Author/s:
Simoneau, G;Levis, B;Cuijpers, P;Ioannidis, JPA;Patten, SB;Shrier, I;Bombardier, CH;de Lima Osório, F;Fann, JR;Gjerdingen, D;Lamers, F;Lotrakul, M;Löwe, B;Shaaban, J;Stafford, L;van Weert, HCPM;Whooley, MA;Wittkampf, KA;Yeung, AS;Thombs, BD;Benedetti, A

Title:
A comparison of bivariate, multivariate random-effects, and Poisson correlated gamma-frailty models to meta-analyze individual patient data of ordinal scale diagnostic tests

Date:
2017-11-01

Citation:
Simoneau, G., Levis, B., Cuijpers, P., Ioannidis, J. P. A., Patten, S. B., Shrier, I., Bombardier, C. H., de Lima Osório, F., Fann, J. R., Gjerdingen, D., Lamers, F., Lotrakul, M., Löwe, B., Shaaban, J., Stafford, L., van Weert, H. C. P. M., Whooley, M. A., Wittkampf, K. A., Yeung, A. S. ,... Benedetti, A. (2017). A comparison of bivariate, multivariate random-effects, and Poisson correlated gamma-frailty models to meta-analyze individual patient data of ordinal scale diagnostic tests. Biometrical Journal, 59 (6), pp.1317-1338. https://doi.org/10.1002/bimj.201600184.

Biometrical Journal **52** (2010) 61, zzz–zzz / DOI: 10.1002/bimj.200100000

A comparison of bivariate, multivariate random-effects, and Poisson correlated gamma-frailty models to meta-analyze individual patient data of ordinal scale diagnostic tests

Gabrielle Simoneau [1], **Brooke Levis** [1,2], **Pim Cuijpers** [3], **John P.A. Ioannidis** [4], **Scott B. Patten** [5], **Ian Shrier** [1,2], **Charles H. Bombardier** [6], **Flavia de Lima Osório** [7], **Jesse R. Fann** [8], **Dwenda Gjerdingen** [9], **Femke Lamers** [10], **Manote Lotrakul** [11], **Bernd Löwe** [12], **Juwita Shaaban** [13], **Lesley Stafford** [14], **Henk C.P.M. van Weert** [15], **Mary A. Whooley** [16], **Karin A. Wittkampf** [15], **Albert S. Yeung** [17], **Brett D. Thombs** [1,2,18,19], and **Andrea Benedetti*** [1,4,20]

[1] Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada, H3A 1A2

[2] Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada, H3T 1E2

[3] Department of Clinical, Neuro and Developmental Psychology, Vrije Universiteit (VU) Amsterdam, Amsterdam, The Netherlands, 1018 HV

[4] Department of Medicine, Department of Health Research and Policy, Department of Statistics, Stanford University, Stanford, California, USA, CA 94305

[5] Departments of Community Health Sciences and Psychiatry, University of Calgary, Calgary, Alberta, Canada, T2N 1N4

[6] Department of Rehabilitation Medicine, University of Washington, Seattle, Washington, USA, WA 98195

[7] Department of Neuroscience and Behavior, Faculty of Medicine of Ribeirão Preto, University of São Paulo, Ribeirão Preto, Brazil, 14049-900

[8] Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, Washington, USA, WA 98195-6560

[9] Department of Family Medicine and Community Health, University of Minnesota, Minneapolis, Minnesota, USA, MN 55455

[10] Department of Psychiatry, EMGO Institute, VU University Medical Center, Amsterdam, The Netherlands, 1081 HL

[11] Department of Psychiatry, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand, 10400

[12] Department of Psychosomatic Medicine and Psychotherapy, University Medical Center Hamburg-Eppendorf and Schön Klinik Hamburg Eilbek, Hamburg, Germany, 20246

[13] Department of Family Medicine, School of Medical Sciences, Universiti Sains Malaysia, Kelantan, Malaysia, 16150

[14] Centre for Women's Mental Health, Royal Women's Hospital, Parkville, Victoria, Australia, 3052

[15] Department of General Practice, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands, 1081 HV

[16] Department of Veterans Affairs Medical Center, San Francisco, California, USA, CA 94121

[17] Depression Clinical and Research Program, Massachussets General Hospital, Boston, Massachussets, USA, MA 02114

[18] Departments of Psychiatry, Educational and Counselling Psychology, and Psychology, McGill University, Montréal, Québec, Canada, H3A 1Y2

[19] Department of Medicine, McGill University, Montréal, Québec, Canada, H4A 3J1

[20] Respiratory Epidemiology and Clinical Research Unit, McGill University Health Centre, Montréal, Québec, Canada, H4A 3J1

*Corresponding author: e-mail: andrea.benedetti@mcgill.ca, Phone: +1-514-934-1934 ext. 32161

Received zzz, revised zzz, accepted zzz

Individual patient data (IPD) meta-analyses are increasingly common in the literature. In the context of estimating the diagnostic accuracy of ordinal or semi-continuous scale tests, sensitivity and specificity are often reported for a given threshold or a small set of thresholds, and a meta-analysis is conducted via a bivariate approach to account for their correlation. When IPD are available, sensitivity and specificity can be pooled for every possible threshold. Our objective was to compare the bivariate approach, which can be applied separately at every threshold, to two multivariate methods: the ordinal multivariate random-effects model and the Poisson correlated gamma-frailty model. Our comparison was empirical, using IPD from 13 studies that evaluated the diagnostic accuracy of the 9-item Patient Health Questionnaire depression screening tool, and included simulations. The empirical comparison showed that the implementation of the two multivariate methods is more laborious in terms of computational time and sensitivity to user-supplied values compared to the bivariate approach. Simulations showed that ignoring the within-study correlation of sensitivity and specificity across thresholds did not worsen inferences with the bivariate approach compared to the Poisson model. The ordinal approach was not suitable for simulations because the model was highly sensitive to user-supplied starting values. We tentatively recommend the bivariate approach rather than more complex multivariate methods for IPD diagnostic accuracy meta-analyses of ordinal scale tests, although the limited type of diagnostic data considered in the simulation study restricts the generalization of our findings.

Key words: Individual patient data; Meta-analysis; Multiple thresholds; Ordinal diagnostic test; Poisson correlated frailty

Supporting Information for this article is available from the corresponding author.

1 Introduction

Diagnostic and screening tests are used to attempt to distinguish between diseased and healthy patients with the true disease status being determined by a gold standard. Diagnostic accuracy studies evaluate the performance of a test with respect to its classification ability. While other measures are also used (Eusebi, 2013), the probability of correctly identifying diseased patients (sensitivity) and healthy patients (specificity) are most commonly used to quantify diagnostic accuracy.

Conventional meta-analyses of diagnostic accuracy have traditionally pooled only one pair of sensitivity and specificity estimates across studies. For ordinal-scale diagnostic tests, where test results fall in multiple, ordered categories, different thresholds may be explicitly defined to classify a result as positive or negative. In this situation, conventional meta-analyses have typically focused on one threshold of interest and produced summary results for that threshold. Two statistically rigorous methods are commonly used in practice for conventional meta-analyses of diagnostic accuracy: the bivariate random-effects model (Chu and Cole, 2006; Reitsma *et al.*, 2005) and its Bayesian counterpart, the hierarchical summary receiver operating characteristic model (HSROC) (Rutter and Gatsonis, 2001). Both methods reflect two important characteristics of such meta-analyses. First, the correlation between sensitivity and specificity across studies is accounted for by pooling the two measures simultaneously (Riley, 2009; Moses *et al.*, 1993; Rutter and Gatsonis, 2001). This correlation arises explicitly when primary studies use a different threshold to define positive and negative test results. If the only source of variability between studies was the chosen threshold, this correlation would be negative because, as the threshold used to identify a likely case becomes stricter, the sensitivity of the diagnostic test decreases while the specificity increases (Moses *et al.*, 1993; Reitsma *et al.*, 2005). Second, heterogeneity between primary studies is to be expected (e.g. through differences in equipment, measurements and population characteristics), and is accounted for by using a mixed model approach or a hierarchical framework. Although it has been demonstrated that the two methods are equivalent in many circumstances (Harbord *et al.*, 2007), they reflect different inferential approaches. While the bivariate random-effects model estimates summary measures of sensitivity and specificity, the HSROC model suggests estimating a summary receiver operating characteristic (sROC)

curve. A receiver operating characteristic (ROC) curve is a plot of all pairs of sensitivity and specificity derived from every possible threshold. The sROC curve has the advantage of describing the overall diagnostic accuracy of a test at the cost of making several parametric assumptions on the ROC curve's shape. Yet, in the absence of covariates, a simple re-parameterization provides an sROC curve from the bivariate approach.

As diagnostic studies of ordinal-scale tests typically report pairs of sensitivity and specificity for more than one threshold, it may be more interesting from a clinical perspective to obtain summary diagnostic accuracy results for all published thresholds (Riley *et al.*, 2014; Dukic and Gatsonis, 2003; Martínez-Camblor, 2014; Steinhauser *et al.*, 2016). A simple approach would be to meta-analyze each threshold separately with a conventional meta-analysis to produce pooled sensitivity and specificity estimates for all published thresholds. However, diagnostic studies are prone to selective reporting of thresholds, where thresholds that perform better within a given dataset are more likely to be published (Levis *et al.*, in press). It has been shown that selective reporting of thresholds biases conventional meta-analyses by exaggerating the accuracy for some thresholds (Levis *et al.*, in press). In addition to numerous other advantages (Riley *et al.*, 2010), using individual patient data (IPD) instead of aggregated data can address the problem of selective cutoff reporting since IPD meta-analyses can include results from primary studies from both published and unpublished thresholds. When IPD are available, sensitivity and specificity can be estimated for all possible thresholds in each study, and an IPD meta-analytic method can then take advantage of the available information to produce pairs of pooled sensitivity and specificity over the entire range of thresholds.

One way to analyze IPD for diagnostic test accuracy is to meta-analyze each threshold separately by applying a conventional meta-analysis as many times as there are thresholds in an ordinal-scale test. However, this approach ignores the within-study correlation between sensitivities and specificities across thresholds, where the within-study correlation now arises because each study carries information not only for one pair of sensitivity and specificity, but for as many pairs as the total number of thresholds considered. Alternatively, a multivariate meta-analytic method can be applied to all thresholds simultaneously, and thus correctly account for data dependencies. Several methods have been proposed to accommodate this complex framework (Hamza *et al.*, 2009; Putter *et al.*, 2010; Dukic and Gatsonis, 2003; Martínez-Camblor, 2014). The multivariate random-effects model (Hamza *et al.*, 2009) suggests estimating a parametric ROC curve using all thresholds simultaneously. A similar method (Dukic and Gatsonis, 2003) also models a summary ROC curve within a Bayesian framework. Departing from these approaches, a Poisson correlated gamma-frailty model (Putter *et al.*, 2010), inspired by a method designed to meta-analyze heterogeneous survival curves, can be adapted to meta-analyze all thresholds of a diagnostic test simultaneously without explicitly imposing constraints to the shape of the underlying ROC curve. Potential advantages of multivariate IPD meta-analyses over conventional approaches to IPD data mainly concern the validity of the inferences since they can accommodate the complex correlation structure arising from the data. The multivariate approach results in more precise estimates of the pooled effects of interest as the method utilizes additional information from the correlated effects, a concept known as borrowing of strength (Riley, 2009; Jackson *et al.*, 2015). However, this advantage can in turn be a disadvantage as more complex modelling techniques require more assumptions about the form of the underlying summary ROC curve and about the correlation structure.

The objective of this study was to compare two statistically rigorous multivariate methods to the conventional bivariate method for IPD meta-analyses of diagnostic accuracy, empirically and via simulations. This work aimed to investigate whether accounting for the within-study correlation across thresholds by using multivariate approaches noticeably improved the validity of the inferences for the pooled sensitivities and specificities over the range of thresholds. In other words, do more complex multivariate methods significantly outperform the simpler yet less theoretically appropriate bivariate method in terms of inferences for the pooled parameters? We used IPD data from a recently completed IPD meta-analysis (Levis *et al.*, in press) and simulated data to address this question. The empirical comparison focused on the applicability of the methods, on the strength of each method in dealing with the complex correlation structure arising

from the data, and on the concordance of the results. The simulations focused on factors that influence the inferences of the methods and on the estimation of the correlation structure.

2 Methods

2.1 Notations

Let i identify a primary study included in the IPD meta-analysis, $i = 1, ..., m$. Let $D = d$ denote the true disease status of a patient as determined by a gold standard test, where "0" stands for healthy and "1" for diseased. We considered diagnostic tests with $J + 1$ ordered categories, where lower categories provided less evidence of the disease. Let Y denote the outcome of the diagnostic test and x_{ij}^d be the number of patients in study i with disease status d with test result falling in category j, $j = 0, ..., J$. Denote $n_i^d = \sum_{k=0}^{J} x_{ik}^d$ as the total number of truly healthy ($d = 0$) and truly diseased ($d = 1$) patients in study i.

In this situation, J thresholds were used to classify a test result as positive or negative. For each threshold j, $j = 1, ..., J$, the sensitivity of a test was defined by $\text{P}(Y \geq j|D = 1)$ and was estimated by $\sum_{k=j}^{J} x_{ik}^1/n_i^1$ within each study i. Similarly, for each threshold j, specificity was defined by $\text{P}(Y < j|D = 0)$ and was estimated by $\sum_{k=0}^{j-1} x_{ik}^0/n_i^0$ within each study i. To match with the notations used in previous methodological papers, we worked with (1-specificity) defined by $\text{P}(Y \geq j|D = 0)$.

2.2 Bivariate random-effects model

The bivariate random-effects model (BREM) (Chu and Cole, 2006; Reitsma *et al.*, 2005; Van Houwelingen *et al.*, 1993) is commonly used for conventional meta-analyses of diagnostic accuracy (Jackson *et al.*, 2011). It meta-analyzes sensitivity and specificity simultaneously for one selected threshold, and thus correctly accounts for the across-study heterogeneity between sensitivity and specificity. With IPD available, the bivariate model can be applied to all thresholds separately to produce pairs of pooled sensitivity and specificity over the range of thresholds. While the model accounts for the correlation between sensitivity and specificity across studies for each threshold, it does not account for the within-study correlation of sensitivity and specificity across thresholds.

The bivariate model exploits the framework of the generalized linear mixed model (Chu and Cole, 2006). For a fixed threshold T, denote θ_{iT}^1 and θ_{iT}^0 as the true (unobservable) study-specific sensitivity and (1-specificity) in study i. Conditional on the random effects u_{iT}^1 and u_{iT}^0, the bivariate model specifies two independent within study models as

$$\begin{aligned} \text{TP}_{iT} &\sim \text{Binomial}\left(n_i^1, \theta_{iT}^1\right), \\ \text{FP}_{iT} &\sim \text{Binomial}\left(n_i^0, \theta_{iT}^0\right), \end{aligned} \tag{1}$$

where $\text{TP}_{iT} = \sum_{j=T}^{J} x_{ij}^1$ and $\text{FP}_{iT} = \sum_{j=T}^{J} x_{ij}^0$ are the number of true positive patients and the number of false positive patients in study i when using threshold T, respectively. Using the canonical logit link, we have

$$\begin{aligned} \text{logit}\left(\mathbb{E}\left(\frac{\text{TP}_{iT}}{n_i^1}\middle| u_{iT}^1\right)\right) &= \theta_{iT}^1 = \bar{\theta}_T^1 + u_{iT}^1, \\ \text{logit}\left(\mathbb{E}\left(\frac{\text{FP}_{iT}}{n_i^0}\middle| u_{iT}^0\right)\right) &= \theta_{iT}^0 = \bar{\theta}_T^0 + u_{iT}^0, \end{aligned}$$

with $\bar{\theta}_T^1$ and $\bar{\theta}_T^0$ being the targeted pooled logit sensitivity and (1-specificity) for threshold T, respectively, where $\text{logit}(x) = \log(x) - \log(1 - x)$. Assuming a bivariate normal distribution for the random effects as

$$\begin{pmatrix} u_{iT}^1 \\ u_{iT}^0 \end{pmatrix} \backsim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{1T}^2 & \rho_T\sigma_{1T}\sigma_{0T} \\ \rho_T\sigma_{1T}\sigma_{0T} & \sigma_{0T}^2 \end{pmatrix}\right),$$

the correlation between sensitivity and specificity for threshold T is estimated by $-\hat{\rho}_T$, for T=1, ..., J. To estimate the parameters of interest $\bar{\theta}_T^1$, $\bar{\theta}_T^0$, σ_{1T}^2, σ_{0T}^2 and ρ_T, the likelihood of the model is approximated by adaptive Gauss-Hermite quadrature (Zhang *et al.*, 2011; Hamza *et al.*, 2009). In R (R Core Team, 2013), the adaptive Gaussian method needs to be carried out with one quadrature point as the dimension of the random-effect parameters is greater than one, which is equivalent to the Laplace approximation. Estimation can be carried out using the function `glmer()` from the `lme4` package (Bates *et al.*) in R or the `PROC NLMIXED` procedure in SAS software (SAS Institute Inc., 2003).

2.3 Ordinal multivariate random-effects model

The ordinal multivariate random-effects model (ordinal model) (Hamza *et al.*, 2009) analyzes pairs of sensitivity and specificity simultaneously for multiple published thresholds, and thus perfectly adapts to the situation where IPD are available. This model is a direct extension of the bivariate approach. It accounts for both the correlation between sensitivity and specificity at each threshold and the correlation of sensitivities and specificities across thresholds.

The ordinal model assumes linearity of sensitivity and (1-specificity) on the logit-scale. The true study-specific logit-transformed sensitivity η_{ij} and (1-specificity) ξ_{ij} for threshold j are modelled as

$$\eta_{ij} = \alpha + u_{\alpha i} + \beta\xi_{ij}, \tag{2}$$

$$\xi_{ij} = \bar{\xi}_j + \Delta_i + \delta_{ij}, \tag{3}$$

where α and β are fixed effects intercept and slope parameters, $u_{\alpha i} \backsim \mathrm{N}(0, \sigma_\alpha^2)$ is a random intercept term, $\bar{\xi}_j$ is the targeted pooled logit-transformed (1-specificity) for threshold j, $\Delta_i \backsim \mathrm{N}(0, \sigma_\Delta^2)$ is a study-level random effect and $\delta_{ij} \backsim \mathrm{N}(0, \sigma_\delta^2)$ is a study- and threshold-specific random effect. The model assumes that the δ_{ij}'s are independent of $u_{\alpha i}$ and Δ_i, and the covariance between $u_{\alpha i}$ and Δ_i is denoted by $\sigma_{\alpha\Delta}$. We note that Equation (2) defines study-specific parametric ROC curves such that we can derive a parametric smooth pooled ROC curve from the model. A random slope effect may also be added to the model in (2) (Hamza *et al.*, 2009).

Previous distributional assumptions yield a multivariate normal distribution for the random effects with a compound symmetric covariance structure as

$$\begin{pmatrix} u_{i\alpha} \\ \Delta_i + \delta_{i1} \\ \Delta_i + \delta_{i2} \\ \vdots \\ \Delta_i + \delta_{i,J-1} \end{pmatrix} \backsim N\left(\mathbf{0}, \begin{pmatrix} \sigma_\alpha^2 & \sigma_{\alpha\Delta} & \cdots & \cdots & \sigma_{\alpha\Delta} \\ \sigma_{\alpha\Delta} & \sigma_\Delta^2 + \sigma_\delta^2 & \sigma_\Delta^2 & \cdots & \sigma_\Delta^2 \\ \vdots & \sigma_\Delta^2 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \sigma_\Delta^2 \\ \sigma_{\alpha\Delta} & \sigma_\Delta^2 & \cdots & \sigma_\Delta^2 & \sigma_\Delta^2 + \sigma_\delta^2 \end{pmatrix}\right), \tag{4}$$

where $\mathbf{0}$ denotes a vector of zeros of length J. Given $(\alpha, \beta, u_{\alpha i}, \Delta_i, \delta_{ij})$, the observed number of patients $\boldsymbol{x}_i^d = (x_{i0}^d, ..., x_{iJ}^d)$ with disease status d follows two independent multinomial distributions, $\boldsymbol{x}_i^d \backsim \mathrm{Multi}(n_i^d, \boldsymbol{\pi}_i^d)$, with $\boldsymbol{\pi}_i^d = (\pi_{i0}^d, ..., \pi_{iJ}^d)$ defined as

$$\pi_{ij}^1 = \mathrm{expit}(\eta_{ij}) - \mathrm{expit}(\eta_{i,j-1}),$$
$$\pi_{ij}^0 = \mathrm{expit}(\xi_{ij}) - \mathrm{expit}(\xi_{i,j-1}),$$

for $j = 1, .., J$, where $\mathrm{expit}(x) = \mathrm{logit}^{-1}(x)$. The proportional odds logit model is used to link the probability parameters $\boldsymbol{\pi}_i^0$ to the linear predictor as shown in Equation (3). The pooled logit sensitivity for threshold j, $\bar{\eta}_j$, is derived using Equation (2) as

$$\bar{\eta}_j = \alpha + \beta\bar{\xi}_j. \tag{5}$$

The ordinal model estimates the correlation between neighboring logit (1-specificities) within study i as

$$\text{corr}(\xi_{ik}, \xi_{il}) = \sigma^2_\Delta/(\sigma^2_\Delta + \sigma^2_\delta), \tag{6}$$

for $k, l = 1, \ldots, J$, $k \neq l$, which is directly derived from (4). Similarly, it estimates the correlation between neighboring logit sensitivities within study i as

$$\text{corr}(\eta_{ik}, \eta_{il}) = \frac{\text{cov}(\eta_{ik}, \eta_{il})}{\sqrt{\text{var}(\eta_{ik})}\sqrt{\text{var}(\eta_{il})}} = \frac{\sigma^2_\alpha + 2\beta\sigma_{\alpha\Delta} + \beta^2\sigma^2_\Delta}{\sigma^2_\alpha + \beta^2(\sigma^2_\Delta + \sigma^2_\delta) + 2\beta\sigma_{\alpha\Delta}} \tag{7}$$

for $k, l = 1, \ldots, J$, $k \neq l$, which is derived from (2), (3) and (4). The ordinal model also estimates the correlation between logit sensitivity and logit (1-specificity) at a given threshold k in study i as

$$\text{corr}(\eta_{ik}, \xi_{ik}) = \frac{\text{cov}(\eta_{ik}, \xi_{ik})}{\sqrt{\text{var}(\eta_{ik})}\sqrt{\text{var}(\xi_{ik})}} = \frac{\sigma_{\alpha\Delta} + \beta(\sigma^2_\Delta + \sigma^2_\delta)}{\sqrt{\sigma^2_\alpha + \beta^2(\sigma^2_\Delta + \sigma^2_\delta) + 2\beta\sigma_{\alpha\Delta}}\sqrt{\sigma^2_\Delta + \sigma^2_\delta}}, \tag{8}$$

for $k = 1, \ldots, J$. SAS software can be used to estimate the model's parameters by approximating the log-likelihood through an adaptive Gauss-Hermite quadrature with five quadrature points using `PROC NLMIXED` (Hamza *et al.*, 2009).

2.4 Poisson correlated gamma-frailty model

The Poisson correlated gamma-frailty model (Poisson model) was first introduced in the context of meta-analyses of heterogeneous survival curves (Fiocco *et al.*, 2009a), and has been adapted to meta-analyze diagnostic accuracy studies with multiple thresholds (Putter *et al.*, 2010). This multivariate model correctly accounts for the correlation between sensitivity and specificity across studies at each threshold, and for their correlation within-study across thresholds.

Define $\text{P}^d(Y \geq j) = \text{P}(Y \geq j \mid D = d)$. Generally, sensitivity and (1-specificity) can be expressed in terms of survival probabilities as

$$\begin{aligned}\text{P}^d(Y \geq j) &= \text{P}^d(Y \geq 0)\text{P}^d(Y \geq 1 \mid Y \geq 0) \ldots \text{P}^d(Y \geq j \mid Y \geq j-1) \\ &= 1(1-\lambda^d_1)\ldots(1-\lambda^d_j) = \prod_{k=1}^{j}(1-\lambda^d_k).\end{aligned}$$

respectively for $d = 1$ and $d = 0$, where $\lambda^d_k = \text{P}(Y \geq k \mid Y \geq j-1 \cap D = d)$ are discrete hazards. Define λ^d_{ij} as the study-specific hazards. Modelling of λ^d_{ij} is straightforward using the framework of survival analysis. Define r^d_{ij} as the number of patients with disease status d in study i for which $Y \geq j-1$ and y^d_{ij} as the number of patients with $D = d$ in study i for which the test result falls in category $j-1$. While the two previous methods model y^d_{ij} as binomial counts, the Poisson model specifies $y^d_{ij} \backsim \text{Poi}(\lambda^d_{ij} r^d_{ij})$ where r^d_{ij} can be thought as the number of "person-time at risk before time j" and y^d_{ij} as the number of events occurring at "time j". It is reasonable to associate the notion of time in survival analysis to the ordered categories in diagnostic accuracy studies.

Between- and within-study correlations are introduced in the previous framework by incorporating correlated frailties Z^d_j to the specification of the hazards λ^d_{ij}. Define $\lambda^d_{ij} = \lambda^d_j \text{Z}^d_{ij}$, where λ^d_j is the target pooled hazard for threshold j, $j = 1, ..., J$, $d = 0, 1$. A multivariate gamma distribution models the frailties $(\text{Z}^1_{i1}, \ldots, \text{Z}^1_{iJ}, \text{Z}^0_{i1}, \ldots, \text{Z}^0_{iJ})$, specified as

$$\text{E}(\text{Z}^d_j) = 1, \;\; \text{var}(\text{Z}^d_j) = \xi^d, \;\; \text{corr}(\text{Z}^1_j, \text{Z}^0_j) = \rho_{\text{dis}}, \;\; \text{corr}(\text{Z}^d_j, \text{Z}^d_k) = \rho^{|j-k|}_{\text{thres}}. \tag{9}$$

Details on the construction of this multivariate gamma distribution have been thoroughly discussed elsewhere (Putter *et al.*, 2010; Fiocco *et al.*, 2009a,b).

The frailties entirely express the variability and correlation of sensitivities and (1-specificities) over the range of thresholds. The correlation parameter ρ_{thres} characterizes a first order auto-regressive correlation structure to model sensitivities (likewise for specificities) across thresholds such that sensitivities for neighboring thresholds are more correlated than sensitivities from distant thresholds. The parameter ρ_{dis} specifies the correlation between sensitivity and (1-specificity) at any given threshold. By construction of the gamma distribution, it follows that $\text{corr}(\text{Z}_j^1, \text{Z}_k^0) = \rho_{\text{dis}}\rho_{\text{thres}}^{|j-k|}$ specifies the correlation between sensitivity and (1-specificity) for distinct threshold values.

Conditional on the frailties $(\text{Z}_{i1}^1, \ldots, \text{Z}_{iJ}^1, \text{Z}_{i1}^0, \ldots, \text{Z}_{iJ}^0)$, the number of patients Y_{ij}^d with disease status d and test result falling in category $j-1$ in study i is modelled as

$$Y_{ij}^d \mid \text{Z}_{ij}^d \backsim \text{Poi}(\mu_{ij}^d \text{Z}_{ij}^d), \text{ with } \mu_{ij}^d = \lambda_j^d r_{ij}^d,$$

for $d = 0, 1$, where r_{ij}^d is the number of patients with disease status d and $Y \geq j-1$, in study i. Estimation of the hazards λ_j^d and of the correlation parameters can be carried out with a two-stage approach, using composite likelihood (Lindsay, 1988). In the first stage, the fact that the margins Y_{ij}^d have negative binomial distributions with shape μ_{ij}^d and scale $1/\xi^d$ is used to estimate the parameters λ_j^d and the variances ξ^d. In the second stage, maximum likelihood estimates of ρ_{dis} and ρ_{thres} can be obtained (Putter *et al.*, 2010; Fiocco *et al.*, 2009b,a), using the estimates of λ_j^ds, ξ^1 and ξ^0 from the first stage. A parametric bootstrap (Wehrens *et al.*, 2000; Fiocco *et al.*, 2009b) is used to recover sensible standard errors for the estimated parameters and the corresponding sensitivities and specificities.

The negative binomial distributions in the first stage can be fitted using `glm.nb()` from the `MASS` package (Venables and Ripley, 2002) in R, with a log-link and $\log(r_{ij}^d)$ added as an offset. Maximization of the log-likelihoods in the second stage can be carried out with `optimize`, also in R.

3 Application to the 9-item Patient Health Questionnaire

The 9-item Patient Health Questionnaire (PHQ-9) (Kroenke and Spitzer, 2002) is a screening tool for major depressive disorder (MDD). Test scores range from 0 to 27, and higher scores indicate more severe symptoms of depression. A score of 10 or greater has been recommended as the threshold for identifying probable depression (Kroenke *et al.*, 2001; Wittkampf *et al.*, 2007; Gilbody *et al.*, 2007b; Kroenke and Spitzer, 2002; Spitzer *et al.*, 1999). Despite this, a recent conventional meta-analysis found that various thresholds were reported in the literature for the diagnostic accuracy of this test (Manea *et al.*, 2012).

For the present study, we used IPD data from 13 of 16 primary studies included in a recently published conventional meta-analysis of diagnostic accuracy of the PHQ-9 questionnaire (Manea *et al.*, 2012; Levis *et al.*, in press). Studies were eligible for the original meta-analysis (Manea *et al.*, 2012) if they (1) defined MDD according to standard classification systems; (2) used a validated diagnostic interview for MDD as the reference standard; and (3) provided sufficient data to calculate 2x2 contingency tables. In this work, we focused on thresholds 7 to 14 of the PHQ-9 questionnaire. A standard weighting procedure was used in the analysis for studies where sampling procedures were used, for instance, when only a random subset of patients with negative screens was administered a diagnostic interview (Levis *et al.*, in press). The bivariate random-effects model, the ordinal multivariate random-effects model and the Poisson correlated gamma-frailty model were applied to the PHQ-9 IPD dataset for this subset of thresholds.

Table 1 summarizes the IPD dataset for thresholds 7 to 14 of the PHQ-9 screening tool. The sample sizes varied considerably across studies, ranging between 96 and 1024, and the number of cases of depression within each study was always small relative to the number of non-cases. Figure 1 (A) shows the empirical ROC curve for each of the 13 studies. The study-specific curves varied over a wide range, suggesting that there was substantial heterogeneity across studies.

Table 2 shows the estimated pooled sensitivities and specificities from the three methods, along with 95% confidence intervals. Figure 1 (B) shows the pooled ROC curve for thresholds 7 to 14 obtained with each method, along with confidence bands. Overall, the BREM and the ordinal model were generally comparable whereas the Poisson model was often significantly different, especially for sensitivity estimates. Estimates of sensitivities, ranging between 0.50 and 0.97 across the 8 thresholds, were different across the three methods for each threshold. The Poisson model systematically estimated lower sensitivities for each threshold in comparison to the ordinal and bivariate models, whereas the bivariate model produced the highest estimates of sensitivity. For example, with threshold 9, sensitivity was estimated as 0.89 (95% CI: 0.79-0.95), 0.88 (95% CI: 0.80-0.93) and 0.80 (95% CI: 0.73-0.87) by the BREM, the ordinal model and the Poisson approach, respectively. Estimates of specificities were more similar across the three methods, ranging from 0.73 to 0.96 across the studied thresholds. In terms of precision, the ordinal approach systematically produced tighter confidence intervals than the two other models, which was due to the parametric assumptions in equations (2) and (3) (Putter *et al.*, 2010). While the difference in confidence interval widths was not important for specificities, the width of the confidence intervals for sensitivity was 11% and 5% smaller, on average, with the ordinal approach than with the BREM and the Poisson model, respectively. For sensitivities, confidence intervals produced by the Poisson approach were 5% tighter, on average, than those produced by the BREM. The opposite relationship was observed for specificities, where confidence intervals produced by the Poisson model were wider than those produced by the BREM by a factor of 30%.

Table 3 and Table 4 show the parameters estimated by the bivariate approach and the ordinal model, respectively. Pooled sensitivity and (1-specificity) were estimated on the logit-scale for thresholds 7 to 14. The standard errors of the pooled logit sensitivities across thresholds were more stable with the ordinal model, ranging between 0.29 and 0.30, than with the bivariate approach, for which the standard errors ranged between 0.23 and 0.72. With the bivariate model, as the threshold value increased, the standard errors of the logit sensitivity decreased. The standard errors of the pooled logit (1-specificities) across thresholds were very stable with both methods, ranging between 0.12 and 0.15.

Table 5 shows the parameter estimates from the Poisson model. For diseased patients, estimated hazards increased with the threshold value, roughly meaning that if a truly diseased patient has not failed at threshold j, his chance of failing at threshold $j + 1$ increased, since higher scores indicated more severe symptoms of depression. For healthy patients, hazard estimates were stable across the set of thresholds considered, as truly healthy patients were more likely to fail at lower thresholds, since smaller scores indicated less severe symptoms of depression.

The three models explicitly assumed different correlation structures to characterize the relation of sensitivities (and specificities) across thresholds. The BREM assumed a correlation of zero by meta-analyzing all thresholds separately. Following (6) and (7), the ordinal model estimated the correlation of sensitivities (and specificities) across thresholds, for any pair of thresholds, as 0.99 (0.96), which appeared unusually high. The Poisson model characterized the correlation between sensitivities (and specificities) across thresholds via an autoregressive correlation structure, with the autoregressive correlation parameter estimated as 0.74 (SE: 0.11) from Table 5. This meant that the correlation between neighboring thresholds was 0.74, and that this correlation decreased as thresholds were further apart.

The three models also considered the correlation between sensitivity and specificity across studies at each threshold. For each threshold T, the BREM estimated the correlation between sensitivity and specificity with ρ_{T} in Table 3. These correlations were moderately high, ranging from 0.27 to 0.77. Following (8), the ordinal model estimated this correlation as 0.55 for any given threshold, which was consistent with the correlation estimates from BREM. The Poisson model estimated this correlation as 0.39 (SE: 0.17).

While the BREM was readily applicable in terms of data manipulation and computational time, the implementation of the Poisson model and the ordinal model was not straightforward. The meta-analysis of the PHQ-9 IPD dataset with the BREM took 1.85 seconds on a 2.9 GHz Core i7 MacBook Pro with 8 GB of RAM. The estimation of the Poisson parameters took 126 seconds on the same machine, with 110 seconds taken for the estimation of the correlation parameters ρ_{thres} and ρ_{dis}. This computational time

did not even include the estimation of the standard errors via parametric bootstrap, which took over 15 h with 500 bootstrap samples on the same machine. The meta-analysis of the PHQ-9 dataset with the ordinal model was even more computationally intensive, taking over 48 h on a similar Windows PC machine. The ordinal model was very sensitive to user-supplied starting values for the curve parameters α and β in (2), the pooled logit (1-specificities) $\bar{\xi}_j$ in (3) and the variance and covariance parameters σ^2_α, σ^2_Δ, σ^2_δ and $\sigma_{\alpha\Delta}$ in (4). With the PHQ-9 meta-analysis, the model failed to converge when the set of supplied starting values was 0.5 unit below or above the final estimated values presented in Table 4 (see SAS example in Supplementary Material).

4 Simulations

As we highlighted in the empirical comparison, the ordinal model was very sensitive to user-supplied starting values, which may explain why Putter *et al.* encountered convergence failures (Putter *et al.*, 2010). For this reason, the ordinal model was considered unsuitable for our simulation studies. We therefore present results only for the BREM and the Poisson model.

Data were simulated, roughly mimicking the PHQ-9 data meta-analysis. We considered a meta-analysis of 13 independent studies of a 9-category diagnostic ordinal-scale test, corresponding to 8 meaningful thresholds. We generated correlated sensitivities and specificities following a data generating mechanism used in Putter *et al.* (Putter *et al.*, 2010). For the 8 thresholds, we set the overall sensitivity/specificity to 0.94/0.74, 0.91/0.79, 0.88/0.83, 0.84/0.87, 0.79/0.89, 0.74/0.91, 0.67/0.93 and 0.57/0.95, respectively labeled as $1, \ldots, 8$, following the pooled sensitivities and specificities found with the PHQ-9 IPD dataset. After logit-transformation of the overall sensitivities and (1-specificities), we added random noises to obtain study-specific sensitivities and specificities for each threshold and for each study. The random noises were simulated from a multivariate normal distribution centered around zero (as in Putter *et al.*, 2010), with correlation structure defined as in (9). In plain words, this covariance matrix defined decreasingly correlated sensitivities (specificities) across thresholds and equally correlated sensitivity and specificity across studies for identical thresholds. Referring to (9), we varied the strength of the correlation by setting $\rho_{\text{dis}}/\rho_{\text{thres}}$ to 0.25/0.25, 0.25/0.75, 0.75/0.25 and 0.75/0.75. We also tuned the variability of the frailties by setting ξ^1/ξ^0 to 0.1/0.05 and 0.25/0.1, which simulated more or less heterogenous sensitivities and specificities across studies. The number of cases within each study n_{i1} was simulated from a Normal distribution centered around n_1 with a standard deviation of sd_1. The parameters n_1/sd_1 were set to 50/40 and 100/80. The number of non-cases n_{i0} was simulated from a Normal distribution centered around 400 with a standard deviation of 150. The parameters of these Normal distributions were chosen to mimic the variation in sample sizes found in the PHQ-9 IPD dataset. Within each study, the diagnostic test data were generated as realizations of two multinomial distributions, for the cases and the non-cases, where the probabilities were derived from the simulated study-specific sensitivities and specificities (Putter *et al.*, 2010). All data simulation steps were performed in R, using the `mvtnorm` (Genz *et al.*, 2014) and `copula` (Hofert *et al.*, 2016) packages.

We characterized the impact of each data-generating parameter on inferences of the pooled sensitivity and specificity using an approach proposed by Chipman *et al.* (Chipman *et al.*, 2015). In the context of this simulation study, the measured outcomes of interest were the bias and root mean squared error (RMSE) of the estimated pooled sensitivity and specificity for each threshold, as well as the coverage of the derived confidence intervals. The factors investigated were: the threshold value (8-level factor), the strength of the correlation across thresholds ρ_{thres} (2-level factor, 0.25 labeled as "low", 0.75 labeled as "high"), the strength of the correlation between sensitivity and (1-specificity) across studies at a given threshold ρ_{dis} (2-level factor, 0.25 labeled as "low", 0.75 labeled as "high"), the size of the between-study heterogeneity through the variability of the frailties ξ^1/ξ^0 (2-level factor, 0.1/0.05 labeled as "low", 0.25/0.1 labeled as "high") and the sample size of cases (2-level factor, $n_1 = 50$ and $n_1 = 100$). These factors yielded to $2 \times 2 \times 2 \times 2 = 16$ simulation scenarios, each considering the 8 possible thresholds. For each scenario,

we generated 1,000 datasets, thus resulting in a total of 16,000 replications.

Figure 2 shows the effect of each factor on the mean bias and RMSE of logit sensitivity (left column) and logit specificity (right column). The BREM (black lines) and the Poisson model (light grey lines) were compared. The threshold value had the largest influence on the magnitude of the bias. Sensitivities for thresholds 4 and 5, corresponding to true overall sensitivities of 0.79 and 0.84, respectively, were estimated with less bias than for smaller or larger thresholds, which corresponded to true overall sensitivities closer to 1 and 0.5, respectively. For example, thresholds 1 and 8, corresponding to true sensitivities of 0.94 and 0.57, respectively, were estimated with more bias than all other thresholds for both the BREM and the Poisson model (top-left figure). This threshold effect was less dramatic for specificities. Overall, the magnitude of the bias induced by the threshold value was very similar between the BREM and the Poisson model. The amount of between-study heterogeneity had a comparable impact on the magnitude of the bias of sensitivity and specificity: as expected, with both methods, sensitivity and specificity were estimated with more bias when underlying sensitivities and specificities were more heterogeneous (*high*) across studies. For both the BREM and the Poisson approach, the correlation between sensitivity and specificity at each threshold ρ_{dis} as well as the correlation of sensitivity (specificity) across thresholds ρ_{thres} had a similar impact on the estimates of sensitivity and specificity. For example, all scenarios with ρ_{dis} set to either 0.25 or 0.75 estimated sensitivity with a higher RMSE (bottom-left figure) compared to estimates of specificity (bottom-right figure). Nevertheless, both models failed to estimate unbiased sensitivity and specificity with either low or high correlations ρ_{dis} and ρ_{thres}. The effect of the sample size of diseased subjects was similar between the BREM and the Poisson model for the two metrics considered.

Figure 3 shows the effect of each factor of the mean coverage of the derived confidence intervals for logit sensitivity (left plot) and logit specificity (right plot). For sensitivity, the Poisson model exhibited coverage rates closer to the 95% nominal rate than the BREM. Yet, both methods exhibited under-coverage compared to the 95% nominal rate, where the mean coverage varied between 60% and 96%. As for the previous analysis, the threshold value had the largest influence on the coverage. The mean coverage for sensitivity at threshold 8, corresponding to true overall sensitivity of 0.57, was approximately 60% with the BREM and 70% with the Poisson model. The amount of between-study heterogeneity had a comparable impact on the resulting coverage rates with the two methods: the mean coverage rates were closer to the nominal rate when underlying sensitivities were less heterogeneous (*low*) across studies. The strength of the correlation at or across thresholds and the sample size of diseased subjects had a modest effect on the coverage rates. For specificity, confidence intervals estimated with the Poisson model were extremely conservative, with the mean coverage above 99% for any levels of each factor. The BREM still exhibited under-coverage with specificities, where the effect of each factor was similar on sensitivity and specificity.

Figure 4 further explores the relation between the threshold value (i.e. the associated true sensitivity and specificity) and the mean bias in sensitivity and specificity. The two methods produced more accurate estimates of sensitivity and specificity when the true values were in the range $0.80 - 0.90$. Both methods overestimated sensitivity when the true sensitivity was close to 1 and underestimated sensitivity when the true sensitivity was smaller than 0.80. Similarly, both methods underestimated specificity when the true specificity was smaller than 0.90, but slightly overestimated specificity when the true value was close to 1. Overall, the mean bias in sensitivity was larger with the BREM compared to the Poisson model for all threshold values while the mean bias in specificity was larger with the Poisson model compared to the BREM.

Our simulation study also investigated the ability of the two methods to estimate the correlation between sensitivity and specificity across thresholds, and at each threshold. Because the estimation of the correlation parameters ρ_{thres} and ρ_{dis} was very computationally intensive with the Poisson model, we focused our comparison on one chosen scenario. Table 6 shows results of this comparison. On average, the BREM estimated the correlation of sensitivity and (1-specificity) across studies close to the true value 0.25 with most thresholds. However, the BREM model failed to estimate sensible correlations (failure rate) more often as true sensitivity decreased (or true specificity increased). Specifically, a failure occurred when the model estimated the variance of the specificity random effect by zero, such that the corresponding correlation

estimate was not a number. The failure of the model did have an impact on the estimates of sensitivity and specificity. Moreover, for each threshold, approximately 50% of the correlation parameters were estimated between -1 and -0.99 or 0.99 and 1, the extremum of the spectrum of possible values for the correlation parameter. The Poisson model underestimated the correlation between sensitivity and (1-specificity) at each threshold denoted by ρ_{dis}. We were unable to correctly estimate the correlation parameter ρ_{thres} in all simulated datasets. In fact, maximization of the second stage log-likelihood as mentioned in §2.4 always estimated ρ_{thres} with a value around 0.001, which corresponded to the lower bound of the optimization interval.

5 Discussion

Methods to meta-analyze IPD for diagnostic accuracy of ordinal or quasi-continuous scale tests are not well-established. The focus has often been on methods to meta-analyze published results from diagnostic accuracy studies which report a pair of sensitivity and specificity for one or more thresholds, but not necessarily for the same set of thresholds across studies (Riley *et al.*, 2014; Dukic and Gatsonis, 2003). However, it was recently shown that relying on such published results can produce biased estimates due to selective cutoff reporting, which can be addressed using IPD meta-analysis (Levis *et al.*, in press). Our comparison, empirical and via simulation studies, focused on three statistically rigorous methods to meta-analyze IPD of diagnostic accuracy studies.

The application to the PHQ-9 IPD dataset focused on the analysis on a subset of clinically relevant thresholds. This choice was motivated by previous meta-analyses of the PHQ-9 (Levis *et al.*, in press; Manea *et al.*, 2012) which included the standard cutoff 10 and further extended their analysis to the subset of clinically relevant thresholds 7 to 15. However, threshold 15 could not be included in our empirical comparison because of the computational complexity of the ordinal model, which was in part attributable to the SAS software version 9.3 used for the data analysis. Nevertheless, this constituted a major drawback of the ordinal model as all IPD information could not be exploited even when it was available.

Our empirical comparison using the PHQ-9 IPD dataset highlighted several differences between the three models. From a practical perspective, the meta-analyses performed with the BREM and the ordinal model found that the PHQ-9 had an excellent diagnostic accuracy for the set of considered thresholds whereas the application with the Poisson model found a substantially weaker accuracy of the test, with sensitivity differing by approximately 10% at each threshold. Thus, it is possible that the choice of modelling might even affect whether one would recommend or not the diagnostic test for clinical application. At each of the 8 investigated thresholds, the estimates of specificity were very similar across the three models. Stability of the specificity estimates was likely due to the large number of non-cases in each study. Estimates of sensitivity were less stable, and the Poisson model estimated significantly different sensitivities compared to the BREM and the ordinal model across the 8 thresholds. The BREM produced less precise estimates of sensitivity: the variance of the pooled sensitivity estimates decreased as the threshold value was less stringent. As the BREM used logistic regression to estimate the model parameters, the standard errors of the regression coefficients depended on the term $\hat{p}_j(1-\hat{p}_j)$, $j = 1, .., J$, where $\hat{p}_j$ was linked to sensitivity at threshold j following (1). This produced higher standard errors at the extremes of $\hat{p}$ i.e. for more stringent thresholds where sensitivity was closer to 0 or 1 (Hosmer Jr and Lemeshow, 2004). Overall, the two multivariate methods, which analyzed all thresholds simultaneously, produced smaller standard errors for sensitivity and specificity compared with the BREM. This was an advantage of the two methods, and can be explained by the fact that the two multivariate methods borrowed strength across thresholds (Riley, 2009; Jackson *et al.*, 2015).

Our empirical comparison further emphasized how each method accounted for the correlation structure induced by the data. Within studies, sensitivities and specificities across thresholds were derived from data on cases and non-cases, respectively, using different rules, implying that as the threshold value increased,

sensitivity decreased while specificity increased. This relationship explicitly induced a correlation between sensitivity and specificity across studies, which was correctly accounted for by each method. With this relationship, sensitivities and specificities were also each correlated across thresholds, and it was reasonable to think that the correlation between neighboring thresholds was larger compared to between non-consecutive thresholds (Hamza *et al.*, 2009). The choice of a realistic correlation structure was an asset of the two multivariate approaches, and the first order auto-regressive correlation structure assumed in the Poisson model may be more realistic than the compound symmetric correlation structure imposed by the ordinal model. Still, the ordinal method could be extended to model an auto-regressive correlation. The BREM failed to model this correlation by meta-analyzing pairs of sensitivity and specificity separately by threshold.

The ordinal multivariate model was not suitable for simulations. Although the theoretical specification of the ordinal model was appealing, previous work (Putter *et al.*, 2010) highlighted the sensitivity of the model to starting values and the complexity of the method. Sensitivity of the model to starting values was also a burden in this work, both in the empirical application and in the simulation study, where this issue prevented us from using the ordinal model in the simulations. Had the ordinal model been suitable for simulations, it may have performed well as the true sensitivity and (1-specificity) were defined almost linearly on the logit-scale in the simulation study, which was one of the model's assumptions. The application of the ordinal model to the PHQ-9 meta-analysis was highly computationally intensive and complex both in terms of computational time and sensitivity to user-supplied values, which would make the model inaccessible to researchers with limited statistical knowledge (see SAS example in Supplementary Material). The computational burden of ordinal regression models was also highlighted in other similar situations (Dukic and Gatsonis, 2003).

Our simulations thus only compared the performance of the BREM and the Poisson model. Overall, the two methods of analysis performed similarly in terms of mean bias and RMSE of sensitivity and specificity. The true underlying sensitivity and specificity, characterized by the threshold value factor, had a more important impact on the bias in sensitivity and specificity than any other factor. More interestingly, the strength of the correlation between sensitivity (and specificity) across thresholds ρ_{thres} did not have a important impact on the bias or RMSE of the estimated sensitivity and specificity when comparing the BREM to the Poisson model. In terms of coverage of the resulting confidence intervals, the BREM always exhibited under-coverage for both sensitivity and specificity while confidence intervals with the Poisson model were extremely conservative for specificity and exhibited under-coverage for sensitivity. This suggested that ignoring the within-study correlation across thresholds, and thus meta-analyzing thresholds separately, may not have a dramatic impact on the accuracy of the estimates, but may lead to invalid inferences. These findings should be investigated further in a simulation study with a wider range of scenarios investigated.

Moreover, our simulation study investigated how well the correlation structure induced by data of diagnostic accuracy was estimated. Both the BREM and the Poisson model sporadically failed to estimate sensible correlation parameters. The generalizability of this finding was limited given that we only considered one simulation scenario. Also, with the BREM, we only focused on one estimation method, the default Gauss Hermite Quadrature with one quadrature point implemented in R (Zhang *et al.*, 2011). In certain cases, increasing the number of quadrature points or using an alternate estimation method (e.g. Penalized Quasi-Likelihood or Bayesian approaches) may have produced more accurate results, especially in terms of the estimation of the correlation parameters (c.f. Table 6).

The generalizability of the results found in both the empirical comparison and the simulation study was limited to a specific type of diagnostic data. Studies in the PHQ-9 IPD dataset were very heterogeneous (c.f. Figure 1 (A)), and the number of cases used to make inferences on sensitivities was always considerably smaller than the number of non-cases in each study. In the empirical comparison, the differences between the three models found in terms of sensitivity estimates may have been less striking had the disease been very prevalent.

Our work was an extension of the methodological paper by Putter *et al.* (Putter *et al.*, 2010) in which

they introduced the Poisson correlated gamma-frailty model in the context of meta-analyses of diagnostic accuracy with multiple thresholds. The authors empirically compared their newly introduced Poisson model to the ordinal random-effects model and to the bivariate random-effects model (Reitsma *et al.*, 2005), which relied on the normal approximation to the exact binomial distribution. However, the bivariate model based on the exact binomial distribution, which was used in our work, is preferable to its normal approximation (Hamza *et al.*, 2008). Putter *et al.* emphasized that applying the BREM at each threshold separately did not guaranteed the resulting pooled sensitivities and specificities to be monotone (Putter *et al.*, 2010). By defining sensitivity and specificity as an accumulation of positive hazards, the Poisson model had the advantage of guaranteeing monotonicity of the resulting sensitivities and specificities. However, in principle, the Poisson model may estimate hazards greater than 1, which would result in negative estimates of the pooled sensitivity and specificity. In practice, both problems are likely to be rare. In our simulation study, in none of the 16,000 simulation replications did we observe a non-monotonic behaviour of sensitivity or specificity with the BREM nor did the Poisson model yield negative estimates of sensitivity and specificity.

Putter *et al.* (Putter *et al.*, 2010) presented an empirical comparison of these methods using the well-known CAGE IPD dataset (Aertgeerts *et al.*, 2004) for which diagnostic accuracy data of the 5-category CAGE ordinal-scale test were available for 10 studies. Our work involved a larger and more complex IPD dataset, where results from a subset of 9 categories out of 28 possible categories were analyzed. Putter *et al.* showed non-significant differences between the estimation of sensitivity and specificity across the three methods whereas, using the PHQ-9 IPD dataset, we found more striking differences, especially for the Poisson model which systematically estimated lower sensitivity for each threshold compared to the two other methods.

Our work focused on diagnostic accuracy of ordinal-scale tests, but the methods and results presented here also apply to continuous-scale tests. For continuous scale-tests, one can consider a pre-specified set of J thresholds to classify the test results as positive or negative, and define the outcome Y accordingly. For example, consider using glycated haemoglobin (HbA1c) in mmol/mol, a continuous biomarker that roughly takes values between 30 and 120, to detect the presence of diabetes. It would be possible to consider the set of thresholds $\{31, 33, 35, ..., 119\}$ (or a finer or coarser grid), and directly apply the methods presented above.

6 Conclusion

A recurrent motivation for proposing multivariate alternatives to the BREM when results from multiple thresholds are available is that the BREM does not correctly account for the correlation of sensitivity and specificity across thresholds. Our simulation study showed that the BREM and the Poisson model were very similar in terms of accuracy and efficiency of the estimates. Based on the findings described in this work, the BREM has the advantage of being simple to understand and easy to implement, while the loss in efficiency is minimal compared to the alternative multivariate approaches. However, alternatives should be considered to construct valid confidence intervals for the pooled estimates with the BREM. Thus, we tentatively recommend the bivariate approach rather than more complex multivariate methods for IPD diagnostic accuracy meta-analyses, although more research is needed with different scenarios to determine if there is a similarly minimal loss of efficiency with the bivariate models in other types of data and simulations.

Acknowledgements The authors thank three anonymous reviewers and the co-Editors for their comments and suggestions that led to an improved paper. We acknowledge the Canadian Institutes of Health Research (CIHR) for sponsoring this work (KRS-134297). Ms. Levis was supported by a CIHR Frederick Banting and Charles Best Canadian Graduate Scholarships doctoral award. Collection of data for the primary study by Fann *et al.* was supported by grant RO1 HD39415 from the US National Center for Medical Rehabilitation Research. Collection of data for the primary study by Gjerdingen *et al.* was supported by grants from the National Institutes of Mental Health (R34

MH072925, K02 MH65919, P30 DK50456). Dr. Lamers has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement # PCIG12-GA-2012-334065. Dr. Stafford received PhD scholarship funding from the University of Melbourne. The AIM study by Williams *et al.* was funded by NINDS (R01 NS 38529). Dr. Thombs was supported by an Investigator Award from the Arthritis Society. Dr. Benedetti was supported by an Fondation de Recherche du Québec - Santé (FRQS) researcher salary award and an FRQS Chercheur Boursier award.

Conflict of Interest

The authors have declared no conflict of interest.

References

Aertgeerts, B., Buntinx, F., and Kester, A. (2004) The value of the CAGE in screening for alcohol abuse and alcohol dependence in general clinical populations: a diagnostic meta-analysis. *Journal of clinical epidemiology*, **57** (1), 30–39.

Azah, M., Shah, M., Juwita, S., Bahri, I., Rushidi, W., and Jamil, Y. (2005) Validation of the Malay version brief Patient Health Questionnaire (PHQ-9) among adult attending family medicine clinics. *International Medical Journal*.

Bates, D., Sarkar, D., Bates, M.D., and Matrix, L. (????) lme4: Linear mixed-effects models using Eigen and S4. URL `http://CRAN.R-project.org/package=lme4`.

Chipman, H., Ranjan, P., and Al-Ahmad, F. (2015) Simulation studies for statistical procedures: Why can't we practice what we preach? 43rd Annual Meeting of the Statistical Society of Canada.

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

de Lima Osório, F., Vilela Mendes, A., Crippa, J.A., and Loureiro, S.R. (2009) Study of the discriminative validity of the PHQ-9 and PHQ-2 in a sample of brazilian women in the context of primary health care. *Perspectives in psychiatric care*, **45** (3), 216–227.

Dukic, V. and Gatsonis, C. (2003) Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds. *Biometrics*, **59** (4), 936–946.

Eusebi, P. (2013) Diagnostic accuracy measures. *Cerebrovasc Dis*, **36** (4), 267–272.

Fann, J.R., Bombardier, C.H., Dikmen, S., Esselman, P., Warms, C.A., Pelzer, E., Rau, H., and Temkin, N. (2005) Validity of the Patient Health Questionnaire-9 in assessing depression following traumatic brain injury. *The Journal of head trauma rehabilitation*, **20** (6), 501–511.

Fiocco, M., Putter, H., and Van Houwelingen, J. (2009a) Meta-analysis of pairs of survival curves under heterogeneity: A Poisson correlated gamma-frailty approach. *Statistics in medicine*, **28** (30), 3782–3797.

Fiocco, M., Putter, H., and Van Houwelingen, J. (2009b) A new serially correlated gamma-frailty process for longitudinal count data. *Biostatistics*, **10** (2), 245–257.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., and Hothorn, T. (2014) *mvtnorm: Multivariate Normal and t Distributions*. URL `http://CRAN.R-project.org/package=mvtnorm`, R package version 1.0-2.

Gilbody, S., Richards, D., and Barkham, M. (2007a) Diagnosing depression in primary care using self-completed instruments: UK validation of PHQ–9 and CORE–OM. *Br J Gen Pract*, **57** (541), 650–652.

Gilbody, S., Richards, D., Brealey, S., and Hewitt, C. (2007b) Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): a diagnostic meta-analysis. *Journal of general internal medicine*, **22** (11), 1596–1602.

Gjerdingen, D., Crow, S., McGovern, P., Miner, M., and Center, B. (2009) Postpartum depression screening at well-child visits: validity of a 2-question screen and the PHQ-9. *The Annals of Family Medicine*, **7** (1), 63–70.

Gräfe, K., Zipfel, S., Herzog, W., and Lowe, B. (2004) Screening for psychiatric disorders with the Patient Health Questionnaire (PHQ). Results from the German validation study. *Diagnostica*, **50** (4), 171–181.

Hamza, T., Arends, L., van Houwelingen, H., and Stijnen, T. (2009) Multivariate random effects meta-analysis of diagnostic tests with multiple thresholds. *BMC medical research methodology*, **9** (1), 73.

Hamza, T.H., van Houwelingen, H.C., and Stijnen, T. (2008) The binomial distribution of meta-analysis was preferred to model within-study variability. *Journal of clinical epidemiology*, **61** (1), 41–51.

Harbord, R.M., Deeks, J.J., Egger, M., Whiting, P., and Sterne, J.A. (2007) A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics*, **8** (2), 239–251.

Hofert, M., Kojadinovic, I., Maechler, M., and Yan, J. (2016) *copula: Multivariate Dependence with Copulas*. URL `http://CRAN.R-project.org/package=copula`, R package version 0.999-13.

Hosmer Jr, D. and Lemeshow, S. (2004) *Applied logistic regression*, John Wiley & Sons.

Jackson, D., Riley, R., and White, I.R. (2011) Multivariate meta-analysis: Potential and promise. *Statistics in Medicine*, **30** (20), 2481–2498.

Jackson, D., White, I.R., Price, M., Copas, J., and Riley, R.D. (2015) Borrowing of strength and study weights in multivariate and network meta-analysis. *Statistical methods in medical research*, p. 0962280215611702.

Kroenke, K. and Spitzer, R. (2002) The PHQ-9: a new depression diagnostic and severity measure. *Psychiatr Ann*, **32** (9), 1–7.

Kroenke, K., Spitzer, R.L., and Williams, J.B. (2001) The PHQ-9. *Journal of general internal medicine*, **16** (9), 606–613.

Lamers, F., Jonkers, C.C., Bosma, H., Penninx, B.W., Knottnerus, J.A., and van Eijk, J.T.M. (2008) Summed score of the Patient Health Questionnaire-9 was a reliable and valid method for depression screening in chronically ill elderly patients. *Journal of clinical epidemiology*, **61** (7), 679–687.

Levis, B., Benedetti, A., Levis, A., Ioannidis, J., Shrier, I., Cuijpers, P., Gilbody, S., Kloda, L., McMillan, D., Patten, S., Steele, R., Ziegelstein, R., Bombardier, C., de Lima Osório, F., Fann, J., Gjerdingen, D., Lamers, F., Lotrakul, M., Loureiro, S., Löwe, B., Shaaban, J., Stafford, L., van Weert, H., Whooley, M., Williams, L., Wittkampf, K., Yeung, A., and Thombs, B. (in press) Selective cutoff reporting in studies of diagnostic test accuracy: A comparison of traditional and individual patient data meta-analyses of the Patient Health Questionnaire-9 depression screening tool. *American Journal of Epidemiology*.

Lindsay, B.G. (1988) Composite likelihood methods. *Contemporary mathematics*, **80** (1), 221–39.

Lotrakul, M., Sumrithe, S., and Saipanish, R. (2008) Reliability and validity of the Thai version of the PHQ-9. *BMC psychiatry*, **8** (1), 1.

Manea, L., Gilbody, S., and McMillan, D. (2012) Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. *Canadian Medical Association Journal*, **184** (3), E191–E196.

Martínez-Camblor, P. (2014) Fully non-parametric receiver operating characteristic curve estimation for random-effects meta-analysis. *Statistical methods in medical research*, p. 0962280214537047.

Moses, L., Shapiro, D., and Littenberg, B. (1993) Combining independent studies of a diagnostic test into a summary roc curve: Data-analytic approaches and some additional considerations. *Statistics in medicine*, **12** (14), 1293–1316.

Putter, H., Fiocco, M., and Stijnen, T. (2010) Meta-analysis of diagnostic test accuracy studies with multiple thresholds using survival methods. *Biometrical Journal*, **52** (1), 95–110.

R Core Team (2013) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

Reitsma, J., Glas, A., Rutjes, A., Scholten, R., Bossuyt, P., and Zwinderman, A. (2005) Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of clinical epidemiology*, **58** (10), 982–990.

Riley, R. (2009) Multivariate meta-analysis: the effect of ignoring within-study correlation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **172** (4), 789–811.

Riley, R., Lambert, P., Abo-Zaid, G. *et al.* (2010) Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ*, **340**.

Riley, R.D., Takwoingi, Y., Trikalinos, T., Guha, A., Biswas, A., Ensor, J., Morris, R.K., and Deeks, J.J. (2014) Meta-analysis of test accuracy studies with multiple and missing thresholds: a multivariate-normal model. *J Biomed Biostat*, **5**, 100 196.

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

SAS Institute Inc. (2003) *SAS/STAT Software, Version 9.3*, Cary, NC. URL `http://www.sas.com/`.

Spitzer, R.L., Kroenke, K., Williams, J.B., Group, P.H.Q.P.C.S. *et al.* (1999) Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. *Jama*, **282** (18), 1737–1744.

Stafford, L., Berk, M., and Jackson, H.J. (2007) Validity of the hospital anxiety and depression scale and Patient Health Questionnaire-9 to screen for depression in patients with coronary artery disease. *General hospital psychiatry*, **29** (5), 417–424.

Steinhauser, S., Schumacher, M., and Rücker, G. (2016) Modelling multiple thresholds in meta-analysis of diagnostic test accuracy studies. *BMC Medical Research Methodology*, **16** (1), 97.

Thombs, B.D., Ziegelstein, R.C., and Whooley, M.A. (2008) Optimizing detection of major depression among patients with coronary artery disease using the patient health questionnaire: data from the heart and soul study. *Journal of General Internal Medicine*, **23** (12), 2014–2017.

Van Houwelingen, H.C., Zwinderman, K.H., and Stijnen, T. (1993) A bivariate approach to meta-analysis. *Statistics in medicine*, **12** (24), 2273–2284.

Venables, W.N. and Ripley, B.D. (2002) *Modern Applied Statistics with S*, Springer, New York, 4th edn.. URL `http://www.stats.ox.ac.uk/pub/MASS4`, iSBN 0-387-95457-0.

Wehrens, R., Putter, H., and Buydens, L.M. (2000) The bootstrap: a tutorial. *Chemometrics and intelligent laboratory systems*, **54** (1), 35–52.

Williams, L.S., Brizendine, E.J., Plue, L., Bakas, T., Tu, W., Hendrie, H., and Kroenke, K. (2005) Performance of the PHQ-9 as a screening tool for depression after stroke. *Stroke*, **36** (3), 635–638.

Wittkampf, K., van Ravesteijn, H., Baas, K., van de Hoogen, H., Schene, A., Bindels, P., Lucassen, P., van de Lisdonk, E., and van Weert, H. (2009) The accuracy of Patient Health Questionnaire-9 in detecting depression and measuring depression severity in high-risk groups in primary care. *General hospital psychiatry*, **31** (5), 451–459.

Wittkampf, K.A., Naeije, L., Schene, A.H., Huyser, J., and van Weert, H.C. (2007) Diagnostic accuracy of the mood module of the Patient Health Questionnaire: a systematic review. *General hospital psychiatry*, **29** (5), 388–395.

Yeung, A., Fung, F., Yu, S.C., Vorono, S., Ly, M., Wu, S., and Fava, M. (2008) Validation of the Patient Health Questionnaire-9 for depression screening among Chinese Americans. *Comprehensive psychiatry*, **49** (2), 211–217.

Zhang, H., Lu, N., Feng, C., Thurston, S.W., Xia, Y., Zhu, L., and Tu, X.M. (2011) On fitting generalized linear mixed-effects models for binary responses using different statistical packages. *Statistics in Medicine*, **30** (20), 2562–2572.

7 Figures

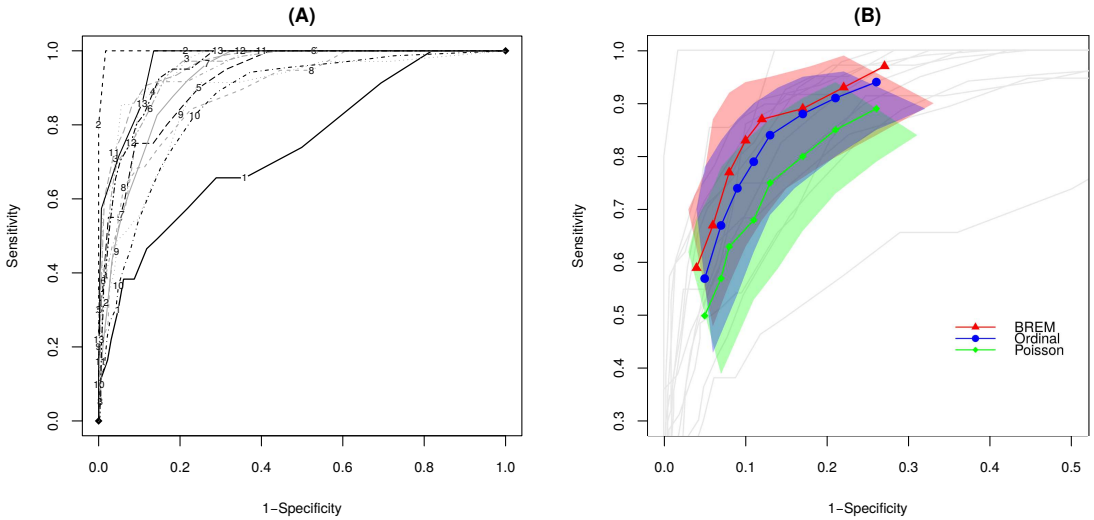

Figure 1 **(A)** Individual ROC curves for the 13 studies investigating the diagnostic accuracy of the PHQ-9 questionnaire. Each line represents a study-specific empirical ROC curve based on estimated sensitivity and specificity for threshold 0 to 27. The study numbers found in Table 1 indicate which ROC curve represents which study. **(B)** Pooled ROC curves for thresholds 7 to 14 with the BREM (red line with triangles), the ordinal (blue line with circles) and the Poisson (green line with diamonds) approaches, superimposed over the study-specific ROC curves (lightgrey). Corresponding confidence bands are shown in red, blue and green overlapping shaded areas for the BREM, the ordinal and the Poisson approaches, respectively.

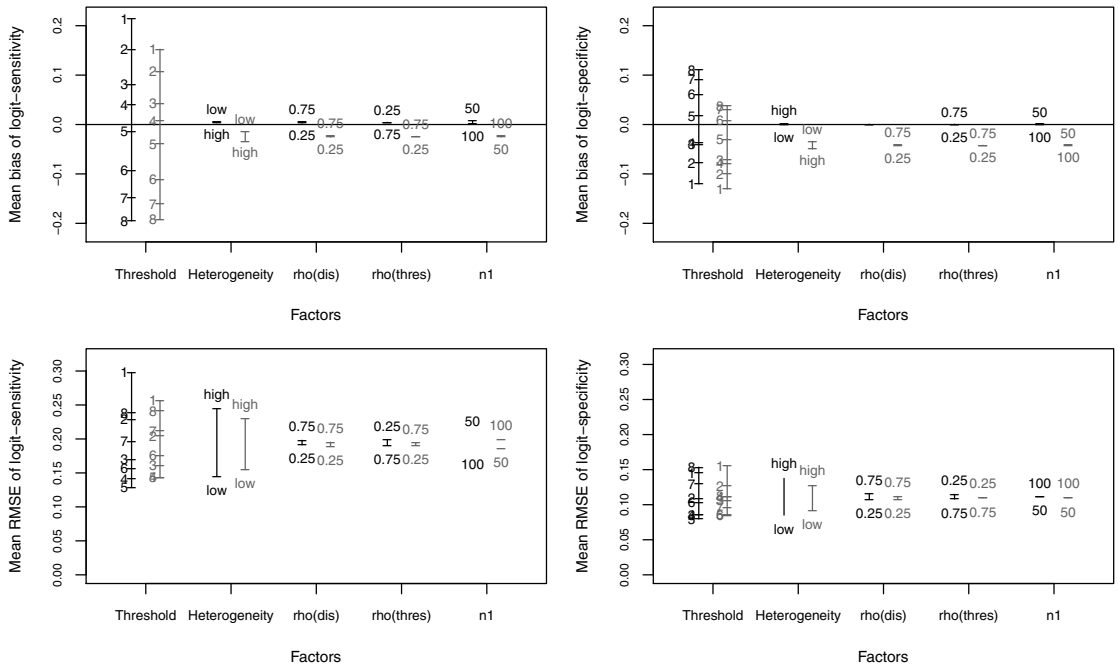

Figure 2 Effect of 5 factors (listed below) on the mean bias and RMSE of logit sensitivity (left column) and logit specificity (right column) with the BREM (black) and Poisson model (light grey). The 5 factors are: the threshold value (*Threshold*) denoted by 1 to 8 where the threshold value corresponds to the true sensitivity/specificity 0.94/0.74, 0.91/0.79, 0.88/0.83, 0.84/0.87, 0.79/0.89, 0.74/0.91, 0.67/0.93 and 0.57/0.95, respectively; the degree of between-study heterogeneity (*Heterogeneity*); the correlation between sensitivity and specificity at each threshold (*rho(dis)*); the correlation between sensitivities and specificities across thresholds (*rho(thres)*); and the sample size of the diseased patients.

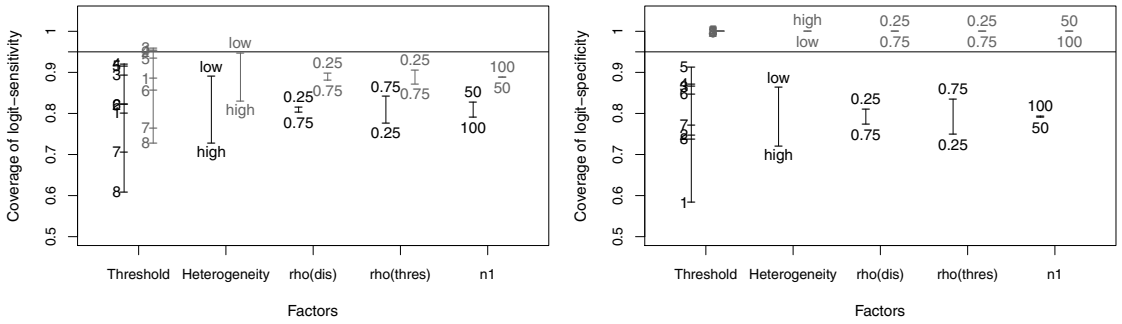

Figure 3 Effect of 5 factors (listed below) on the confidence interval coverage of logit sensitivity (left column) and logit specificity (right column) with the BREM (black) and Poisson model (light grey). The 5 factors are: the threshold value (*Threshold*) denoted by 1 to 8 where the threshold value corresponds to the true sensitivity/specificity 0.94/0.74, 0.91/0.79, 0.88/0.83, 0.84/0.87, 0.79/0.89, 0.74/0.91, 0.67/0.93 and 0.57/0.95, respectively; the degree of between-study heterogeneity (*Heterogeneity*); the correlation between sensitivity and specificity at each threshold (*rho(dis)*); the correlation between sensitivities and specificities across thresholds (*rho(thres)*); and the sample size of the diseased patients.

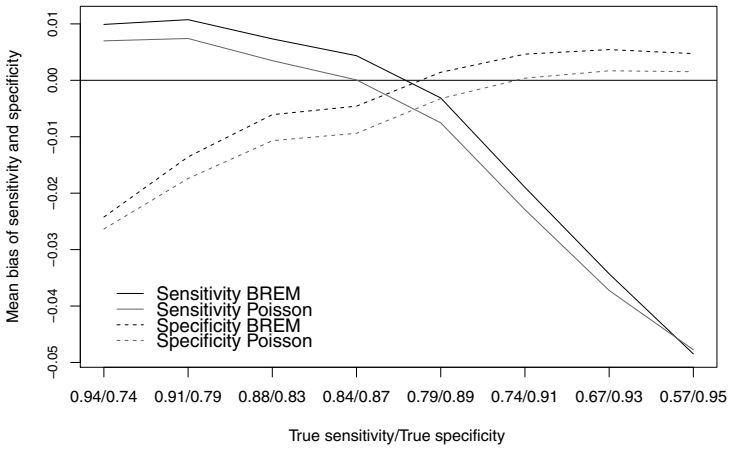

Figure 4 Effect of the threshold value on the mean bias of sensitivity (thick lines) and specificity (dashed lines) with the BREM (black lines) and Poisson model (grey lines).

8 Tables

Table 1 Number of truly MDD/non-MDD patients with PHQ-9 score less or equal to 7, between 8 and 13, or equal or larger than 14 in 13 studies

		PHQ-9 Score								
Reference	Study ID	$\leqslant$7	8	9	10	11	12	13	$\geqslant$14	Total
Azah *et al.*, 2005	1	9/100	3/17	1/6	3/7	0/6	3/3	3/4	8/7	30/150
de Lima Osório *et al.*, 2009	2	0/105	0/4	0/6	0/0	6/1	6/1	3/0	45/0	60/117
Fann *et al.*, 2005	3	5/53	0/5	1/7	1/5	0/6	7/4	4/4	27/6	45/90
Gilbody *et al.*, 2007a	4	0/38	2/6	1/3	0/2	0/2	1/1	1/2	31/6	36/60
Gjerdingen *et al.*, 2009	5	4/348	1/14	0/18	1/4	2/7	1/5	0/5	11/17	20/418
Gräfe *et al.*, 2004	6	1/290	1/22	0/28	0/17	3/19	6/17	3/9	57/48	71/450
Lamers *et al.*, 2008	7	7/87	5/27	9/17	13/25	14/28	16/27	17/21	196/102	277/334
Lotrakul *et al.*, 2008	8	2/169	1/32	2/21	1/9	0/4	1/9	3/6	9/10	19/260
Stafford *et al.*, 2007	9	12/136	4/7	0/1	1/3	1/4	1/0	2/1	14/6	35/158
Thombs *et al.*, 2008	10	70/677	18/25	15/20	18/16	9/11	12/12	14/5	68/34	224/800
Williams *et al.*, 2005	11	3/158	4/12	3/16	6/7	7/4	6/5	8/3	69/5	106/210
Wittkampf *et al.*, 2009	12	4/236	0/21	2/19	4/18	2/10	4/8	4/15	57/31	77/358
Yeung *et al.*, 2008	13	0/106	0/3	0/6	0/4	0/2	1/4	1/8	35/14	37/147
	Total	117/2503	39/195	34/168	48/117	44/104	65/96	63/83	627/286	1037/3552

Table 2 Estimates (95% CI) of the pooled sensitivity and specificity from the three methods

	Bivariate approach		Ordinal approach		Poisson approach	
Threshold	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
7	0.97 (0.90-0.99)	0.73 (0.67-0.78)	0.94 (0.89-0.96)	0.74 (0.68-0.78)	0.89 (0.84-0.94)	0.74 (0.69-0.79)
8	0.93 (0.84-0.97)	0.78 (0.74-0.82)	0.91 (0.85-0.95)	0.79 (0.74-0.83)	0.85 (0.79-0.91)	0.79 (0.74-0.83)
9	0.89 (0.79-0.95)	0.83 (0.80-0.87)	0.88 (0.80-0.93)	0.83 (0.79-0.86)	0.80 (0.73-0.87)	0.83 (0.79-0.87)
10	0.87 (0.74-0.94)	0.88 (0.85-0.90)	0.84 (0.74-0.90)	0.87 (0.84-0.89)	0.75 (0.66-0.83)	0.87 (0.83-0.90)
11	0.83 (0.68-0.92)	0.90 (0.88-0.92)	0.79 (0.69-0.87)	0.89 (0.87-0.91)	0.68 (0.59-0.78)	0.89 (0.86-0.93)
12	0.77 (0.63-0.87)	0.92 (0.90-0.94)	0.74 (0.62-0.83)	0.91 (0.89-0.93)	0.63 (0.53-0.74)	0.92 (0.89-0.94)
13	0.67 (0.56-0.77)	0.94 (0.92-0.95)	0.67 (0.54-0.78)	0.93 (0.91-0.95)	0.57 (0.46-0.68)	0.93 (0.91-0.96)
14	0.59 (0.48-0.70)	0.96 (0.94-0.97)	0.57 (0.43-0.70)	0.95 (0.94-0.96)	0.50 (0.39-0.62)	0.95 (0.93-0.97)

Table 3 Parameter estimates (standard errors) obtained by the bivariate approach

Threshold	Logit sensitivity	Logit (1-specificity)	Correlations ρ_T
7	3.55 (0.72)	-1.01 (0.15)	0.64
8	2.63 (0.48)	-1.29 (0.13)	0.57
9	2.13 (0.41)	-1.62 (0.12)	0.61
10	1.89 (0.42)	-1.97 (0.14)	0.27
11	1.62 (0.44)	-2.21 (0.13)	0.46
12	1.23 (0.35)	-2.44 (0.14)	0.76
13	0.72 (0.25)	-2.71 (0.14)	0.77
14	0.37 (0.23)	-3.07 (0.15)	0.57

For thresholds $j = 7, .., 14$, sensitivity$_j$=expit(Logit sensitivity$_j$) and specificity$_j$=1-expit(Logit $(1 -$ specificity$_j$)).

Table 4 Parameter estimates (standard errors) obtained by the ordinal approach

Threshold	Logit sensitivity	Logit (1-specificity)	Slope/ Intercept		Correlations	
7	2.69 (0.30)	-1.02 (0.13)	α	3.95 (0.28)	σ^2_α	0.67 (0.32)
8	2.35 (0.29)	-1.30 (0.13)	β	1.23 (0.06)	σ^2_Δ	0.18 (0.08)
9	1.98 (0.29)	-1.60 (0.13)			σ^2_δ	0.007 (0.002)
10	1.63 (0.29)	-1.89 (0.13)			$\sigma_{\alpha\Delta}$	0.006 (0.12)
11	1.34 (0.29)	-2.12 (0.13)				
12	1.05 (0.29)	-2.35 (0.13)				
13	0.70 (0.29)	-2.64 (0.13)				
14	0.30 (0.29)	-2.97 (0.14)				

For thresholds $j = 7, .., 14$, sensitivity$_j$=expit(logit sensitivity$_j$) and specificity$_j$=1-expit(logit $(1 -$ specificity$_j$)).

Table 5 Parameter estimates (standard errors) obtained by the Poisson approach

Threshold	Hazards Diseased	Hazards Healthy	Frailty Variances		Correlations	
7	0.05 (0.020)	0.19 (0.025)	ξ^1	0.71 (0.238)	ρ_{thres}	0.74 (0.113)
8	0.04 (0.018)	0.18 (0.025)	ξ^0	0.11 (0.027)	ρ_{dis}	0.39 (0.170)
9	0.06 (0.025)	0.21 (0.029)				
10	0.06 (0.026)	0.23 (0.035)				
11	0.08 (0.032)	0.19 (0.030)				
12	0.07 (0.030)	0.19 (0.031)				
13	0.10 (0.038)	0.22 (0.037)				
14	0.12 (0.045)	0.24 (0.041)				

For thresholds $j = 7, .., 14$, sensitivity$_j = \prod_{k=1}^{j}(1 -$ Hazard Diseased$_k$) and sensitivity$_j = 1 - \prod_{k=1}^{j}(1-$ Hazard Healthy$_k$).

Table 6 Bias, SDs and rate of failure to estimate the correlation parameters of 1000 simulated datasets with the bivariate and the Poisson models in the scenario where ρ_{thres} and ρ_{dis} were both set to 0.25, the variability of the frailties ξ^1/ξ^0 was set to 0.1/0.05 (low variability) and the sample size of the diseased patients was set to $n_1 = 50$

Bivariate model			
True sensitivity/ True specificity	True correlation	Mean (SD)	Failure rate[1]
0.94/0.74	0.25	0.21 (0.75)	1.5 %
0.91/0.79	0.25	0.19 (0.74)	3.4 %
0.88/0.83	0.25	0.20 (0.75)	4.9 %
0.84/0.87	0.25	0.22 (0.72)	9.3 %
0.79/0.89	0.25	0.26 (0.69)	12.7 %
0.74/0.91	0.25	0.23 (0.69)	16.5 %
0.67/0.93	0.25	0.28 (0.67)	20.3 %
0.57/0.95	0.25	0.26 (0.70)	20.5 %
Poisson model			
	True correlation	Mean (SD)	Failure rate
ρ_{dis}	0.25	0.17 (0.14)	0 %
ρ_{thres}	0.25	NA[2]	NA

1 A failure occurred when the model estimated at least one of the specificity random effects' variance by zero, such that the corresponding correlation coefficient was not a number.

2 NA: Not Applicable

