from several thresholds, however, for just one single diagnostic test.[22;23]

While introducing our model, we proposed to estimate the random effects covariance matrix (8) in its full unrestricted form. However, this might not always be necessary and restricting variances to the same value or covariances to zero might result in improved fits. Fits for different matrices could be compared by the BIC and by the -2 Log Likelihoods (-2LogL) of nested models, however, only if exact maximum likelihood estimates (e.g., by Gaussian quadrature) are calculated. In case of our diabetes example we achieve the best results in terms of BIC indeed not for the full model with 14 parameters, but for a smaller model with 11 parameters (BIC=184,276.45). It is of interest here that the quadrivariate model which closely approximates the bivariate

Cochrane model in terms of the random effects matrix is judged inferior with respect to the BIC (BIC=184,283.96).

On the other hand, some limitations of the model should be pointed out. First, though the PQL method for parameter estimation was more robust than Gaussian quadrature, there are still some problems concerning numerical robustness. This was expected, because the number of estimated parameters is large in the quadrivariate model, especially as compared to the number of observations, i.e. the numbers of sensitivities and specificities across studies and tests. Models without random effects like copula-based ones as proposed in our previous work[5;7] could be an alternative. In any case, the 'Cochrane' model which is simpler from a statistical viewpoint, behaved well in the simulation, especially concerning robustness, and is a good alternative when the quadrivariate model has converge problems.

It should be noted that our model assumes only the two standard aggregated four-fold tables to be available from each single study. Especially it does not need individual proband data where the three binary outcomes for each individual (result for tests 1 and 2, and the true disease status) would be explicitly given. We do not consider this a real limitation of our model, because in our experience individual-proband data are rarely accessible. On the other hand, if such information were actually available we could introduce an additional hierarchical (that is, proband) level to adequately adjust for within-proband correlation. The resulting, more complex model would still be a quadrivariate GLMM.

Thinking further, methods for comparing more than two diagnostic tests while fully accounting for correlations between tests are definitely needed. For example, in a subsample of larger studies in Takwoingi et al.[8], only one third of all studies compared two tests, but two thirds compared two or more tests. As such, network meta-analyses of diagnostic tests or multiple-test (not multiple-treatment) comparisons will be a fruitful area in future research.

References

1. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology* 2005; **58**(19):982-990. DOI: 10.1016/j.jclinepi.2005.02.022

2. Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *Journal of Clinical Epidemiology* 2006; **59**(12):1331-1332. DOI: 10.1016/j.jclinepi.2006.06.011
3. Chen Y, Liu Y, Ning J, Nie L, Zhu H, Chu H. A composite likelihood method for bivariate meta-analysis in diagnostic systematic reviews. *Statistical Methods in Medical Research* 2014 Dec 14.
4. Zapf A, Hoyer A, Kramer K, Kuss O. Nonparametric meta-analysis for diagnostic accuracy studies. *Statistics in Medicine* 2015; **34**(29)3831-3841. DOI: 10.1002/sim.6583
5. Kuss O, Hoyer A, Solms A. Meta-analysis for diagnostic accuracy studies: A new statistical model using beta-binomial distributions and bivariate copulas. *Statistics in Medicine* 2014; **33**(1):17-30. DOI: 10.1002/sim.5909
6. Chu H, Nie L, Cole SR, Poole C. Meta-analysis of diagnostic accuracy studies accounting for disease prevalence: alternative parameterizations and model selection. *Statistics in Medicine* 2009; **28**(18):2384-99. DOI: 10.1002/sim.3627
7. Hoyer A, Kuss O. Statistical methods for meta-analysis of diagnostic tests accounting for prevalence - A new model using trivariate copulas. *Statistics in Medicine* 2015; **34**(11):1912-24. DOI: 10.1002/sim.6463
8. Takwoingi Y, Leeflang MM, Deeks JJ. Empirical Evidence of the Importance of Comparative Studies of Diagnostic Test Accuracy. *Annals of Internal Medicine* 2013; **158**(7):544-54. DOI: 10.7326/0003-4819-158-7-201304020-00006
9. Tatsioni A, Zarin DA, Aronson N, Samson DJ, Flamm CR, Schmid C, Lau J. Challenges in Systematic Reviews of Diagnostic Technologies. *Annals of Internal Medicine* 2005; **142**(12):1048-1055.
10. Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM; Cochrane Diagnostic Test Accuracy Working Group. Systematic Reviews of Diagnostic Test Accuracy. *Annals of Internal Medicine* 2008; **149**(12):889-897.
11. Bennett CM, Guo M, Dharmage SC. HbA(1c) as a screening tool for detection of Type 2 diabetes: a systematic review. *Diabetic Medicine* 2007; **24**(4):333-43.
12. Kodama S, Horikawa C, Fujihara K, Hirasawa R, Yachi Y, Yoshizawa S, Tanaka S, Sone Y, Shimano H, Iida KT, Saito K, Sone H. Use of high-normal levels of haemoglobin A_1C and fasting plasma glucose for diabetes screening and for prediction: a meta-analysis. *Diabetes/ Metabolism Research and Reviews* 2013; **29**(8):680-692. DOI: 10.1002/dmrr.2445
13. Siadaty MS, Philbrick JT, Heim SW, Schectman JM. Repeated-measures modeling improved comparison of diagnostic tests in meta-analysis of dependent studies. *Journal of Clinical Epidemiology* 2004; **57**(7):698-711. DOI: 10.1016/j.jclinepi.2003.12.007

14. Siadaty MS, Shu J. Proportional odds ratio model for comparison of diagnostic tests in meta-analysis. *BMC Medical Research Methodology* 2004; **4:27**.
15. Trikalinos TA, Hoaglin DC, Small KM, Terrin N, Schmid CH. Methods for the joint meta-analysis of multiple tests. *Research Synthesis Methods* 2014; **5**(4):294-312. DOI: 10.1002/jrsm.1115
16. Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. 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, 2010. Available from: http://srdta.cochrane.org/
17. Menten J, Lesaffre E. A general framework for comparative Bayesian meta-analysis of diagnostic studies. *BMC Medical Research Methodology* 2015; 15:70. DOI: 10.1186/s12874-015-0061-7
18. Picano E, Bedetti G, Varga A, Cseh E. The comparable diagnostic accuracies of dobutamine-stress and dipyridamole-stress echocardiographies: a meta-analysis. *Coronary Artery Disease* 2000. **11**(2):151-159.
19. Menke J. Bivariate random-effects meta-analysis of sensitivity and specificity with SAS PROC GLIMMIX. *Methods of Information in Medicine* 2010; **49**(1):54-62, 62-64. DOI: 10.3414/ME09-01-0001
20. American Diabetes Association. Classification and diagnosis of diabetes. *Diabetes Care* 2015; **38**(Suppl. 1):S8 - S16.
21. World Health Organization. *Use of glycated haemoglobin (HbA1c) in the diagnosis of diabetes mellitus.* 2011. (Available at http://www.who.int/cardiovascular_diseases/report-hba1c_2011_edited.pdf), [Assessed 10 November 2015]
22. Riley RD, Takwoingi Y, Trikalinos T, Guha A, Biswas A, Ensor J, Morris RK, Deeks JJ. Meta-Analysis of Test Accuracy Studies with Multiple and Missing Thresholds: A Multivariate-Normal Model. *J Biomet Biostat* 2014; **5:3**.
23. Martínez-Camblor P. Fully non-parametric receiver operating characteristic curve estimation for random-effects meta-analysis. *Statistical Methods in Medical Research* 2014; 0(0):1-16. (Epub ahead of print)

Table 1. Type 2 diabetes dataset from Bennett et al.[11] and Kodama et al.[12], 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 2. 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

		Estimated model									
True ΔSe and ΔSp	True corr	SN		SI		SL		CM		CA	
		ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp
-10%/5%	none	-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%/-10%	none	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

Table 3. 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

		Estimated model									
True ΔSe and ΔSp	True corr	SN		SI		SL		CM		CA	
		ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp
-10%/5%	none	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%/-10%	none	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

Table 4. 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 ΔSe and ΔSp	True corr	Estimated model									
		SN		SI		SL		CM		CA	
		ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp
-10%/5%	none	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%/-10%	none	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

Table 5. Results using the different GLMMs

Model	Sensitivity HbA1c [95% CI] (in %)	Specificity HbA1c [95% CI] (in %)	Sensitivity FPG [95% CI] (in %)	Specificity FPG [95% CI] (in %)	Difference of sensitivities [95% CI] (in pp)	Difference of specificities [95% CI] (in pp)
GLMM Gaussian Quadrature (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]	1.1 [-0.6; 2.8]	-4.4 [-5.1; -3.6]
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]	-1.0 [-7.8; 5.8]	-3.1 [-8.2; 2.0]

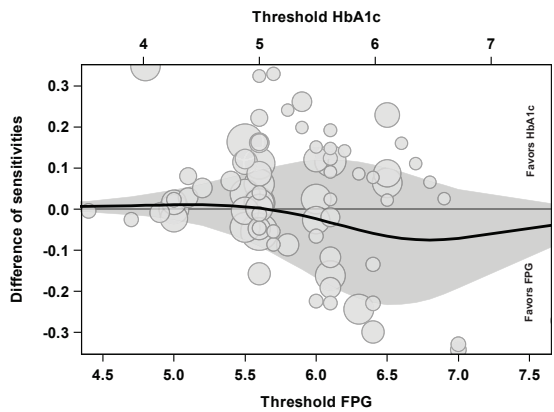

Figure 1. Estimated difference of sensitivities with respect to different thresholds

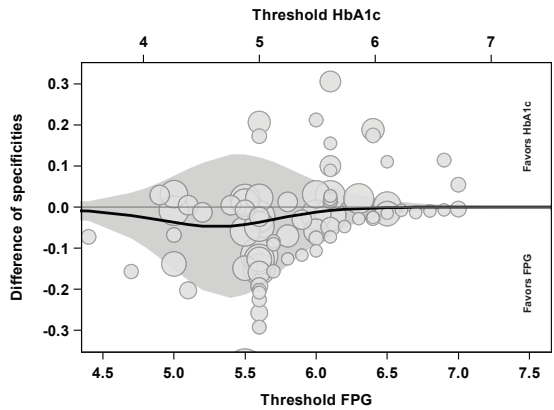

Figure 2. Estimated difference of specificities with respect to different thresholds

Supporting Web Materials for Hoyer and Kuss: Meta-analysis for the comparison of two diagnostic tests to a common gold standard: A generalized linear mixed model approach

In this supplementary material the complete diabetes data set is given. The bold entries are the original included thresholds.

Study	Threshold HbA_{1c}	TP1	FN1	FP1	TN1	Threshold FPG	TP2	FN2	FP2	TN2
Badings et al.	**5.9**	**574**	**262**	**682**	**1389**	**5.5**	**445**	**190**	**524**	**8216**
Choi et al.	5.0	617	18	7735	1005	**6.4**	**633**	**203**	**465**	**1606**
	5.1	607	28	7123	1617	.	.	.	.	.
	5.2	600	35	6302	2438	.	.	.	.	.
	5.3	581	54	5331	3409	.	.	.	.	.
	5.4	563	72	4318	4422	.	.	.	.	.
	5.5	550	85	3356	5384	.	.	.	.	.
	5.6	522	113	2473	6267	.	.	.	.	.
	5.7	**489**	**146**	**1774**	**6966**	.	.	.	.	.
	5.8	457	178	1206	7534	.	.	.	.	.
	5.9	429	206	813	7927	.	.	.	.	.
	6.0	393	242	568	8172	.	.	.	.	.
	6.2	332	303	280	8460	.	.	.	.	.
	6.6	236	399	70	8670	.	.	.	.	.
Li et al.	**6.2**	**36**	**13**	**95**	**998**	**5.6**	**33**	**16**	**120**	**973**
	.	.	.	.	.	6.1	19	30	21	1072
Schöttker et al.	**5.7**	**338**	**29**	**2376**	**4060**	**5.6**	**266**	**101**	**1389**	**5047**
Tahrani et al.	**6.0**	**16**	**25**	**10**	**147**	**5.6**	**21**	**20**	**25**	**132**
	.	.	.	.	.	6.1	13	28	8	149
Wang et al.	**6.0**	**424**	**192**	**121**	**2112**	**5.6**	**612**	**4**	**1281**	**952**
Hu et al.	**6.1**	**644**	**151**	**286**	**1217**	5.5	735	60	687	816
	.	.	.	.	.	**6.1**	**648**	**147**	**293**	**1210**
	.	.	.	.	.	7.0	433	362	0	1503
Zhang et al.	6.0	57	7	12	32	**6.5**	**57**	**7**	**6**	**38**
	6.4	**50**	**14**	**4**	**40**	.	.	.	.	.
	6.5	47	17	3	41	.	.	.	.	.
	7.0	40	24	1	43	.	.	.	.	.
Zhou et al.	**5.6**	**176**	**102**	**768**	**1286**	**6.1**	**206**	**72**	**823**	**1231**
	6.5	69	209	187	1867	.	.	.	.	.
Kim et al.	**6.1**	**72**	**16**	**46**	**258**	**6.1**	**75**	**13**	**35**	**269**
	.	.	.	.	.	7.0	49	39	0	304
Nakagami et al.	**5.3**	**89**	**26**	**302**	**1382**	**6.1**	**74**	**41**	**79**	**1605**
	5.5	72	43	126	1558	.	.	.	.	.
	5.6	65	50	82	1602	.	.	.	.	.
Salmasi et al.	**6.1**	**23**	**7**	**5**	**109**	**6.1**	**16**	**14**	**21**	**93**
Glümer et al.	5.8	207	45	3161	2704	5.5	229	23	2745	3120
	5.9	196	56	2557	3308	**6.1**	**198**	**54**	**721**	**5144**
	6.0	**181**	**71**	**1988**	**3877**	6.3	191	61	416	5449
Anand et al., South Asia	**5.9**	**25**	**2**	**45**	**243**	**5.7**	**24**	**3**	**60**	**228**
Anand et al., China	**5.9**	**12**	**2**	**25**	**268**	**5.7**	**12**	**2**	**59**	**234**
Anand et al., Europe	**5.9**	**13**	**6**	**35**	**260**	**5.7**	**9**	**10**	**40**	**255**
Jesudason et al.	3.9	54	0	450	1	3.0	54	0	451	0
	4.7	54	0	406	45	4.7	54	0	347	104
	5.6	46	8	88	363	5.6	43	11	64	387
	5.7	**43**	**11**	**62**	**389**	**6.0**	**40**	**14**	**24**	**427**
	6.2	23	31	4	447	6.4	32	22	4	447
	6.8	12	42	0	451	7.7	17	37	0	451
Tavintharan et al.	5.9	20	1	30	60	5.8	11	10	9	81
	6.0	19	2	28	62	5.9	11	10	5	85
	6.1	17	4	14	76	6.0	11	10	2	88
	6.2	**17**	**4**	**11**	**79**	**6.1**	**10**	**11**	**2**	**88**
	6.3	17	4	11	79	6.2	9	12	2	88
	6.4	16	5	9	81	6.3	9	12	2	88
	6.5	13	8	9	81	6.4	8	13	1	89
	6.6	12	9	4	86	6.5	8	13	1	89
	6.7	11	10	3	87	6.6	4	17	1	89
	6.8	11	10	3	87	6.7	4	17	0	90
	6.9	11	10	2	88	6.8	4	17	0	90
	7.0	10	11	2	88	6.9	4	17	0	90

	.	.	.	.	.	7.0	4	17	0	90
Ko et al.	**5.5**	**575**	**52**	**1270**	**980**	**5.6**	**554**	**73**	**469**	**1781**
	6.1	486	141	477	1773	5.8	534	93	351	1899
Papoz et al.	5.0	111	1	406	83	4.4	111	1	440	49
	5.5	110	2	264	225	5.0	104	8	367	122
	6.0	**100**	**12**	**108**	**381**	5.6	95	17	264	225
	6.5	78	34	24	465	**6.4**	**77**	**35**	**103**	**386**
	7.0	60	52	10	479	7.0	57	55	29	460
	.	.	.	.	.	7.8	41	71	5	484
Choi et al.	5.0	861	34	4439	611	**4.8**	**555**	**340**	**1667**	**3383**
	5.1	837	58	4050	1000	.	.	.	.	.
	5.2	793	102	3525	1525	.	.	.	.	.
	5.3	740	155	2934	2116	.	.	.	.	.
	5.4	687	208	2288	2762	.	.	.	.	.
	5.5	**610**	**285**	**1692**	**3358**	.	.	.	.	.
	5.6	532	363	1167	3883	.	.	.	.	.
	5.7	455	440	773	4277	.	.	.	.	.
	5.8	376	519	465	4585	.	.	.	.	.
	5.9	298	597	268	4782	.	.	.	.	.
	6.0	235	660	167	4883	.	.	.	.	.
	6.2	136	759	66	4984	.	.	.	.	.
	6.6	46	849	5	5045	.	.	.	.	.
Heianza et al.	**5.7**	**184**	**154**	**638**	**5265**	**5.6**	**262**	**76**	**1418**	**4485**
Law et al.	5.5	72	9	213	120	**5.6**	**22**	**59**	**25**	**308**
	5.6	70	11	189	144	.	.	.	.	.
	5.7	66	15	163	170	.	.	.	.	.
	5.8	**58**	**23**	**129**	**204**	.	.	.	.	.
	5.9	50	31	99	234	.	.	.	.	.
	6.0	40	41	77	256	.	.	.	.	.
Mukai et al.	4.8	285	10	1532	155	4.9	286	9	1507	180
	5.1	262	33	1275	412	5.1	269	26	1293	394
	5.2	253	42	1171	516	5.2	258	37	1151	536
	5.3	234	61	1007	680	5.4	242	53	909	778
	5.5	**195**	**100**	**718**	**969**	5.5	219	76	746	941
	5.6	169	126	588	1099	**5.6**	**199**	**96**	**580**	**1107**
	5.7	138	157	476	1211	5.8	16	279	403	1284
	5.9	91	204	278	1409	5.9	136	159	239	1448
Soulimane et al., Denmark	5.0	108	6	4190	626	5.0	111	3	4094	722
	5.5	91	23	2071	2745	5.6	95	19	1878	2938
	5.7	**74**	**40**	**1156**	**3660**	5.5	99	15	2215	2601
	6.0	48	66	337	4479	**6.0**	**80**	**34**	**771**	**4045**
	6.4	7	107	0	4816	6.5	36	78	144	4672
	.	.	.	.	.	6.8	14	100	48	4768
Soulimane et al., Australia	5.0	184	2	5645	181	5.0	177	9	4719	1107
	5.5	164	22	2389	3437	5.5	154	32	2272	3554
	5.7	**145**	**41**	**1107**	**4719**	5.6	153	33	1864	3962
	6.0	84	102	233	5593	**6.0**	**121**	**65**	**641**	**5185**
	6.4	4	182	0	5826	6.5	63	123	117	5709
	.	.	.	.	.	6.8	20	166	0	5826
Soulimane et al., France	5.0	88	4	3212	480	5.0	86	6	2511	1181
	5.5	74	18	1440	2252	5.5	69	23	1108	2584
	5.7	**61**	**31**	**738**	**2954**	**5.6**	**69**	**23**	**886**	**2806**
	6.0	35	57	185	3507	6.0	51	41	258	3434
	6.4	4	88	0	3692	6.5	16	76	37	3655
	.	.	.	.	.	6.8	7	85	0	3692
Cederberg et al.	**5.7**	**21**	**43**	**36**	**284**	**5.6**	**14**	**50**	**24**	**296**
Nakagami et al.	5.1	49	8	441	691	5.1	49	8	633	499
	5.2	**42**	**15**	**318**	**814**	**5.6**	**35**	**22**	**198**	**934**
	5.3	32	25	138	994	.	.	.	.	.
	5.4	26	31	87	1045	.	.	.	.	.
Sato et al.	**5.5**	**392**	**267**	**1130**	**5015**	**5.6**	**541**	**118**	**2116**	**4029**
	5.0	596	63	4083	2062	6.1	334	325	460	5685
	6.0	137	522	146	5999	.	.	.	.	.
Inoue et al.	**5.5**	**187**	**181**	**1112**	**8562**	**5.6**	**328**	**40**	**2411**	**7263**
Inoue et al.	**5.8**	**9**	**8**	**37**	**395**	**5.6**	**15**	**2**	**71**	**361**
	.	.	.	.	.	6.1	9	8	17	415
Norberg et al.	5.5	122	42	89	215	**6.1**	**82**	**82**	**33**	**271**
	5.7	**88**	**76**	**39**	**265**	.	.	.	.	.
Takahashi et al.	**5.6**	**52**	**13**	**37**	**79**	**6.1**	**39**	**26**	**29**	**87**
Ko et al.	**6.1**	**22**	**22**	**35**	**129**	**6.1**	**19**	**25**	**20**	**144**
Mannucci et al.	**6.6**	**79**	**1**	**689**	**223**	**7.0**	**75**	**5**	**686**	**226**
Wiener et al.	**6.9**	**114**	**64**	**20**	**203**	6.0	160	18	76	147
	7.4	90	88	4	219	**6.9**	**139**	**39**	**27**	**196**
	7.6	73	105	0	223	.	.	.	.	.
Tanaka et al.	**5.9**	**135**	**43**	**96**	**592**	**7.0**	**93**	**85**	**0**	**688**
	6.5	87	91	14	674	.	.	.	.	.

Supporting Web Materials for Hoyer and Kuss: Meta-analysis for the comparison of two diagnostic tests to a common gold standard: A generalized linear mixed model approach

0.1 Bias

Table 0.1: 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 ΔSe and ΔSp	True corr	Estimated model									
		SN		SI		SL		CM		CA	
		ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp
0%/0%	none	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%	none	-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%	none	-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%	none	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%	noen	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%	none	-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%	none	-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%	none	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%	none	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

0.2 Coverage

Table 0.2: 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 ΔSe and ΔSp	True corr	Estimated model									
		SN		SI		SL		CM		CA	
		ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp
0%/0%	none	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%	none	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%	none	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%	none	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%	none	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%	none	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%	none	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%	none	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%	none	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

0.3 Number of converged runs

Table 0.3: 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 ΔSe and ΔSp	True corr	Estimated model									
		SN		SI		SL		CM		CA	
		ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp
0%/0%	none	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%	none	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%	none	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%	none	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%	none	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%	none	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%	none	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%	none	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%	none	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

Supporting Web Materials for Hoyer and Kuss: Meta-analysis for the comparison of two diagnostic tests to a common gold standard: A generalized linear mixed model approach

In this supplementary material the SAS code to fit the generalized linear mixed model 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;
```