Network meta-analysis of diagnostic test accuracy studies identifies and ranks the optimal diagnostic tests and thresholds for healthcare policy and decision making

Rhiannon K. Owen[1]
Nicola J. Cooper[1]
Terence J. Quinn[2]
Rosalind Lees[2]
Alex J. Sutton[1]

[1]Deaprtment of Health Sciences, University of Leicester
[2]Institute of Cardiovascular and Medical Sciences, University of Glasgow

Corresponding author:
Dr Rhiannon K Owen,
Department of Health Sciences,
College of Medicine, Biological Sciences and Psychology,
University of Leicester,
Leicester,
LE1 7RH
Tel: +44(0)116 229 7295, Email: Rhiannon.owen@le.ac.uk

The Complex Reviews Support Unit (CRSU) is funded by the National Institute for Health Research (project number 14/178/29). The views and opinions expressed herein are those of the authors and do not necessarily reflect those of the NIHR, NHS or the Department of Health.

Abstract

Objective:

Network meta-analyses have extensively been used to compare the effectiveness of multiple interventions for healthcare policy and decision-making. However, methods for evaluating the performance of multiple diagnostic tests are less established. In a decision-making context, we are often interested in comparing and ranking the performance of multiple diagnostic tests, at varying levels of test thresholds, in one simultaneous analysis.

Study design and setting:

Motivated by an example of cognitive impairment diagnosis following stroke, we synthesized data from 13 studies assessing the efficiency of two diagnostic tests: Mini-Mental State Examination (MMSE) and Montreal Cognitive Assessment (MoCA), at two test thresholds: MMSE <25/30 and <27/30, and MoCA <22/30 and <26/30. Using Markov Chain Monte Carlo (MCMC) methods, we fitted a bivariate network meta-analysis model incorporating constraints on increasing test threshold, and accounting for the correlations between multiple test accuracy measures from the same study.

Results:

We developed and successfully fitted a model comparing multiple tests/threshold combinations while imposing threshold constraints. Using this model, we found that MoCA at threshold <26/30 appeared to have the best true positive rate, whilst MMSE at threshold <25/30 appeared to have the best true negative rate.

Conclusion:

The combined analysis of multiple tests at multiple thresholds allowed for more rigorous comparisons between competing diagnostics tests for decision making.

Keywords:

Network meta-analysis; meta-analysis; diagnostic test accuracy; multiple tests; multiple thresholds

1. Background

Evidence-based healthcare evaluations, as endorsed by the National Institute for Health and Care Excellence (NICE) in the UK and similar organisations worldwide, have highlighted the crucial role systematic reviews, including meta-analysis where appropriate, have to play in the decision making process to answer clinically relevant questions such as whether a technology works, for whom and how it compares with alternatives (1). Such evidence based evaluations are important to the decision making process within the area of diagnostic test performance as early diagnosis of disease can lead to more successful treatment than if treatment is delayed.

Diagnostic test accuracy is defined by Leeflang *et al* (2) as the ability of a test to distinguish between patients with a specified target condition and those without, and the results are usually expressed in terms of sensitivity (i.e. the proportion of people with the condition correctly detected by the test) and specificity (i.e. the proportion of people without the condition correctly detected by the test). The dependence between these outcomes (i.e. sensitivity and specificity) adds an additional complexity that makes evidence synthesis of diagnostic test accuracy data more complicated than for intervention studies. The dependence between outcomes can occur in one of two ways. For meta-analyses of studies evaluating a single pair of sensitivity and specificity, the dependence between sensitivity and specificity occurs across studies for differing thresholds (3). For meta-analyses of studies evaluating multiple tests and/or multiple thresholds, and therefore, multiple pairs of sensitivity and specificity, the dependence between outcomes occurs both within and across studies. Another additional issue includes correctly estimating the joint conditional performance of multiple tests when they are used in combination (4). A number of methods have been proposed to meta-analyse studies reporting single pairs of sensitivity and specificity data including independent meta-analyses of sensitivity and specificity (5), meta-analysis of diagnostic odds ratios (5), summary receiver operating curve (sROC) regression modelling (6), hierarchical sROC model

(7), and bivariate meta-analysis models (7, 8). In cases where the test threshold is suspected or known to vary between studies, the latter 2 approaches (which have been shown to be equivalent (9)) are the most appropriate as they allow for this variability in the analysis. However, neither of these *incorporates* test threshold information explicitly into the analysis, which hampers accurate prediction of test performance at particular thresholds; a fact which limits the clinical applicability of results, at least in contexts where test threshold can be explicitly specified. Work extending the hierarchical sROC / bivariate approach, allowing for data from multiple thresholds for the same tests from the same study to be synthesised, was described some time ago (10) but is rarely used in practice and suffers from the same limitation that explicit threshold value information is not included in the analysis. Alternative approaches have since been described (3, 11-14) including a generalization of the bivariate model to include the use of multivariate random effects (11), and the use of Poisson-correlated gamma frailty models (12). For both of these approaches, the number of thresholds across all studies has to be identical, which is often impractical in a meta-analysis setting. Multivariate regression models (13) and linear mixed effects models (3) have also been proposed. These methods consist of a two-stage approach, however, estimation of the uncertainty from stage-one of these analyses are ignored and may lead to unrealistic results. More recently, Hoyer *et al* (14) proposed a bivariate time-to-event model for interval censored data incorporating random effects. This approach overcomes the limitations described above, and provides a flexible framework to account for various distributions of the underlying diagnostic marker. However, a limitation of this model is the potential constraint of the proportional hazards assumptions when meta-analysing ROC curves from studies which report a single threshold.

In comparison to interventional research, methodology for meta-analysing diagnostic test accuracy data has relatively recently been adopted by Cochrane. As such, the development of guidance, and best practice statements are ever evolving. However, at present, the preferred Cochrane methods are suited to meta-analysing diagnostic test accuracy of single tests (using the statistical methods outlined above). From a clinical and decision-making perspective, often interest lies in assessing the performance of multiple diagnostic tests with multiple thresholds in one simultaneous analysis, since it addresses clinically relevant questions such as: which test at which threshold is most effective, or most cost-effective (15). In interventional research, when multiple competing healthcare interventions are of interest, network meta-analyses (NMA) have been used extensively to compare and identify the “best” intervention(s). However, methods for evaluating the performance of multiple diagnostic tests are less established. Network meta-analyses of healthcare interventions are commonly used to synthesise data from several clinical trials in similar patient populations with the aim to evaluate multiple interventions that may not have been compared otherwise. This approach combines both direct information (obtained from head-to-head trials) and indirect information (obtained from trials that share a common comparator) to obtain relative treatment effects for all interventions whilst maintaining randomisation. Combining direct and indirect information in this way assumes an additive relationship between treatment effects. In a diagnostic test accuracy setting, this framework has previously been adopted to model the difference or relative risk in sensitivity and specificity (16), as well as the relative Diagnostic Odds Ratio between two tests (16-18). Throughout this paper, network meta-analyses are used to describe the synthesis of diagnostic test accuracy data from a network of diagnostic tests that have been evaluated in the same study and thus, the same individuals. Our approach differs to the framework commonly used for healthcare interventions in that our approach models the absolute sensitivity and specificity of each test incorporating random effects to allow for heterogeneity, as well as similarities between data belonging to the same study, and the same tests within studies (i.e. to account for multiple thresholds within tests).

Through careful consideration of a motivating example, this paper sets out an approach to network meta-analysis of diagnostic test accuracy studies which allows for both the incorporation of multiple tests and multiple explicit threshold values, potentially reported by the same studies. Section 2 introduces the motivating example. Section 3 discusses the visual representation of diagnostic test accuracy data, highlighting where appropriate, key assumptions that are revisited throughout the remainder of this article. Section 4 describes each of the proposed meta-analysis models. Section 5 presents the results from the analysis of the motivating example and Section 6 concludes this article with a discussion.

2. Motivating example

Cognitive impairment is highly prevalent in stroke survivors and it can lead to increased mortality, disability, and institutionalization. Early detection of cognitive impairment is an essential step in the efficient management of patient care. In the UK, screening for cognitive impairment is recommended by governing bodies such as the National Institute for Health and Care Excellence (NICE) and the Royal College of Physicians (RCP). However, although cognitive assessment is recommended in various clinical guidelines, there is no consensus on how to efficiently diagnose patients, where differing guidelines recommend differing tests and thresholds (19). A robust synthesis of the evidence in one simultaneous analysis has the potential to provide an evidence base where currently best clinical practice is primarily opinion based.

In a study by Lees *et al* (20), the authors investigated the test accuracy of multiple screening tests for the diagnosis of cognitive impairment and dementia in stroke patients. Cognitive impairment is an umbrella term that encompasses any objective memory and thinking problem. It includes, but is not limited to, the clinical syndrome of dementia. Gold standard assessment for cognitive impairment is a detailed examination of various aspects of cognition (neuropsychological battery (NPB)), while there is no consensus on the 'gold' standard ante-mortem diagnosis of dementia, dementia diagnosis is currently made according to clinical criteria which can often be informed by data from NPB. In the Lees et al review, as the index tests of interest were 'screening' tests, the authors decided to include both cognitive impairment on NPB and clinical dementia diagnosis as their reference standard. The authors were able to synthesize data from 13 diagnostic test accuracy studies for two key screening tests; Folstein's Mini Mental State Examination (MMSE) and Montreal Cognitive Assessment (MoCA). MMSE and MoCA examinations are based on a scoring system, where a higher value indicates a more desired response from the participant. Thus a disease-positive response on a lower test threshold would suggest a more severe case of cognitive impairment. The authors focused on 2 disparate cut points per test for dementia diagnosis; MMSE <25/30, MMSE <27/30, MoCA <22/30, MoCA <26/30. In this article, each of the test-threshold combinations are treated as independent tests. Figure 1 illustrates the network of comparative studies. The nodes represent each of the test-threshold combinations of interest. The dashed interconnecting lines illustrate that a comparative study examining both tests in the same patient population exists. Typically networks of treatment comparisons are presented alongside network meta-analyses of healthcare interventions, where the interconnecting lines represent direct evidence on the relative differences between treatments. Figure 1 differs to a network of treatment comparisons for network meta-analyses of healthcare interventions whereby the interconnecting lines represent a comparative study illustrating tests that have been undertaken in the same cohort of patients and thus, there is a within-study dependence between pairs of sensitivity and specificity for these tests. Table 1 shows the extracted data from the original papers. The choice of test thresholds were determined by the most commonly reported test-threshold combinations in the published literature at the time of publication. The reference standard tests included NPB and clinical diagnosis of dementia (and the authors assumed that these were perfect, as do we in this paper). In this example, Lees *et al* (20) pooled data from studies using both reference tests. Data from 12, 5, 4, and 6 studies were pooled in separate bivariate meta-analyses for MMSE<25/30, MMSE<27/30, MoCA<26/30 and MoCA<22/30, respectively.

Figure 1: Network of comparative studies

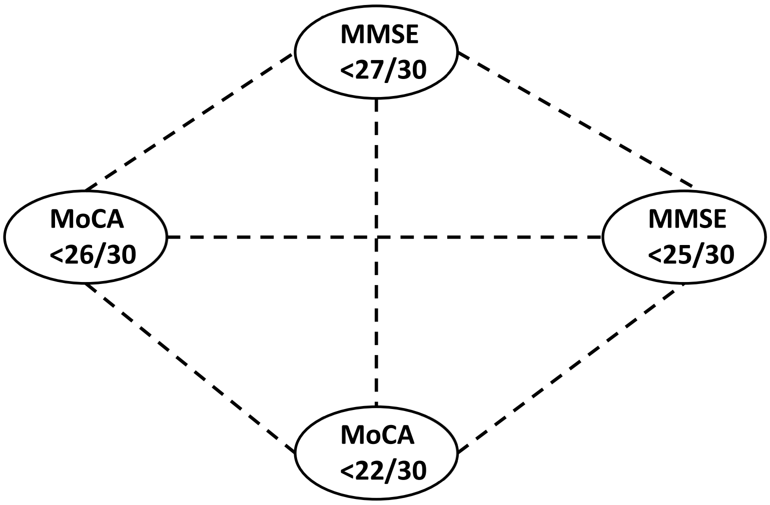


Table 1: Diagnostic test accuracy data obtained from the original papers

Study Author	Test	TP	FP	FN	TN	Sensitivity (95%CI)	Specificity (95%CI)
Blake 2002	MMSE<25	19	10	12	71	0.61 [0.42,0.78]	0.88[0.78,0.94]
Bour 2010	MMSE<25	21	29	1	143	0.95 [0.77,1.00]	0.83 [0.77,0.88]
Cumming 2010	MMSE<25	48	22	10	69	0.83 [0.71,0.91]	0.76 [0.66,0.84]
Cumming 2013	MMSE<25	21	4	17	17	0.55 [0.38,0.71]	0.81 [0.58,0.95]
	MMSE<27	35	10	3	11	0.92 [0.79,0.98]	0.52 [0.30,0.74]
	MoCA<26	39	12	0	9	1.00 [0.91,1.00]	0.43 [0.22,0.66]
	MoCA<22	30	5	9	16	0.77 [0.61,0.89]	0.76 [0.53,0.92]
de Koning 1998	MMSE<25	44	51	11	178	0.80 [0.67,0.90]	0.78 [0.72,0.83]
Dong 2010	MMSE<25	52	10	8	45	0.87 [0.75,0.94]	0.82 [0.69,0.91]
	MMSE<27	45	17	3	26	0.94 [0.83,0.99]	0.60 [0.44,0.75]
	MoCA<22	54	13	6	42	0.90 [0.79,0.96]	0.76 [0.63,0.87]

Dong 2012	MMSE<25	28	32	4	65	0.88 [0.71,0.96]	0.67 [0.57,0.76]
	MoCA<22	53	65	7	114	0.88 [0.77,0.95]	0.64 [0.56,0.71]
Godefroy 2011	MMSE<25	45	2	19	29	0.70 [0.58,0.81]	0.94 [0.79,0.99]
	MMSE<27	55	12	9	19	0.86 [0.75,0.93]	0.61 [0.42,0.78]
	MoCA<26	60	20	4	11	0.94 [0.85,0.98]	0.35 [0.19,0.55]
	MoCA<22	48	4	16	28	0.75 [0.63,0.85]	0.88 [0.71,0.96]
Grace 1995	MMSE<25	20	9	26	46	0.43 [0.29,0.59]	0.84 [0.71,0.92]
Morris 2012	MMSE<25	21	3	15	10	0.58 [0.41,0.74]	0.77 [0.46,0.95]
	MMSE<27	30	8	6	5	0.83 [0.67,0.94]	0.38 [0.14,0.68]
Pendlebury 2012	MMSE<25	11	3	8	69	0.58 [0.33,0.80]	0.96 [0.88,0.99]
	MMSE<27	15	15	4	57	0.79 [0.54,0.94]	0.79 [0.68,0.88]
	MoCA<26	19	39	0	33	1.00 [0.82,1.00]	0.46 [0.34,0.58]
	MoCA<22	13	11	6	61	0.68 [0.43,0.87]	0.85 [0.74,0.92]
Salvadori 2013	MoCA<26	78	46	2	29	0.97 [0.91,1.00]	0.39 [0.28,0.51]
	MoCA<22	73	18	7	58	0.91 [0.83,0.96]	0.76 [0.65,0.85]
Srikanth 2006	MMSE<25	4	4	4	67	0.50 [0.16,0.84]	0.94 [0.86,0.98]

3. Illustrating diagnostic test accuracy data

This section highlights some of the key principles of diagnostic test accuracy data through the use of visual representations. Figures 2 and 3 illustrate the relationship between the thresholds of the same study for MMSE and MoCA, respectively. The nodes represent the observed sensitivity and specificity, colour-coded for each threshold; and the corresponding number represents the exact threshold used. The interconnecting lines illustrate the ROC curve for each study that reports multiple threshold values. One of the properties of an ROC curve is that higher test thresholds for a positive outcome must, mathematically, have an increased sensitivity but decreased specificity. For the remainder of this article, this property will be referred to as a threshold assumption. In this example, MMSE and MoCA are scored out of 30 points, with a point deducted for each error. Lower scores therefore suggest greater impairment. If the ‘test positive’ threshold is lowered then at the lower threshold the test is more specific and less sensitive, and thus should lie in the lower left hand side of the ROC space. From Figure 2, it is evident that there is a large amount of heterogeneity in sensitivities and specificities reported between the studies identified by Lees *et al* (20), where datapoints with lower thresholds for MMSE lie towards the top left hand side of the ROC space, above that of higher thresholds.

Figure 2: Relationship of increasing test thresholds within MMSE

Figure 4 illustrates schematically the relationship between the threshold assumption and that of heterogeneity. Heterogeneity between studies can be described as the difference in ROC curves *between* studies, shifting the ROC curve in a Southeast or Northwest direction from the summary ROC. The threshold assumption shifts the observed sensitivity and specificity in a Northeast or Southwest direction *along* the study specific ROC curve.

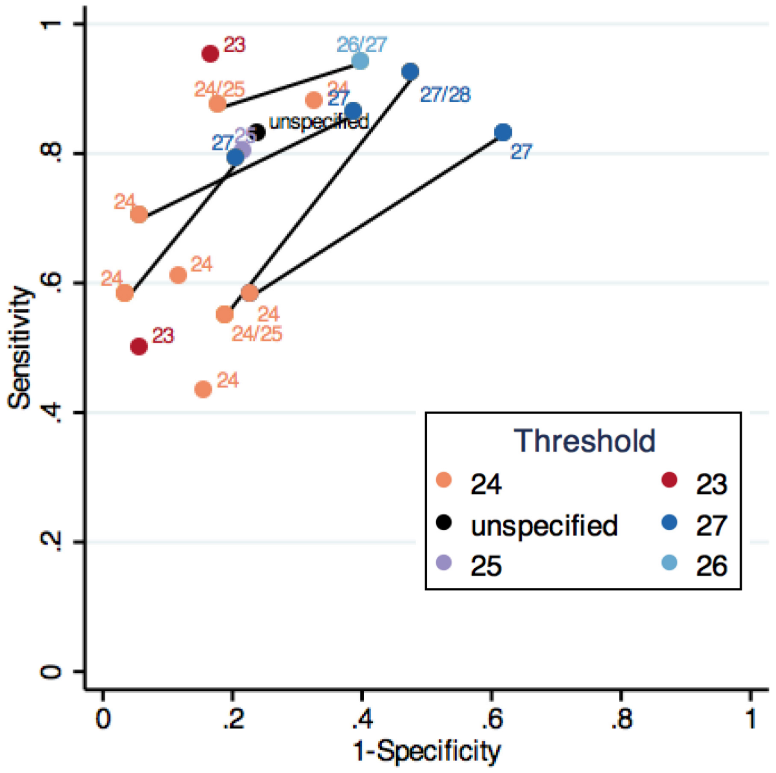


Figure 3: Relationship of increasing test thresholds within MoCA

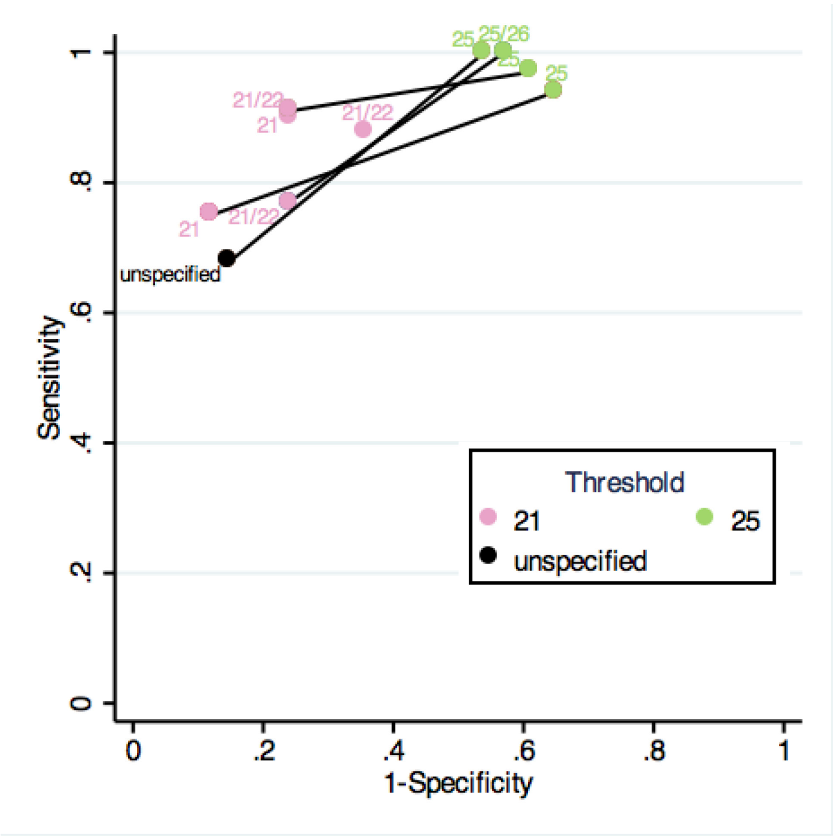
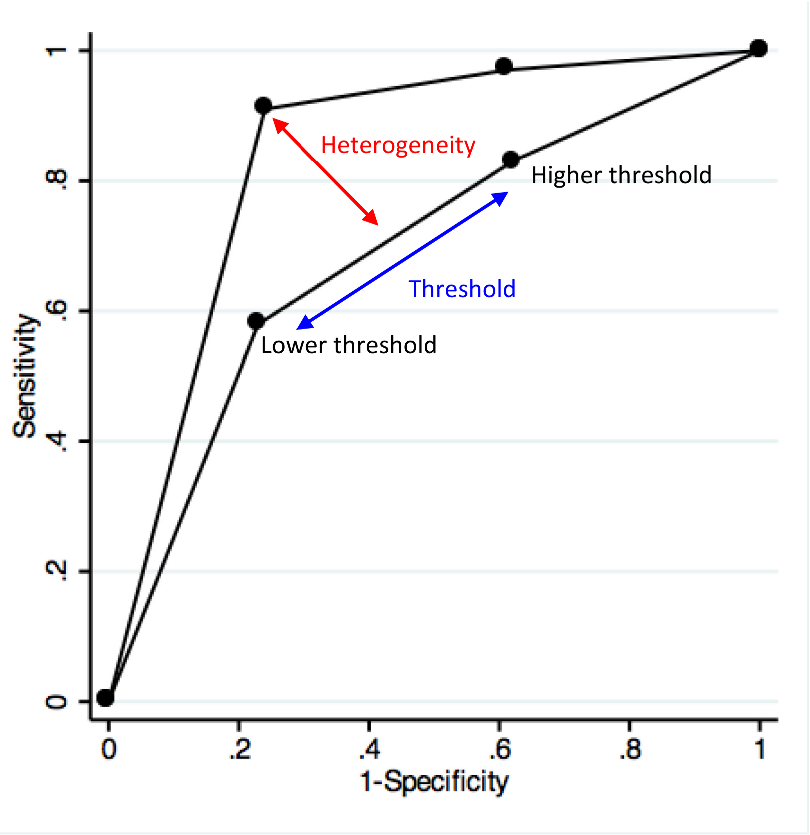


Figure 4: Schematic relationship of the impact of heterogeneity in test performance due to threshold and between study heterogeneity. The labels "Lower threshold" and "Higher threshold" apply specifically to the given example.



4. Models and estimation

4.1. Network meta-analysis for diagnostic test accuracy studies

Building on the bivariate meta-analysis model described by Reitsma *et al* (8) and Sutton *et al* (15), let the observed number of true positives, tp_i, for the i^{th} observation be considered as a binomial count from a sample of disease positive individuals, pos_i. This information allows for estimation of the diagnostic sensitivity, $sens_i$, which is associated with the rate of accurate detection of diseased individuals. Likewise, let the number of true negatives, tn_i, be considered as a binomial count from a sample of disease negative individuals, neg_i. This information allows for estimation of the specificity of the test, $spec_i$, which is associated with the rate of accurate detection of non-diseased individuals, such that:

$$tp_i \sim Binomial(sens_i, pos_i)$$
$$tn_i \sim Binomial(spec_i, neg_i)$$

(1)

Logistic regression models can be used to specify sensitivity and specificity. Across studies, the sensitivity and specificity of each test are likely to be correlated. To account for this across-study dependence the (logit) sensitivity, $sens_i$, and specificity, $spec_i$, of observation *i*, where observation *i* represents each pair of sensitivities and specificities for each study, are drawn from a bivariate normal distribution with mean equal to the pooled test accuracy estimates of sensitivity, $\mu_{i_{sens}}$, and specificity, $\mu_{i_{spec}}$, with between-observation covariance matrix $\mathbf{\Sigma}$:

$$\begin{pmatrix} logit(sens_i) \\ logit(spec_i) \end{pmatrix} \sim Normal\left[\begin{pmatrix} \mu_{i_{sens}} \\ \mu_{i_{spec}} \end{pmatrix}, \mathbf{\Sigma}\right]$$
$$\mathbf{\Sigma} = \begin{pmatrix} \tau_{sens}^2 & \rho\tau_{sens}\tau_{spec} \\ \rho\tau_{sens}\tau_{spec} & \tau_{spec}^2 \end{pmatrix}, \tag{2}$$

where τ_{sens} and τ_{spec} denote the between-observation standard deviation in logit transformed sensitivity and specificity, and ρ denotes the between-observation correlation. In order to model the inherent correlations between multiple sensitivity and specificity data-pairs from the same study, a variance component model can be used (21) such that diagnostic test-threshold combinations are considered fixed effects, while study, and study-specific diagnostic test are considered random effects. Study and study-specific diagnostic test are nested within each observation. The model is given by:

$$\mu_{i_{sens}} = \beta_{j,k} + c_{l(i)} + d_{l(i),j(i)}$$
$$\mu_{i_{spec}} = \delta_{j,k} + e_{l(i)} + f_{l(i),j(i)} \tag{3}$$

where β and δ denote fixed effects of sensitivity and specificity due to diagnostic test, *j*, and test-threshold, *k*, respectively. Parameters *c* and *d*, and *e* and *f,* denote random effects of sensitivity and specificity due to study, *l*, and the interaction between study, *l*, and diagnostic test, *j*, respectively. Non-informative prior distributions were specified for the test-specific and threshold-specific accuracy parameters on the logit scale (with more information given in technical appendix A.1). All models were estimated in a Bayesian framework using Markov Chain Monte Carlo (MCMC) simulation, and implemented in the WinBUGS 1.4.3 software (22). Example WinBUGS code is given in Appendix A.2.

4.2. Network meta-analysis incorporating threshold constraints

Building on the network meta-analysis framework of 4.1, constraints can be specified on the underlying threshold parameters, assuming that overall, higher test-thresholds have an increased sensitivity but decreased specificity. We applied these constraints using a method described by Owen *et al* (23), where the authors place constraints on increasing doses of an intervention for comparative effectiveness research. In this paper, imposing constraints allows information to be borrowed between thresholds within a test, and thus potentially increase the precision in the estimated test accuracy for decision-making (which is further discussed in Section 6). An alternative approach is to specify the underlying cumulative distribution function of the tests (3, 12, 14). This approach is flexible to multiple and different thresholds. However, specifying constraints on the summary estimates of the accuracy measures for increasing test thresholds, as described in this paper, makes fewer assumptions about the distributional form of these estimates and rather this approach simply allows the summary estimates to be greater than or equal to, or less than or equal to, the reference threshold.

5. Analysis of the motivating example

In this section, we synthesized diagnostic test accuracy data across all tests and thresholds of interest. This includes MMSE at threshold <25 and <27, and MoCA at threshold <22 and <26. We illustrate estimates of diagnostic test accuracy using a number of models each with different heterogeneity and correlation assumptions, as described in the technical appendix (Appendix A.1). We extend each of these models to incorporate constraints on increasing thresholds.

Table 2 gives the model fit statistics for each of these models. Using measures of DIC, it is apparent that there is very little difference in model fit across all models considered (a difference in DIC > 5 is considered an important difference). For this reason, the most simplistic model, assuming a common heterogeneity and correlation parameter across models was used for illustration. The results from the remaining models are given in Appendix A.3.

Table 2: Model fit statistics

Model	Posterior between-observation SD(s)		Correlation parameter(s)	DIC
	Sensitivity	**Specificity**		

	MMSE	MoCA	MMSE	MoCA	MMSE	MoCA	
Assuming common heterogeneity and correlation parameters across tests							
Without threshold constraints	0.28 (SD:0.21)		0.17 (SD:0.13)		-0.17 (SD:0.70)		281.59
With threshold constraints	0.29 (SD:0.21)		0.16 (SD:0.13)		-0.16 (SD:0.70)		281.23
Assuming common heterogeneity and test-specific correlation parameters							
Without threshold constraints	0.29 (SD:0.22)		0.17 (SD:0.13)		-0.14 (SD:0.71)	-0.01 (SD:0.71)	281.64
With threshold constraints	0.29 (SD:0.21)		0.17 (SD:0.13)		-0.20 (SD:0.70)	0.02 (SD:0.71)	281.18
Assuming test-specific heterogeneity and common correlation parameters across tests							
Without threshold constraints	0.39 (SD:0.27)	0.29 (SD:0.21)	0.45 (SD:0.38)	0.22 (SD:0.19)	-0.28 (SD:0.68)		282.94
With threshold constraints	0.40 (SD:0.27)	0.28 (SD:0.20)	0.44 (SD:0.38)	0.21 (SD:0.19)	-0.28 (SD:0.68)		282.82

5.1. Assuming common heterogeneity and correlation parameters across tests

This section illustrates the results from a model assuming a common heterogeneity and correlation parameter across multiple, diverse tests. Section 5.1.1 presents estimates of test accuracy from a model without threshold constraints and section 5.1.2 presents results from a model with threshold constraints.

5.1.1. Model without threshold constraints

Table 3 displays the results of the summary test accuracy measures, relative rankings and probabilities that each test was the most accurate overall in terms of the true positive rate (sensitivity) and true negative rate (specificity). Combining diagnostic test accuracy data for all test-threshold combinations illustrated that MoCA<26 appeared to have the optimal true positive rate (sensitivity: 0.97, 95%CrI: 0.94, 0.99) for 99% of MCMC iterations. There appeared to be little difference between MMSE<27 and MoCA<22 in terms of sensitivity, with each of these tests ranking in second (95%CrI: 2,3) and third place (95%CrI: 2,4), respectively. However, MoCA<22 appeared to have a better true negative rate (specificity: 0.77, 95%CrI: 0.67, 0.85) compared to MMSE<27 (specificity: 0.58, 95%CrI: 0.45, 0.70), and subsequently ranked in second place for specificity (Rank: 2, 95%CrI: 1,2). MMSE<25 appeared to have the optimal true negative rate overall (specificity: 0.84, 95%CrI: 0.79, 0.89), ranking in first place for 97% of model iterations.

5.1.2. Model with threshold constraints

Incorporating threshold constraints on increasing thresholds marginally reduced the between-observation standard deviation for specificity (Table 2). However, including threshold constraints had very little impact on the posterior point estimates which mirror those of the unconstrained model (Table 3). Overall, incorporating threshold constraints appeared to marginally increase precision in the effect estimates. Subsequently, MoCA<26 ranked in first place for sensitivity for 100% of MCMC iterations. Incorporating threshold constraints reduced the variability within-studies. The estimated within-study standard deviation for sensitivity was 0.62 (SD: 0.31) compared to 0.63 (SD: 0.30) from the unconstrained model. Similarly, the estimated within-study standard deviation for specificity was 0.42 (SD: 0.21) compared to 0.44 (SD: 0.20) from the unconstrained model. However, the variability within-tests within a study marginally increased when incorporating threshold constraints. For sensitivity, the estimated standard deviation from the model incorporating threshold constraints was 0.30 (SD: 0.22) compared to 0.28 (SD: 0.22) from the unconstrained model. For specificity, the estimated standard deviation was 0.27 (SD: 0.20) compared to a standard deviation of 0.25 (SD: 0.17) for the unconstrained model.

Figure 5 displays the posterior point estimates and 95% credible regions in ROC space. Whilst MoCA<26 and MMSE<25 rank in first place for sensitivity and specificity, respectively, joint and equal consideration of these diagnostic measures would suggest that MoCA<22 appears to have the optimal diagnostic accuracy overall. Table 4 gives the estimated mean difference in sensitivity (top right) and specificity (bottom left) between each of the tests. In comparison to MoCA<22, the estimated sensitivity gained by receiving the optimal true positive test (MoCA<26) is

0.14 with corresponding 95% credible interval (0.07,0.25). Similarly, the estimated specificity gained from receiving the optimal true negative test (MMSE<25) is 0.07 with corresponding 95% credible interval (-0.01,0.18). As these intervals are close to, or span, the point of no difference, there is no strong evidence to suggest that MoCA<22 loses efficiency in accurate diagnosis.

Table 3: Posterior point estimates and 95% credible intervals obtained from a network meta-analysis model assuming a common between-observation standard deviation and correlation parameter

Test	Sensitivity (95%CrI)	Specificity (95%CrI)	Rank best sensitivity (95%CrI)	p(Best) sensitivity	Rank best specificity (95%CrI)	p(Best) specificity
Without threshold constraints						
MMSE<25	0.72 (0.61, 0.82)	0.84 (0.79, 0.89)	4 (3,4)	0	1 (1,2)	0.97
MMSE<27	0.89 (0.81, 0.95)	0.58 (0.45, 0.70)	2 (2,3)	0.01	3 (3,3)	0
MoCA <22	0.82 (0.70, 0.91)	0.77 (0.67, 0.85)	3 (2,4)	0	2 (1,2)	0.03
MoCA <26	0.97 (0.94, 0.99)	0.35 (0.23, 0.48)	1 (1,1)	0.99	4 (4,4)	0
With threshold constraints						
MMSE<25	0.73 (0.62, 0.82)	0.84 (0.79, 0.88)	4 (3,4)	0	1 (1,2)	0.96
MMSE<27	0.90 (0.81, 0.95)	0.58 (0.44, 0.70)	2 (2,3)	0	3 (3,3)	0
MoCA <22	0.83 (0.71, 0.91)	0.77 (0.67, 0.86)	3 (2,4)	0	2 (1,2)	0.04
MoCA <26	0.98 (0.94, 0.99)	0.35 (0.22, 0.47)	1 (1,1)	1	4 (4,4)	0

Figure 5: Posterior point estimates and 95% credible regions in ROC space obtained from a model incorporating threshold constraints and assuming a common heterogeneity and correlation parameter across tests

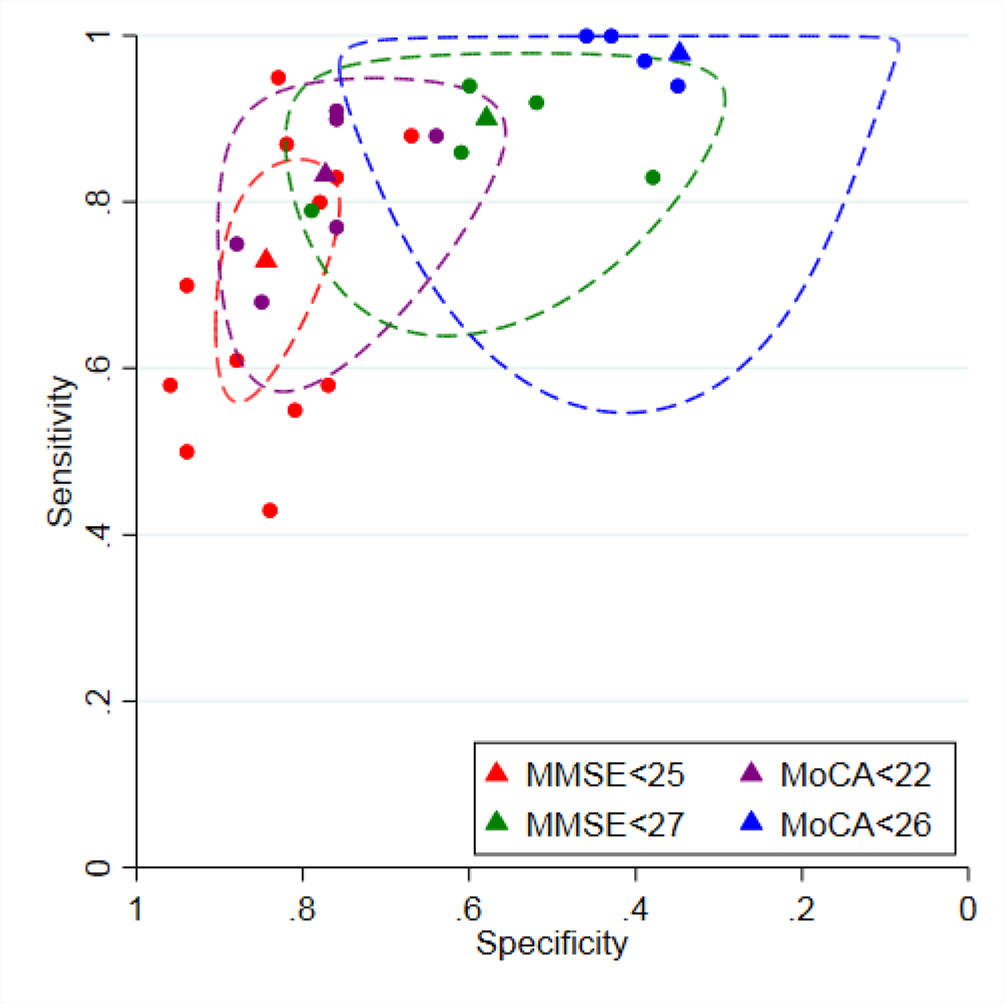


Table 4: Estimated mean difference (95%CrI) in sensitivity (top right) and specificity (bottom left) between each test-threshold combination (row – column) obtained from a model incorporating threshold constraints and assuming a common heterogeneity and correlation parameter across tests

	MMSE<25	MMSE<27	MoCA<22	MoCA<26
MMSE<25	-	0.17 (0.08,0.26)	0.10 (-0.01,0.22)	0.25 (0.15,0.35)
MMSE<27	0.26 (0.15,0.39)	-	-0.07 (-0.18,0.03)	0.08 (0.02,0.16)
MoCA<22	0.07 (-0.01,0.18)	-0.19 (-0.33,-0.06)	-	0.14 (0.07,0.25)
MoCA<26	0.49 (0.38,0.61)	0.23 (0.08,0.37)	0.42 (0.31,0.52)	-

Above the leading diagonal gives estimates of the mean difference (row – column) in sensitivity (95%CrI), and below the leading diagonal gives estimates of the mean difference in specificity (95%CrI)

6. Discussion

In this paper, we propose a unified network meta-analysis framework for synthesizing diagnostic test accuracy data which allows for both the incorporation of multiple tests and multiple explicit thresholds of the same tests. We further developed this framework to incorporate constraints on increasing test thresholds such that estimates for higher test thresholds had an increased sensitivity but decreased specificity compared to lower thresholds of the same test. In this way, by departing from the usually conducted separate analyses of different tests, the combined analysis of multiple tests/threshold allows for more detailed and rigorous comparisons between competing tests. For example in the dementia example presented, we found that MoCA<26 had the optimum true positive rate, and MMSE<25 had the optimum true negative rate of diagnosing cognitive impairment following a stroke. While joint and equal consideration of sensitivity and specificity suggested that MoCA<22 appeared to have the optimal diagnostic test accuracy overall.

From a decision makers', clinicians', and patients' perspective, interest lies in both the true positive rate (sensitivity) and true negative rate (specificity) of a diagnostic test in order to efficiently manage patient care. Therefore, unlike the analysis considered here, consideration to the relative weighting of sensitivity and specificity will be required (i.e. the relative health implications of a false negative compared to a false positive result). In order to make robust decisions regarding the optimal diagnostic test in terms of clinical-effectiveness and/or cost-effectiveness, a fully comprehensive clinical or economic decision model will need to be developed incorporating potential treatment plans and longer-term follow-up.

Further extensions to the network meta-analysis framework described in this paper could include incorporating meta-regression methods. Both observation or study-level covariates could easily be included in the bivariate component of the model as described by Reitsma *et al (8)*, with the aim to explain some of the heterogeneity between observations or between studies. Further work could also look to derive a set of inconsistency equations to assess consistency between different sources of information i.e. comparative versus non-comparative studies.

In clinical practice, a sequential approach to testing is often used. Thus stroke patients may receive a brief diagnostic test initially, and those identified as remaining ‘at risk’ may receive further, more comprehensive, testing such as NPB or clinical diagnosis. In this example, it may be reasonable to administer MoCA<26 initially to ensure that the optimal number of true negative patients are identified and discharged from further routine cognitive assessment. Alternatively, MoCA<22 may be used initially to ensure that the maximum number of true positive and true negative patients are identified. The utility of a staged and triaged approach to diagnostic testing in this setting remains unanswered and provides an opportunity for further work, as does the further development of statistical methods to evaluate the performance of sequences of tests taking lack of dependence into account (4, 24).

A potential limitation of our approach is that it treats different test and threshold combinations as separate tests for the purposes of the analysis, and thus full sROC plots are not estimated across different thresholds, as has been done elsewhere (3, 10-14), however methods to estimate full sROC curves have not taken into account explicit threshold values until very recently (14, 25), or allow for constraints to be placed on the heterogeneity which can be attributed to threshold differences. Combining a comparative framework such as the model described in this paper, with the flexible approach of modelling the distributional form of multiple test thresholds such as the model described by Hoyer *et al* (14, 25), whereby both the model parameters and distributional forms could be estimated simultaneously, is an opportunity for further work.

Further, if studies do not report the threshold values used, or the thresholds cannot be explicitly expressed, as is often the case for tests involving the interpretation of some sort of diagnostic image etc, then the performance of a test at different thresholds cannot be estimated via the model presented. However this is not an argument for not using the valuable extra information when thresholds are known, and more generally there is no reason why the same methodology should be used in these two different contexts (despite this being the case historically).

Table 5 describes the similarities and differences between a number of additional approaches in the current literature to synthesise diagnostic test accuracy data from multiple tests (16, 26-31). All of these methods extend the bivariate model of Reitsma *et al* (8) to synthesise data for two or more tests. Trikalinos *et al* (26), Hoyer and Kuss (27), and Cheng (31) make use of multinomial distributions to model the within-study variation for multiple tests, whilst Dimou *et al* (30) makes use of multivariate normal distributions on logit sensitivities and specificities to account for within-study covariances. If full cross-tabulations, i.e. the full response array across all competing tests, are available, then the correlation between tests can be taken in to account as in Trikalinos *et al* (26), Dimou et al (30), and Cheng (31). These approaches have the advantage of appropriately modelling the within-study covariance structure, however, partial or full cross-tabulations are required for a sufficient number of studies in order to adequately model the correlation between tests (31). A limitation of these approaches is that as the number of competing tests increases, the number of parameters to be estimated by the model rapidly increases, which can result in issues with model convergence (31). Cheng (31) further considers a multivariate extension of the HSROC model by Rutter and Gatsonis (7), and explores the use of beta-binomial marginals and multivariate Gaussian copulas. The author found that use of the beta-binomial marginal and multivariate Gaussian copulas produced less biased estimates of the posterior mean summary points compared to the multivariate extension of the bivariate model and HSROC model, however this approach appeared to be computationally expensive. Nyaga *et al* (28, 29) use a two stage hierarchical model based on the logit transformed sensitivity and specificity (28), and a one-stage approach based on the beta-binomial distribution modelling on the probability scale (29). Both models include shared random effects to induce study level correlations. This approach is most similar to the model we describe in this paper. However, our model further incorporates shared random effects on test to account for multiple thresholds and applies constraints on increasing test-thresholds. Similarly to the model outlined in this paper, many of the approaches described in Table 5 adopt an arm-based approach and model the absolute measures of test accuracy (26-31). Menten and Lesaffre (16) developed a contrast-based approach, which models both the direct (or head-to-head) comparisons of diagnostic tests as well as indirect comparisons through a common diagnostic test. This approach directly models the relative logit sensitivities and specificities between multiple tests. This model can be further extended to a hierarchical latent class model to account for imperfect gold standards. A contrast-based approach works well if all studies evaluate all diagnostic tests. In the case of mixed reporting, further assumptions regarding the missingness of data is required. In a comparison between the contrast-based approach of Menten and Lesaffre (16) and a hierarchical arm-based approach of Nyaga *et al* (28), Nyaga *et al* (28) argue that an arm-based approach may be more appealing than contrast-based approaches since it allows for a more straightforward interpretation of the parameters, makes use of all available data resulting in increased precision, and adopts a more natural variance-covariance structure. In choosing an approach to synthesise diagnostic test accuracy data for multiple tests, the user needs to consider a number of factors related to the decision question. The first of which is the available data; if partial, or full cross-tabulations are available for a sufficient number of studies, a multivariate approach may be preferred to adequately model the within-study correlation structures (16, 26, 30, 31). If there are a number of test-thresholds of interest, a multivariate approach to the HSROC model (31), or an extension to the multinomial approach as described by Hoyer and Kuss (27) may be most appropriate. However, if there are many competing diagnostic tests, the number of additional parameters to be estimated by a multivariate model may be too large causing issues with model convergence. In this instance, a hierarchical model (28, 29), including the model described in this paper, may be preferred.

Furthermore, in the case of many competing tests and multiple test thresholds, our approach using shared random effects and constraints on increasing test thresholds may be considered. Indeed, a comprehensive comparison of all of the approaches described in Table 5, including an assessment of model performance under different criteria and different reporting structures, is still required and would be a valuable addition to the current literature.

In this paper, we evaluated a number of models each with different assumptions regarding the heterogeneity parameters and correlation parameters. The first model assumed common heterogeneity and correlation parameters across tests, the second model assumed common heterogeneity and test-specific correlation parameters, and the third model assumed test-specific heterogeneity parameters and a common correlation parameter. It is worth noting that allowing both heterogeneity and correlation parameters to be test-specific leads to unidentifiability of the covariance matrix, and thus causes issues with model convergence.

In conclusion, this paper proposes a number of network meta-analysis models for synthesizing diagnostic test accuracy data. The proposed frameworks allow for the analysis of multiple tests at multiple thresholds together with the option to incorporate constraints on increasing test thresholds. It could be argued that constraints on threshold effects should be applied to all models with explicit threshold information regardless of model fit, due to the implicit threshold assumption which by definition must be satisfied (but is not imposed by previous models commonly fitted). Incorporating this information through the use of constraints has the potential to more appropriately attribute variability between results to (genuine) threshold effects and better explain heterogeneity between studies.

Table 5: Approaches to synthesising diagnostic test accuracy data of multiple tests

Reference		Model Description	Type of model	Available data	Number of tests	Multiple thresholds per test	Imperfect GS
(26)	*Trikalinos, T. A., (2014). Research Synthesis Methods.*	Multinomial model approximated by multivariate normal distribution, modelling the joint TPR and FPR. Correlations between tests are modelled as random parameters.	Arm-based	Full cross-tabulations	2 index tests + GS	No	No
(16)	*Menten, J. and Lesaffre, E. (2015). BMC Medical Research Methodology.*	Hierarchical model which is partly based on contrasts between transformed sensitivity and specificity (similar to that of NMA for interventions) Adds in allowance for imperfect GS by modelling response pattern across tests as multinomial: latent class models	Contrast-based	Full cross-tabulations	3 index tests + multiple GS	No	Yes
(27)	*Hoyer, A. and Kuss, O. (2016). Statistical Methods in Medical Research.*	Multinomial model similar to that of Trikalinos *et al* (21) but does not account for correlations between tests since full cross-tabulations are not used. This model can be extended to account for multiple test thresholds.	Arm-based	2x2 tables for each test versus GS only	2 index tests + GS from each study	Yes	No
(28, 29)	*Nyaga, V. N., et al. (2016a,b).*	Two-stage hierarchical model based on logit transformed sensitivity and	Arm-based	2 x 2 tables for each test versus GS only	11 tests (in example	No	No

	Statistical Methods in Medical Research.	specificity. Shared random effects are specified to induce study level correlations. One stage approach modelling directly on the probability scale (using beta-binomial distribution) without logit transformation.			dataset) Between 1 and 6 tests per study.		
(30)	*Dimou, N. L, et al. (2016). Statistics in Medicine.*	Multivariate normal distribution with closed form formulae for the within-study covariance matrix (needs full cross-tabulations). Full reporting then used to impute when only 2 x 2 table information is presented. Correlation between tests are estimated and "plugged in".	Arm-based	Full cross-tabulation and 2x2 tables for each test versus GS only	2 index tests + GS (Not all studies have to report both tests)	No	No
(31)	*Cheng, W. (2016). Repository Library, Brown University (Doctoral Thesis)*	Multinomial model with decomposition of test and study specific effects (Chapter 2) Multivariate extension of the HSROC model (Chapter 3) Beta-binomial marginals and multivariate Gaussian copulas (Chapter 4) All models account for study-type specific effects and within study-type random effects	Arm-based	Full cross-tabulations, partially crossed-tabulations, 2x2 tables for each test versus GS only	3 index tests + GS	No Yes No	No

GS: Gold standard

References

1. National Institute for Health and Care Excellence. Guide to the Methods of Technology Appraisal 2013. NICE Process and Methods Guides. London2013.
2. Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM, Group CDTAW. Systematic reviews of diagnostic test accuracy. Ann Intern Med. 2008;149(12):889-97.
3. Steinhauser S, Schumacher M, Rucker G. Modelling multiple thresholds in meta-analysis of diagnostic test accuracy studies. BMC Med Res Methodol. 2016;16(1):97.
4. Novielli N, Sutton AJ, Cooper NJ. Meta-Analysis of the Accuracy of Two Diagnostic Tests Used in Combination: Application to the Ddimer Test and the Wells Score for the Diagnosis of Deep Vein Thrombosis. Value in Health. 2013;16(4):619-28.
5. Deeks JJ, Altman DG. Effect Measures for Meta–Analysis of Trials with Binary Outcomes, in Systematic Reviews in Health Care: Meta-Analysis in Context, Second Edition (eds M. Egger, G. D. Smith and D. G. Altman). London, UK: BMJ Publishing Group; 2001.
6. Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. Medical Decision Making. 1993;13(4):313-21.
7. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta‐analysis of diagnostic test accuracy evaluations. Statistics in medicine. 2001;20(19):2865-84.
8. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of clinical epidemiology. 2005;58(10):982-90.
9. Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics. 2007;8(2):239-51.
10. Dukic V, Gatsonis C. Meta‐analysis of Diagnostic Test Accuracy Assessment Studies with Varying Number of Thresholds. Biometrics. 2003;59(4):936-46.
11. Hamza TH, Arends LR, van Houwelingen HC, Stijnen T. Multivariate random effects meta-analysis of diagnostic tests with multiple thresholds. BMC medical research methodology. 2009;9(1):73.
12. Putter H, Fiocco M, Stijnen T. Meta‐Analysis of Diagnostic Test Accuracy Studies with Multiple Thresholds using Survival Methods. Biometrical Journal. 2010;52(1):95-110.
13. Riley RD, Takwoingi Y, Trikalinos T, Guha A, Biswas A, Ensor J, et al. Meta-analysis of test accuracy studies with multiple and missing thresholds: a multivariate-normal model. J Biomed Biostat. 2014;5:196.
14. Hoyer A, Hirt S, Kuss O. Meta‐analysis of full ROC curves using bivariate time‐to‐event models for interval‐censored data. Research synthesis methods. 2017.
15. Sutton AJ, Cooper NJ, Goodacre S, Stevenson M. Integration of meta-analysis and economic decision modeling for evaluating diagnostic tests. Medical Decision Making. 2008;28(5):650-67.

16. Menten J, Lesaffre E. A general framework for comparative Bayesian meta-analysis of diagnostic studies. BMC Med Res Methodol. 2015;15:70.
17. Suzuki S, Moro-oka T, Choudhry NK. The conditional relative odds ratio provided less biased results for comparing diagnostic test accuracy in meta-analyses. Journal of clinical epidemiology. 2004;57(5):461-9.
18. Siadaty MS, Philbrick JT, Heim SW, Schectman JM. Repeated-measures modeling improved comparison of diagnostic tests in meta-analysis of dependent studies. Journal of clinical epidemiology. 2004;57(7):698-711.
19. Lees R, Fearon P, Harrison JK, Broomfield NM, Quinn TJ. Cognitive and mood assessment in stroke research: focused review of contemporary studies. Stroke. 2012;43(6):1678-80.
20. Lees R, Selvarajah J, Fenton C, Pendlebury ST, Langhorne P, Stott DJ, et al. Test accuracy of cognitive screening tests for diagnosis of dementia and multidomain cognitive impairment in stroke. Stroke. 2014:STROKEAHA. 114.005842.
21. DuMouchel W. Repeated measures meta-analysis. Bulletin of the International Statistical Institute, Session 51, Tome LVII, Book 1: 285-288. 1997.
22. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and computing. 2000;10(4):325-37.
23. Owen RK, Tincello DG, Keith RA. Network meta-analysis: development of a three-level hierarchical modeling approach incorporating dose-related constraints. Value Health. 2015;18(1):116-26.
24. Novielli N, Cooper NJ, Sutton AJ. Evaluating the Cost-Effectiveness of Diagnostic Tests in Combination: Is It Important to Allow for Performance Dependency? Value in Health. 2013;16(4):536-41.
25. Schneider A, Linde K, Reitsma JB, Steinhauser S, Rücker G. A novel statistical model for analyzing data of a systematic review generates optimal cutoff values for fractional exhaled nitric oxide for asthma diagnosis. Journal of clinical epidemiology. 2017;92:69-78.
26. Trikalinos TA, Hoaglin DC, Small KM, Terrin N, Schmid CH. Methods for the joint meta‐analysis of multiple tests. Research synthesis methods. 2014;5(4):294-312.
27. Hoyer A, Kuss O. Meta-analysis for the comparison of two diagnostic tests to a common gold standard: A generalized linear mixed model approach. Statistical methods in medical research. 2016:0962280216661587.
28. Nyaga VN, Aerts M, Arbyn M. ANOVA model for network meta-analysis of diagnostic test accuracy data. Statistical methods in medical research. 2016:0962280216669182.
29. Nyaga VN, Arbyn M, Aerts M. Beta-binomial analysis of variance model for network meta-analysis of diagnostic test accuracy data. Statistical Methods in Medical Research.2016:0962280216682532.
30. Dimou NL, Adam M, Bagos PG. A multivariate method for meta‐analysis and comparison of diagnostic tests. Statistics in medicine. 2016;35(20):3509-23.
31. Cheng W. Network Meta-Analysis of Diagnostic Accuracy Studies: Brown University; 2016.