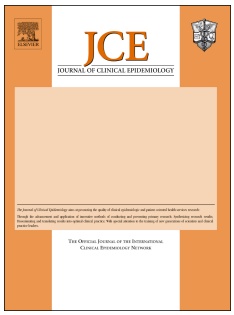
Journal Pre-proof



A new method for synthesizing test accuracy data outperformed the bivariate method

Luis Furuya-Kanamori, Polychronis Kostoulas, Suhail A.R. Doi

PII: S0895-4356(20)31219-1

DOI: https://doi.org/10.1016/j.jclinepi.2020.12.015

Reference: JCE 10376

To appear in: *Journal of Clinical Epidemiology*

Received Date: 23 July 2020

Revised Date: 4 November 2020

Accepted Date: 9 December 2020

Please cite this article as: Furuya-Kanamori L, Kostoulas P, Doi SAR, A new method for synthesizing test accuracy data outperformed the bivariate method, *Journal of Clinical Epidemiology* (2021), doi: https://doi.org/10.1016/j.jclinepi.2020.12.015.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2020 Published by Elsevier Inc.

Journal Pre-proof

Original article

A new method for synthesizing test accuracy data outperformed the bivariate method

Luis Furuya-Kanamori[1], Polychronis Kostoulas[2], and Suhail A. R. Doi[3*]

[1]Research School of Population Health, College of Health & Medicine, Australian National University, Canberra, Australia

[2] Faculty of Public Health, School of Health Sciences, University of Thessaly, Karditsa, Greece

[3]Laboratory for Clinical Epidemiology Methods (LabCEM), Department of Population Medicine, College of Medicine, QU Health, Qatar University, Doha, Qatar

Running title: SCS method for diagnostic meta-analysis

***Correspondence to**
Suhail A. Doi MBBS, MMed, MClinEpid, PhD, FRCP(Edin)
Department of Population Medicine, College of Medicine, Qatar University Drive
P.O. Box 2713, Doha, Qatar
T: +975 66001271
E: s.doi@gmx.net

Journal Pre-proof

Journal Pre-proof

Abstract

Objectives: This paper outlines the development of a new method (split component synthesis; SCS) for meta-analysis of diagnostic accuracy studies and assesses its performance against the commonly used bivariate random effects model.

Methods: The SCS method summarises the study-specific natural logarithm of the diagnostic odds ratios (ln(DOR)), which mainly reflects test discrimination rather than threshold effects, and then splits the summary ln(DOR) into its component parts, logit of sensitivity and logit of specificity. Performance of the estimator under the SCS method was assessed through simulation and compared against the bivariate random effects model estimator in terms of bias, mean squared error (MSE), and coverage probability across varying degrees of between-studies heterogeneity.

Results: The SCS estimator for the DOR, Se, and Sp were less biased and had smaller MSE than the bivariate model estimators. Despite the wider width of the 95% confidence intervals under the bivariate model, the latter had a poorer coverage probability compared to that under the SCS method.

Conclusion: The SCS estimator outperforms the bivariate model estimator and thus represents an improvement in our approach to diagnostic meta-analyses. The SCS method is available to researchers through the *diagma* module in Stata and the *SCSmeta* function in **R**.

Keywords: diagnostic odds ratio; diagnostic accuracy, performance; hierarchical; bivariate; meta-analysis

Journal Pre-proof

Key findings

- A new method is outlined that implements a unified approach to meta-analysis of diagnostic accuracy studies (the SCS method)
- Traditional bivariate methods for meta-analysis of sensitivity and specificity pairs have both more error and poorer error estimation than the new SCS method reported in this paper

What this adds to what is known

- Meta-analysis of diagnostic accuracy studies should start with the unified construct (diagnostic odds ratios) and not sensitivity and specificity pairs
- Better quality evidence can be generated through diagnostic accuracy meta-analyses by changing our approach to such meta-analyses

What is the implication/what should change now

- The new SCS method should be adopted for meta-analysis of diagnostic accuracy studies
- Researchers can access this method through the *diagma* module in Stata and *SCSmeta* function in R.

Journal Pre-proof

Journal Pre-proof

Introduction

As first stated by David Sackett and endorsed widely, evidence-based medicine (EBM) aims to provide the best care for patients through conscientious, explicit, and judicious use of clinical evidence [1]. To ensure the best available evidence in clinical diagnosis, the performance of diagnostic tests need to be properly established. Such evaluations usually involve multiple studies whose results are synthesised to produce a summary estimate of test performance.

When initially implemented, meta-analyses of diagnostic accuracy studies generally pooled sensitivity (Se), specificity (Sp), positive (pLR) or negative (nLR) likelihood ratios. However, this approach lost support because it did not account for the correlation between Se/Sp or pLR/nLR resulting in impossible values when summary LRs were converted into Se or Sp [2;3]. This led to the increasing uptake of a method proposed by Moses and Littenberg of combining independent studies of diagnostic tests into a summary ROC (sROC) curve [4;5]. They had proposed that the study-specific logit-transformed Se (logit(Se)) and Sp (logit(Sp)) be used to fit a linear regression model to estimate the natural logarithm of the summary diagnostic odds ratio (ln(DOR)), which could be used as a single overall indicator of diagnostic accuracy when transformed back to the natural scale.

The Moses-Littenburg method did not, however, preserve the two-dimensional nature of the underlying data and therefore the pooled Se and Sp were not available. For this reason their approach was eventually replaced by the bivariate model, proposed by Reitsma et al. in 2005 [6], which produced summary estimates of Se and Sp. This bivariate modelling approach produced equivalent results to the hierarchical summary receiver operating characteristic (HSROC) model described by Rutter and Gatsonis in 2001 [7] and the empirical Bayes approach introduced by Macaskill in 2004 [8]. Chu and Cole [9] proposed an extension to the bivariate model of Reitsma et al. in 2006, which was a generalised linear mixed model that utilized a statistical modelling approach for sparse data (instead of the continuity correction) and was postulated to perform better in the situation of low cell

84 counts. The bivariate model was eventually adopted over the univariate or Moses-Littenburg
85 approaches [10;11] and today is the most commonly used method for diagnostic meta-analysis [12].
86 An issue with bivariate models are that the inputs into the model are the study-specific pairs
87 of Se and Sp, and the latter can demonstrate heterogeneity across studies either due to systematic
88 differences or dissimilarity in test thresholds or both. The bivariate approach to such heterogeneity is
89 to assume random effects within the modelling assumptions and the latter will typically be
90 approximations at best and are hard to verify [13]. The new-style [14] random effects assumption
91 underpinning the bivariate models may, to some extent, explain why performance deteriorates when
92 systematic error related between-study heterogeneity increases and when number of studies decrease
93 [15]. Another issue is that some of the between study variability could be due to some degree of
94 threshold variability [16] and while the bivariate approach takes the negative correlation between Se
95 and Sp into account when modelling Se/Sp pairs, such a correlation may also be artefactual due to
96 systematic error (study biases), spectrum effects or implicit variations in thresholds when tests are
97 interpreted differently. Of note, a pre-requisite and an implicit assumption to any diagnostic meta-
98 analysis is that there is a similar threshold for the test of interest across all studies and the meta-
99 analysis output is, therefore, for this fixed threshold.
100 In contrast, the DOR and area under the curve (AUC) are solely indices of test discrimination
101 whose maximum values indicate absolute discrimination between diseased and non-diseased states.
102 There has been some suggestion that the DOR could also vary across thresholds [17;18] in which
103 case the shape of the ROC curve may become asymmetrical. This is thought to depend on the
104 underlying distribution of test results in patients with and without the target condition [18].
105 Regardless of the latter observations, in most cases the diagnostic test measures tend to be normally
106 distributed within diseased and non-diseased subpopulations and it has been shown that for a very
107 wide range of choices of the threshold there is almost the same value of the sum of the logit(Se) and
108 logit(Sp) [19;20]. In other words, the ln(DOR) (and thus the DOR) is almost invariant under choice

of the threshold with the most widely seen test scores in medicine. It follows therefore that while Se/Sp pairs reflect both test discrimination and threshold effects, the DOR and AUC are relatively immune to threshold effects and therefore are better candidates for synthesis in meta-analysis.

The four key measures discussed above make up an integral part of a unified diagnostic performance metric and can be related to each other [21;22] as follows:

$$DOR = \frac{Se_t}{(1-Se_t)} \times \frac{Sp_t}{(1-Sp_t)} \qquad [1.1]$$

$$\ln(DOR) = \text{logit}(Se_t) + \text{logit}(Sp_t) \qquad [1.2]$$

$$\text{logit}(AUC) = \ln(DOR)/2 \quad (\text{if DOR} \geq 1) \qquad [1.3]$$

where *t* represents a particular test threshold for the Se/Sp pair. Thus, for a heterogeneous set of studies, it makes sense to synthesize the DOR and derive summary estimates of Se/Sp by splitting the summary DOR into its component parts [23]. This follows from expressions 1.1 and 1.2, and once the Se and Sp are derived, the LRs can also be calculated. In addition, the DOR can be converted into the AUC given expression 1.3..

We utilise these principles for the development of a new method for the meta-analysis of the results of diagnostic accuracy studies. In this paper, we outline the development of the new method and compare its performance (through simulation) against the bivariate model in terms of bias, mean squared error (MSE), and coverage, especially when there is systematic error leading to between-studies heterogeneity.

Development of the new method

This new method (henceforth called the split component synthesis [SCS] method) starts off with the meta-analysis of the DOR across studies using a robust inverse variance heterogeneity (IVhet) model of meta-analysis [24] that is known to maintain performance characteristics under considerable heterogeneity [25] and has no assumptions on the outcome distribution. Although other

meta-analytical models can be used (e.g. the random effects model [26]) this is strongly discouraged as it will result in a rapid drop off in coverage of the confidence interval of the summary DOR and is associated with a larger MSE [25]. Once the summary ln DOR and its standard error are obtained, the summary ln DOR needs to be split into its component logit Se and logit Sp.

The principle behind the splitting of the DOR is that when the DOR in a study changes due to systematic or random error the Se and/or Sp in the same study will move in the same direction. What is needed is to determine from a set of such studies (all of which have been presumably subject to varying degrees of systematic and/or random error) what summary pair of logit Se/Sp corresponds to the summary ln DOR. To do this we use ordinary least squares (OLS) regression of study-specific logit Se or logit Sp on the centred ln DOR (i.e. study-specific ln DOR - summary ln DOR), and this will produce an intercept equal to the desired summary logit Se or logit Sp. This procedure makes sense because OLS regression minimises the presumed error that leads to the varying Se and Sp. Of note, if some of the studies utilize a different threshold, they can be picked up as they will have a different intercept from the rest of the studies on the regression plot (assuming this is not overshadowed by the extent of systematic error). Since the OLS regression is used in a predictive modelling approach this obviates any concern regarding regression dilution and though the dependent and independent variables are correlated, this is not a problem because variance estimates from the regression are not of interest.

The next step is to determine the variance of the summary logit Se and logit Sp. The OLS regression itself does not provide any information about the variance of the summary logit Se and logit Sp. However, from expression 1.2, the ln DOR is the sum of these two estimates; hence the sum of the variance of these two estimates equals the variance of the ln DOR. The variance of the summary logit Se and logit Sp are obtained by splitting the variance of the summary ln DOR. The split is not done equally as the variance of the proportions is dependent on the size of the proportions

and thus the split is done as explained in expression-7 and expression-8 in supplementary material S1.

The summary DOR, Se and Sp, and their variances using the SCS method are now available. In a similar fashion, the summary LRs can be obtained by proportioning the variance of the ln DOR based on their proportional absolute values over the sum of the absolute values for the ln pLR and ln nLR. The summary AUC is estimated from expression 1.3, whereas the standard error of the logit AUC is half of the standard error of the ln DOR. The specific steps involved in the SCS method are provided in the supplementary material S1.

Summary ROC plot

The sROC plot from the SCS method is created for the summary DOR by selecting several Se values across its range and computing its paired Sp according to the following expression which is a rearrangement of expression 1.1 as follows:

$$Se = \left(DOR(1-Sp)\right) / \left(\left(DOR(1-Sp)\right) + Sp\right)$$

The summary Se and Sp intersection point is indicated on the ROC curve as a solid square and its confidence interval indicated by a shaded rectangle whose upper and lower boundaries represent the confidence limits of the Se and left and right boundaries the confidence limit of 1-Sp. The confidence limits of the DOR are also indicated on the plot. Individual study Se / Sp pairs are indicated on the plot as open circles with size proportional to the inverse of the variance of the study ln DOR (Figure 1A and 1B).

Simulation

Data generation

The aim of the simulation was to generate the 4-cell structure of the data for each study (*tp, fp, fn,* and *tn*). To do this a true value of Sp and Se were assigned as well as the study diseased

population size and non-diseased population size. From these 4 parameters, the true cell counts were obtained. The Se was then subjected to repeated draws from a beta distribution with parameters *tp* and *fn*; while the Sp was subjected to repeated draws from a beta distribution with parameters *tn* and *fp*. This was sufficient to introduce random error, but to add in systematic error the 4-cell counts were divided by a positive scaled parameter with increasing value greater than 1 across runs. After application of the scale parameter, the Se and Sp were drawn from the beta distributions with rescaled values of *tp* & *fn*, or *tn* & *fp* respectively. The scale variable was derived by a transformation of the bias variance whose computation has been described previously [25].

In this simulation, 10 levels of bias variance and therefore 10 levels of increasing values of the scale variable were set. Therefore, the simulation was conducted in 10 runs with run 1 representing random error alone (scale parameter = 1) and runs 2 – 10 having increasing value of the scale parameter and thus additional systematic error. One thousand meta-analyses were simulated in each run, each containing 10 studies, although the range of studies was from 6-10 as the beta distribution in Stata reports a missing value at certain extremes which allowed us to examine performance under varying study numbers per meta-analysis. The population size of diseased and non-diseased in each study was drawn from a uniform distribution between 35 and 175 to mirror the sample sizes reported in such studies [27]. The simulation protocol for the data generation is detailed within the Stata code in the supplementary material S2.

Performance comparison between the SCS method and the bivariate model

For each level of heterogeneity, summary DOR, Se, and Sp estimated by the SCS method and the bivariate model (using the generalised linear mixed model approach [9]) were compared based on mean absolute estimation error squared (bias squared), MSE, width of the confidence interval, and coverage probability as we have previously described [25]. The actual degree of systematic error in each run was estimated by the median between studies variance (tau^2) computed for each meta-

analysis. The Stata codes used for the performance comparison are provided in the supplementary material S3–S5.

The simulation study results revealed that the SCS estimator (for DOR, Se, and Sp) was less biased (Figure 2A) and had a smaller MSE than the bivariate model estimator (Figure 2B). Despite the wider width of the 95% confidence intervals under the bivariate model (Figure 2C), it had a poorer coverage probability of the confidence interval compared to that under the SCS method (Figure 2D).

When extensive heterogeneity was introduced (i.e. median tau^2 >1), there was a substantial drop in performance for the bivariate model with a significant increase in type I error of up to 35%. The SCS method coverage probability remained stable bot under extensive heterogeneity and increased sample sizes of 200 – 2000 per simulated study (Supplementary material S6 & S7). The simulation was repeated 19 times with meta-analyses including different pairs of Se/Sp (DORs from 0.1 – 0.9 in steps of 0.1, DOR of 1 and DORs 2 – 10 in steps of 1) and the performance comparisons remained similar (results not shown).

Application to data from a published meta-analysis

A diagnostic meta-analysis by Wacker et al. [28] examined the performance of procalcitonin in differentiation of septic patients (i.e. sepsis, severe sepsis, or septic shock) from those with a systemic inflammatory response syndrome of non-infectious origin. The performance of procalcitonin was examined using the SCS method and the bivariate model (using the generalised linear mixed model approach). The analysis was conducted in Stata MP-64 version 14, College Station, TX using the *midas* module [29] for the bivariate model while the SCS method was run using a new Stata module created with this paper (*diagma)* [30] as well as the *SCSmeta* R function with code given in supplementary material S8 [31].

The meta-analysis included 31 datasets (3244 participants) and all the estimates from the SCS method were more conservative than with the bivariate method. Both methods had similar summary Se and Sp; however, the DOR (8 and 13), AUC (0.73 and 0.85) were very different across the two methods (Table 1 and Figure 1) and this was expected because the bivariate analysis computes the DOR from its components instead of the proper vice-versa sequence. .

Discussion

In this paper, we introduce the SCS method and demonstrate that its performance under systematic error was superior to that of the bivariate method currently being used. This is probably because the SCS method starts off with input of the DOR and has no modelling assumptions while inputs are the Se/Sp pairs and random effects are assumed under the hierarchical and bivariate models [6;13]. Of note, the SCS method had smaller bias and MSE and the coverage was kept to nominal levels despite a narrower width of the confidence interval.

The discriminative capacity of a diagnostic test can be summarised by two main measures (that are mostly independent of threshold) – the DOR and the AUC [19;20]. The larger these values are for a test, the more discrimination it has between diseased and non-diseased individuals. The main difference between the DOR and the AUC is in their ranges and interpretation. The DOR ranges between 0 and ∞ while the AUC ranges between 0 and 1. Nevertheless, their pragmatic ranges are between 1 and ∞ for the DOR and 0.5 and 1 for the AUC. Given the relationship between ln(DOR) and logit(AUC) (as shown in expression 1.3), ln(DOR) can be transformed into the logit(AUC) and vice versa. While the DOR is an index of test discrimination, it can be partitioned into several other indices of test performance (i.e. Se, Sp, pLR, and nLR) [23]. For every DOR there are many (sometimes infinite) values for these indices because they are threshold dependant unlike the DOR. It is therefore not appropriate to meta-analyse these four measures in a univariate or

bivariate analysis when there is systematic error between studies, because the distinction between variation due to systematic error and variation due to implicit variation in thresholds gets blurred. This is one reason why the bivariate model fails to achieve optimal performance when there is heterogeneity [13;15]. The bivariate method models the correlation both between and within study between Se and Sp [32;33]. Riley et al have shown that the bivariate method may produce an increased precision of results compared to a method that does not consider such correlations although such benefits are likely to be marginal at best [32]. We do not demonstrate this benefit with the bivariate method within our simulations probably because we no longer simulate the way the data will be analysed which was a weakness in previous studies. The performance estimates from the simulation in this paper do not confirm this benefit.

The improvement proposed here is to meta-analyse the DOR and then partition it into its component parts. One limitation that may arise in this approach is when false positive (or false negative) values equal 0 leading to unidentifiable DOR and bias in the SCS method. In such instances, the continuity correction is utilized – as is the case with the classical meta-analysis of OR – which may introduce some bias associated with it [34]. Nevertheless, even in this case, the MSE and coverage are still better than the corresponding estimates under the bivariate model. Hence, the overall model performance remains better for the SCS method, even though some theoretical bias may be – and is – introduced in this case.

One form of variability across diagnostic studies is the spectrum effect which implies that disease symptomatology and severity or characteristics of patients can affect Se and Sp, or both. In this situation, research has shown that bias may not occur because the Se and Sp move in opposite directions and the pLR and nLR within the subgroup may remain similar to the overall pLR/nLR [35]. This means that spectrum effects may not alter the discrimination of a test but will act as if the threshold has changed and studies may cluster at different points on the SROC plot based on the disease/patient spectrum. The meta-analysis will therefore produce an average Se and Sp across all

patient spectrums. For this reason, Moons et al. [36] have suggested that Se and Sp may have no direct diagnostic meaning because they vary across patient populations and subgroups within populations and support the argument that post-test probabilities remain stable as discussed above. It is our view that there is an advantage in pursuing Se and Sp over and above post-test probabilities, and that is to determine, for an average subject of the types represented in the trials, what the expected false-positive and false-negative rates are likely to be. It is therefore important to obtain summary estimates of the component parts of the DOR (Se, Sp, pLR and nLR) for decision making.

The suggestion by some researchers that these components of the DOR are affected by prevalence [37] is actually reflected by three types of problems: a spectrum effect at different levels of prevalence, low precision in low prevalence states or different implicit thresholds at different levels of prevalence [38]. These are all problems related to systematic error and, in the simulations performed, a relationship with study level prevalence could also be demonstrated in some of the heterogeneous meta-analyses. The latter probably reflects systematic error mimicking spectrum effects. By extension of this logic, meta-analysis of predictive values [39] will have a similar issue as these measures are not solely characteristics of the test, but instead reflect the prevalence in the study population. This also has implications for the trivariate synthesis of Se, Sp, and prevalence [40] and thus all these concepts linked to prevalence of disease need to be reconsidered when contemplating meta-analysis of diagnostic accuracy studies.

The sROC plot, given that we assume that the meta-analysis is of studies at a common threshold, will always be symmetric since multiple Se and Sp values are computed from a single summary DOR to create the curve as shown in Figure 1A. Thus, asymmetric ROC curves do not occur with the SCS method and it is important to point out that if the test threshold varies across studies, the SCS method can still be used to synthesize the DOR, but not its Se/Sp components. A way to check for varying thresholds is to look at the scatter of study points on the sROC plot and if the studies cluster at different points on the sROC curve (as opposed to a scatter around one point on

the curve), it is likely that thresholds are varying. The latter is however not very sensitive as the scatter patterns may look similar with systematic error or spectrum effects, If the SCS method is used to synthesize components of the DOR when there are thought to be spectrum effects, it results in an average across the spectrum of disease.

Another issue with diagnostic meta-analyses is publication bias. We noticed very high levels of asymmetry under simulated heterogeneity and this is not surprising since Begg [41] concluded, based on data from Deeks et al., [42] that the validity of tests of publication bias is compromised when the DOR is high, cut-off value is extreme and prevalence of disease is low; reflecting the fact that these features tend to lead to extreme 2x2 tables with low cell frequencies in which the undesired correlation between DOR and its variance is most apparent. We therefore advocate caution in concluding that asymmetry indicates publication bias when these circumstances are present. The method for publication bias incorporated into the *diagma* module is not P-value driven [43] but nevertheless depends on the variance of the study DOR and suffers from the same issues flagged by Begg when used for diagnostic meta-analyses. We support the recommendation by Begg of first examining the results for heterogeneity and then the reasons behind the heterogeneity as the preferred approach for making sense of the data [41].

In conclusion, our results suggest that the new SCS method represents an improvement in our approach to meta-analysis of diagnostic accuracy studies given that it is associated with less MSE and better coverage (despite a smaller width of the confidence interval) than is seen with the commonly used bivariate and related hierarchical models. To make the SCS method accessible to researchers, we have developed the *diagma* module [30] which is available in Stata (type *ssc install diagma* in the command window) and the *SCSmeta* function in R [31].

Journal Pre-proof

Acknowledgement

Author CRediT statement
Luis Furuya-Kanamori: Methodology, Formal analysis, Software, Writing - Review & Editing, Data curation. **Polychronis Koustoulas:** Formal analysis, Software, Writing - Review & Editing. **Suhail Doi:** Conceptualization, Methodology, Supervision, Formal analysis, Writing - Original Draft, Writing - Review & Editing , Funding acquisition.

Funding

This work was made possible by Program Grant #NPRP10-0129-170274 from the Qatar National Research Fund (a member of Qatar Foundation) to Suhail A. Doi. The findings herein reflect the work, and are solely the responsibility of the authors. All authors had full access to all the data in the study and the corresponding author (SD) had final responsibility for the decision to submit for publication and is the guarantor of this study.
LFK was supported by an Australian National Health and Medical Research Council Fellowship (APP1158469).

Conflicts of interest

The authors do not have any conflicts of interest to declare.

Journal Pre-proof

References

1. Knottnerus, J.A. & Tugwell, P. (2017) Evidence-based medicine: achievements and prospects. *J Clin Epidemiol*, **84,** 1-2.

2. Zwinderman, A.H. & Bossuyt, P.M. (2008) We should not pool diagnostic likelihood ratios in systematic reviews. *Stat Med*, **27,** 687-97.

3. Shapiro, D.E. (1995) Issues in combining independent estimates of the sensitivity and specificity of a diagnostic test. *Acad Radiol*, **2 Suppl 1,** S37-47; discussion S65-9, S83.

4. Moses, L.E., Shapiro, D., & Littenberg, B. (1993) Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Stat Med*, **12,** 1293-316.

5. Littenberg, B. & Moses, L.E. (1993) Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. *Med Decis Making*, **13,** 313-21.

6. Reitsma, J.B., Glas, A.S., Rutjes, A.W., Scholten, R.J., Bossuyt, P.M., & Zwinderman, A.H. (2005) Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol*, **58,** 982-90.

7. Rutter, C.M. & Gatsonis, C.A. (2001) A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med*, **20,** 2865-84.

8. Macaskill, P. (2004) Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. *J Clin Epidemiol*, **57,** 925-32.

9. Chu, H. & Cole, S.R. (2006) Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *J Clin Epidemiol*, **59,** 1331-2; author reply 1332-3.

10. Dinnes, J., Mallett, S., Hopewell, S., Roderick, P.J., & Deeks, J.J. (2016) The Moses-Littenberg meta-analytical method generates systematic differences in test accuracy compared to hierarchical meta-analytical models. *J Clin Epidemiol*, **80,** 77-87.

11. Harbord, R.M., Whiting, P., Sterne, J.A., Egger, M., Deeks, J.J., Shang, A., & Bachmann, L.M. (2008) An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. *J Clin Epidemiol*, **61,** 1095-103.

12. Ochodo, E.A., Reitsma, J.B., Bossuyt, P.M., & Leeflang, M.M. (2013) Survey revealed a lack of clarity about recommended methods for meta-analysis of diagnostic accuracy data. *J Clin Epidemiol*, **66,** 1281-8.

13. Begg, C.B. (2008) Meta-analysis methods for diagnostic accuracy. *J Clin Epidemiol*, **61,** 1081-2; discussion 1083-4.

14. Hodges, J.S. (2013) Random effects old and new. In *Richly Parameterized Linear Models. Additive, Time Series, and Spatial Models Using Random Effects* (Hodges, J.S., ed), pp. 285-302. Chapman and Hall/CRC, USA.

15. Diaz, M. Performance measures of the bivariate random effects model for meta-analyses of

391 diagnostic accuracy. Computational Statistics & Data Analysis 83, 82-90. 2015.

392 16. Naaktgeboren, C.A., Ochodo, E.A., Van Enst, W.A., de Groot, J.A.H., Hooft, L., Leeflang,
393 M.M.G., Bossuyt, P.M., Moons, K.G.M., & Reitsma, J.B. (2016) Assessing variability in
394 results in systematic reviews of diagnostic studies. *BMC Med Res Methodol*, **16,** 6.

395 17. Irwig, L., Macaskill, P., Glasziou, P., & Fahey, M. (1995) Meta-analytic methods for
396 diagnostic test accuracy. *J Clin Epidemiol*, **48,** 119-30; discussion 131-2.

397 18. Glas, A.S., Lijmer, J.G., Prins, M.H., Bonsel, G.J., & Bossuyt, P.M. (2003) The diagnostic
398 odds ratio: a single indicator of test performance. *J Clin Epidemiol*, **56,** 1129-35.

399 19. Edwards, J.H. Some Taxonomic Implications of a Curious Feature of the Bivariate Normal
400 Surface. Br J Prev Soc Med 20[1], 42-43. 1966.

401 20. Hasselblad, V. & Hedges, L.V. (1995) Meta-analysis of screening and diagnostic tests. *Psychol*
402 *Bull*, **117,** 167-78.

403 21. Suzuki, S., Moro-oka, T., & Choudhry, N.K. (2004) The conditional relative odds ratio
404 provided less biased results for comparing diagnostic test accuracy in meta-analyses. *J Clin*
405 *Epidemiol*, **57,** 461-9.

406 22. Walter, S.D. & Sinuff, T. (2007) Studies reporting ROC curves of diagnostic and prediction
407 data can be incorporated into meta-analyses using corresponding odds ratios. *J Clin Epidemiol*,
408 **60,** 530-4.

409 23. Simel, D.L., Easter, J., & Tomlinson, G. (2013) Likelihood ratios, sensitivity, and specificity
410 values can be back-calculated when the odds ratios are known. *J Clin Epidemiol*, **66,** 458-60.

411 24. Doi, S.A., Barendregt, J.J., Khan, S., Thalib, L., & Williams, G.M. (2015) Advances in the
412 meta-analysis of heterogeneous clinical trials I: The inverse variance heterogeneity model.
413 *Contemp Clin Trials*, **45,** 130-8.

414 25. Doi, S.A.R. & Furuya-Kanamori, L. (2020) Selecting the best meta-analytic estimator for
415 evidence-based practice: a simulation study. *Int J Evid Based Healthc*, **18,** 86-94.

416 26. DerSimonian, R. & Laird, N. (1986) Meta-analysis in clinical trials. *Control Clin Trials*, **7,**
417 177-88.

418 27. Bachmann, L.M., Puhan, M.A., Ter Riet, G., & Bossuyt, P.M. Sample sizes of studies on
419 diagnostic accuracy: literature survey. BMJ 332[7550], 1127-1129. 2006. British Medical
420 Journal Publishing Group.

421 28. Wacker, C., Prkno, A., Brunkhorst, F.M., & Schlattmann, P. (2013) Procalcitonin as a
422 diagnostic marker for sepsis: a systematic review and meta-analysis. *Lancet Infect Dis*, **13,**
423 426-35.

424 29. MIDAS: Stata module for meta-analytical integration of diagnostic test accuracy studies.
425 Dwamena, B. 2007. https://ideas.repec.org/c/boc/bocode/s456880.html, Boston College
426 Department of Economics. Statistical Software Components.

427 30. DIAGMA: Stata module for DIAGnostic Meta-Analysis using the split component synthesis

method. Furuya-Kanamori, L. and Doi, S. A. R. 2020. https://ideas.repec.org/c/boc/bocode/s458815.html, Boston College Department of Economics, Statistical Software Components.

31. SCSmeta: The Split Component Synthesis function for meta-analysis of diagnostic test accuracy studies. Kostoulas, P., Furuya-Kanamori, L., and Doi, S. A. R. 2020. https://rpubs.com/polyvet/SCSMeta.

32. Riley, R.D., Abrams, K.R., Lambert, P.C., Sutton, A.J., & Thompson, J.R. (2007) An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Stat Med*, **26,** 78-97.

33. Riley, R.D., Abrams, K.R., Sutton, A.J., Lambert, P.C., & Thompson, J.R. (2007) Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC Med Res Methodol*, **7,** 3.

34. Sweeting, M.J., Sutton, A.J., & Lambert, P.C. (2004) What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Stat Med*, **23,** 1351-75.

35. Goehring, C., Perrier, A., & Morabia, A. (2004) Spectrum bias: a quantitative and graphical analysis of the variability of medical diagnostic test performance. *Stat Med*, **23,** 125-35.

36. Moons, K.G. & Harrell, F.E. (2003) Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies. *Acad Radiol*, **10,** 670-2.

37. Leeflang, M.M., Bossuyt, P.M., & Irwig, L. (2009) Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. *J Clin Epidemiol*, **62,** 5-12.

38. Li, J. & Fine, J.P. (2011) Assessing the dependence of sensitivity and specificity on prevalence in meta-analysis. *Biostatistics*, **12,** 710-22.

39. Leeflang, M.M., Deeks, J.J., Rutjes, A.W., Reitsma, J.B., & Bossuyt, P.M. (2012) Bivariate meta-analysis of predictive values of diagnostic tests can be an alternative to bivariate meta-analysis of sensitivity and specificity. *J Clin Epidemiol*, **65,** 1088-97.

40. Chu, H., Nie, L., Cole, S.R., & Poole, C. (2009) Meta-analysis of diagnostic accuracy studies accounting for disease prevalence: alternative parameterizations and model selection. *Stat Med*, **28,** 2384-99.

41. Begg, C.B. (2005) Systematic reviews of diagnostic accuracy studies require study by study examination: first for heterogeneity, and then for sources of heterogeneity. *J Clin Epidemiol*, **58,** 865-6.

42. Deeks, J.J., Macaskill, P., & Irwig, L. (2005) The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol*, **58,** 882-93.

43. Furuya-Kanamori, L., Xu, C., Lin, L., Doan, T., Chu, H., Thalib, L., & Doi, S.A.R. (2020) P value-driven methods were underpowered to detect publication bias: analysis of Cochrane review meta-analyses. *J Clin Epidemiol*, **118,** 86-92.

466
467

Journal Pre-proof

468 **Table 1.** Summary estimates using the split component synthesis method and the bivariate model for
469 31 datasets that assessed procalcitonin as a diagnostic marker for sepsis in critically ill patients
470

	Split component synthesis method	**Bivariate model**
Sensitivity	0.72 (0.66-0.78)	0.77 (0.72-0.81)
Specificity	0.74 (0.68-0.80)	0.79 (0.74-0.84)
Positive likelihood ratio	2.82 (2.07-3.82)	3.70 (2.95-4.63)
Negative likelihood ratio	0.37 (0.28-0.50)	0.29 (0.24-0.36)
Diagnostic odds ratio	7.57 (4.93-11.61)	12.56 (8.82-17.88)
Area under the curve	0.73 (0.69-0.77)	0.85 (0.81-0.88)
Between-study heterogeneity (I-squared): 66.3%		

471

Journal Pre-proof

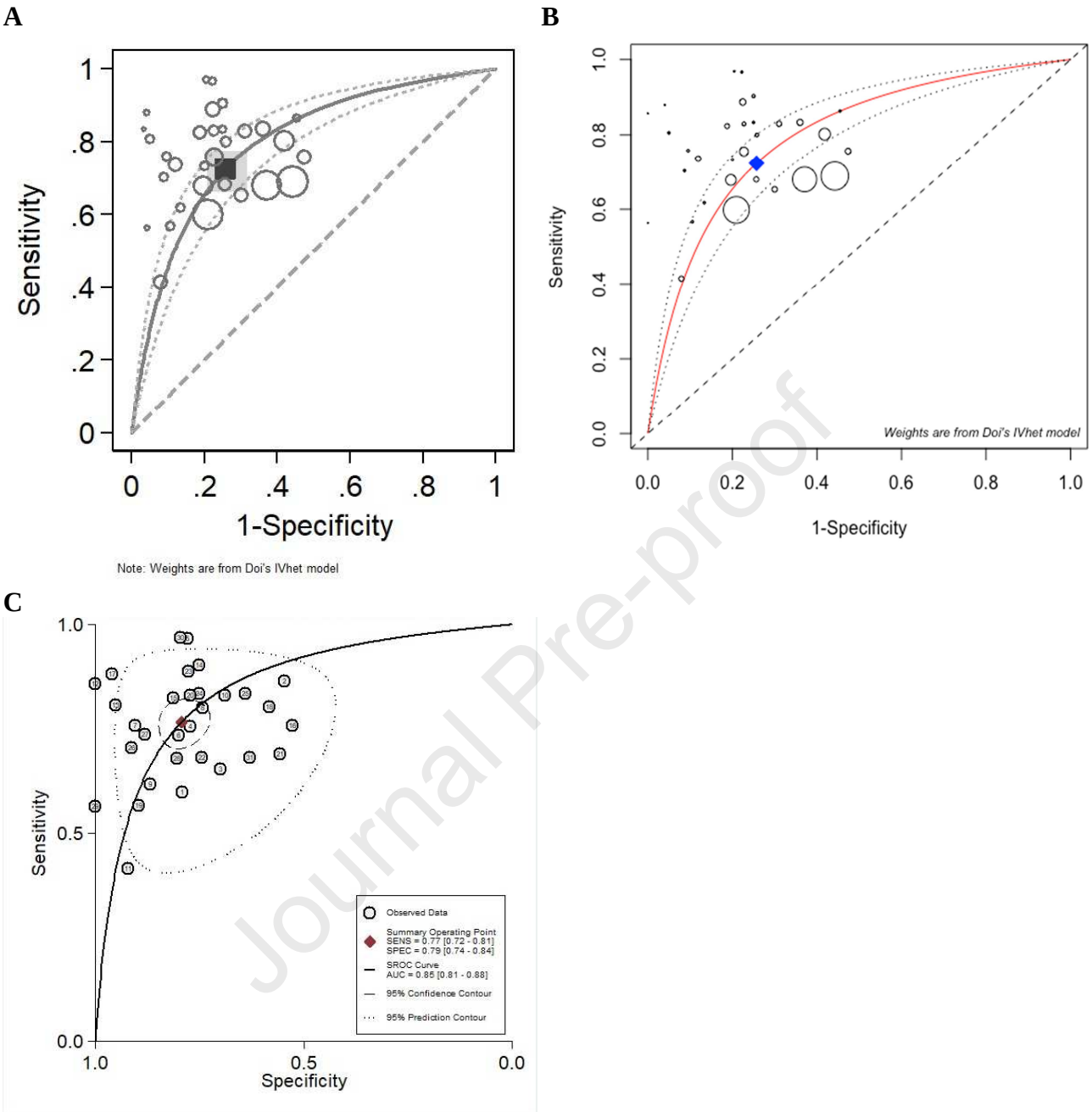


Figure 1. Summary ROC plots using the SCS method (A, *diagma* in Stata and B, *SCSmeta* in R) and the bivariate model (C, *midas* in Stata) for 31 datasets that assessed procalcitonin as a diagnostic marker for sepsis in critically ill patients

Journal Pre-proof

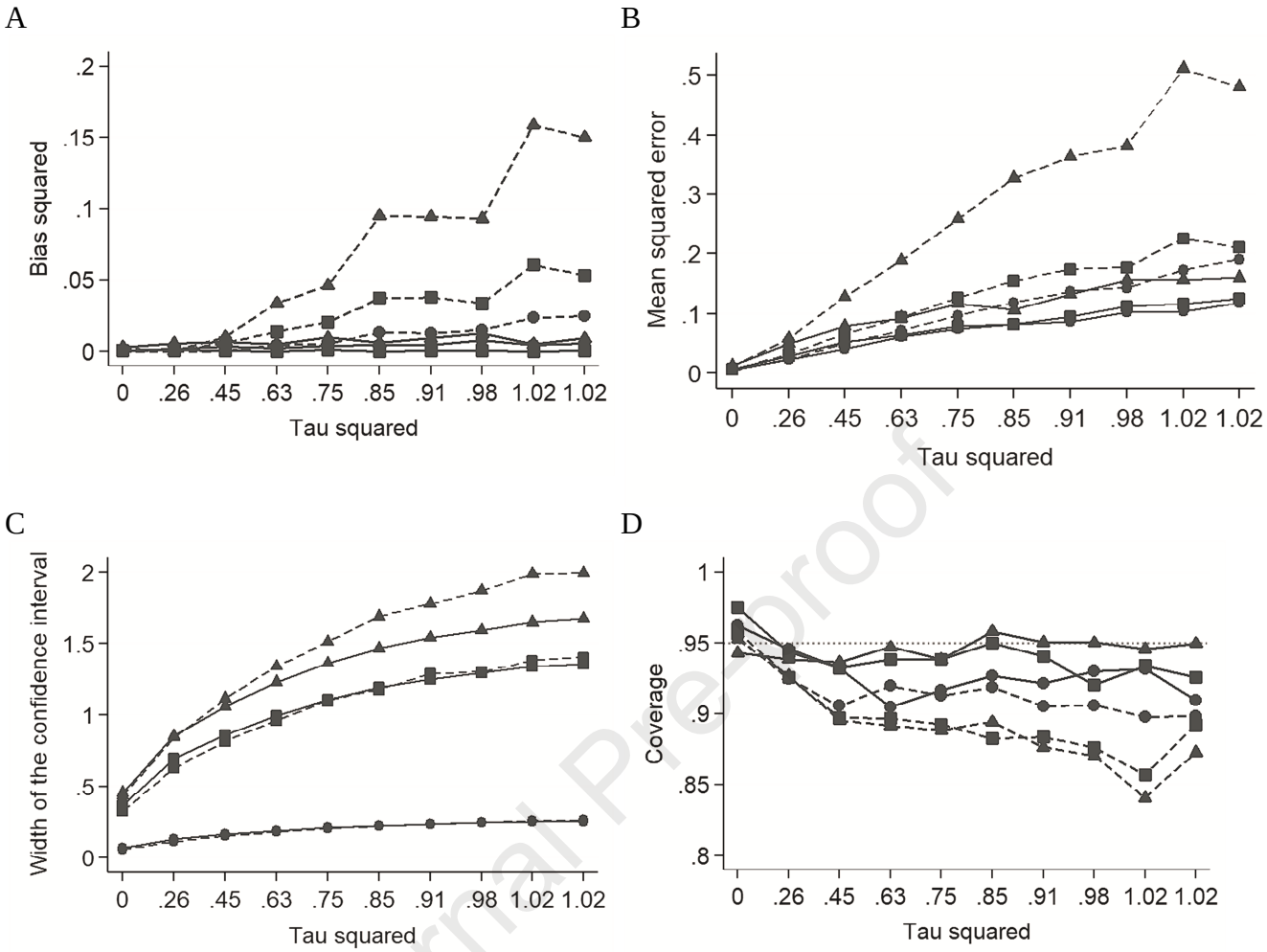


Figure 2. Performance comparison of diagnostic odds ratio (triangle), sensitivity (circle), and specificity (square) between the split component synthesis method (continuous) and the bivariate model (dashed)

Journal Pre-proof

Journal Pre-proof

Highlights

- Meta-analysis of diagnostic studies are currently undertaken by synthesizing sensitivity (Se) and specificity (Sp) pairs from included studies
- Bivariate linear mixed models are the most popular method for synthesis
- The Se/Sp pairs are components of the diagnostic odds ratio (DOR) but are threshold variant unlike the DOR
- This paper proposes that synthesis models should start from the DOR and then split this into its components rather than the reverse as is currently done
- The new method describes how this split can be achieved and has two main strengths:It is free of modelling assumptions and circumvents threshold effects at the synthesis stage
- Performance measures demonstrate a lower mean squared error and nominal coverage for the new method with smaller CI width all of which are poorer with the traditional approach
- Software is available to run the new method

Journal Pre-proof

Journal Pre-proof

Author CRediT statement

Luis Furuya-Kanamori: Methodology, Formal analysis, Software, Writing - Review & Editing, Data curation. **Polychronis Koustoulas:** Formal analysis, Software, Writing - Review & Editing. **Suhail Doi:** Conceptualization, Methodology, Supervision, Formal analysis, Writing - Original Draft, Writing - Review & Editing , Funding acquisition.

Journal Pre-proof

Journal Pre-proof

There are no conflicts of interest

Journal Pre-proof