Received: 30 March 2020 | Revised: 14 October 2020 | Accepted: 24 October 2020

DOI: 10.1002/bimj.202000091

Biometrical Journal

CASE STUDY

Meta-analysis of diagnostic accuracy studies with multiple thresholds: Comparison of different approaches

Antonia Zapf[1] | Christian Albert[2,3] | Cornelia Frömke[4] | Michael Haase[2,3] | Annika Hoyer[5] | Hayley E. Jones[6] | Gerta Rücker[7]

[1] Department of Medical Biometry and Epidemiology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

[2] Faculty of Medicine, Otto-von-Guericke University, Magdeburg, Germany

[3] Diaverum Renal Services Germany, MVZ Am Neuen Garten, Potsdam, Germany

[4] Department of Information and Communication, Faculty for Media, Information and Design, University of Applied Sciences and Arts Hannover, Hannover, Germany

[5] Department of Statistics, Ludwig-Maximilians-University Munich, Munich, Germany

[6] Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

[7] Institute of Medical Biometry and StatisticsFaculty of Medicine and Medical Center – University of Freiburg, Freiburg, Germany

Correspondence
Antonia Zapf, Department of Medical Biometry and Epidemiology, University Medical Center Hamburg-Eppendorf, Martinistr. 52, 20251 Hamburg, Germany.
Email: a.zapf@uke.de

RR Reproducible Research

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract

Methods for standard meta-analysis of diagnostic test accuracy studies are well established and understood. For the more complex case in which studies report test accuracy across multiple thresholds, several approaches have recently been proposed. These are based on similar ideas, but make different assumptions. In this article, we apply four different approaches to data from a recent systematic review in the area of nephrology and compare the results. The four approaches use: a linear mixed effects model, a Bayesian multinomial random effects model, a time-to-event model and a nonparametric model, respectively. In the case study data, the accuracy of neutrophil gelatinase-associated lipocalin for the diagnosis of acute kidney injury was assessed in different scenarios, with sensitivity and specificity estimates available for three thresholds in each primary study. All approaches led to plausible and mostly similar summary results. However, we found considerable differences in results for some scenarios, for example, differences in the area under the receiver operating characteristic curve (AUC) of up to 0.13. The Bayesian approach tended to lead to the highest values of the AUC, and the nonparametric approach tended to produce the lowest values across the different scenarios. Though we recommend using these approaches, our findings motivate the need for a simulation study to explore optimal choice of method in various scenarios.

KEYWORDS

bivariate endpoint, meta-analysis, multiple thresholds, sensitivity, specificity

1 | INTRODUCTION

Diagnostic tests are often based on biomarkers, psychiatric scales, or risk scores that are measured on a continuous, discrete, or ordinal scale. These tests can be implemented at varying diagnostic thresholds (i.e. value such that results, say, greater than or equal to the threshold are called 'positive' and those less than the threshold 'negative'). The aim of diagnostic test accuracy (DTA) studies is to estimate the accuracy of such diagnostic tests. In systematic reviews, the results from two or more DTA studies are combined. In a primary DTA study, each individual provides two pieces of information: their test value (on any scale) and their true disease status (with/without the target condition, for example, a disease, herein referred to as 'diseased' and 'disease-free'), which is usually assumed to be measured without error by a reference or 'gold standard' test. All individuals in a study together therefore provide empirical distributions of test values in each disease group. However, instead of reporting the individual values or distributions, publications of primary studies usually report only a few two-by-two tables, or pairs of sensitivity (true positive fraction) and specificity (true negative fraction) corresponding to a small number of thresholds. Researchers also often display receiver operating characteristic (ROC) curves, which are plots of sensitivity against 1 − specificity across all possible thresholds, and/or the area under the ROC curve (AUC). For an overview of these and further diagnostic accuracy measures, see, for example, Eusebi (2013).

There are well-established methods for meta-analysis of DTA, namely the hierarchical summary receiver operating characteristic curve model proposed by Rutter and Gatsonis (2001) and the bivariate model proposed by Reitsma et al. (2005) and amended by Chu and Cole (2006). These approaches are fully equivalent if no covariates are considered (Harbord, Deeks, Egger, Whiting, & Sterne, 2007). Though both models allow for heterogeneity and correlation resulting from variation in threshold across studies, they do not use the actual numerical values of these thresholds. This means it is unknown what threshold value any pooled estimate (or point on the summary ROC curve) corresponds to. Further, these approaches can only accommodate a single two-by-two table (i.e. test accuracy measures at a single threshold) from each study, whereas primary studies often report at multiple thresholds. Choosing just one of these pairs of data to input into the meta-analysis would lead to a heavy loss of information (Trikalinos et al., 2012).

Whereas the *Cochrane Handbook for DTA Reviews* currently recommends using standard meta-analysis methodology separately for each reported threshold (Macaskill, Gatsonis, Deeks, Harbord, & Takwoingi, 2010), more advanced and specialized statistical methods to address these issues through a unified analysis are now available. There are many approaches for the analysis of full ROC curves. However, we have not considered those with relevant limitations, such as, for example, the requirement of identical thresholds across studies (for a detailed discussion, see Hoyer, Hirt, & Kuss, 2018) and have therefore chosen the approaches from Hoyer et al. (2018), Jones, Gatsonsis, Trikalinos, Welton, and Ades (2019), Frömke, Kirstein, and Zapf (2020), and Steinhauser, Schumacher, and Rücker (2016). The desirability of such approaches has also been expressed in a *Cochrane review* (Heazell et al., 2019) and in a recent simulation study (Vogelgesang, Schlattmann, & Dewey, 2018).

These more advanced approaches all view the problem from the following perspective: in each study, test results have different distributions among diseased and disease-free persons. The cumulative distribution function (cdf) of test results in the disease-free individuals defines the specificities across the full range of possible thresholds, while the cdf in the diseased individuals defines the full range of false negative fractions (= 1 – sensitivities). The task of a meta-analysis of DTA studies is to use the available data to estimate a 'summary' cdf in each of the two populations, which provides summary sensitivities and specificities across a sensible range of thresholds.

The starting point for this article was a data challenge on the dataset used here (see Section 2) at a conference. The aim of the data challenge was to apply four such approaches to a case study dataset and to compare the results. We found that results were quite heterogeneous. We wanted to share this knowledge first and then systematically investigate it in a second step by means of a simulation study (see Sections 5 and 6). In the following section the case study data and the underlying medical question are explained. After giving an overview of the individual approaches in Section 3, the results are presented in Section 4 and discussed in Section 5. We finish with some concluding remarks and suggested next steps.

2 | CASE STUDY META-ANALYSIS

The case study meta-analysis comes from the field of nephrology. The aim was to evaluate the diagnostic accuracy of neutrophil gelatinase-associated lipocalin (NGAL) as a test for acute kidney injury (AKI). Haase-Fielitz, Haase, and Devarajan

TABLE 1 Number of primary studies in the different scenarios

Sampling material	AKI	Severe AKI	RR
Plasma	18	16	12
Urine	12	10	9

(2014) systematically reviewed the utility of NGAL for the diagnosis of AKI. The review found 58 relevant articles overall, and the authors calculated raw (un-weighted) mean sensitivity and specificity across all studies in each clinical setting (cardiac surgery versus critical care/emergency department) and with urine and blood as sampling material. Despite this consideration of influencing factors, the results were extremely heterogeneous (within one scenario study-specific AUC values varied between 0.50 and 0.99). The authors identified variation in thresholds and in reference standard (based on different AKI definitions) as key potential reasons for heterogeneity. Therefore, Albert et al. (2020b) conducted a successional meta-analysis, contacted the authors of identified primary studies and asked for additional information and data to allow exploration of this. The authors of 26 primary studies responded and provided 30 reassessed datasets. The quality assessment using the QUADAS-2 tool (Whiting et al., 2011) demonstrated overall a low risk of bias and high applicability (Albert et al., 2020b).

Using a consensus AKI definition based on the RIFLE criteria (risk, injury, failure, loss, end-stage renal disease), three event types were distinguished: AKI, severe AKI and renal replacement therapy (RRT). Whereas the outcome measures AKI and severe AKI are based on consensus classification criteria (Bellomo, Ronco, Kellum, Mehta, & Palevsky, 2004), the assessment of RRT as outcome measure is limited by the absence of a consented gold standard on whether and when to initiate RRT. Therefore, clinical practice variability may contribute to pronounced variability of thresholds (Klein et al., 2018). Data were stratified by these three alternative 'reference standards', which can clearly be expected to lead to differences in estimates of sensitivity and specificity. Some studies provided data relating to two or three of these reference standards. Data were also stratified by whether NGAL was measured in urine or in blood (in the following, referred to as sampling material). In secondary analyses, the effect of the clinical setting (see above) and of the use or non-use of the urine output criteria for the assessment of the AKI stage was investigated. However, these secondary analyses will not be considered here. Therefore, in total we consider six separate but related datasets: the number of contributed individual datasets according to each endpoint is displayed in Table 1. Our meta-analyses for these six scenarios were performed separately.

For each primary study, sensitivity and specificity estimates were requested and provided for three different thresholds: such that (1) the estimated sensitivity is at least 95%; (2) the estimated specificity is at least 95%; and (3) the estimated sum of sensitivity and specificity is maximized (Youden criterion for 'optimal' threshold; Schisterman & Perkins, 2007). NGAL levels peak approximately 6 h after kidney tubular injury and followed a dose–response curve with respect to severity of injury (Haase-Fielitz et al., 2009; Mishra et al., 2005). Accordingly, the thresholds for which data were provided had a wide range (up to 10,000 ng/mL), independently of the event type, since a specific number of patients included in the outcome measure 'severe AKI' overlap with 'AKI' and to a lesser extend with 'RRT'. Albert et al. (2020b) found that the median threshold were lower for urine than for plasma and all threshold concentrations increased with increasing AKI severity or requirement of RRT. Figure 1 shows weighted NGAL threshold concentrations separately for the endpoints and sample materials. The number of outliers was low and represented studies with low patient numbers.

3 | METHODS

In this section, the different approaches are briefly summarized; for full details we refer to the original publications. As far as possible we use a common notation for all approaches.

3.1 | Notation

We index the studies in each dataset by $k = 1, \dots, N$ and the true disease state by $d = 0$ (disease-free individuals) and $d = 1$ (diseased individuals). In study k, status group d, there are n_{kd} participants. We index these individual study participants by $s = 1, \dots, n_{kd}$. The actual continuous test result of participant s in status group d in study k is denoted by

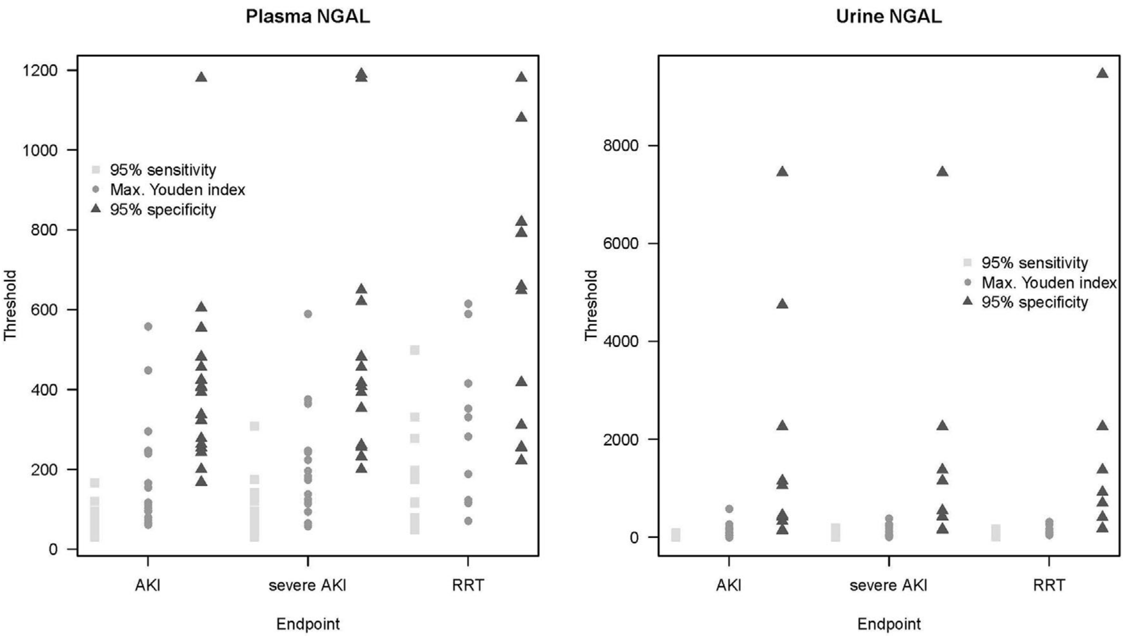

FIGURE 1 Distribution of the thresholds (in ng/mL) at which data were provided across studies, for the three endpoints and three criteria for plasma (left panel) and urine (right panel). Each boxplot includes the number of contributing studies (see Table 1). the right panel, four extreme values are not displayed: 4742 and 7445 for AKI, 7445 also for severe AKI, and 9453 for RRT (all for the criterion of 95% specificity)

X_{kds}. We index the thresholds at which data are available in study k by $i = 1, \ldots, t_k$ (where $t_k = 3$ for all k in the present case study) and denote the numerical threshold values that these correspond to by c_{ki}.

The sensitivity (se) is equivalent to the true positive fraction (tpf), while the specificity (sp) is equivalent to the true negative fraction (tnf), leading accordingly to $1 - sp$ as the equivalent of the false positive fraction (fpf). Further, we use the definition of the logit function $\text{logit}(x) = \log(x) - \log(1 - x)$.

3.2 | The random effects model by Steinhauser et al. (2016)

The model by Steinhauser is a two-stage random effects model. At the study level, for each (in the present study log-transformed) value of the threshold, the observed (true or false) negative fraction is transformed using a suitable quantile function f (in the present study, the logit function). At the meta-analysis level, a linear mixed effect model is fitted to the resulting values across studies. The model used in this study (originally called DIDS*, standing for Different random Intercepts and Different random Slopes, Steinhauser et al., 2016) is given by

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

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

Here, the logit transformation in combination with log-transforming the threshold values (c_{ki}) corresponds to the assumption of underlying log-logistic distributions for X_{k0s} and X_{k1s}. Here $\widehat{sp}_{ki}$ and $\widehat{se}_{ki}$ denote the crude estimates of (i.e. observed values of) specificity and the sensitivity at threshold c_{ki} 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, which are assumed to follow a common four-dimensional normal distribution, that is, to be correlated across studies, and $\in_{ki}$ and δ_{ki} represent within-study random errors. Each data point is weighted with the inverse variance of the respective logit-transformed proportion. Back-transformation of the fixed effects part of the model equations provides the model-based distribution functions for disease-free and diseased individuals, from which a model-based summary ROC curve with pointwise confidence regions is obtained. The area under the curve (AUC) is obtained by numerical integration based on the trapezoidal rule. An optimal threshold, defined as a threshold where the Youden index is maximized, is identified as the point where the densities intersect.

The model was implemented in the R package diagmeta (Rücker, Steinhauser, Kolampally, & Schwarzer, 2018) in the free software environment R (R Development Core Team, 2008). For more details of the modelling, we refer to the original article (Steinhauser et al., 2016).

3.3 | The Bayesian model by Jones et al. (2019)

Jones et al. (2019) propose a Bayesian model with multinomial likelihoods and (similarly to Steinhauser et al., 2016, DIDS* model, described above) four sets of random effects. Study-specific estimates of sensitivity and specificity are first used to derive the number of test results in each of the diseased and disease-free populations that fell below the lowest threshold c_{k1}, between each pair of thresholds c_{ki} and c_{ki+1} (for $i = 1, \ldots, t_{k-1}$), and above the highest threshold c_{kt_k}. That is, test results are categorized into $t_k + 1$ ($= 4$ for each study k, in this case study) groups, in each of the diseased and disease-free groups. A multinomial likelihood is assumed for each set of four values.

We denote the underlying true and false positive fractions at the ith threshold in study k by tpf_{ki} and fpf_{ki}. The model specifies that

$$\text{logit}\ (fpf_{ki}) = \frac{\mu_{k0} - g(c_{ki})}{\sigma_{k0}}$$

$$\text{logit}\ (tpf_{ki}) = \frac{\mu_{k1} - g(c_{ki})}{\sigma_{k1}}\ ,$$

where $g()$ is a function that transforms test results in the disease-free population to approximately Logistic(μ_{k0}, σ_{k0}) and test results in the diseased population to approximately Logistic (μ_{k1}, σ_{k1}), where μ_{kd} and σ_{kd} are mean and scale parameters in status group d. For example, if $g()$ is the natural logarithm, this corresponds to assuming log-logistic distributions for test results in each status group (in common with the Steinhauser et al., 2016, model, as described above).

Jones et al. (2019) demonstrate that it is not necessary to pre-specify the transformation function, $g()$: we can assume simply that g() is in the set of Box–Cox transformation functions and estimate the Box–Cox transformation parameter from the data. Computation time is substantially reduced if g() can be pre-specified, however. In this case study, we fitted two versions of the model to each of the six datasets: (i) with $g()$ set to the natural logarithm, (ii) the extended version in which the Box–Cox transformation parameter, λ, is estimated. A value of $\lambda = 0$ corresponds to $g() = \log()$. In this paper, we present results from the extended version of the model if the 95% credible interval around λ did not include 0, and results from the $g() = \log()$ model otherwise. For one of the six analyses (AKI measured in plasma), the extended version of the model did not converge; so we present the $g() = \log()$ results.

The model assumes four sets of normally distributed random effects: $\mu_{k0},\ \mu_{k1}, \log(\sigma_{k0}),\ \log(\sigma_{k1})$, thereby allowing for heterogeneity across studies in both the 'average' and spread of test results in the diseased and disease-free populations. A quadrivariate normal distribution can be fitted to these.[11] However, in two previous case studies, Jones et al. (2019) found no benefit in terms of model fit (after penalizing for complexity) from doing so, relative to treating the four sets of random effects as independent. In this paper, we show results from this simplified version of the model only.

The AUC is calculated by simulation within the model code as the probability that $y_1 \geq y_0$, where y_1 and y_0 are logistic distributions with means and log-scale parameters set at the means of the random effects. The model was implemented in WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000), and the code is available in Jones et al. (2019).

3.4 | The time-to-event model by Hoyer et al. (2018)

The central assumption of the model proposed by Hoyer et al. (2018) is that diagnostic test values can be considered as interval-censored as we only know if these test values (and how many of them) lie above or below a predefined threshold. For modelling these interval-censored diagnostic test values in the diseased and disease-free populations, three different distributions are assumed: the Weibull, log-normal and log-logistic distribution. In line with the class of time-to-event models, the 'events' of interest are testing positive or negative in the diseased or disease-free populations, respectively.

Furthermore, the diagnostic test values are considered as a 'timescale'. Consequently, sensitivity is our event probability in the diseased population (and, vice versa, 1-specificity in the disease-free population). To arrive at an accelerated failure time model with a unified linear predictor, the outcome, that is the diagnostic test values, are log-transformed. The final model equations are then

$$\log(x_{k0}) = b_0 + \varepsilon_0 + u_{k0}$$

$$\log(x_{k1}) = b_1 + \varepsilon_1 + u_{k1}$$

with

$$\begin{pmatrix} u_{k0} \\ u_{k1} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \rho\sigma_0\sigma_1 \\ \rho\sigma_0\sigma_1 & \sigma_1^2 \end{pmatrix}\right],$$

where b_1 and b_0 are location parameters after log-transforming the outcome, whereas ε_1 and ε_0 are error terms with distributions of the log-transformed diagnostic test values x_{k1} and x_{k0} of the diseased and disease-free populations, respectively. Study-specific random effects u_{k1} and u_{k0} are assumed to have a bivariate normal distribution with variances σ_0^2 and σ_1^2 and correlation parameter ρ. These random effects are added to the location parameters after log-transformation and are used to account for between-study heterogeneity and potential between-study correlations. Note there are only two sets of random effects, rather than four in the Steinhauser et al. (2016) and Jones et al. (2019) models. Finally, the resulting survival functions are used to predict sensitivities and specificities at several thresholds.

Results shown in this article are based on the assumption of an underlying Weibull distribution for X_{k0} and X_{k1}, as this version of the model performed best in simulation studies (Hoyer et al., 2018).

The model was implemented in `SAS 9.3` (SAS Institute Inc., Cary, NC, USA). Corresponding source code is available in Hoyer et al. (2018).

3.5 | The nonparametric model by Frömke et al. (2020)

Rather than assuming parametric distributions of test results in the diseased and disease-free groups (as each of the three approaches described above does), the approach proposed by Frömke et al. (2020) is nonparametric, that is does not require any assumptions about distributions. The approach is an extension of a nonparametric method proposed for analysis of diagnostic studies with repeated measures (Brunner & Zapf, 2013; Konietschke & Brunner, 2009) to a meta-analysis of diagnostic accuracy studies with multiple thresholds (Frömke et al., 2020).

The AUC is equal to the relative effect $p = P(X_{k0s} < X_{k1s}) + \frac{1}{2}P(X_{k0s} = X_{k1s})$. To estimate the AUC, all measurements X_{kds} are replaced by their global mid-ranks R_{kds} and the mean rank $\bar{R}_{.d.}$ is calculated of all individuals with status group d over all studies. Then the AUC is estimated by

$\widehat{AUC} = \frac{1}{2} + \frac{1}{n_{..}}(\bar{R}_{.1.} - \bar{R}_{.0.})$, where $n = \sum_{k=1}^{N} \sum_{d\in[0,1]} n_{kd}$ is the total number of diseased and disease-free participants across all studies. Instead of individual patient data X_{kds}, needed for the calculation of the ranks, only aggregate information is available. However, for each study, the number of diseased and disease-free individuals less or greater than the study-specific thresholds c_{ki} is known. Based on this information, we generated data from a one-point distribution, such that $X_{kds} = c_{k1} - 1$ for all

$X_{kds} < c_{k1}, X_{kds} = \frac{(c_{ki}+c_{ki+1})}{2}$ if $c_{ki} < X_{kds} < c_{ki+1}$ with (for $i = 2, \ldots, t_{k-1}$), and $X_{kds} = c_{kt_k} + 1$ for all $X_{kds} > c_{kt_k}$.

For a specific threshold, sensitivity and specificity are estimated in the same way as the AUC, but with transformed data. In order to estimate the sensitivity, the observations of the disease-free individuals are replaced by a one-point distribution at the chosen threshold. Likewise, specificity is obtained by replacing the observations of the individuals with the disease by a one-point distribution (Lange & Brunner, 2012). By means of the asymptotic equivalence theorem, it can be shown that $\sqrt{N}(\hat{p} - p) \sim N(0, \sigma^2)$. Therefore, the standard Wald confidence interval for the AUC as well as for sensitivity and specificity can be computed. Applying the logit transformation to the confidence interval, the limits stay within the interval $[0, 1]$.

TABLE 2 The estimated AUC for the six scenarios and the four approaches

Endpoint	Sampling material	Steinhauser et al. (2016)	Jones et al. (2019)	Hoyer et al. (2018)	Frömke et al. (2020)
AKI	Plasma	0.75	0.81	0.75	0.68
	Urine	0.69	0.74	0.63	0.65
Severe AKI	Plasma	0.77	0.81	0.80	0.77
	Urine	0.74	0.79	0.69	0.73
RRT	Plasma	0.84	0.86	0.87	0.86
	Urine	0.76	0.82	0.72	0.71

TABLE 3 Variation between the four approaches, depending on the scenario (AKI/severe AKI/RRT and sampling material urine/plasma) and on the criterion (95% sensitivity, maximum Youden index, 95% specificity)

Endpoint	Sampling material	95% sensitivity		Maximum Youden index		95% specificity	
		Parameter	Range	Parameter	Range	Parameter	Range
AKI	Plasma	Threshold Specificity	47–76 17– 27%	Threshold Sensitivity Specificity	119–191 63–75% 54–79%	Threshold Sensitivity	255–397 17–41%
	Urine	Threshold Specificity	1–15 2–26%	Threshold Sensitivity Specificity	51–111 56–73% 64–74%	Threshold Sensitivity	308–733 10–24%
Severe AKI	Plasma	Threshold Specificity	53–76 17–28%	Threshold Sensitivity Specificity	177–231 64–74% 73–82%	Threshold Sensitivity	341–420 26–44%
	Urine	Threshold Specificity	5–24 11–28%	Threshold Sensitivity Specificity	94–142 61–72% 68–76%	Threshold Sensitivity	318–881 14–36%
RRT	Plasma	Threshold Specificity	71–181 23–63%	Threshold Sensitivity Specificity	171–374 73–87% 71—94%	Threshold Sensitivity	371–612 25–62%
	Urine	Threshold Specificity	10–48 22–48%	Threshold Sensitivity Specificity	87–121 70–83% 60–73%	Threshold Sensitivity	465–809 16–28%

The model was implemented in in the free software environment R (R Development Core Team, 2008) and corresponding source code is available in Frömke et al. (2020).

4 | RESULTS

It should be noted in advance that the results in the article by Albert et al. (2020b) differ from the results in this article because these authors used the log-normal distribution, whereas we use the Weibull distribution. The individual ROC curves and summary ROC curves resulting from the different approaches are illustrated for all scenarios in Figure 2 (including the confidence region of the Steinhauser model) and in Figure 3 (including the individual ROC curves). Even if the curves are partly quite different, it is noticeable that the confidence region of the Steinhauser model includes all other curves almost completely (NB: no confidence regions can be given for the other models). The differences regarding the curves are also reflected in the corresponding AUCs, which are provided for all six scenarios in Table 2. The differences in AUCs for NGAL measured in plasma range from 0.04 for severe AKI and RRT to 0.13 for AKI. The Jones et al. approach (2019) tended to lead to the highest values of AUC, and the Frömke et al. approach (2020) to the lowest values.

The entire results are given in the Supporting Information (Table A1–A3), with a summary provided in Table 3, which shows that these results sometimes differed considerably across methods. For the criterion of 95% sensitivity, the threshold and the estimated specificity of the four approaches were quite similar for plasma and urine for the endpoints AKI and

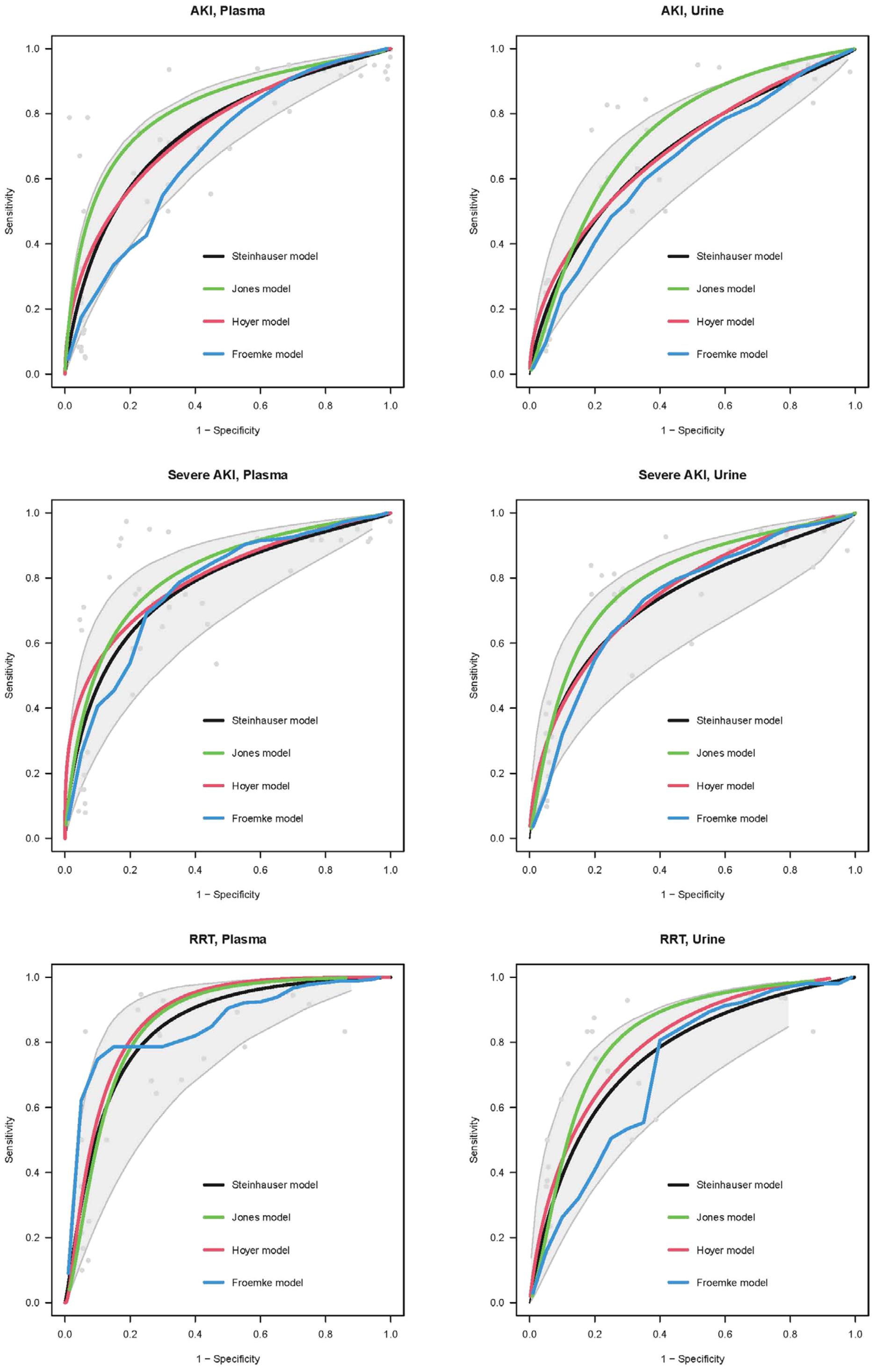

FIGURE 2 Summary ROC curves of the four approaches in the six scenarios and the confidence region for the Steinhauser model as grey area. The pairs of sensitivity and 1 – specificity of the individual studies are displayed as grey dots

severe AKI. For the endpoint of RRT, the threshold and therefore also the estimated specificity was considerably higher than for AKI and severe AKI.

The estimated summary sensitivity and specificity to maximize the Youden index were quite similar across methods for AKI and severe AKI, as well as for NGAL measured in urine and in plasma. Thresholds were lower for urine than for plasma. The variability across methods was quite small regarding the threshold as well as estimated sensitivity and

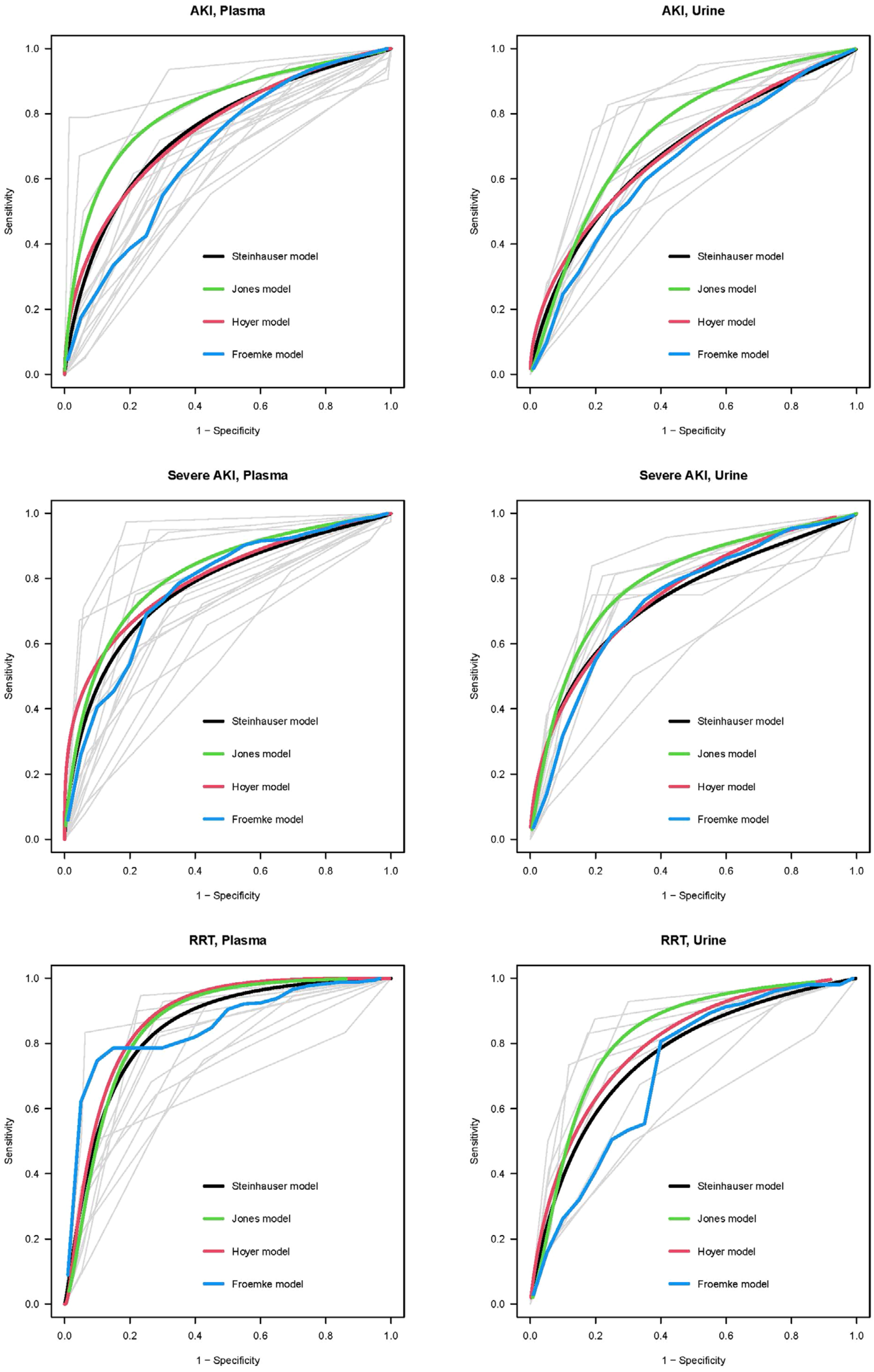

FIGURE 3 Summary ROC curves of the four approaches in the six scenarios and the ROC curves of the individual studies as grey lines

specificity for severe AKI and NGAL measured in plasma. However, for RRT as the endpoint and NGAL measured in plasma, the variability was remarkably high, and also the estimated sensitivity and specificity were higher than in the other scenarios. Regarding the criterion of 95% specificity, the variability in thresholds was very high. The estimated sensitivity was in the same range for AKI plasma, severe AKI urine and plasma and RRT urine, but remarkably higher for RRT plasma.

5 | DISCUSSION

In this paper, we presented, compared and applied four recently proposed approaches to meta-analyse data from DTA studies providing information on sensitivity and specificity at several thresholds. All four approaches use the full data from these studies and thus are promising additions to the toolbox of meta-analysis of DTA studies.

As an example, we showed results from applying each approach to data evaluating the diagnostic accuracy of NGAL as a test for AKI. From a medical point of view, the results of our case study can be summarized as follows: The relatively large differences between the lowest and the highest AUC estimates in the presented approaches (0.68–0.81 for AKI measured in plasma) may translate into withholding or not withholding diagnostic or therapeutic measures (Albert et al., 2018, 2020a). A biomarkers applicability as a screening test will call for high sensitivity. For the corresponding criterion of at least 95% sensitivity, the result was (across all four meta-analysis methods) an estimated specificity below 30% for AKI and severe AKI. For RRT, the estimated specificity was higher, but quite different between the analysis approaches (23–63% for plasma, 22–48% for urine). From a statistical point of view, in our case study we found considerable differences between the approaches in summary results for some data scenarios.

While Jones et al. (2019), Steinhauser et al. (2016), and Hoyer et al. (2018) are parametric approaches, the approach from Frömke et al. (2020) is a non-parametric one. The three parametric approaches make different assumptions about the distributional form of the underlying test values. As applied in this article, the model of Steinhauser et al. (2019) assumes log-logistic distributions, although alternatively an assumption of logistic, log-normal or normal distributions is possible. The model of Jones et al. (2019) assumes some Box–Cox transformation of continuous test results (e.g. log), where the Box–Cox transformation does not need to be pre-specified. In this paper, for the model of Hoyer et al. (2018) a Weibull distribution is assumed, but alternatively it is possible to assume log-normal or log-logistic distributions within the Hoyer framework. Other differences include the number of random effects (4 or 2), the choice of parameters that are random effects (which are not linear functions of each other) and the likelihoods (normal or multinomial) (see also Jones et al., 2019). A summary of some of the characteristics of the four approaches is provided in Table 4.

For the original analysis presented in Albert et al. (2020a), the model of Hoyer et al. (2018) assuming a log-normal distribution was applied. Slight differences between results given in the present article and the original publication are caused by varying distributional assumptions. However, confidence intervals of estimated sensitivities, specificities and AUC overlap each other, indicating that there is no evidence for a difference between them. We decided to use the Weibull distribution in this article because it performed best in simulation studies compared to other distributions (Hoyer et al., 2018).

Our case study dataset is a real-life example that highlights that it may be difficult for analysts to choose between the recently published approaches. No systematic comparison of the approaches has been published so far. A limitation of our analysis is that not all possible approaches have been included. However, the other proposed approaches each have specific disadvantages, such that their general applicability is more limited. Several of these models rely on each study reporting data at an identical number of thresholds (Bipat, Zwinderman, Bossuyt, & Stoker, 2007; Hamza, Arends, Van Houwelingen, & Stijnen, 2009; Poon, 2004; Putter, Fiocco, & Stijnen, 2010). Other approaches ignore the precise threshold values (Dukic & Gatsonis, 2003; Martinez-Camblor, 2017) or are fixed-effect models (Riley et al., 2014). Due to their disadvantages, we decided to not include these models in the present case study. We instead used only approaches that can cope with the features of the underlying dataset and, in addition, do not require the same number of thresholds.

We note that our case study data were somewhat artificial, with the same number of thresholds and same requirement for each threshold (that gave 95% sensitivity, 95% specificity, and maximized the Youden index, respectively) in all primary studies. This specific setting resulted from the original study's aim to identify urinary and plasma NGAL cut-off concentrations with high sensitivity or high specificity to complement the identification of patients at high kidney risk in clinical research and practice (Albert et al., 2020b, 2021).

In practice, primary studies in a systematic review may report at varied numbers of thresholds and perhaps with even more variability in threshold values.

In this article, the events AKI, severe AKI and RRT were analysed separately. An alternative approach would be to define the true state not dichotomously but in the form of a four-level score (no AKI, AKI, severe AKI, RRT). However, none of the four approaches investigated here is suitable for such an analysis.

Based on just one case study dataset, we cannot identify the reasons for the differences in summary results across methods or generalize our results to other datasets, although we have discussed some of the features that vary across the approaches. Nevertheless, given that standard approaches to meta-analysis of DTA data across multiple and varying

TABLE 4 Characteristics of the four approaches regarding requirements, procedure, results and specific advantages

	Steinhauser et al. (2016)	Jones et al. (2019)	Hoyer et al. (2018)	Frömke et al. (2020)
Requirements and procedure				
Aggregated data sufficient	X	X	X	X
One-step approach		X	X	
Zero cells allowed	(X)[a]	X	X	X
Random effects model	X	X	X	
Results				
Summary ROC curve	X	X	X	X
with confidence bands	X			
AUC	X	X	X	X
with confidence interval	X		X	X
Specific thresholds	X	X	X	X
with confidence intervals	(X)[2]	X	X	X
corresp. sens. and spec.	X	X	X	X
with confidence intervals	X	X		
Other measures available (PPV, NPV, DOR)	X	X	X	X
with confidence intervals		X		X
Measure of heterogeneity (like I^b) regarding diagnostic accuracy / thresholds	no; standard deviation, correlation, graphics, and prediction intervals recommended			
Specific advantages				
No convergence problems regarding the method				X
Inclusion of study-level covariates possible (i.e. meta-regression)		X	X	

[a] In the case of zero cells, a continuity correction is necessary.

[b] Thresholds are presented with confidence intervals, if the normal transformation is used, but not for the logit transformation.

thresholds lead to a heavy loss of information and likely biased results (see Introduction), we recommend using one of the presented approaches for data of this type.

In a recent study by Benedetti et al. (2020), in which the results of meta-analyses based on published and on individual level data were compared, the differences between the models from Steinhauser and Jones were much smaller. The question therefore arises whether the differences detected here might be due to some peculiarities specific to our case study dataset. In addition, it can be seen that the confidence region of the Steinhauser model almost completely includes the other summary ROC curves, so that it cannot be excluded that the differences are simply due to random error. The differences regarding the model from Frömke are perhaps driven by the use of ranks instead of original observations. Therefore, this approach is insensitive to outliers.

To assess which approach can be recommended in which situation, we are currently conducting a systematic simulation study. We are investigating the effect of a number of factors, including the true distributions of the test results, the number of studies and thresholds, and the effect of any outliers. We hope that the results observed in the present case study may then be explained.

6 CONCLUSION

We compared four existing statistical approaches to analyse multiple-threshold information from DTA studies by applying them to a real-life meta-analysis of the diagnostic accuracy of NGAL for the detection of acute kidney injury with three thresholds per study. The results show quite large heterogeneity of the approaches in some data scenarios, with differences in estimated AUC of up to 0.13. However, all summary ROC curves lie within the confidence region of the Steinhauser model and therefore the differences are perhaps only random variations. We recommend use of these approaches in practice. However, at this stage it is unclear which approach should be preferred in any given situation. Therefore, currently a simulation study is conducted to investigate the advantages and limitations of the different approaches. However, this case report makes aware that the results one gets – and this is true for most statistical analyses – are not the final truth. This, in turn, makes it clear how important it is to pre-specify methods that readers can be confident the method was not chosen based on results.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.
This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

REFERENCES

Albert, C., Albert, A., Bellomo, R., Kropf, S., Devarajan, P., Westphal, S., … Haase-Fielitz, A. (2018). Urinary neutrophil gelatinase-associated lipocalin-guided risk assessment for major adverse kidney events after open-heart surgery. *Biomarkers Medicine*, *12*, 975–985.

Albert, C., Haase, M., Albert, A., Zapf, A., Braun-Dullaeus, R. C., & Haase-Fielitz, A. (2021). Biomarker-guided risk assessment for acute kidney injury: Time for clinical implementation? *Annals of Laboratory Medicine*, *41*(1), 1–15.

Albert, C., Haase, M., Albert, A., Kropf, S., Bellomo, R., Westphal, S., … Haase-Fielitz, A. (2020a). urinary biomarkers may complement the Cleveland score for prediction of adverse kidney events after cardiac surgery: A pilot study. *Annals of Laboratory Medicine*, *40*, 131–141.

Albert, C., Zapf, A., Haase, M., Röver, C., Pickering, J. W., Albert, A., … Haase-Fielitz, A. (2020b). Neutrophil gelatinase-associated lipocalin measured on clinical laboratory platforms for the prediction of acute kidney injury and the associated need for dialysis therapy: A systematic review and meta-analysis. *American Journal of Kidney Diseases*, *76*, 826–841.e1.

Bellomo, R., Ronco, C., Kellum, J. A., Mehta, R. L., & Palevsky, P. (2004). Acute renal failure – definition, outcome measures, animal models, fluid therapy and information technology needs: The Second International Consensus Conference of the Acute Dialysis Quality Initiative (ADQI) Group. *Critical Care*, *8*, R204–R212.

Benedetti, A., Levis, B., Rücker,, G., Jones, H. E., Schumacher, M., Ioannidis, J. P. A., & Thombs, B. DEPRESsion Screening Data (DEPRESSD) Collaboration. (2020).An empirical comparison of three methods for multiple cutoff diagnostic test meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) depression screening tool using published data vs individual level data. *Research Synthesis Methods*, *11*, 833–848. https://doi.org/10.1002/jrsm.1443

Bipat, S., Zwinderman, A. H., Bossuyt, P. M. M., & Stoker, J. (2007). Multivariate random-effects approach: For meta-analysis of cancer staging studies. *Academic Radiology*, *14*, 974–984.

Brunner, E., & Zapf, A. (2013). Nonparametric ROC analysis for diagnostic trials. In N. Balakrishnan (Ed.), *Handbook of methods and applications of statistics in clinical trials, Vol. 2: Planning, analysis, and inferential methods* (pp. 471–483). Hoboken, NJ: Wiley.

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

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

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

Frömke, C., Kirstein, M., & Zapf, A. (2020). A nonparametric approach for meta-analysis of diagnostic accuracy studies with multiple cut-offs. Manuscript in Preparation.

Haase-Fielitz, A., Bellomo, R., Devarajan, P., Bennett, M., Story, D., Matalanis, G., … Haase, M. (2009). The predictive performance of plasma neutrophil gelatinase-associated lipocalin (NGAL) increases with grade of acute kidney injury. *Nephrology Dialysis Transplantation*, *24*, 3349–3354.

Haase-Fielitz, A., Haase, M., & Devarajan, P. (2014). Neutrophil gelatinase-associated lipocalin as a biomarker of acute kidney injury: A critical evaluation of current status. *Annals of Clinical Biochemistry*, *51*(Pt 3), 335–351.

Hamza, T. H., Arends, L. R., Van Houwelingen, H. C., & Stijnen, T. (2009). Multivariate random effects meta-analysis of diagnostic tests with multiple thresholds. *BMC Medical Research Methodology*, *9*, 73.

Heazell, E. P., Hayes, D. J. L., Whitworth, M., Takwoingi, Y., Bayliss, S. E., & Davenport, C. (2019). Biochemical tests of placental function versus ultrasound assessment of fetal size for stillbirth and small-for-gestational-age infants. *Cochrane Database of Systematic Reviews*, *5*, 1465–1858.

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

Hoyer, A., Hirt, S., & Kuss, O. (2018). Meta-analysis of full ROC curves using bivariate time-to-event models for interval-censored data. *Research Synthesis Methods*, *9*(1), 62–72.

Jones, H. E., Gatsonsis, C. A., Trikalinos, T. A., Welton, N. J., & Ades, A. E. (2019). Quantifying how diagnostic test accuracy depends on threshold in a meta-analysis. *Statistics Medicine*, *38*(24), 4789–4803.

Klein, S. J., Brandtner, A. K., Lehner, G. F., Ulmer, H., Bagshaw, S. M., Wiedermann, C. J., & Joannidis, M. (2018). Biomarkers for prediction of renal replacement therapy in acute kidney injury: A systematic review and meta-analysis. *Intensive Care Medicine*, *44*, 323–336.

Konietschke, F., & Brunner, E. (2009). Nonparametric analysis of clustered data in diagnostic trials: Estimation problems in small sample sizes. *CSDA*, *53*, 730–741.

Lange, K., & Brunner, E. (2012). Sensitivity, specificity and ROC-curves in multiple reader diagnostic trials— A unified, nonparametric approach. *Statistical Methodology*, *9*, 490–490.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS — a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics Computing*, *10*, 325–337.

Macaskill, P., Gatsonis, C., Deeks, J. J., Harbord, R., & Takwoingi, Y. (2010). Chapter 10: Analysing and Presenting Results. In J. J. Deeks, P. M. Bossuyt, & C. Gatsonis (Eds.), *Cochrane handbook for systematic reviews of diagnostic test accuracy, Version 1.0. The Cochrane Collaboration*. Retrieved from https://methods.cochrane.org/sites/methods.cochrane.org.sdt/files/public/uploads/Chapter-%20-%20Version%201.0.pdf.

Martínez-Camblor, P. (2017). Fully non-parametric receiver operating characteristic curve estimation for random-effects meta-analysis. *Statistical Methods of Medical Research*, *26*, 5–20.

Mishra, J., Dent, C., Tarabishi, R., Mitsnefes, M. M., Ma, Q., Kelly, C., & Devarajan, P. (2005). Neutrophil gelatinase-associated lipocalin (NGAL) as a biomarker for acute renal injury after cardiac surgery. *Lancet*, *365*, 1231–1238.

Poon, W. Y. (2004). A latent normal distribution model for analysing ordinal responses with applications in meta-analysis. *Statistics Medicine*, *23*, 2155–2172.

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

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org.

Reitsma, J. B., Glas, A. S., Rutjes, A. W.S., Scholten, R. J.P.M., Bossuyt, P. M., & Zwinderman, A. H. (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology*, *58*(10), 982–990.

Riley, R. D. (2014). Meta-analysis of test accuracy studies with multiple and missing thresholds: A multivariate-normal model. *Journal of Biometrics Biostatistics*, *05*(3), 196.

Rücker, G., Steinhauser, S., Kolampally, S., & Schwarzer, G. (2018). diagmeta: Meta-analysis of diagnostic accuracy studies with several cutpoints. R package version 0.2-0. Retrieved from https://CRAN.R-project.org/package=diagmeta.

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

Schisterman, E. F., & Perkins, N. (2007). Confidence intervals for the Youden index and corresponding optimal cut-point. *Communications in Statistics: Simulations and Computations*, *36*, 549–563.

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

Trikalinos, T. A., Balion, C. M., Coleman, C. I., Griffith, L., Santaguida, P. L., Vandermeer, B., & Fu, R. (2012). Chapter 27: Meta-analysis of test performance when there is a 'gold standard'. *Journal of General Internal Medicine*, *27*, S56–S66.

Vogelgesang, F., Schlattmann, P., & Dewey, M. (2018). The evaluation of bivariate mixed models in meta-analyses of diagnostic accuracy studies with SAS, Stata and R. *Methods of Information Medicine*, *57*(3), 111–119.

Whiting, P. F., Rutjes AWS, Westwood ME et al. (2011). QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine*, *155*, 529–536.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Zapf A, Albert C, Frömke C, et al. Meta-analysis of diagnostic accuracy studies with multiple thresholds: Comparison of different approaches. *Biometrical Journal*. 2021;1–13. https://doi.org/10.1002/bimj.202000091