Full title: An empirical comparison of three methods for multiple cut-off diagnostic test meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) depression screening tool using published data versus individual level data

Short title: Three approaches to meta-analyze the PHQ-9

Authors:

Benedetti Andrea[1, 2], Levis Brooke[1,3], Rücker Gerta[4], Jones Hayley E[5], Schumacher Martin[4], Ioannidis John P. A.[6], Thombs Brett[3], and the DEPRESSion Screening Data (DEPRESSD) Collaboration

1. Department of Epidemiology, Biostatistics & Occupational Health, McGill University, Canada
2. Centre for Outcomes Research and Evaluation, McGill University Health Centre, Canada
3. Lady Davis Research Institute, SMBD Jewish General Hospital, Canada
4. Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg, Germany
5. Population Health Sciences, Bristol Medical School, University of Bristol, UK
6. Meta-Research Innovation Center at Stanford (METRICS), and Departments of Medicine, Health Research and Policy, Biomedical Data Science, and Statistics, Stanford University, Stanford, CA, USA

Corresponding author:

Andrea Benedetti, PhD
Address: 5252 boul de Maisonneuve,
Office #3D.59
Montréal, QC H4A 3S5
Canada
Email: andrea.benedetti@mcgill.ca
Phone number: (514) 934 1934 ext. 32161

The DEPRESSD COLLABORATION:

(Kira E. Riehm, Lady Davis Institute for Medical Research, Montréal, Québec, Canada; Nazanin Saadat, Lady Davis Institute for Medical Research, Montréal, Québec, Canada; Alexander W. Levis, McGill University, Montréal, Québec, Canada; Marleine Azar, McGill University, Montréal, Québec, Canada; Danielle B. Rice, McGill University, Montréal, Québec, Canada; Ying Sun, Lady Davis Institute for Medical Research, Montréal, Québec, Canada; Ankur Krishnan, Lady Davis Institute for Medical Research, Montréal, Québec, Canada; Chen He, McGill University, Montréal, Québec, Canada; Yin Wu, McGill University, Montréal, Québec, Canada; Parash Mani Bhandari, McGill University, Montréal, Québec, Canada; Dipika Neupane, McGill University, Montréal, Québec, Canada; Mahrukh Imran, Lady Davis Institute for Medical Research, Montréal, Québec, Canada; Jill Boruff, McGill University, Montréal, Québec, Canada; Pim Cuijpers, Vrije Universiteit, Amsterdam, the Netherlands; Simon Gilbody, University of York, Heslington, York, UK; Lorie A. Kloda, Concordia University, Montréal, Québec, Canada; Dean McMillan, University of York, Heslington, York, UK; Scott B. Patten, University of Calgary, Calgary, Alberta, Canada; Ian Shrier, McGill University, Montréal, Québec, Canada; Roy C. Ziegelstein, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA; Dickens H. Akena, Makerere University College of Health Sciences, Kampala, Uganda; Bruce Arroll, University of Auckland, Auckland, New Zealand; Hamid R. Baradaran, Endocrinology Research Center, Institute of Endocrinology and Metabolism, Iran University of Medical Sciences, Tehran, Iran; Charles H. Bombardier, University of Washington, Seattle, Washington, USA; Peter Butterworth, The University of Melbourne, Melbourne, Australia; Gregory Carter, University of Newcastle, New South Wales, Australia; Marcos H. Chagas, University of São Paulo, Ribeirão Preto, Brazil; Juliana C. N. Chan, The Chinese University of Hong Kong, Hong Kong Special Administrative Region, China; Rushina Cholera, University of North Carolina at Chapel Hill School of Medicine, Chapel Hill, North Carolina, USA; Neerja Chowdhary, Clinical practice, Mumbai, India; Kerrie Clover, University of Newcastle, New South Wales, Australia; Yeates Conwell, University of Rochester Medical Center, Rochester, New York, USA; Janneke M. de Man-van Ginkel, University Medical Center Utrecht, Utrecht, The Netherlands; Jaime Delgadillo, University of Sheffield, Sheffield, UK; Jesse R. Fann, University of Washington, Seattle, Washington, USA; Felix H. Fischer, Charité - Universitätsmedizin Berlin, Berlin, Germany; Daniel Fung, Duke-NUS Medical School, Singapore; Bizu Gelaye, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA; Felicity Goodyear-Smith, University of Auckland, Auckland, New Zealand; Patricia A. Harrison, City of Minneapolis Health Department, Minneapolis, Minnesota, USA; Martin Harter, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; Thomas Hyphantis, University of Ioannina, Ioannina, Greece; Masatoshi Inagaki, Shimane University, Shimane, Japan; Khalida Ismail, King's College London Weston Education Centre, London, UK; Nathalie Jetté, Ichan School of Medicine at Mount Sinai, New York, New York, USA; Mohammad E. Khamseh, Endocrinology Research Center, Institute of Endocrinology and Metabolism, Iran University of Medical Sciences, Tehran, Iran; Kim M. Kiely, University of New South Wales, Sydney, Australia; Femke Lamers, Amsterdam UMC, Amsterdam, the Netherlands ; Shen-Ing Liu, Mackay Memorial Hospital, Taipei, Taiwan; Manote Lotrakul, Mahidol University, Bangkok, Thailand; Sonia R. Loureiro, University of São Paulo, Ribeirão Preto, Brazil; Bernd Löwe, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; Laura Marsh, Baylor College of Medicine, Houston and Michael E. DeBakey Veterans Affairs Medical Center, Houston, Texas, USA; Anthony McGuire, St. Joseph's College, Standish, Maine, USA; Sherina

Mohd Sidik, Universiti Putra Malaysia, Serdang, Selangor, Malaysia; Tiago N. Munhoz, Federal University of Pelotas, Pelotas, Brazil; Flávia L. Osório, University of São Paulo, Ribeirão Preto, Brazil; Vikram Patel, Harvard Medical School, Boston, Massachusetts, USA; Brian W. Pence, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA; Katrin Reuter, Group Practice for Psychotherapy and Psycho-oncology, Freiburg, Germany; Alasdair G. Rooney, University of Edinburgh, Edinburgh, Scotland, UK; Iná S. Santos, Federal University of Pelotas, Pelotas, Brazil; Juwita Shaaban, Universiti Sains Malaysia, Kelantan, Malaysia; Abbey Sidebottom, Allina Health, Minneapolis, Minnesota, USA; Lesley Stafford, Royal Women's Hospital, Parkville, Australia; Sharon C. Sung, Duke-NUS Medical School, Singapore; Alyna Turner, University of Newcastle, New South Wales, Newcastle, Australia; Christina M. van der Feltz-Cornelis, University of York, York, UK; Henk C. van Weert, Amsterdam University Medical Centers, Location AMC, Amsterdam, the Netherlands; Paul A. Vöhringer, Universidad de Chile, Santiago, Chile; Jennifer White, Monash University, Melbourne, Australia; Mary A. Whooley, Veterans Affairs Medical Center, San Francisco, California, USA; Kirsty Winkley, King's College London, Waterloo Road, London, UK ; Mitsuhiko Yamada, National Center of Neurology and Psychiatry, Tokyo, Japan; Yuying Zhang, The Chinese University of Hong Kong, Hong Kong Special Administrative Region, China)

ABSTRACT (250 words)
Selective cut-off reporting in primary diagnostic accuracy studies with continuous or ordinal data may result in biased estimates when meta-analyzing studies. Collecting individual participant data (IPD) and estimating accuracy across all relevant cut-offs for all studies can overcome such bias but is labour-intensive.
We meta-analyzed the diagnostic accuracy of the Patient Health Questionnaire-9 (PHQ-9) depression screening tool. We compared results for two statistical methods proposed by Steinhauser and by Jones to account for missing cut-offs, with results from a series of bivariate random effects models (BRM) estimated separately at each cut-off. We applied the methods to a dataset that contained information only on cut-offs that were reported in the primary publications, and to the full IPD dataset that contained information for all cut-offs for every study. For each method, we estimated pooled sensitivity and specificity and associated 95% confidence intervals for each cut-off and area under the curve (AUC).
The full IPD dataset comprised data from 45 studies, 15,020 subjects and 1,972 cases of major depression, and included information on every possible cut-off.
When using data available in publications, using statistical approaches out-performed the BRM applied to the same data.
AUC was similar for all approaches when using the full IPD dataset, though pooled estimates were slightly different.
Overall, using statistical methods to fill in missing cut-off data recovered the receiver operating characteristic (ROC) curve from the *full IPD dataset* well when using only the *published* subset. All methods performed similarly when applied to the full IPD dataset.

KEYWORDS: individual participant data; meta-analysis; diagnostic accuracy; bivariate random effects model

INTRODUCTION

In clinical practice, screening tests are used to attempt to distinguish between diseased and non-diseased patients. The diagnostic accuracy of screening tests is assessed in research studies by comparing screening test results to diagnostic classifications based on a reference standard. Typically, sensitivity (the proportion of truly diseased patients who test positive on the screening test) and specificity (the proportion of truly non-diseased patients who test negative on the screening test) are reported.[1]

When the screening test is an ordinal or continuous measure, multiple cut-off thresholds may be evaluated, and sensitivity and specificity depend on the cut-off used to define a positive versus negative screening result. While some studies present diagnostic accuracy results for all possible cut-offs, most studies present estimates for either a single cut-off or for a subset of possible cut-offs. Some authors report results for a pre-identified "standard" cut-off or set of cut-offs. Other authors only report accuracy results for a cut-off or set of cut-offs that perform well in their study, (e.g., high combined sensitivity and specificity) and do not report accuracy results for other cut-offs, even when the other cut-offs are considered standard.[2,3]

Different approaches have been used to meta-analyze diagnostic accuracy studies when some studies do not provide data for all possible cut-offs. One approach has considered cut-offs separately. In this approach, researchers have restricted their meta-analyses to one or two pre-selected cut-offs[4] or to a range of pre-selected cut-offs. For each cut-off meta-analyzed, only results from primary studies that published results for that cut-off have been included.[5,6] If cut-offs reported in some primary studies were selected to maximize accuracy estimates, however, then the pooled accuracy estimates in the meta-analyses will tend to overestimate accuracy compared to what would occur in practice.[2] A second approach involves estimating accuracy based on a single cut-off per primary study, even if different studies used different cut-offs.[7] This approach would amplify biases from selective cut-off reporting and the resulting pooled estimate or 'summary Receiver Operating Characteristic (ROC) curve' is not useful clinically.[8] A third set of approaches has considered correlations across cut-offs and modeled all cut-offs of a measure simultaneously,[9,10] but these methods cannot estimate accuracy reliably under certain missing cut-off data patterns within primary studies.

One way to overcome bias due to selective cut-off reporting in primary diagnostic accuracy studies is to collect individual participant data (IPD) from researchers who conducted original primary studies.[2,11] Collecting IPD however, is labour-intensive, as it requires substantial time to identify and obtain original data, clarify data-related issues with data providers, and generate a consistent data format across studies.[11] These patient-level data can then be used to estimate accuracy across all possible cut-off thresholds for all studies, but the best way to do this is still controversial. Conventionally, this type of data may be analyzed by a series of bivariate random effects models (BRM), fit at each cut-off separately. The BRM model is based on a number of two-by-two tables coming from independent studies. The BRM requires no assumptions about the association between cut-off and sensitivity or specificity. However, analyzing data at multiple cut-offs separately via the BRM assumes independence when in fact, these analyses are not independent: they are based on the same data and ignore within-study correlation across cut-offs. Indeed, because of this specific limitation the BRM cannot be considered the "gold standard" approach and this has led to the development of new approaches.[9,10,12-15]

Formatted: English (Canada)

[illegible] available subset of data based on assumptions about the distributional form of the test result, and

may overcome bias due to selective cut-off reporting. Several approaches have been proposed. Steinhauser et al.[13,16] recently proposed a novel approach to pooling estimates of sensitivity and specificity based on estimating the parameters of an assumed distribution of an underlying continuous marker in diseased and non-diseased subjects using a linear mixed effects model. This approach allows for differing numbers of cut-offs in each primary study and accounts for the complex dependencies in the data. The approach assumes that the underlying continuous marker arises from a logistic distribution and that there is a linear association between cut-off and logit(sensitivity) and logit(specificity). Hoyer [ref], following Putter [ref], proposed using time-to-event methods for analyzing and comparing the distribution functions. It is a one-step approach, accounting for interval censoring for individuals whose values lie between two discrete thresholds. Jones[17] suggested a related approach to that of Steinhauser et al. using a generalised non-linear mixed model, with multinomial likelihoods. The Jones model allows some flexibility in the functional form of the association between the cut-offs and logit(sensitivity) and logit(specificity), and thereby in the distribution of the underlying continuous marker, via a Box-Cox transformation. These two approaches also make different assumptions about the nature of between-study variability in test accuracy. [illegible] of thresholds per study. These approaches may also be used when information on all cut-offs is available. In this work, we consider both the Steinhauser and Jones approaches.

The overarching aim of the present study was to compare results from applying these
The overarching aim of the present study was to compare results from applying these approaches
The overarching aim of the present study was to compare results from applying these approaches
The overarching aim of the present study was to compare results from applying these approaches
The overarching aim of the present study was to compare results from applying these approaches
The overarching aim of the present study was to compare results from applying these approaches
The overarching aim of the present study was to compare results from applying these approaches
The overarching aim of the present study was to compare results from applying these approaches
The overarching aim of the present study was to compare results from applying these approaches
The overarching aim of the present study was to compare results from applying these approaches
The overarching aim of the present study was to compare results from applying these approaches

The overarching aim of the present study was to compare results from applying these approaches for meta-analyzing diagnostic accuracy studies in which reported results may be biased due to selective cut-off reporting (i.e., meta-analysis of only the published data) to approaches that use the *full IPD dataset*. To do this, we compared a series of approaches to meta-analyze the diagnostic accuracy of the Patient Health Questionnaire-9 (PHQ-9) depression screening tool. Specifically, we compared results from the Steinhauser and Jones approaches, which use statistical methods to account for missing cut-offs and were applied to data for published cut-offs only (as would be the case in a conventional, non-IPD meta-analysis), to results from a series of separate bivariate random effects models (BRM) that were estimated using IPD at each possible cut-off (i.e., using data for all cut-offs for all studies). In addition, to better understand whether any observed differences were due to using full versus published data as opposed to differences in models (Steinhauser et al., Jones et al., BRM), we compared results when the Steinhauser and Jones approaches were also applied to the *full IPD dataset* with data for all relevant cut-offs for every study.

Formatted: Font: Italic

Formatted: Font: Italic

METHODS

This is a secondary analysis that uses data from an IPD meta-analysis of the diagnostic accuracy of the PHQ-9 for screening to detect major depression.[3,18] Detailed methods of the IPD meta-analysis were registered in PROSPERO (CRD42014010673), and a protocol was published.[3]

We summarize the methods below. For the present study, the three approaches were conducted independently by three different research teams blind to the results from the other approaches.

Identification of Eligible Studies

Datasets from articles in any language were eligible for inclusion if they included diagnostic classifications for current Major Depressive Disorder (MDD) or Major Depressive Episode (MDE) based on a validated diagnostic interview conducted within two weeks of PHQ-9 administration among participants ≥18 years not recruited from youth or psychiatric settings. Datasets where not all participants were eligible were included if primary data allowed selection of eligible participants. For defining major depression, we considered MDD or MDE based on the Diagnostic and Statistical Manual of Mental Disorders (DSM), or MDE based on the International Classification of Diseases (ICD). If more than one was reported, we prioritized DSM over ICD, and DSM MDE over DSM MDD. Across all studies, there were only 23 discordant diagnoses that depended on classification prioritization (0.1% of participants).

Search Strategy and Study Selection

A medical librarian searched Medline, Medline In-Process & Other Non-Indexed Citations via Ovid, PsycINFO, and Web of Science from January 2000 through February 7, 2015, using a peer-reviewed search strategy.[19] We also reviewed reference lists of relevant reviews and queried contributing authors about non-published studies. Search results were uploaded into RefWorks (RefWorks-COS, Bethesda, MD, USA). After de-duplication, unique citations were uploaded into DistillerSR (Evidence Partners, Ottawa, Canada) for storing and tracking search results.

Two investigators independently reviewed titles and abstracts for eligibility. If either deemed a study potentially eligible, full-text article review was done by two investigators, independently, with disagreements resolved by consensus, including a third investigator as necessary. Translators were consulted to evaluate titles, abstracts and full-text articles for languages other than those for which team members were fluent.

Data Contribution and Synthesis

Authors of eligible datasets were invited to contribute de-identified primary data. Participant-level data included major depression status and PHQ-9 scores. When datasets included appropriate statistical weighting to reflect sampling procedures, we used the provided weights. For studies where sampling procedures merited weighting, but the original study did not weight, we constructed appropriate weights using inverse selection probabilities. The same sampling weighting was used in all analyses presented.

Data Used in the Present Study

In addition to the inclusion and exclusion criteria described above, we further required that included studies for this analysis published diagnostic accuracy results for at least one cut-off threshold, since the purpose of the present study was to compare methods of estimating results with published data versus IPD. Therefore, we did not consider any datasets that we had retrieved in the IPD project but for which no published data existed. Similarly, we did not consider any data from published studies for which the IPD could not be retrieved. For the eligible data, we constructed a dataset comprised of 2 x 2 tables (true positives, false positives, true negatives, false negatives) for only published cut-offs for each study, and we refer to this as the *published dataset*. We refer to the dataset that included results for all cut-offs for each eligible study, rather than just published cut-offs, as the *full IPD dataset*.

Ethical Approval

This study involved secondary analysis of anonymized previously collected data. As such, the Research Ethics Committee of the Jewish General Hospital declared that this project did not

require research ethics approval. For each included dataset, the authors confirmed that the original study received ethics approval and that all participants provided informed consent.

Differences Between IPD Dataset, Primary Datasets and Published Results

For some studies, data in the *published dataset* and *full IPD dataset* used in the present study differ from the data included in the originally published primary study reports. First, we applied the inclusion criteria for the main IPD meta-analysis to all subjects consistently. That is, in some primary studies, only some participants met the inclusion criteria for the main IPD meta-analysis. As an example, we required administration of the PHQ-9 index test and reference standard within a two-week period and only included participants aged 18 or older recruited from non-psychiatric settings. For some primary studies, we included participants who met these criteria, and excluded participants who did not. Second, we used one consistent outcome definition: major depression. Some primary studies reported accuracy results for depression diagnoses broader than major depression, such as "major + minor depression" or "any depressive disorder." We restricted our depression variable to major depression diagnoses as the reference standard diagnosis. Third, for studies where sampling procedures merited weighting, but the original study did not weight, we constructed appropriate weights using inverse selection probabilities. This occurred, for instance, when all patients with positive screens, but only a random subset of patients with negative screens, were administered a diagnostic interview. Fourth, as part of our data verification process, we compared published participant characteristics and diagnostic accuracy results with results obtained using the raw datasets. In cases where primary data that we received from investigators and original publications were discrepant, we identified and corrected errors in consultation with the original primary study investigators.

For the published data set, after applying the above exclusions and corrections, we estimated sensitivity and specificity for the cut-offs that were reported in the primary studies.

Statistical Analyses

First, to provide a baseline for what is often done in conventional meta-analyses, we estimated pooled sensitivity and specificity by applying the bivariate binomial-normal random effects model (BRM), similar to that of Chu and Cole[20], as described in Riley et al.,[20-23] to the *published* dataset for cut-offs 5-15 separately. The PHQ total score ranges from 0 to 27. A cut-off of 10 is generally used to classify someone as possibly depressed, requiring further investigation. While we used data from all cut-offs in the modelling process, we only present results for cut-offs from 5-15 as we believe these represent a range of potentially clinically useful cut-offs. For each cut-off, pooled sensitivity and specificity were estimated using data from only the studies that published diagnostic accuracy results for the cut-off.

Second, we addressed our main objective, which was to compare three approaches to estimating diagnostic accuracy in the context of missing outcomes for some cut-offs in published results: (1) the BRM with the *full IPD dataset*; (2) the Steinhauser model with the *published dataset*; and (3) the Jones model with the *published dataset*.

Third, to elucidate whether differences in results may have reflected differences in modelling approaches, rather than in use of the *published dataset* versus the *full IPD dataset*, we compared results obtained when the BRM, Steinhauser and Jones approaches were all applied to the *full IPD dataset*.

Each approach was applied separately by the 3 different teams blind to the results from the other teams. The team applying the BRM to the *full IPD dataset* (AB, BL, BDT) used the same modelling approach that they applied to the *published dataset*. The teams that applied the Steinhauser (GR, MS) and Jones (HEJ) models were initially only provided access to the *published*

dataset and conducted their analyses blind to the *full IPD dataset*. They were only given the *full IPD dataset* once results from the *published dataset* had been delivered.

The BRM is a two-stage random effects meta-analytic approach that models sensitivity and specificity at the same time, accounting for the inherent correlation between them and for the precision of estimates within studies. As the model is fitted separately at each cut-off, this approach does not account for the correlation across cut-offs within the same study, nor does it make any assumptions about the shape of the association between cut-off and sensitivity or specificity and in fact does not explicitly use cut-off information. The area under the full curve (AUC) was obtained by numerical integration based on the trapezoidal rule, and a 95% confidence interval for the AUC was estimated via bootstrap, resampling at the study and individual level.

The Steinhauser approach is a two-stage random effects model.[13] At the study level, for each observed cut-off, the observed values of specificity and 1 - sensitivity (corresponding to the proportion of negative test results for each group of participants) are logit-transformed. For the meta-analysis, the resulting values are fitted across studies using a linear mixed effects model. For the analyses in this paper we used model DIDS* ("Different random Intercepts and Different random Slopes").[13] It is given by

$$\text{logit}(\widehat{sp}_{kt}) = \alpha_0 + a_{0k} + (\beta_0 + b_{0k}) \log(c_{kt}) + \epsilon_{kt}$$

$$\text{logit}(1 - \widehat{se}_{kt}) = \alpha_1 + a_{1k} + (\beta_1 + b_{1k}) \log(c_{kt}) + \delta_{kt}$$

where $\widehat{sp}_{kt}$ and $\widehat{se}_{kt}$ denote the observed values of specificity and sensitivity at threshold c_{kt} in study k, α_0 and α_1 are fixed intercepts, and β_0 and β_1 are fixed slopes for the disease-free and the diseased individuals, respectively. The terms a_{0k}, a_{1k}, b_{0k}, b_{1k} denote random intercepts and slopes that are assumed to follow a common multivariate normal distribution, reflecting the correlation across studies, and ϵ_{kt} and δ_{kt} represent within-study random errors. Each data point (i.e., $\widehat{sp}_{kt}$ or $\widehat{se}_{kt}$ for all studies k and all thresholds c_{kt}) is weighted with the inverse variance of the respective logit-transformed proportion. In contrast to the original implementation[13], the published R package diagmeta[24] explicitly accounts for estimation uncertainty in the first stage.

This model was selected as it showed the smallest AIC under all models that allowed the variance to differ between the groups.[13] The model-based distribution functions for non-diseased and diseased individuals are obtained by back-transformation of the fixed effects part of the model. This also provides a model-based ROC curve and 95% confidence interval. The area under the full curve (AUC) is obtained by numerical integration based on the trapezoidal rule.

The Jones et al.[17] approach models the numbers of diseased and non-diseased individuals with test results above each reported threshold using (conditional) binomial likelihoods. This is equivalent to multinomial likelihoods for the full categorisation of test results in each study (number with result < cut-off 1, number with result between cut-off 1 and cut-off 2, etc.). No normal approximations are therefore required. A one stage approach is used. The model assumes that test results or some monotonic transformation of test results, g(.), (e.g., the natural logarithm) has a logistic distribution, in each of the diseased and non-diseased populations:

$$g(y_{ijk}) \sim Logistic(\mu_{jk}, \sigma_{jk})$$

where y_{ijk} is the test result, with i denoting the individual, j=0,1 denoting the disease group, and k denoting the study. Here, μ_{jk} and σ_{jk} are the mean and scale parameters for disease group j and study k. Since g(.) is monotonic, it follows then that:

$$\text{logit}(1 - sp_{kt}) = \frac{\mu_{0k} - g(c_{kt})}{\sigma_{0k}}$$

$$\text{logit}(sp_{kt}) = \frac{\mu_{1k} - g(c_{kt})}{\sigma_{1k}}$$

where c_{kt} is the cut-off for t=1, …. T_k.

The means and log-scale parameters of these distributions are assumed to be normally distributed random effects across studies. The 'summary' sensitivity and specificity at each threshold is estimated by evaluating the equations above at the means of these four sets of random effects. Rather than pre-specifying the transformation, g(.), of test results that has an approximate logistic distribution, the Jones approach can estimate a Box-Cox transformation parameter λ from the data, simultaneously to performing the meta-analysis. For example, a value of λ = 1 corresponds to the identity function (such that underlying test results have a symmetric, logistic distribution) while λ = 0 corresponds to the natural logarithm (such that test results have a log-logistic distribution). We estimated a separate Box-Cox transformation parameter for the depressed and non-depressed populations. The Jones approach was fitted in a Bayesian framework.

At each step and for each approach, we estimated pooled sensitivity and specificity and associated 95% confidence intervals (credible intervals for the Jones approach) for each cut-off, as well as the AUC across the full range of possible cut-offs (0 to 27). We compared point estimates, confidence interval widths, and AUC between methods and datasets.

All BRM analyses were run in R (R Version 3.4.1 and R Studio Version 1.0.143) using the lme4 package.[25-27] The Steinhauser model was implemented in the R package diagmeta (Version 0.3-1[24]) in the software environment R, Version 3.6.1.[26] The Jones model was fitted using Bayesian statistical software, WinBUGS.[28] Example code for each approach is provided in the Supplemental Materials.

RESULTS

Search Results and Inclusion of Primary Datasets

Of 5,248 unique titles and abstracts identified from the database search, 5,039 were excluded after title and abstract review and 113 after full-text review, leaving 96 eligible articles with data from 69 unique participant samples, of which 55 (80%) contributed datasets (Figure 1). Authors of included studies contributed data from three unpublished studies. We excluded 13 datasets for the present study that did not publish diagnostic accuracy results for any PHQ-9 cut-offs, leaving a total of 45 studies. (See Figure 1).

Among the 45 studies, 20 (44%) used the Structured Clinical Interview for DSM Disorders (SCID), 11 (24%) used the Mini International Neuropsychiatric Interview (MINI), 8 (18%) used the Composite International Diagnostic Interview (CIDI) and 6 (13%) used other types of diagnostic interviews.

Description of Included Studies

Table 1 shows the numbers of studies, subjects and true cases of major depression for the *full IPD* dataset and for the *published dataset*, for clinically relevant cut-offs 5 to 15. The *full IPD dataset* comprised data from 45 studies, 15,020 subjects and 1,972 cases of major depression, and included information on every possible cut-off. When applying the BRM to the published dataset, the size of the *published dataset* varied by cut-off with as few as 8 studies (2,007 participants and 397 major depression cases) when the cut-off was 14. At the most common cut-off of 10, the *published* dataset included data from 37 studies with 13,375 participants and 1,738 major depression cases. The total number of studies in the *published* dataset was 45. Note that the

Formatted: Font: Italic

published data set is a subset of the *full IPD data-set* in that the full IPD has information on every cut-off for all eligible studies, whereas the published data includes information only on cut-offs which were investigated in the primary studies. (See Table 1). Figure 2 displays observed sensitivity and specificity for each cut-off, by primary study, and indicates which cut-offs were published for each study, providing an impression of the mixture of study-level distributions of PHQ-9 values over all studies. In the Supplemental Materials we present the overall distribution of PHQ9 scores in depressed and non-depressed subjects (Figure S1) and fitted distributions for each approach when using the *published* or *full IPD datasets* (Figure S2).

Sensitivity and Specificity

All sensitivity and specificity results with 95% confidence intervals (BRM and Steinhauser approaches) or credible intervals (Jones approach) for clinically relevant cut-offs 5-15 for the *published dataset* and *full IPD dataset* are shown in the Supplemental Materials for each approach (Tables S1-S3).

Figure 3 compares the BRM, Jones and Steinhauser approaches applied to the *published dataset* with the BRM approach applied to the *full IPD dataset*. The left-hand panel shows that applying the BRM to the *published dataset* underestimates sensitivity for lower cut-offs and overestimates it for higher cut-offs (average absolute difference: 0.06) compared to the BRM with the *full IPD dataset*. Moreover, sensitivity appears to increase with increasing cut-offs for some sections of the curve, which is logically impossible. Compared to the BRM applied to the *full IPD dataset*, both the Steinhauser and Jones approaches applied to the *published dataset* estimate slightly lower sensitivity for all cut-offs (average absolute difference: 0.02, range: 0.01-0.03 for both approaches).

The right-hand panel demonstrates that specificity estimated from the BRM applied to the *published dataset* is higher, but much closer to that estimated with the BRM applied to the *full IPD dataset* than sensitivity (average absolute difference: 0.02, range: 0.00-0.07). The Steinhauser and Jones approaches applied to the *published dataset* estimated specificity very similarly to that estimated from the BRM applied to the *full IPD dataset*.

Figure 4 compares the Jones and Steinhauser approaches applied to the *full IPD dataset* with the BRM approach applied to the *full IPD dataset*. The Steinhauser approach (blue) estimates of sensitivity (left side) and specificity (right side) at lower cut-offs are lower than the BRM with the degree of difference decreasing as the cut-off increases. For both sensitivity and specificity, the magnitude by which estimates are lower goes from 4% at the lowest cut-off to 0% at the highest. The Jones approach (green) generates estimates of sensitivity that are similar to the BRM at lower cut-offs but become lower as the cut-off increases. The magnitude by which estimates are lower goes from 0% at the lowest cut-off to 4% at the highest. The Jones approach had quite similar specificity across the range of cut-offs, with a maximal difference of 3%.

The Steinhauser approach and the Jones approach produced similar estimates of sensitivity when applied to the *published* vs. *full IPD dataset*, differing (0.01 to -0.03 for the Steinhauser approach; and -0.01 to 0.01 for the Jones approach). For both approaches, the differences were larger for specificity. Estimates of specificity were higher applying the Steinhauser approach to the *published dataset*, than for the *full IPD dataset*, and this difference decreased as the cut-off increased (0.05 to 0). For the Jones approach, estimates of specificity were lower when applied to the *published dataset*, than to the *full IPD dataset*, and this difference decreased as the cut-off increased (-0.04 to 0.01). (See Supplemental Tables S3 and S4).

Confidence/Credible Interval Width

Supplemental Figures S3 and S4 present estimated sensitivity and specificity and 95% confidence or credible intervals by cut-off, for the BRM, Jones and Steinhauser methods applied to the *full IPD dataset* and *published dataset*, respectively.

For the BRM and Steinhauser approaches, for sensitivity, confidence intervals estimated using the *full IPD dataset* were narrower than those estimated using the *published* data, as expected (5.5% and 4.7% narrower, respectively), whereas the credible intervals estimated via the Jones approach were of similar width for both datasets (1% average difference across all cut-offs), indicating that the inclusion of the additional data increased precision minimally. For specificity, confidence or credible interval widths were more similar when using the *full IPD dataset* or *published dataset* (within 2%, 1% and 1% for the BRM, Steinhauser and Jones approaches, respectively).

When using the *full IPD dataset*, sensitivity confidence interval widths were largely similar between the three approaches and the interval widths increased as the cut-off increased (and number of cases decreased) for all approaches.

ROC Curves and AUC

Figure 5 compares the ROC curves for the BRM, Steinhauser and Jones methods using the *published dataset* as compared to the BRM on the *full IPD* dataset. For the *published dataset*, the AUC was 0.90 (95% confidence interval (CI): 0.84, 0.92) for the BRM; 0.89 (95% CI: 0.86, 0.92) for the Steinhauser approach, and 0.89 (95% credible interval not available) for the Jones approach. The BRM approach applied to the *published dataset* produced an empirical ROC curve that deviated substantially from that obtained when using the *full IPD dataset*. The Jones and Steinhauser methods produced generally similar ROCs as compared to that using the BRM on the *full IPD dataset*, with slight levelling off observed at lower cut-offs, and slightly lower AUCs.

Figure 6 compares the ROC curves for the three approaches applied to the *full IPD dataset*. With the *full IPD dataset*, the ROC for the Steinhauser approach was slightly lower at lower cut-offs as compared to that for the BRM approach. The ROC for the Jones approach on the other hand, was a bit lower at upper cut-offs. The AUC for the *full IPD dataset* was very similar for all three approaches: 0.91 (95% CI: 0.89, 0.94) for the BRM, 0.89 (95% credible interval not available) for the Jones approach and 0.88 (95% CI: 0.85, 0.90) for the Steinhauser approach.

Agreeing with what we observed in the IPD, the Jones model fitted to the published data only estimated the Box-Cox transformation parameter (λ) to be 0.57, 95% Credible Interval 0.45 to 0.69 in the non-depressed population, indicating right-skew and 0.82, 95% CI (0.50-1.10) in the depressed population. Constraining λ to equal 1 (i.e., assumed symmetrical underlying distributions) gave very similar results across the cut-offs of clinical interest (not shown).

DISCUSSION

When attempting to meta-analyze diagnostic accuracy scores, relying only on information from cut-offs presented in the original primary studies results in biased estimates of sensitivity and specificity at some cut-offs and ROC curves that display logically impossible shapes.[2] One option to estimate pooled sensitivity and specificity across all cut-offs not biased by selective cut-off reporting is to collect IPD, and thus have and analyse information from every study at every cut-off. A less costly and labour-intensive solution is to account for information from missing cut-offs statistically. In this work, we have empirically compared these methods using IPD from 45 studies on the diagnostic accuracy of the PHQ-9 depression screening tool.

One advantage of the Steinhauser and Jones methods is that IPD are not required. These models estimate the sensitivity and specificity across all cut-offs in all studies, regardless of which data were reported in the primary studies. As such, our primary comparison was between these methods applied to data that would have been collected via conventional aggregate data meta-analysis to a bivariate random effects model applied to IPD that included information on all relevant cut-offs. To clarify which differences were due to data availability versus differences in the specifications of the models, we further compared these approaches when applied to the IPD.

We found that when using data available in publications, using statistical methods to estimate accuracy of the PHQ-9 out-performed the BRM applied to the same data. The Steinhauser and Jones methods resulted in sensible ROC curves, close to those obtained via The

Steinhauser and Jones methods resulted in sensible ROC curves, close to those obtained via the same approaches or the BRM on the *full IPD dataset*, with only small diversions at low cut-offs. Confidence (or credible) intervals were on average slightly narrower using the Steinhauser and Jones methods than those estimated via the BRM applied to the published dataset: this is to be expected, since these modelling approaches 'borrow strength' from data relating to other cut-offs.

When we applied the different methods to the *full IPD dataset*, we found that the Steinhauser method estimated slightly lower sensitivity and specificity at low cut-offs than the BRM and that the difference decreased as the cut-off increased (4% at the most for sensitivity and 5% at the most for specificity). On the other hand, the Jones method estimated slightly lower sensitivity at higher cut-offs (4% at the most), and the degree of difference decreased as the cut-off decreased; the Jones approach estimated slightly higher specificity at lower cut-offs (3% at the most), and the difference decreased as the cut-off increased. Overall, the ROC curve estimated via the Jones method was quite similar to the ROC curve estimated via the BRM, with only small diversions at high cut-offs, while the Steinhauser estimated ROC curve had larger diversions at lower cut-offs. As compared to the BRM, confidence interval widths were narrower at most cut-offs when the Steinhauser method was used, and much narrower when the Jones method was used. This seems intuitive, since these approaches simultaneously model all of the data, whereas the BRM approach involves performing meta-analysis separately at each cut-off.

Steinhauser et al. [13] evaluated their method via a simulation study and found that large values of sensitivity and specificity were underestimated, and vice versa. Consistent with their findings, we observed lower estimates at low cut-offs for sensitivity when using the Steinhauser approach as compared to the BRM on the *full IPD dataset*, and also that that difference decreased as the sensitivity increased. However, for specificity, it was less clear.

The Steinhauser, Jones and BRM approaches make a number of different assumptions that could have resulted in differences in the ROC curves estimated from the *full IPD dataset*:

First, the Steinhauser approach starts by assuming that the distribution of the underlying continuous marker (here, the ordinal PHQ-9 score) is normally or logistically distributed in both diseased and non-diseased persons. While the validity of this assumption in our data is questionable, given that non-diseased subjects have a very skewed PHQ-9 score distribution, relaxing this assumption (assuming instead underlying log-logistic distributions in both populations) did not change results appreciably (not shown). The Jones model assumes that some unspecified Box-Cox transformation of test results has a logistic distribution.

Second, the Steinhauser model's distributional assumptions correspond to assuming the cut-off has a linear association with logit(sensitivity) and logit(specificity). The Jones approach is more flexible, assuming only that logit(sensitivity) and logit(specificity) have a linear relationship

Formatted: Font: Italic

Formatted: Font: Italic

with some unspecified Box-Cox transformation of the cut-off. In contrast, the BRM does not assume any parametric form for these relationships.

Third, the BRM is based on the binomial distribution, whereas the Steinhauser method first transforms estimates of sensitivity and specificity, and then uses a linear mixed model. The Jones method is based on the multinomial distribution, the generalisation of the binomial distribution to model test result data across more than one threshold. Where individual studies report data at only one threshold, the model reduces to binomial distributions. Fourth, the Steinhauser approach makes a continuity correction to deal with sensitivities and specificities at 0 or 1. This is not required for the Jones or BRM approaches. Fifth, the Steinhauser approach weights studies by the inverse variance. The BRM, and Jones approaches do not need explicit weighting. Finally, the Jones and Steinhauser approaches account for the correlation between sensitivities and specificities across cut-offs within studies, while the BRM does not.

We do not believe that the different assumptions made by the approaches with respect to the complicated dependencies in the data are the source of the differences in the estimated ROC curve. Simoneau et al. compared the BRM method to an approach proposed by Putter et al. that takes into account the complicated dependence structure empirically in a subset of this data and via a limited simulation study.[10,29] In that work, they found little impact of within-study correlation across cut-offs on bias in the estimated sensitivity and specificity. However, coverage of confidence intervals in the BRM was affected. [29] Indeed, the differences in confidence interval widths that we observed are likely due to this difference.

We hypothesize that the most likely explanation for the small differences between estimates derived from the BRM vs. those derived from the Steinhauser and Jones methods applied to the full IPD is that while the fit of the BRM assumes no parametric association between cut-off and sensitivity or specificity, the Steinhauser and Jones methods both maximize a global fit, and assume a parametric form between the cut-off and logit(sensitivity) and logit(specificity). In this case, the Steinhauser approach results in a curve for sensitivity that may fit better at higher cut-offs (where most of the data is for those with major depression) but perhaps fits less well at lower cut-offs. The Jones approach may do a better job at recovering the true curve because of the ability to tailor the transformation to the data.

Although the BRM model is recommended in the Cochrane Handbook for Diagnostic Test Accuracy Systematic Reviews,[30] it should not be regarded as the gold standard for analyzing data of this type. The BRM model ignores within-study correlation across cut-offs, leading to the development of new approaches.[9,10,12-14] These novel approaches may be preferable when analyzing diagnostic test accuracy data from the full range of cut-offs (sometimes called "full ROC curves"). This is because these methods analyze all of the data at once, make explicit use of the cut-offs and account for the complex dependencies that exist, although these also come with additional assumptions that may or may not be met.

One limitation of this work is that we compared only three possible approaches. While ~~S~~some approaches were not useable given the patterns of missing data in our data set~~.~~,[9,10,15] the approach proposed by Hoyer et al.[12] could have been used. ~~However~~Nevertheless, the three data analysis teams operated independently, and results for the *published dataset* were produced before access to the *full IPD dataset* was provided to the Jones et al. and Steinhauser et al. groups. Also, we have focussed here on the case of ordinal measurements. In the case of a continuous biomarker, the BRM could be used by considering several given cut-offs. For the Steinhauser approach, the R package diagmeta[24] allows entering ordinal or continuous data at the participant level. The data is transformed to study-level data (number of true positives, false positives, false negatives, and

Formatted: Highlight

Formatted: Highlight

Formatted: Highlight

true negatives) for each cut-off of interest, or even all observed values. Similarly, the Jones approach is easily applicable to continuous measures, by fitting the same model using the continuous likelihoods directly.

This work was based on an empirical comparison of several methods. As such, an important limitation is that we do not know the truth. Only a simulation study could address this limitation. However, we have applied three different methods to a large, IPD dataset and compared the results when the *full IPD dataset* was used, or when only the subset that included cut-offs that could have been collected via conventional aggregate data meta-analysis was used.

Formatted: Highlight

Overall, we have shown that the Steinhauser and Jones methods using statistical methods to fill in missing cut-off data recovered the ROC curve from the *full IPD dataset* well when using only the *published* subset. How best to analyze IPD that include data from every cut-off remains unknown: different approaches make different assumptions that may or may not be met in any given data set.

In this work we considered as the *full IPD dataset* data that included information on every relevant cut-off from every study. While this is valuable information that removes bias due to selective cut-off reporting, the IPD collected also includes individual patient information such as age, sex, and comorbidities, that can be used in important ways. For example, IPD permit the evaluation of diagnostic accuracy in key subgroups. Moreover, collecting IPD permitted us to standardize eligibility criteria for study participants, restrict our study populations to subjects who received the screening questionnaire and diagnostic interview within two weeks of each other and to use a consistent outcome definition. All of these may reduce heterogeneity across studies and ensure that the most accurate estimates of diagnostic accuracy are obtained.

Details of Contributors

AB, BL, GR, HEJ, and BDT contributed to the conception and design of the study, participated in the data analysis, and helped to draft the manuscript. MS helped to draft the manuscript and provided critical revisions. JPAI contributed to the conception and design of the study and provided critical revision on the manuscript. All authors read and approved the final manuscript.

Funding information

This work was funded by the Canadian Institute of Health Research (CIHR; KRS-134297, PCG-155468, PJT-162206. AB and BDT were supported by the Fonds de recherche de Québec – Santé researcher salary awards. BL was supported by a CIHR Frederick Banting and Charles Best Canada Graduate Scholarship doctoral award. GR is funded by the German Research Foundation (DFG), grant no. RU 1747/1-2. HEJ was supported by a Medical Research Council (MRC) Career Development Award in Biostatistics (MR/M014533/1).

Data Availability

The data that support the findings of this study are available upon reasonable request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions in agreements with individual data contributors.

Acknowledgements

None.

Conflicts of Interest

The authors declare that there is no conflict of interest.

Highlights

What is already known?

There are several approaches that might be appropriate to meta-analyze data that consists of diagnostic accuracy at several cut-offs per study, when these data may suffer from selective cut-off reporting.

What is new?

We compared the usual approach (fitting separate bivariate random effects models at each cut-off) to methods that used statistical models to account for information from missing cut-offs.

Findings

Using statistical methods to fill in missing cut-off data recovered the receiver operating characteristic (ROC) curve from the *full IPD dataset* well when using only the *published* subset. All methods performed similarly when applied to IPD that included data from every cut-off.

References

1. Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. *BMJ.* 1994;308(6943):1552.
2. Levis B, Benedetti A, Levis AW, et al. Selective Cutoff Reporting in Studies of Diagnostic Test Accuracy: A Comparison of Conventional and Individual-Patient-Data Meta-Analyses of the Patient Health Questionnaire-9 Depression Screening Tool. *Am J Epidemiol.* 2017;185(10):954-964.
3. Thombs BD, Benedetti A, Kloda LA, et al. The diagnostic accuracy of the Patient Health Questionnaire-2 (PHQ-2), Patient Health Questionnaire-8 (PHQ-8), and Patient Health Questionnaire-9 (PHQ-9) for detecting major depression: protocol for a systematic review and individual patient data meta-analyses. *Systematic reviews.* 2014;3:124.
4. Brennan C, Worrall-Davies A, McMillan D, Gilbody S, House A. The Hospital Anxiety and Depression Scale: a diagnostic meta-analysis of case-finding ability. *Journal of psychosomatic research.* 2010;69(4):371-378.
5. Manea L, Gilbody S, McMillan D. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. *CMAJ.* 2012;184(3):E191-196.
6. Moriarty AS, Gilbody S, McMillan D, Manea L. Screening and case finding for major depressive disorder using the Patient Health Questionnaire (PHQ-9): a meta-analysis. *General hospital psychiatry.* 2015;37(6):567-576.
7. Mitchell AJ, Meader N, Symonds P. Diagnostic validity of the Hospital Anxiety and Depression Scale (HADS) in cancer and palliative settings: a meta-analysis. *Journal of affective disorders.* 2010;126(3):335-348.
8. Deeks JJ, Bossuyt P, Gastonis C. Cochrane handbook for systematic reviews of diagnostic test accuracy, Version 1.0.0, The Cochrane Collaboration. In: srdta.cochrane.org; 2013.
9. Hamza TH, Arends LR, van Houwelingen HC, Stijnen T. Multivariate random effects meta-analysis of diagnostic tests with multiple thresholds. *BMC Med Res Methodol.* 2009;9:73.
10. Putter H, Fiocco M, Stijnen T. Meta-analysis of diagnostic test accuracy studies with multiple thresholds using survival methods. *Biometrical journal Biometrische Zeitschrift.* 2010;52(1):95-110.
11. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ.* 2010;340:c221.
12. Hoyer A, Hirt S, Kuss O. Meta-analysis of full ROC curves using bivariate time-to-event models for interval-censored data. *Res Synth Methods.* 2018;9(1):62-72.
13. Steinhauser S, Schumacher M, Rücker G. Modelling multiple thresholds in meta-analysis of diagnostic test accuracy studies. *BMC Med Res Methodol.* 2016;16(1):97.
14. Riley R, Takwoingi Y, Trikalinos T, et al. Meta-analysis of test accuracy studies with multiple and missing thresholds: a multivariate normal model *J Biomed Biostat.* 2014;5(100).
15. Ensor J, Deeks JJ, Martin EC, Riley RD. Meta-analysis of test accuracy studies using imputation for partial reporting of multiple thresholds. *Res Synth Methods.* 2018;9(1):100-115.
16. Rücker G, Steinhauser S, Schumacher M. RE: "Selective Cutoff Reporting in Studies of Diagnostic Test Accuracy: A Comparison of Conventional and Individual-Patient-Data Meta-Analyses of the Patient Health Questionnaire - 9 Depression Screening Tool". *Am J Epidemiol.* 2017;186(7):894.
17. Jones HE, Gatsonis CA, Trikalinos TA, Welton NJ, Ades AE. Quantifying how diagnostic test accuracy depends on threshold in a meta-analysis. *Statistics in Medicine.* In Press.

18. Levis B, Benedetti A, Thombs BD, Collaboration DESD. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. *BMJ.* 2019;365:l1476.
19. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement. *J Clin Epidemiol.* 2016;75:40-46.
20. Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *J Clin Epidemiol.* 2006;59(12):1331-1332; author reply 1332-1333.
21. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol.* 2005;58(10):982-990.
22. Riley R, Dodd S, Craig J, Thompson J, Williamson P. Meta-analysis of diagnostic test studies using individual patient data and aggregate data. *Statistics in Medicine.* 2008;27(6111):6136.
23. Riley RD, Abrams KR, Sutton AJ, Lambert PC, Thompson JR. Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC MedRes Methodol.* 2007;7:3.
24. Rücker G, Steinhauser S, Kolampally S, Schwarzer G. diagmeta: Meta-analysis of diagnostic accuracy studies with several cutpoints. *R package.* 2019.
25. RStudio Team. *RStudio: Integrated development for R.* Boston, MA: RStudio, Inc.; 2015.
26. R Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing; 2018.
27. Bates D, Machler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software.* 2015;67(1):1-48.
28. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing.* 2000;10:325-337.
29. Simoneau G, Levis B, Cuijpers P, et al. 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 Biometrische Zeitschrift.* 2017;59(6):1317-1338.
30. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. In: Deeks JJ, Bossuyt P, Gatsonis C, eds.: The Cochrane Collaboration; 2013.