SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Statistical Methods in Medical Research
2018, Vol. 27(5) 1410–1421

DOI: 10.1177/0962280216661587
journals.sagepub.com/home/smm
SAGE

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

Annika Hoyer and Oliver Kuss

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 across single studies and between the two tests. 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 naturally generalizes the common standard bivariate model of meta-analysis for a single diagnostic test. If information on several thresholds of the tests is available, the quadrivariate model can be further generalized to yield a comparison of full receiver operating characteristic (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

1 Introduction

Statistical methods for the meta-analysis of diagnostic studies have 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 exist only a few approaches that allow 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 until 2013. 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

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

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

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 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.

Previously proposed methods, however, are not without problems. For example, 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. The approach by Trikalinos et al.[15] can only be applied if both individual and aggregated proband data are available. 'Aggregated' means here that only the four-fold tables of both tests are reported. In case of individual data, the fully cross-classified tables of results across both tests are given. The Diagnostic Test Accuracy Working Group of the Cochrane Collaboration suggests a meta-regression extension of the bivariate model including a binary covariate for the tests to compare.[16] However, this model does not account for potential correlations between the two tests probably compromising the statistical properties of the method. Only recently, Dimou et al.[17] proposed a model relying on ideas from multivariate meta-analysis. This model comes with the challenge that data imputation steps have to be performed for within-proband correlations when studies do not report the full information of the individual participants.

In the following, a new model is presented which compensates for the disadvantages of earlier approaches. It computes differences of sensitivities and specificities while accounting for correlations between tests across studies and heterogeneities across studies. The model is a natural quadrivariate extension of the standard bivariate model for meta-analysing 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.

2 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 include 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 publications.[11,12] 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.

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 two single pairs 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.

It should be noted that ideally the single studies would report individual data on the two diagnostic test results and the true disease status for each proband. In practice, however, information in this extensive way is rarely reported. In most cases, systematic reviews of diagnostic test accuracy studies report aggregated data in form of two four-fold tables as in our example data set. This implicates that in this situation it is not possible to account for within-proband correlation.

Table 1. Type 2 diabetes data set: First test: HbA_{1c}, second test: Fasting plasma glucose (FPG).

Studies from Bennett et al.[11] and Kodama et al.[12]	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

3 Statistical methods

3.1 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 ith 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:

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

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

To model potential across study correlation and heterogeneity of sensitivity and specificity, a generalized linear mixed model (GLMM) 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 Section 1, 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.

3.2 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 ith study and the jth 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 ith study and the jth test are binomially distributed, given the sensitivities and the specificities of test j and study i

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

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

The corresponding logit transformations are

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

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

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

The four variance parameters $\sigma_{\phi_1}^2, \sigma_{\psi_1}^2, \sigma_{\phi_2}^2, \sigma_{\psi_2}^2$ 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 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 $\hat{\nu}_j$.

We want to emphasize again that the model as described here only needs aggregated data, that is, two four-fold tables from the single studies. Admittedly this comes with the restriction that within-proband correlations of test results are assumed to be zero. However, if such information would be available from the single studies, our model could be easily generalized by introducing an additional hierarchical level. The resulting, more complex model would still be a quadrivariate GLMM.

3.3 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 are 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

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

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 receiver operating characteristic (ROC) curves from the single studies. As such we propose here also a method for the meta-analysis of differences of ROC curves which can be seen as a simple alternative of existing methods for the meta-analysis of single ROC curves as proposed by various authors.[18–21]

4 Simulation

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

4.1 Setting

Being inspired by our two example meta-analyses[11,12] and another data set from cardiology,[22] the following parameters were varied:

(1) 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.
(2) The true association between sensitivities and specificities of both tests: To this task, we assumed the following three random effect matrices (as in equation (8)), here given as their corresponding correlation matrices

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

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

Γ_{none} assumes that sensitivities and specificities across studies and even between the two tests 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.

4.2 Data generation

After combining the design parameters we got 27 different simulation scenarios. For each of them, 1000 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 study size. These choices were based on different meta-analyses reported in practice, for example by Kodama et al.[12] or Menke.[23] 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 used to generate the final numbers of true positives and true negatives from binomial distributions.

4.3 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 parameters in the random effects covariance matrix. Parameter estimation via the maximum likelihood principle in GLMMs is complicated by the fact that integrals which cannot 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. 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 three implementations:

- PQL using the logit link (PROC GLIMMIX);
- PQL 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 non-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.

We emphasize that we only generated data from our quadrivariate model itself. As such, in the simulation we are only comparing different estimation methods for our proposed model. Especially, we do not aim for a comparison of our model to the 'Cochrane' method.

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.

4.4 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 to 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.

4.4.1 Bias

In terms of bias all estimation methods performed nearly similar, except in a few situations. Averaging over the different correlation structures, the overall bias from the model using Gaussian quadrature was slightly higher

Table 2. Bias (multiplied by 100) for the differences of sensitivity and specificity on the [0, 1]- scale.

		Estimated model									
		SN		SI		SL		CM		CA	
True ΔSe and ΔSp	True corr	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp
−10%/5%	None	−0.22	−0.40	−0.16	−0.14	0.01	−0.14	0.05	−0.16	0.10	−0.22
	Negative	−0.37	0.38	0.52	−0.16	0.27	−0.01	0.29	0.08	0.17	0.21
	Mixed	−0.07	0.19	0.48	−0.17	0.20	−0.05	0.26	−0.05	0.24	−0.02
5%/−10%	None	0.31	−0.04	0.35	0.22	−0.06	−0.18	−0.03	−0.26	−0.03	−0.26
	Negative	−0.07	−0.23	0.07	0.15	0.14	−0.22	0.04	−0.43	0.01	−0.44
	Mixed	−0.06	−0.28	−0.65	0.19	0.06	−0.13	−0.08	−0.18	−0.09	−0.26

CA: Cochrane model using the PQL and the logit link; CM: Cochrane model using GQ and the logit link; 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; SN: GLMM using GQ; ΔSe: difference of sensitivities; ΔSp: difference of specificities.

Table 3. Empirical coverage (in %) for the 95% confidence intervals for the differences of sensitivity and specificity on the [0, 1]-scale.

		Estimated model									
		SN		SI		SL		CM		CA	
True ΔSe and ΔSp	True corr	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp	ΔSe	ΔSp
−10%/5%	None	94.0	94.3	91.1	94.3	94.2	92.4	66.3	67.8	83.2	84.1
	Negative	92.0	91.5	93.3	94.0	93.5	92.9	64.2	62.5	78.0	77.2
	Mixed	93.7	94.8	90.5	93.7	91.7	93.0	69.2	68.2	86.6	87.2
5%/−10%	None	91.8	94.4	91.6	92.2	93.6	93.3	59.5	73.8	81.3	84.5
	Negative	92.9	90.4	94.7	91.2	93.1	92.1	54.3	70.8	72.6	81.2
	Mixed	93.8	93.2	88.3	91.0	94.0	95.1	66.6	77.5	88.0	87.8

CA: Cochrane model using the PQL and the logit link; CM: Cochrane model using GQ and the logit link; 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; SN: GLMM using GQ; ΔSe: difference of sensitivities; ΔSp: difference of specificities.

Table 4. Number of converged runs from 1000 simulation runs.

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	609	609	315	315	845	845	925	925	976	976
	Negative	439	439	300	300	757	757	685	685	864	864
	Mixed	566	566	317	317	843	843	911	911	974	974
5%/−10%	None	550	550	154	154	840	840	916	916	976	976
	Negative	408	408	114	114	692	692	674	674	873	873
	Mixed	479	479	145	145	811	811	888	888	969	969

CA: Cochrane model using the PQL and the logit link; CM: Cochrane model using GQ and the logit link; 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; SN: GLMM using GQ; ΔSe: difference of sensitivities; ΔSp: difference of specificities.

compared to the other estimation methods. Referring explicitly to the underlying correlation matrices, the differences of sensitivities and specificities were often 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 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.

4.4.2 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 1000 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 when data were generated from our model.

4.4.3 Convergence

In terms of convergence none of the models reaches 1000 converged runs and worst results were observed in cases with the negative underlying correlation structure. The models using the logit link and PQL (the Cochrane and the quadrivariate model) were always superior to the models using Gaussian quadrature and the quadrivariate model using the identity link was always inferior. This approach seemed to be very 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.

5 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.

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)	72.3 [67.2; 77.4]	80.9 [76.8; 85.1]	73.3 [66.7; 79.9]	84.1 [79.7; 88.5]	−1.0 [−7.8; 5.7]	−3.1 [−8.1; 1.8]
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]
Cochrane model Gaussian quadrature (logit link)	69.6 [64.8; 74.4]	80.4 [76.8; 84.0]	73.6 [69.2; 78.0]	82.2 [78.8; 85.5]	−4.1 [−5.5; −2.6]	−1.8 [−2.2; −1.3]
Cochrane model (PQL, logit link)	69.5 [64.4; 74.1]	80.3 [76.4; 83.7]	73.5 [68.8; 77.7]	82.1 [78.4; 85.3]	−4.1 [−10.6; 2.5]	−1.8 [−6.8; 3.2]

5.1 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 large 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. We also estimated the correlation matrix of the random effects. Relying on the estimated correlation matrix of the model using PQL, the situation corresponds to the mixed correlation structure Γ_{mix} of the simulation study. In case of the GLMM using Gaussian quadrature and the logit link, we had to tune the starting values somewhat to achieve a fully estimated random effects correlation matrix. Nevertheless, based on the satisfying results of that model in this simulation setting, it seems reasonable to trust also in the GQ results for the example data set. Estimates from the GQ and the PQL approach are identical, with slightly smaller GQ confidence intervals. 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. In case of the Cochrane approach, the estimated differences of sensitivities and specificities are somewhat larger compared to the quadrivariate GLMM. In line with the quadrivariate GLMM using PQL and the logit link, FPG is preferred compared to HbA_{1c} which is indicated by a higher sensitivity and specificity. Implementing the Cochrane approach using the PQL method leads to broader confidence intervals as compared to Gaussian quadrature.

5.2 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. We used the quadrivariate GLMM to estimate sensitivities and specificities of HbA_{1c} and FPG with their corresponding 95% confidence intervals at prespecified thresholds on the respective scale and drew a summary ROC curve for each test. The results are illustrated in Figure 1 where the two summary ROC curves with corresponding pointwise 95% confidence intervals and the ROC curves from the single studies are given.

We only use the GLMM with PQL estimation and the logit link, because the simulation has shown that this model performs best in the one-threshold case, and we expect this estimation method also to work well in the case of more thresholds. It can be seen that FPG performs better than HbA_{1c} in case of thresholds which are not too extreme. For high or low thresholds there seems to be no relevant difference in the performance of both tests. At the threshold of 6.5 which is recommended by the American Diabetes Association[24] as well as by the WHO[25] for diagnosing diabetes, FPG performs better in terms of sensitivity when holding constant the specificities of both test. To be concrete, the estimated sensitivity and specificity of HbA_{1c} at an HbA_{1c} value of 6.5 are 29.4% [20.2%; 40.6%] and 98.7% [98.0%; 99.2%], respectively. At the identical specificity value for FPG, the sensitivity of FPG is 38.6% [28.2%; 50.2%], which corresponds roughly to a FPG value of 6.7.

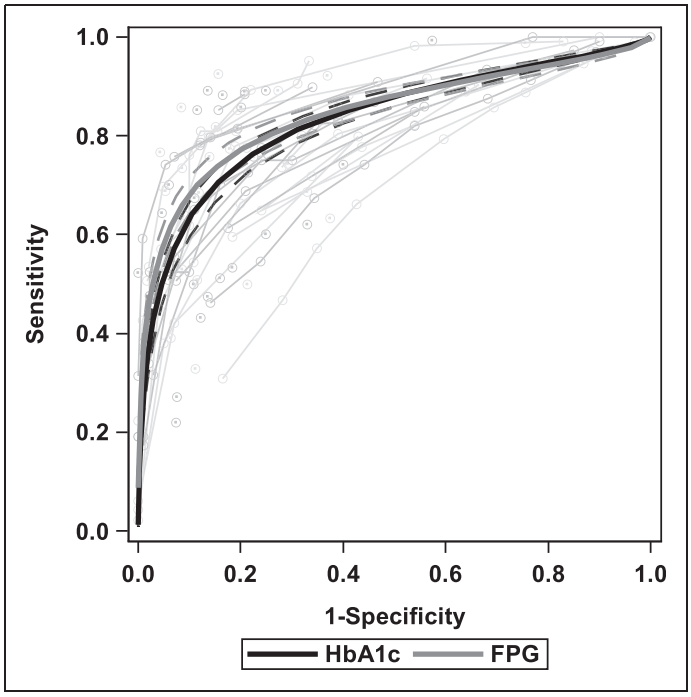

Figure 1. Estimated summary ROC curves of HbA1c (black solid line) and FPG (dark-grey solid line) with 95% confidence intervals (dashed lines) and ROC curves from the 38 single studies.

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.

6 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 potential correlations between tests across studies and heterogeneities across studies. Our model constitutes a quadrivariate GLMM 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 GLMM 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 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 alternative to previous methods that proposed estimation of summary ROC curves while using information from several thresholds, however, for just one single diagnostic test.[18–21]

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 (BIC = 184,286.19), 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. 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 convergence problems.

We emphasize again 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 (i.e. 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 three 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. Only recently, the 'Cochrane' model has been extended to the network meta-analysis situation, allowing comparison of more than two tests by Menten and Lesaffre.[26]

As a reviewer pointed out, another interesting direction for future work would be to compare the area under curves (AUCs) between the two tests. In principle it is easy to compute differences of AUCs by estimating ROC curves for specific values of thresholds and straightforwardly using the trapezoidal rule. Confidence intervals could be estimated with the multivariate delta rule or a simple non-parametric bootstrap approach.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Reitsma JB, Glas AS, Rutjes AW, et al. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 2005; **58**: 982–990.
2. Chu H and Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *J Clin Epidemiol* 2006; **59**: 1331–1332.
3. Chen Y, Liu Y, Ning J, et al. A composite likelihood method for bivariate meta-analysis in diagnostic systematic reviews. *Stat Methods Med Res* 14 December 2014. pii: 0962280214562146.
4. Zapf A, Hoyer A, Kramer K, et al. Nonparametric meta-analysis for diagnostic accuracy studies. *Stat Med* 2015; **34**: 3831–3841.
5. Kuss O, Hoyer A and Solms A. Meta-analysis for diagnostic accuracy studies: a new statistical model using beta-binomial distributions and bivariate copulas. *Stat Med* 2014; **33**: 17–30.
6. Chu H, Nie L, Cole SR, et al. Meta-analysis of diagnostic accuracy studies accounting for disease prevalence: alternative parameterizations and model selection. *Stat Med* 2009; **28**: 2384–2399.
7. Hoyer A and Kuss O. Statistical methods for meta-analysis of diagnostic tests accounting for prevalence – a new model using trivariate copulas. *Stat Med* 2015; **34**: 1912–1924.
8. Takwoingi Y, Leeflang MM and Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. *Ann Intern Med* 2013; **158**: 544–554.
9. Tatsioni A, Zarin DA, Aronson N, et al. Challenges in systematic reviews of diagnostic technologies. *Ann Intern Med* 2005; **142**: 1048–1055.

10. Leeflang MM, Deeks JJ, Gatsonis C, et al., Cochrane Diagnostic Test Accuracy Working Group. Systematic reviews of diagnostic test accuracy. *Ann Intern Med* 2008; **149**: 889–897.
11. Bennett CM, Guo M and Dharmage SC. HbA(1c) as a screening tool for detection of Type 2 diabetes: a systematic review. *Diabet Med* 2007; **24**: 333–343.
12. Kodama S, Horikawa C, Fujihara K, et al. Use of high-normal levels of haemoglobin A_1C and fasting plasma glucose for diabetes screening and for prediction: a meta-analysis. *Diabet/Metab Res Rev* 2013; **29**: 680–692.
13. Siadaty MS, Philbrick JT, Heim SW, et al. Repeated-measures modeling improved comparison of diagnostic tests in meta-analysis of dependent studies. *J Clin Epidemiol* 2004; **57**: 698–711.
14. Siadaty MS and Shu J. Proportional odds ratio model for comparison of diagnostic tests in meta-analysis. *BMC Med Res Methodol* 2004; **4**: 27.
15. Trikalinos TA, Hoaglin DC, Small KM, et al. Methods for the joint meta-analysis of multiple tests. *Res Synth Methods* 2014; **5**: 294–312.
16. Macaskill P, Gatsonis C, Deeks JJ, et al. Chapter 10: analysing and presenting results. In: Deeks JJ, Bossuyt PM, Gatsonis C (eds) *Cochrane handbook for systematic reviews of diagnostic test accuracy version 1.0*. The Cochrane Collaboration, 2010. Available at: http://methods.cochrane.org/sdt/sites/methods.cochrane.org.sdt/files/uploads/Chapter%2010%20-%20Version%201.0.pdf (accessed 19 July 2016).
17. Dimou NL, Adam M and Bagos PG. A multivariate method for meta-analysis and comparison of diagnostic tests. *Stat Med* 2016; **35**: 3509–3523.
18. Dukic V and Gatsonis C. Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds. *Biometrics* 2003; **59**: 936–946.
19. Hamza TH, Arends LR, van Houwelingen HC, et al. Multivariate random effects meta-analysis of diagnostic tests with multiple thresholds. *BMC Med Res Methodol* 2009; **9**: 73.
20. Riley RD, Takwoingi Y, Trikalinos T, et al. Meta-analysis of test accuracy studies with multiple and missing thresholds: a multivariate-normal model. *J Biometr Biostat* 2014; **5**: 3.
21. Martínez-Camblor P. Fully non-parametric receiver operating characteristic curve estimation for random-effects meta-analysis. *Stat Methods Med Res*. Epub ahead of print 28 May 2014. pii: 0962280214537047.
22. Picano E, Bedetti G, Varga A, et al. The comparable diagnostic accuracies of dobutamine-stress and dipyridamole-stress echocardiographies: a meta-analysis. *Coron Artery Dis* 2000; **11**: 151–159.
23. Menke J. Bivariate random-effects meta-analysis of sensitivity and specificity with SAS PROC GLIMMIX. *Methods Inf Med* 2010; **49**: 54–62, 62–64.
24. American Diabetes Association. Classification and diagnosis of diabetes. *Diabetes Care* 2015; **38**: S8–S16.
25. 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 (accessed 10 November 2015).
26. Menten J and Lesaffre E. A general framework for comparative Bayesian meta-analysis of diagnostic studies. *BMC Med Res Methodol* 2015; **15**: 70.