Journal Title Here, 2022, pp. 1–12

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

Hierarchical multinomial processing tree models for meta-analysis of diagnostic accuracy studies

Annamaria Guolo[1,*]

[1]Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241/243, I-35121, Padova, Italy
[*]Corresponding author. annamaria.guolo@unipd.it
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Meta-analysis represents a widely accepted approach for evaluating the accuracy of diagnostic tools in clinical and psychological investigations. This paper investigates the applicability of multinomial tree models recently suggested in the literature under a fixed-effects formulation for assessing the accuracy of binary classification tools, where the study specific disease prevalences are taken into account. The model proposed in this paper extends previous results to a hierarchical structure accounting for the variability between the studies included in the meta-analysis. Interestingly, by exploiting the parameter separability of the complete likelihood function, the resulting hierarchical multinomial tree model is shown to coincide, in its interest parameter component, with the well-known bivariate random-effects model under an exact within-study distribution for the number of true positives and true negatives subjects. The proposal is in line with a latent-trait approach, where inference follows a frequentist point of view. The applicability of the proposed model and its performance with respect to its fixed-effects counterpart and to the approximate bivariate random-effects model based on normality assumptions commonly used in the literature is evaluated in a series of simulation studies. Methods are applied to a real meta-analysis about the accuracy of the confusion assessment method as delirium screening tool.

Key words: Diagnostic test, Prevalence, Random-effects, Sensitivity, Specificity

Introduction

The evaluation of a patient's disease status or the early detection of a certain disorder in clinical and psychological investigations is often reached through very accurate instruments, which are typically expensive, time-consuming, or discomforting. As a consequence, their large scale application is not appealing and the need for simpler, inexpensive, but still accurate classification tools remains an aim of the research.

The accuracy of new proposed classification or diagnostic tools is evaluated through the comparison to a reference test assumed to be unquestionable, also called gold standard. In the last years, meta-analysis of diagnostic studies has been widely accepted as an approach for the assessment of the accuracy of a diagnostic or screening test in identifying a patient's specific status, or, more generally, in distinguishing between diseased and nondiseased patients. A diagnostic study is commonly evaluated in terms of sensitivity, i.e., the conditional probability of testing positive in subjects classified as positive by the reference test, and specificity, i.e., the conditional probability of testing negative in subjects classified as negative by the reference test. As an alternative, a diagnostic test is evaluated using a two-by-two table of agreement between the test results and the reference test results (e.g., Honest and Khan, 2002).

The accuracy of a novel diagnostic test is often assessed using meta-analysis methods (e.g., Jackson et al., 2011). Within this framework, the bivariate hierarchical model (Reitsma et al., 2005; Arends et al., 2008) is currently a well-established technique. It is preferable to the traditional approach based on separate analyses for sensitivity and specificity, which do not account for the correlation between the diagnostic measures of accuracy. In addition, the bivariate hierarchical model improves on the popular proposal in Littenberg and Moses (1993) and in Moses et al. (1993) to construct a summary receiver operating characteristic curve based on the regression of the difference between sensitivity and specificity on their sum, a solution which has been criticised for not providing reliable inferential conclusions (e.g., Rutter and Gatsonis, 2001 ; Arends et al., 2008). The bivariate model has a hierarchical structure accounting for the within-study sampling variability and for the between-study variability arising from differences due, for example, to study design's characteristics. Likelihood inference in this framework is affected by several issues (e.g., Guolo, 2017; Takwoingi et al., 2017). Authors warn against the risk of unreliable conclusions when the sample size is small, as well as the risk of non-convergence of the optimisation algorithms. Computational obstacles, e.g., the need for numerical integration, reduce the appealing of the approach. The mentioned issues leave space to alternative solutions, as,

for example, solutions relaxing likelihood assumptions and relying on simulation strategies (e.g., Guolo, 2017). The use of copulas to account for the correlation between the accuracy measures has been proposed in Kuss et al. (2014). Beta-binomial margins for the numbers of true positives and true negatives are linked by a bivariate copula distribution, resulting a likelihood function in closed-form. Zapf et al. (2015) suggest a non-parametric approach to the analysis, with large flexibility with respect to the correlation structure between the accuracy measures, and no convergence issues.

Chu et al. (2009b) extend the bivariate mixed-effects model to jointly model the information from disease prevalence, sensitivity, and specificity of the diagnostic test. The model retains some computational issues of the bivariate counterpart, see Chen et al. (2015) and Chen et al. (2017a). Pseudo-likelihood solutions have been proposed to overcome such limitations, see Chen et al. (2015), Chen et al. (2017a), and Guolo (2023). The use of trivariate copulas to account for the correlation between the accuracy measures and the disease prevalence has been recently proposed in Hoyer and Kuss (2015) and Nikoloulopoulos (2018), the last one including the model in Chen et al. (2015) as a special case.

This paper investigates the applicability of multinomial tree models, starting from a recent proposal in Botella et al. (2013) within the psychological literature, to assess the accuracy of binary classification tools. In particular, in this paper an extension of the fixed-effects multinomial tree model in Botella et al. (2013) is proposed, which turns out into a hierarchical model accounting for between-study heterogeneity. The resulting hierarchical multinomial tree model gives rise to some interesting results. The log-likelihood function for the proposed model shows a clear separation of the parameters associated to the prevalence of the disease and parameters associated to the diagnostic accuracy measures, a property which makes inferences advantageous. In addition, the resulting log-likelihoods focused on the interest parameters component coincides with that from the traditional bivariate random-effects model under the exact – binomial – distribution for the number of true positives and true negatives within each study included in the meta-analysis. Accordingly, the proposed multinomial tree model approach provides an alternative way to arrive to the usual estimate of sensitivity and specificity given by the classical bivariate model. Such a way has the advantage of explicitly following and describing the path or sequence of steps – latent process – leading to the response of the diagnostic test while incorporating the prevalence of the disease. The performance of the proposed multinomial tree based method is compared to that of its fixed-effects counterpart and to that of the likelihood-based approach for the classical bivariate random-effects model under the normal approximation for transformation of study sensitivity and specificity. The methods are compared under different scenarios, including increasing sample size and increasing correlation between sensitivity and specificity. Scenarios include different transformations of sensitivity and specificity given by logit function, probit function, and cloglog function. The applicability of the competing methods is also evaluated on a meta-analysis about the accuracy of the confusion assessment method as delirium screening tool (Shi et al., 2013).

Methods

Consider a meta-analysis of n diagnostic accuracy studies. Each study i, $i = 1, \ldots, n$, provides information about the number of true positives TP_i, true negatives TN_i, false positives FP_i, and false negatives FN_i, see Table 1. Let D_i^+ be the number of diseases subjects and let D_i^- be the number of non-diseased subjects. The estimates of sensitivity (SE_i) and specificity (SP_i) can be obtained from study i as $\widehat{SE}_i = TP_i/D_i^+$ and $\widehat{SP}_i = TN_i/D_i^-$, respectively. A common evaluation of the accuracy of the studies is in terms of a real-line transformation $\eta_i = g(SE_i)$ and $\xi_i = g(SP_i)$, with $g(\cdot)$ usually chosen to be the logit transformation. In this case, $\eta_i = logit(SE_i) = \log\{SE_i/(1-SE_i)\}$ and $\xi_i = logit(SP_i) = \log\{SP_i/(1-SP_i)\}$. Other choices are possible, as the probit transformation and the cloglog transformation, although rarely adopted. The estimates of η_i and ξ_i represented by $\hat{\eta}_i$ and $\hat{\xi}_i$, respectively, are obtained from the sample counterpart.

Table 1. Two-by-two table of data from the comparison between the test under evaluation and the reference standard.

	Reference test		
Test	Disease status	No disease status	
Positives	TP_i	FP_i	
Negatives	FN_i	TN_i	
	D_i^+	D_i^-	n_i

Likelihood-based approach

The original bivariate random-effects model for meta-analysis of diagnostic accuracy studies follows the formulation developed in Reitsma et al. (2005) and in Arends et al. (2008). The model has a hierarchical structure distinguishing the within-study level, describing variability inside each study included in the meta-analysis, and the between-study level accounting for the heterogeneity among the studies, as a consequence of different study designs' or patients' characteristics.

In line with the traditional approach, consider sensitivity and specificity expressed according to a transformation from the $[0, 1]$ interval to the real line and let η_i and ξ_i, respectively, denote such transformation. Usually, the logit transformation is adopted. At the within-study level, a model is specified for the observed $\left(\hat{\eta}_i, \hat{\xi}_i\right)^\top$ conditionally on the true study specific $(\eta_i, \xi_i)^\top$. A computationally convenient formulation is based on the normal approximation

$$\begin{pmatrix} \hat{\eta}_i \\ \hat{\xi}_i \end{pmatrix} \Bigg| \begin{pmatrix} \eta_i \\ \xi_i \end{pmatrix} \sim N\left(\begin{pmatrix} \eta_i \\ \xi_i \end{pmatrix}, \Gamma_i\right), \tag{1}$$

with known diagonal covariance matrix Γ_i, characterized by non-zero entries estimated in each study given the independence between positive and negative subjects at the within-study level. In case of logit transformation,

$$\Gamma_i = \begin{pmatrix} {D_i^+}^{-1} + \left(D_i^+ - TP_i\right)^{-1} & 0 \\ 0 & {D_i^-}^{-1} + \left(D_i^- - TN_i\right)^{-1} \end{pmatrix}.$$

At the between-study level, the random effects $(\eta_i, \xi_i)^\top$ follow a normal distribution, namely,

$$\begin{pmatrix} \eta_i \\ \xi_i \end{pmatrix} \sim N\left(\mu = \begin{pmatrix} \bar{\eta} \\ \bar{\xi} \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_\eta^2 & \rho\sigma_\eta\sigma_\xi \\ \rho\sigma_\eta\sigma_\xi & \sigma_\xi^2 \end{pmatrix}\right), \tag{2}$$

where $\bar{\eta}$ and $\bar{\xi}$ are the means over the studies, σ^2_η and σ^2_ξ are the between-study variances and ρ is the correlation between η_i and ξ_i. The combination of (1) and (2) gives rise to a normal-normal model, with marginal specification

$$\begin{pmatrix} \hat{\eta}_i \\ \hat{\xi}_i \end{pmatrix} \sim N\left(\begin{pmatrix} \bar{\eta} \\ \bar{\xi} \end{pmatrix}, \Gamma_i + \Sigma \right).$$

The associated log-likelihood function for the whole parameter vector $\theta = \left(\bar{\eta}, \bar{\xi}, \sigma^2_\eta, \sigma^2_\xi, \rho\right)^\top$ is available in closed form and it can be conveniently computed using standard software. Despite the computational advantages of the approach, several studies in the literature have highlighted the drawbacks, mainly related to the risk of unreliable inferential conclusions with few or sparse data. See, e.g., Chu et al. (2006), Guolo (2017), Takwoingi et al. (2017). An alternative within-study model specification considers the exact distribution of observed true positives and false positives as realisations of binomial variables, instead of approximating the estimated transformations $\hat{\eta}_i, \hat{\xi}_i$ through a normal distribution. Namely,

$$TP_i \sim Binom\left(D_i^+, \left(1+\exp^{-\eta_i}\right)^{-1}\right),$$

$$TN_i \sim Binom\left(D_i^-, \left(1+\exp^{-\xi_i}\right)^{-1}\right),$$

see, e.g., Arends et al. (2008) and Hamza et al. (2008). The exact binomial specification for the true positives and false positives combined with the normal specification (2) for the between-study level gives rise to a marginal generalised linear mixed model, with no closed-form expression for the associated likelihood function, namely,

$$\ell(\theta) = \sum_{i=1}^{n} \log \int\int \frac{e^{\eta_i TP_i}}{(1+e^{\eta_i})^{D_i^+}} \times \frac{e^{\xi_i TN_i}}{(1+e^{\xi_i})^{D_i^-}} \phi_2\left(\eta_i, \xi_i; \mu; \Sigma\right) d\eta_i d\xi_i. \quad (3)$$

Numerical integration is needed for likelihood computation and convergence issues can arise, in terms of non-positive definite covariance matrix or estimates of the parameters of the covariance matrix on the boundary of the parameter space. Such a drawback is more relevant in case of small sample size (e.g., Chen et al., 2017b; Guolo, 2017; Takwoingi et al., 2017).

Multinomial processing tree models

Multinomial tree models (MTMs) represent a popular class of models for categorical behavioral data widely used in psychological research as an instrument to investigate cognitive processes (Riefer and Batchelder, 1988; Batchelder and Riefer, 1999; Erdfelder et al., 2009). MTM analysis of categorical data is based on the assumption that the sample frequencies observed for a set of responses follow a multinomial distribution. As a relevant feature of the approach, interest is not only on the probabilities associated to the sample frequencies, but also to the path, or latent processes, leading to a response or behaviour of the cognitive process. Differently from classical modeling of categorical data, e.g., using log-linear models or logit models, MTMs are structured in way to reflect a cognitive process, represented by a sequence of processing stages, each of them resulting in a response category. The path followed by the cognitive process is conveniently represented by a tree with a single root, where each branch is a sequence of potential cognitive stages ending with the response category. The probability associated to each category is given by the sum of the probabilities associated to the branches leading to the same response (Batchelder and Riefer, 1999). Traditionally, data are aggregated across subjects, and then analyzed under the assumption of independently and identically distribution. Hierarchical extensions of the multinomial processing tree model accounting for between-subjects heterogeneity are the latent-trait approach and the beta-multinomial processing tree approach, both introducing random effects associated to the subjects specific parameters, although under different distributional specification. See, for example, Heck et al. (2018). The latent-trait approach in Klauer (2010) considers a probit transformation of the random components at the population level, following a multivariate normal distribution, in this way explicitly incorporating correlation structures. Inference is then performed according to a Bayesian perspective. The beta-multinomial processing tree approach assumes that the random components follow independent beta distributions. Not modelling potential correlations makes the approach less attractive. See also Jobst et al. (2020). Both the hierarchical extensions of the MTMs represent a valid alternative to the latent-class approach (Klauer, 2006; Stahl and Klauer, 2007), where discrete population-level distributions model the between-study variability, with substantial computational effort.

The application of MTMs in meta-analysis of diagnostic accuracy studies has been originally proposed in Botella et al. (2013). According to this interpretation, the cognitive process is the result of the tools used to classify the patients according to their status, with response categories given by the cells in Table 1. Suppose that study i included in the meta-analysis has an associated parameter π_i reflecting the prevalence of the disease. Then, the tree diagram in Figure 1 illustrates the process of assessment of the patients' status, under the assumption of perfect reference standard.

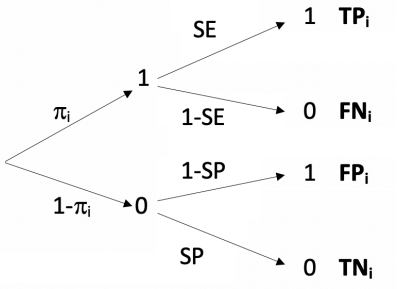

Fig. 1. Tree diagram associated to study i.

The multinomial tree model assumes that the studies are homogeneous and independent, with possible different sample sizes and different prevalences π_i, $i = 1, \ldots, n$, and common diagnostic measures SE and SP. The probability associated to each response or category in study i is

$$p_{TP_i} = \pi_i SE; \qquad p_{FN_i} = \pi_i(1-SE);$$

$$p_{FP_i} = (1-\pi_i)(1-SP); \qquad p_{TN_i} = (1-\pi_i)SP.$$

The number of parameters to be estimated is $n+2$, with $2n-2$ degrees of freedom. As a consequence, the model

needs at least two studies for estimation. The associated log-likelihood function for the whole parameter vector $\theta_{MTM} = (\pi_1, \ldots, \pi_n, SE, SP)^\top$ is

$$\ell(\theta_{MTM}) = \sum_{i=1}^{n} \left[TP_i \log(\pi_i SE) + FN_i \log\{\pi_i (1-SE)\} + FP_i \log\{(1-\pi_i)(1-SP)\} + TN_i \log\{(1-\pi_i) SP\}\right].$$

In this paper we consider a hierarchical extension of the multinomial tree model in Botella et al. (2013) for meta-analysis of diagnostic tests. Let $g(\cdot)$ denote a general link function to translate the test accuracy measure SE and SP to the real line and use a multivariate normal distribution to model the random-effects associated to the transformation of SE and SP. The link function $g(\cdot)$ is not restricted to probit, as in Klauer (2010), but it can be chosen among classical transformations as logit, probit, cloglog. See, for example, Chen et al. (2017b) for an evaluation of the composite likelihood approach for meta-analysis of diagnostic accuracy studies under different link functions. Differently from the latent-trait approach in Klauer (2010), inference will be performed from a frequentist point of view. Let the $g(\cdot)$ link function be

$$g(SE_i) = \eta_i = \log \frac{SE_i}{1-SE_i}, \quad g(SP_i) = \xi_i = \log \frac{SP_i}{1-SP_i}$$

in case of logit transformation,

$$g(SE_i) = \eta_i = \Phi^{-1}(SE_i), \quad g(SP_i) = \xi_i = \Phi^{-1}(SP_i)$$

in case of probit transformation,

$$g(SE_i) = \eta_i = \log\{-\log(1-SE_i)\},$$

$$g(SP_i) = \xi_i = \log\{-\log(1-SP_i)\}$$

in case of cloglog transformation. Then, given the above specifications, the hierarchical version of the MTM model (HMTM) in Botella et al. (2013) has associated log-likelihood function for the whole parameter vector $\theta_{HMTM} = \left(\pi_1, \cdots, \pi_n, \bar{\eta}, \bar{\xi}, \sigma^2_\eta, \sigma^2_\xi, \rho\right)^\top$ equal to

$$\ell_{HMTM}(\theta_{HMTM}) = \sum_{i=1}^{n} \log \int\int \pi_i^{D_i^+} (1-\pi_i)^{D_i^-} \times \eta_i^{TP_i} (1-\eta_i)^{FN_i} (1-\xi_i)^{FP_i} \xi_i^{TN_i} \times \phi_2(\eta_i, \xi_i; \mu; \Sigma)\, d\eta_i d\xi_i, \quad (4)$$

where $\phi_2(\eta_i, \xi_i; \mu; \Sigma)$ is the density function of the bivariate normal distribution for $(\eta_i, \xi_i)^\top$ with mean μ and covariance matrix Σ, as in (2). The log-likelihood function (4) has separable parameters, and it distinguishes a component accounting for disease study prevalences $(\pi_1, \cdots, \pi_n)^\top$ and another component associated to the diagnostic accuracy parameters $\left(\bar{\eta}, \bar{\xi}, \sigma^2_\eta, \sigma^2_\xi, \rho\right)^\top$. Accordingly,

$$\ell_{HMTM}(\theta_{HMTM}) = \ell_{HMTM,1}(\pi_1, \cdots, \pi_n) + \ell_{HMTM,2}\left(\bar{\eta}, \bar{\xi}, \sigma^2_\eta, \sigma^2_\xi, \rho\right).$$

The separability of the parameters in the likelihood function implies that inference on the diagnostic accuracy parameters can be based only on

$$\ell_{HMTM,2}\left(\bar{\eta}, \bar{\xi}, \sigma^2_\eta, \sigma^2_\xi, \rho\right) = \sum_{i=1}^{n} \log \int\int \eta_i^{TP_i} (1-\eta_i)^{FN_i} \times (1-\xi_i)^{FP_i} \xi_i^{TN_i} \times \phi_2(\eta_i, \xi_i; \mu; \Sigma)\, d\eta_i d\xi_i. \quad (5)$$

Whichever the specification of the link function $g(\cdot)$, the two-dimensional integral needs to be solved numerically, for example, via Gauss-Hermite quadrature. Interestingly, the log-likelihood function (5) coincides with the log-likelihood function obtained by substituting the within-study approximate distribution (1) with the exact distribution of TP_i and TN_i, given in (3) with logit link. Differently from the exact likelihood approach, the use of the hierarchical MTM approach allows to follow and describe the path or sequence of steps of the process leading to the response of the diagnostic test under study. In addition, such a process takes into account the within-study prevalences $\pi_i, i = 1, \ldots, n$ of the disease as starting point of the path. Although the presence of prevalences $\pi_i, i = 1, \ldots, n$ makes the dimension of the whole parameter vector increase with the sample size, the special structure of the likelihood function (5) with separable parameters allows a straightforward independent estimation of the prevalences, based on the restricted likelihood

$$\ell_{HMTM,1}(\pi_1, \cdots, \pi_n) = \sum_{i=1}^{n} \log\{\pi_i^{P_i} (1-\pi_i)^{N_i}\}.$$

The estimate of the study-specific disease prevalences can be obtained in closed form,

$$\hat{\pi}_i = \frac{D_i^+}{n_i}$$

as the fraction of positives in each study of dimension n_i, with standard errors given by

$$se(\hat{\pi}_i) = n_i^3 \left(D_i^{+-1} \cdot D_i^{--1}\right).$$

With reference to the diagnostic accuracy parameters, instead, an appropriate evaluation of standard error is via the sandwich method, see Kauermann and Carroll (2001).

Simulation study

The performance of the proposed hierarchical multinomial tree model has been investigated through a series of simulation studies under a variety of scenarios and compared to that of the likelihood-based approach for the bivariate random-effects model under the normal approximation for transformation of study sensitivity and specificity and to that of the fixed-effects MTM in Botella et al. (2013). Comparison with the likelihood-based approach for the bivariate random-effects model under the exact within-study model is not carried out since the resulting likelihood coincides with the interest component likelihood from the hierarchical MTM approach (5).

Data simulation follows a two-stage procedure. First, for given number of studies n, the sample size n_i of each study included in the meta-analysis is generated from a uniform variable on $[50, 200]$ and the number of true positives in each study is generated from a binomial distribution with parameters n_i and a given prevalence of the disease. Then, for each

true positive or true negative in the study, the corresponding classification provided by the test under evaluation is obtained as the result of a binomial distribution with probability of success given by the inverse of the link function $g(\cdot)$. We distinguish logit function, probit function, and cloglog function. The comparison between the data generated at the first step and the data generated at the second step provides the two-by-two Table 1 for study i. From the Table, quantities $\hat{\eta}_i$ and $\hat{\xi}_i$ can be determined according to the chosen link function $g(\cdot)$. Examined scenarios include sample size n varying in $\{10, 25\}$, increasing prevalence of the disease at the population level, ranging in $\{0.08, 0.20, 0.35\}$, increasing correlation between accuracy measures $\rho \in \{0.2, 0.6, 0.8\}$, and large accuracy of the test $(SE, SP)^\top = (0.9, 0.85)^\top$ or smaller accuracy of the test $(SE, SP)^\top = (0.80, 0.92)^\top$. The simulation is based on 1,000 replicates of each scenario. All the methods are implemented in the `R` programming language (R Core Team, 2022). The code is available at https://github.com/annamariaguolo/MTM-meta-analysis.

Maximum likelihood estimation is carried out using the Nelder and Mead algorithm (Nelder and Mead, 1965), with integral evaluation for the MTM model based on the Gauss-Hermite quadrature with 21 nodes. Starting values for the optimization procedure are given by the empirical estimates of sensitivity and specificity and the empirical prevalences for the MTM model. In case of nonconvergence of the optimization algorithm, other solutions have been examined, as the quasi-Newton BFGS algorithm and changes of the starting values of the parameters.

Results

Methods are compared in terms of bias, standard deviation, and average of standard errors of the estimators of the parameters. Empirical coverages for Wald-type confidence intervals at nominal level 0.95 for parameters $\bar{\eta}$ and $\bar{\xi}$ are also reported. Results under nonconvergence were excluded when evaluating the simulations results. The failure rate of the approaches is another criterion used for comparison.

Tables 2-3 report the bias, the standard deviation and the average of standard error for the estimators of the fixed-effects components $\bar{\eta}$ and $\bar{\xi}$, for the estimators of the variance components σ^2_η and σ^2_ξ, and for the estimator of the correlation parameter ρ, under increasing values of ρ, different link functions $g(\cdot)$, large accuracy of the test $(SE, SP)^\top = (0.9, 0.85)^\top$, small number of studies included in the meta-analysis $n = 10$, by distinguishing small and large prevalence of the disease, equal to 0.20 and to 0.35, respectively. Similar results for very small prevalence of the disease (0.08), for low accuracy of the test and for large sample size are reported in the Supplementary Material. The use of the approximate likelihood approach gives rise to more biased estimates of $\bar{\eta}$ and the associated variance component, if compared to the HMTM solution. A similar behaviour is experienced for the estimator of $\bar{\xi}$, although at a lower extent. In addition, the correlation parameter tends to be overestimated. Such a performance of the competing methods is more evident in case of logit link and for small values of the correlation ρ, with biased more pronounced under low prevalence of the disease, see Table 2 for $\pi \in \{0.20, 0.35\}$ and Table S1 for $\pi = 0.08$ in the Supplementary Material. Results tend to be less biased under the probit link. The use of HMTM, conversely, produces almost unbiased estimators, without being substantially affected by changes of the link function, of the values of the correlation or of the prevalence of the disease. The price to pay is a slight increase of the variability associated to the estimates. The use of the fixed-effects MTM approach provides substantially biased estimates of $\bar{\eta}$ and $\bar{\xi}$, whichever the link function and the prevalence of the disease. The difference between the standard deviation and the average standard error confirms the inadequacy of the method.

In case of very small prevalence of disease $\pi = 0.08$ results emphasize a small bias for the HTMTM approach if compared to alternatives, with a larger variability, under all the examined scenarios. Bias of the approximate likelihood approach is comparable to that of the fixed-effects approach, when the interest is on $\bar{\eta}$, while is reduced for $\bar{\xi}$. See Table S1 in the Supplementary Material.

The relative behaviour of the competing methods is maintained under increasing sample size $n = 25$, see the corresponding results in Tables S2-S4 in the Supplementary Material. For the hierarchical approaches, increasing the sample size gives rise to a better performance in terms of bias of the estimators of the fixed-effects parameters, the variance components and the correlations, and in terms of the associated variability, as expected.

Figure 2 reports the empirical coverage of Wald-type confidence intervals at nominal level 95% for the estimators of $\bar{\eta}$ and $\bar{\xi}$ obtained from the approximate likelihood approach and from the HMTM approach, under small and large sample size, for low disease prevalence $\pi = 0.20$. Results from the fixed-effects MTM approach are not reported, as the method results in very poor empirical coverages, whichever the examined scenario. The advantages of HMTM in reaching the target 95% is substantial. The use of the approximate likelihood approach provides empirical coverage probabilities notably below the target level, as a consequence of biased estimates of the parameters, whichever the link function. As the associated standard error reduces with increasing the sample size, it turns out that coverages can be even worse for $n = 25$, see, for example, the results for the estimator of $\bar{\eta}$. The use of HMTM provides empirical coverage probabilities closer to the 95% level, with an improved performance as the sample size increases. Similar results are obtained for larger prevalence of disease, $\pi = 0.35$, see Figure 3 and smaller prevalence of the disease, $\pi = 0.08$, see Figure S1 in the Supplementary Material.

Methods behave substantially differently from a computational point of view. The approximate likelihood does not require computational effort, while the application of the HMTM approach requires numerical integration. This leads also to difference in terms of failure rate. Failure is a consequence of estimates of the correlation parameters on the boundary of the parameter space and/or not positive definite covariance matrix and it is mainly experienced in case of small sample size $n = 10$. Convergence problems affect both the approaches, and the HMTM solution in particular, in case of logit link and small value of the correlation parameter ρ. The result is not unexpected, as previous findings in the literature of meta-analysis of diagnostic accuracy studies confirm the risk of convergence problems for the exact likelihood approach. See, for example, Guolo (2017), Takwoingi et al. (2017), and Chen et al. (2017b). The failure rate of the approximate likelihood solution tends to deeply reduce under the probit link or the cloglog link, while the decay for the HMTM approach is slower. Increasing the sample size helps to eliminate the convergence problems for both the approaches.

When the accuracy of the test is low, namely, $(SE, SP)^\top = (0.80, 0.92)^\top$, the bias is reduced for all the estimators of the parameters, especially under the approximate likelihood

solution. See the results reported in Tables S5-S7 for $n = 10$ and Tables S8-S10 for $n = 25$, under increasing prevalence of the disease. The relative performance is maintained, with results from the approximate normal solution more biased for $\bar{\xi}$ than for $\bar{\eta}$, although at a lower extent. This result is in line with previous findings in Hamza et al. (2008), who highlighted the worse performance of the approximate normal model when sensitivity increases, in a meta-analysis of sensitivity alone.

The empirical coverages probabilities at nominal level 0.95 confirm a satisfactory behaviour of HMTM under all the examined scenario, with values closer to the target level than the likelihood approach, although differences between the approaches are less marked than under the large accuracy case. See Figures S2-S4 in the Supplementary Material. Convergence problems are reduced for both the approaches, if compared to the high accuracy case.

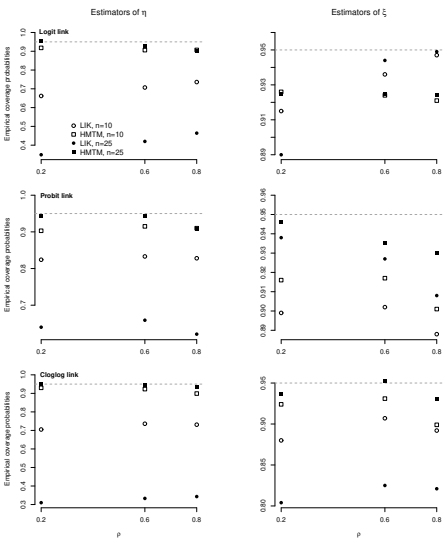

Fig. 2. Empirical coverage probability of Wald-type confidence interval for the estimators of $\bar{\eta}$ and $\bar{\xi}$ obtained from the approximate likelihood approach (LIK) and from the hierarchical MTM approach (HMTM), under increasing ρ, increasing sample size n and different link function. High accuracy of the test. Prevalence of disease $\pi = 0.20$. Dashed line: nominal 95% level.

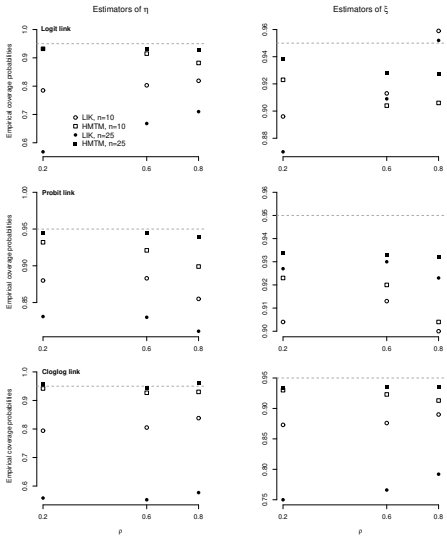

Fig. 3. Empirical coverage probability of Wald-type confidence interval for the estimators of $\bar{\eta}$ and $\bar{\xi}$ obtained from the approximate likelihood approach (LIK) and from the hierarchical MTM approach (HMTM), under increasing ρ, increasing sample size n and different link function. High accuracy of the test. Prevalence of disease $\pi = 0.35$. Dashed line: nominal 95% level.

Table 2. Bias (standard deviation s.d., average standard error s.e.) for the estimators of $\bar{\eta}, \bar{\xi}$ obtained from the fixed-effects MTM approach and for the estimators of $\bar{\eta}, \bar{\xi}, \sigma^2_\eta, \sigma^2_\xi, \rho$ obtained from the approximate likelihood approach (LIK) and the hierarchical MTM approach (HMTM), under increasing ρ and different link function. High accuracy of the test. Sample size $n = 10$. Prevalence of disease $\pi = 0.20$.

Method	ρ	$\bar{\eta}$	$\bar{\xi}$	σ^2_η	σ^2_ξ	ρ
		Bias (s.d., s.e.)	Bias (s.d., s.e.)	Bias (s.d., s.e.)	Bias (s.d., s.e.)	Bias (s.d., s.e.)
Logit link						
MTM	0.2	-0.372 (0.474, 0.255)	-0.146 (0.250, 0.101)			
LIK		-0.622 (0.313, 0.380)	-0.054 (0.237, 0.235)	-0.680 (0.436, 0.342)	-0.123 (0.244, 0.206)	0.066 (0.347, 0.252)
HMTM		-0.029 (0.465, 0.461)	0.022 (0.268, 0.259)	-0.290 (0.697, 0.586)	-0.034 (0.319, 0.268)	0.009 (0.583, 0.399)
MTM	0.6	-0.392 (0.460, 0.252)	-0.158 (0.249, 0.100)			
LIK		-0.574 (0.320, 0.371)	-0.020 (0.232, 0.243)	-0.664 (0.414, 0.339)	-0.061 (0.313, 0.234)	0.190 (0.291, 0.218)
HMTM		-0.021 (0.477, 0.459)	0.016 (0.263, 0.270)	-0.199 (0.674, 0.603)	-0.004 (0.398, 0.290)	0.000 (0.481, 0.304)
MTM	0.8	-0.371 (0.483, 0.256)	-0.153 (0.252, 0.101)			
LIK		-0.551 (0.312, 0.367)	0.024 (0.237, 0.251)	-0.647 (0.434, 0.349)	-0.019 (0.320, 0.258)	0.238 (0.244, 0.180)
HMTM		0.026 (0.467, 0.445)	0.032 (0.274, 0.275)	-0.154 (0.658, 0.627)	0.033 (0.371, 0.318)	0.024 (0.363, 0.235)
Probit link						
MTM	0.2	-0.475 (0.320, 0.107)	-0.285 (0.235, 0.062)			
LIK		-0.243 (0.267, 0.236)	-0.024 (0.207, 0.193)	-0.697 (0.327, 0.130)	-0.131 (0.165, 0.130)	0.049 (0.324, 0.241)
HMTM		-0.001 (0.426, 0.415)	0.015 (0.261, 0.257)	-0.102 (0.768, 0.582)	-0.018 (0.333, 0.268)	-0.002 (0.463, 0.320)
MTM	0.6	-0.489 (0.313, 0.106)	-0.289 (0.228, 0.061)			
LIK		-0.250 (0.258, 0.240)	-0.038 (0.203, 0.192)	-0.679 (0.334, 0.240)	-0.135 (0.169, 0.131)	0.156 (0.251, 0.196)
HMTM		-0.014 (0.392, 0.408)	0.002 (0.255, 0.249)	-0.097 (0.769, 0.581)	-0.029 (0.315, 0.248)	-0.029 (0.345, 0.226)
MTM	0.8	-0.479 (0.329, 0.107)	-0.292 (0.230, 0.061)			
LIK		-0.244 (0.275, 0.240)	-0.043 (0.218, 0.195)	-0.684 (0.341, 0.239)	-0.134 (0.154, 0.132)	0.202 (0.193, 0.155)
HMTM		-0.003 (0.418, 0.391)	0.008 (0.279, 0.258)	-0.138 (0.741, 0.521)	-0.017 (0.300, 0.247)	-0.034 (0.224, 0.125)
Cloglog link						
MTM	0.2	-0.429 (0.265, 0.083)	-0.284 (0.200, 0.043)			
LIK		-0.344 (0.219, 0.211)	-0.138 (0.174, 0.165)	-0.831 (0.234, 0.202)	-0.235 (0.147, 0.109)	0.037 (0.313, 0.230)
HMTM		0.013 (0.408, 0.415)	0.011 (0.264, 0.254)	-0.015 (0.875, 0.634)	0.005 (0.349, 0.268)	-0.027 (0.420, 0.303)
MTM	0.6	-0.432 (0.252, 0.082)	-0.278 (0.193, 0.043)			
LIK		-0.337 (0.203, 0.211)	-0.125 (0.168, 0.164)	-0.826 (0.234, 0.211)	-0.237 (0.147, 0.106)	0.174 (0.228, 0.183)
HMTM		0.004 (0.418, 0.399)	0.026 (0.259, 0.249)	-0.047 (0.860, 0.598)	0.005 (0.360, 0.257)	-0.002 (0.325, 0.216)
MTM	0.8	-0.432 (0.265, 0.083)	-0.281 (0.202, 0.043)			
LIK		-0.334 (0.212, 0.212)	-0.130 (0.173, 0.163)	-0.822 (0.241, 0.219)	-0.238 (0.144, 0.106)	0.224 (0.172, 0.143)
HMTM		-0.005 (0.413, 0.388)	0.010 (0.261, 0.244)	-0.092 (0.849, 0.562)	-0.021 (0.320, 0.236)	-0.014 (0.224, 0.125)

Table 3. Bias (standard deviation s.d., average standard error s.e.) for the estimators of $\bar{\eta}, \bar{\xi}$ obtained from the fixed-effects MTM approach and for the estimators of $\bar{\eta}, \bar{\xi}, \sigma^2_\eta, \sigma^2_\xi, \rho$ obtained from the approximate likelihood approach (LIK) and the hierarchical MTM approach (HMTM), under increasing ρ and different link function. High accuracy of the test. Sample size $n = 10$. Prevalence of disease $\pi = 0.35$.

Method	ρ	$\bar{\eta}$ Bias (s.d., s.e.)	$\bar{\xi}$ Bias (s.d., s.e.)	σ^2_η Bias (s.d., s.e.)	σ^2_ξ Bias (s.d., s.e.)	ρ Bias (s.d., s.e.)
Logit link						
MTM	0.2	-0.415 (0.422, 0.188)	-0.158 (0.263, 0.112)			
LIK		-0.432 (0.328, 0.367)	-0.093 (0.240, 0.246)	-0.553 (0.460, 0.372)	-0.121 (0.252, 0.213)	0.057 (0.338, 0.252)
HMTM		-0.017 (0.422, 0.426)	0.013 (0.271, 0.271)	-0.144 (0.681, 0.557)	-0.012 (0.362, 0.285)	0.016 (0.556, 0.365)
MTM	0.6	-0.378 (0.425, 0.191)	-0.164 (0.258, 0.111)			
LIK		-0.369 (0.330, 0.354)	-0.054 (0.249, 0.242)	-0.559 (0.487, 0.355)	-0.110 (0.285, 0.218)	0.164 (0.270, 0.213)
HMTM		0.024 (0.437, 0.413)	0.010 (0.283, 0.271)	-0.144 (0.651, 0.550)	-0.025 (0.345, 0.288)	-0.002 (0.444, 0.291)
MTM	0.8	-0.388 (0.439, 0.190)	-0.151 (0.259, 0.112)			
LIK		-0.356 (0.331, 0.355)	-0.015 (0.235, 0.248)	-0.549 (0.490, 0.366)	-0.078 (0.290, 0.236)	0.241 (0.245, 0.182)
HMTM		0.039 (0.466, 0.416)	0.027 (0.291, 0.281)	-0.041 (0.727, 0.602)	0.020 (0.367, 0.325)	0.021 (0.358, 0.201)
Probit link						
MTM	0.2	-0.491 (0.317, 0.080)	-0.291 (0.240, 0.068)			
LIK		-0.193 (0.276, 0.255)	-0.041 (0.203, 0.190)	-0.563 (0.359, 0.253)	-0.152 (0.154, 0.122)	0.041 (0.328, 0.243)
HMTM		-0.006 (0.396, 0.399)	0.002 (0.260, 0.260)	-0.048 (0.728, 0.588)	-0.024 (0.322, 0.267)	-0.018 (0.448, 0.306)
MTM	0.6	-0.485 (0.315, 0.080)	-0.281 (0.230, 0.068)			
LIK		-0.185 (0.281, 0.253)	-0.036 (0.194, 0.187)	-0.569 (0.362, 0.254)	-0.156 (0.155, 0.123)	0.139 (0.245, 0.193)
HMTM		0.001 (0.399, 0.393)	0.009 (0.256, 0.260)	-0.078 (0.737, 0.547)	-0.030 (0.327, 0.258)	-0.016 (0.327, 0.219)
MTM	0.8	-0.489 (0.334, 0.080)	-0.277 (0.238, 0.069)			
LIK		-0.195 (0.290, 0.250)	-0.032 (0.201, 0.187)	-0.590 (0.347, 0.240)	-0.158 (0.159, 0.124)	0.180 (0.191, 0.150)
HMTM		-0.027 (0.391, 0.372)	0.024 (0.255, 0.259)	-0.162 (0.667, 0.493)	-0.032 (0.318, 0.242)	-0.029 (0.230, 0.126)
Cloglog link						
MTM	0.2	-0.439 (0.255, 0.062)	-0.287 (0.201, 0.048)			
LIK		-0.298 (0.231, 0.223)	-0.152 (0.169, 0.163)	-0.726 (0.280, 0.222)	-0.246 (0.144, 0.106)	0.047 (0.321, 0.235)
HMTM		0.014 (0.403, 0.400)	0.021 (0.269, 0.265)	-0.013 (0.819, 0.610)	0.026 (0.378, 0.284)	-0.015 (0.412, 0.304)
MTM	0.6	-0.438 (0.265, 0.062)	-0.280 (0.199, 0.048)			
LIK		-0.291 (0.234, 0.219)	-0.141 (0.167, 0.161)	-0.733 (0.278, 0.218)	-0.250 (0.140, 0.107)	0.147 (0.241, 0.181)
HMTM		-0.009 (0.384, 0.387)	0.017 (0.262, 0.254)	-0.082 (0.758, 0.565)	-0.009 (0.333, 0.258)	-0.006 (0.316, 0.217)
MTM	0.8	-0.429 (0.265, 0.063)	-0.290 (0.190, 0.048)			
LIK		-0.278 (0.228, 0.219)	-0.148 (0.165, 0.162)	-0.725 (0.283, 0.225)	-0.245 (0.139, 0.104)	0.198 (0.168, 0.137)
HMTM		0.006 (0.399, 0.379)	0.005 (0.263, 0.253)	-0.092 (0.787, 0.515)	-0.014 (0.330, 0.237)	-0.013 (0.205, 0.125)

Application

Delirium is an acute confusional state with varying disturbances of cognition, memory, attention, behaviour, and orientation. It is often observed in early stages of the hospitalization for acute and chronic diseases, especially in the elderly (Lipowski, 1987; Rai et al., 2014). Since delirium has been associated with unfavorable outcomes, early recognition and prompt treatment is crucial to decrease the risk of morbidity and/or mortality. Shi et al. (2013) perform a meta-analysis of diagnostic accuracy of the confusion assessment method, which is one of the most widely used delirium screening tool (Inouye et al., 1990) used by non-psychiatrically trained clinicians (nurses, general practitioners, ...) to identify and recognize delirium quickly. The confusion assessment method can be applied to verbal and nonverbal (e.g., mechanically ventilated) patients. It is considered as an alternative to the golden standard diagnostic criterion, namely, the Diagnostic and Statistical Manual of Mental Disorders IV, which cannot be easily applied to daily bedside practice. Table 4 reports the data for 20 studies included in the meta-analysis.

Table 4. Data for the confusion assessment method example (Shi et al., 2013). TP=true positives, FP=false positives, FN=false negatives, TN=true negatives.

Study	TP	FP	TN	FN
1	21	4	43	3
2	35	2	104	40
3	77	0	16	3
4	16	3	80	1
5	27	0	93	3
6	22	0	19	3
7	33	3	70	13
8	26	8	41	6
9	19	0	75	6
10	22	2	76	2
11	137	11	36	42
12	225	12	706	60
13	14	3	344	62
14	9	2	131	12
15	15	0	71	2
16	15	0	37	2
17	39	13	64	16
18	15	0	58	6
19	56	6	59	5
20	34	0	29	3

The HMTM and the standard approximate bivariate model are applied to the data under a logit specification of the relationship between sensitivity and specificity of the diagnostic tool. The choice is motivated by the reduced value of the Akaike Information Criterion associated to the logit specification if compared to the probit and the cloglog alternatives. Results are reported in Table 5.

No convergence problem has been experienced for both the approaches. The hierarchical HMTM provides larger estimates of the variance components and a smaller value of the correlation between sensitivity and specificity, at the price of a slightly larger standard error as a consequence of the increased complexity of the model if compared to the approximate likelihood approach. Globally, sensitivity and specificity of the competing methods are close, with specificity larger than sensitivity. In particular, the HMTM approach

Table 5. Results for the confusion assessment method example (Shi et al., 2013). Estimates and standard error (in parentheses) of the parameters of the HMTM and the approximate bivariate random-effects model and estimate (and standard error) of sensitivity and specificity of the diagnostic tool.

	HMTM	Approximate likelihood
$\bar{\eta}$	1.44 (0.25)	1.38 (0.23)
$\bar{\xi}$	3.67 (0.30)	3.25 (0.27)
σ^2_η	1.17 (0.49)	1.06 (0.47)
σ^2_ξ	1.45 (0.35)	1.06 (0.27)
ρ	-0.10 (0.20)	-0.21 (0.18)
SE	0.81 (0.04)	0.80 (0.04)
SP	0.98 (0.01)	0.96 (0.01)

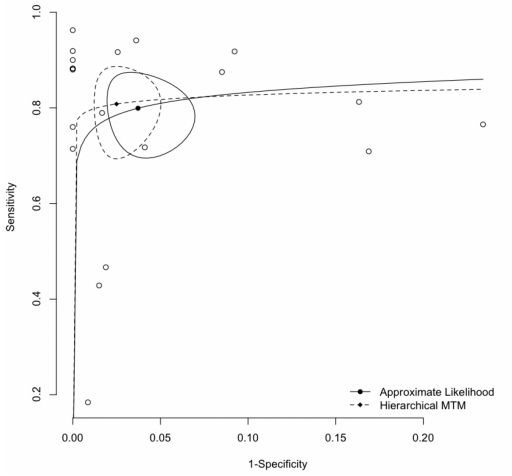

Fig. 4. Summary ROC curves, estimated sensitivity and 1-specificity, and associated 95% confidence regions from approximate likelihood and HMTM approach for the confusion assessment method example. (Shi et al., 2013).

provides an estimate of sensitivity and specificity equal to 0.81 (standard error 0.04) and 0.98 (standard error 0.01), respectively. The likelihood approach for the bivariate model provides an estimate of sensitivity and specificity equal to 0.80 (standard error 0.04) and to 0.96 (standard error 0.01), respectively. The mean of the estimated prevalences is 0.37, with an estimated standard error equal to 0.04.

Figure 4 reports the summary ROC curves from each method, the estimated sensitivity and 1-specificity, together with the associated 95% confidence region. The summary ROC curve is computed by fitting the regression line of the mean logit transformed sensitivity on the mean logit transformed 1-sensitivity, and then transforming it to the ROC space. See Arends et al. (2008) for alternative choices of the summary ROC curve. Differences among the summary ROC curves and among the 95% confidence regions reflect the slight differences in the estimated sensitivity and specificity from the approaches. Actually, the interpretation of the SROC curve as a way to summarize the results of a meta-analysis of diagnostic accuracy studies must be cautious. As Arends et al. (2008) underline, the SROC curve cannot be interpreted as an average of the study-specific ROC curves. It is a graphical representation of the relationship between $(\eta_i, \xi_i)^T$, with a shape that can even be very different from that of the study-specific ROC curves.

Conclusions

This paper explores the use of multinomial tree models, an instrument often adopted in psychological research to investigate cognitive processes, to carry out meta-analysis of diagnostic accuracy studies. An extension of the fixed-effects model developed in Botella et al. (2013) is proposed in order to define a hierarchical structure accounting for heterogeneity among studies included in the meta-analysis. In this way, the model meets the random-effects formulation commonly adopted in meta-analysis and allows to properly distinguish within-study and between-study heterogeneity. No restrictions are made on the link functions applied to sensitivity and specificity as accuracy measures of the test. Inference is then performed from a frequentist perspective, in contrasts to the standard latent-trait approach usually based on Bayesian solutions (Klauer, 2010). Simulation studies under a variety of scenarios show that the HMTM likelihood-based approach is preferable to the classical likelihood solution constructed on the approximate normal distribution for the sensitivity and specificity of the test in terms of accuracy of the inferential results, especially in case of small number of studies included in the meta-analysis. An expected improvement is confirmed with respect to the multinomial tree model approach based on a fixed-effects formulation, originally developed in Botella et al. (2013). The only disadvantage is represented by the need of numerical integration, which can be easily solved via quadrature methods.

According to the proposed HMTM approach, study-specific disease prevalences are explicitly taken into account and inference can be performed straightforwardly, by exploiting the parameters' separability of the complete likelihood function. Note that accounting for the potential dependence of the test performance measures on the disease prevalence is relevant when the severity of the disease status can vary across studies, in order to provide generalizable measures of test accuracy, see, e.g., Leeflang et al. (2009, 2013). This makes sense when dealing with cohort studies, while, when studies included in the meta-analysis are based on a case-control design, the prevalence is fixed and the estimation of the disease prevalence is not representative, see, e.g., Chu et al. (2009b) and Chen et al. (2015).

Interestingly, the paper shows that the associated likelihood function for the parameters of interest expressing the accuracy measures coincides with the likelihood function obtained under a classical bivariate random-effects approach to meta-analysis of diagnostic tests under an exact specification of the distribution for the true positives and the true negatives classified by the test under study (Arends et al., 2008). The framework from which the model is obtained, instead, is different, as using the multinomial tree approach allows to define, describe, and follow the path starting from the prevalence of the disease to the results of the application of the test under evaluation. Accordingly, deeper investigations are possible, by defining deviations from the structure of the tree with respect to that examined in this paper. For example, an interesting modification of the tree structure may account for imperfect reference standard, including different sensitivities and specificities for the test under study and the reference. The increased number of elements (branches) in the tree is expected to make the tree diagram more complex, as suggested in Botella et al. (2013) for the fixed-effects model, although the resulting associated likelihood function might be naturally defined by following the path on the branches to the final test outputs. How the deal with complications of the resulting model and likelihood in case of imperfect reference and how they compare to existing results in the literature (Chu et al., 2009a, Dendukuri et al., 2012, van Smeden et al., 2014; see Chapter 9.5 in Macaskill et al., 2023 and Chapter 10.8 in Takwoingi et al., 2023, for detailed descriptions) represents an interesting future direction of the present work.

The proposed approach considers the information from each study given in terms of true/false positives and true/false negatives, which turns out into one pair of sensitivity and specificity. A more complex situation arises when test performance is evaluated at multiple thresholds. In this case, the likelihood approach based on normal approximation has been examined in the literature (Riley et al., 2014; To and Guolo, 2021), with suggestions for reducing computational difficulties. A multinomial tree model approach in this framework would entail a complex tree structure with multiple branches. This would imply additional computational difficulties, given the likely increasing number of integrals to be computed.

Acknowledgments

The Author is grateful to Dr. Jessica Battagello for helpful discussion.

Conflict of Interest: None declared.

Funding

None declared.

Data availability

Software in the form of `R` code, together with the dataset, is available at https://github.com/annamariaguolo/MTM-meta-analysis.

References

1. Arends, L., Hamza, T., van Houwelingen, H., Heijenbrok-Kal, M., Hunink, M., and Stijnen, T. (2008). Bivariate random effects meta-analysis of ROC curves. *Med Decis Making*, **28**: 621–638.
2. Batchelder, W. and Riefer, D. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychon B Rev*, **6**: 57–86.
3. Botella, J., Huang, H., and Suero, M. (2013). Multinomial tree models for assessing the status of the reference in studies of the accuracy of tools for binary classification. *Front Psychol*, **4**: 694.
4. Chen, Y., Liu, Y., Ning, J., Cormier, J., and Chu, H. (2015). A hybrid model for combining case-control and cohort studies in systematic reviews of diagnostic tests. *Appl Statistic*, **64**: 469–489.
5. Chen, Y., Liu, Y., Chu, H., Ting, M., and Schmid, C. (2017a). A simple and robust method for multivariate meta-analysis of diagnostic test accuracy. *Stat Med*, **36**: 105–121.
6. Chen, Y., Liu, Y., Ning, J., Nie, L., Zhu, H., and Chu, H. (2017b). A composite likelihood method for bivariate meta-analysis in diagnostic systematic reviews. *Stat Methods Med Res*, **26**: 914–930.
7. Chu, H., Chen, S., and Louis, T.A. (2009a). Random effects models in a meta-analysis of the accuracy of two diagnostic

tests without a gold standard. *J Am Stat Assoc*, **104**: 512–523.

8. Chu, H. and Cole, S. (2006). Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *J Clin Epidemiol*, **59**: 1331–1333.
9. Chu, H., Nie, L., Cole, S., and Poole, C. (2009b). Meta-analysis of diagnostic accuracy studies accounting for disease prevalence: alternative parameterizations and model selection. *Stat Med*, **28**: 2384–2399.
10. Dendukuri, N., Schiller, I., Joseph, L., and Pai. M. (2012). Bayesian meta-analysis of the accuracy of a test for tuberculous pleuritis in the absence of a gold standard reference. *Biometrics*, **68**: 1285–1293.
11. Erdfelder, E., Auer, T., Hilbig, B., Aßfalg, A., Moshagen, M., and Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *J Psychol*, **217**: 108–124.
12. Guolo, A. (2017). A double SIMEX approach for bivariate random-effects meta-analysis of diagnostic accuracy studies. *BMC Med Res Methodol*, **17**: 6.
13. Guolo, A. (2023). Approximate likelihood and pseudo-likelihood inference in meta-analysis of diagnostic accuracy studies accounting for disease prevalence and study design. *Stat Med*, **42**: 4602–4617.
14. Hamza, T., van Houwelingen, H., and Stijnen, T. (2008). The binomial distribution of meta-analysis was preferred to model within-study variability. *J Clin Epidemiol*, **61**: 41–51.
15. Heck, D., Arnold, N., and Arnold, D. (2018). Treebugs: An r package for hierarchical multinomial- processing-tree modeling. *Behav Res Methods*, **50**: 264–284.
16. Honest, H. and Khan, K. (2002). Reporting of measures of accuracy in systematic reviews of diagnostic literature. *BMC Health Serv Res*, **2**: 4.
17. Hoyer, A. and Kuss, O. (2015). Meta-analysis of diagnostic tests accounting for disease prevalence: a new model using trivariate copulas. *Stat Med*, **3**, 1912–1924.
18. Inouye, S., van Dyck, C., Alessi, C., Balkin, S., Siegal, A., and Horwitz, R. (1990). Clarifying confusion: the confusion assessment method. a new method for detection of delirium. *Ann Intern Med*, **113**: 941–948.
19. Jackson, D., Riley, R., and White, I. (2011). Multivariate meta-analysis: Potential and promise. *Stat Med*, **30**: 2481–2498.
20. Jobst, L., Heck, D., and Moshagen, M. (2020). A comparison of correlation and regression approaches for multinomial processing tree models. *J Math Psychol*, **98**: 102400.
21. Kauermann, G. and Carroll, R. (2001). A note on the efficiency of sandwich covariance matrix estimation. *J Am Stat Assoc*, **96**: 1387–1396.
22. Klauer, K. (2006). Hierarchical multinomial processing tree models: a latent-class approach. *Psychometrika*, **71**: 7–31.
23. Klauer, K. (2010). Hierarchical multinomial processing tree models: a latent-trait approach. *Psychometrika*, **75**: 70–98.
24. Kuss, O., Hoyer, A., and Solms, A. (2014). Meta-analysis for diagnostic accuracy studies: a new statistical model using beta-binomial distributions and bivariate copulas. *Stat Med*, **33**, 17–30.
25. Leeflang, M., Bossuyt, P. and Irwig, L. (2009). Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. *J Clin Epidemiol*, **62**: 5–12.
26. Leeflang, M., Rutjes, A., Reitsma, J., Hooft, L. and Bossuyt, P. (2013). Variation of a test's sensitivity and specificity with disease prevalence. *Can Med Ass J*, **185**: E537–E544.
27. Lipowski, Z. (1987). Delirium (acute confusional states). *J Amer Medical Assoc*, 258:1789–1792.
28. Littenberg, B. and Moses, L. (1993). Estimating diagnostic-accuracy from multiple conflicting reports: a new meta-analytic method. *Med Decis Making*, **13**: 313–321.
29. Macaskill, P., Takwoingi, Y., Deeks, J.J., and Gatsonis, C. (2023). Understanding meta-analysis. In Deeks, J.J., Bossuyt, P.M., Leeflang, M.M., Takwoingi, Y., Editors. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*, Version 2.0 (updated July 2023). Cochrane. Chapter 9.
30. Moses, L., Shapiro, D., and Littenberg, B. (1993). Combining independent studies of a diagnostic test into a summary roc curve: data-analytic approaches and some additional consideration. *Stat Med*, **12**: 1293–1316.
31. Nelder, J. and Mead, R. (1965). A simplex algorithm for function minimization. *Scand J Stat*, **7**: 308–313.
32. Nikoloulopoulos A. (2018). Hybrid copula mixed models for combining case-control and cohort studies in meta-analysis of diagnostic tests. *Stat Methods Med Res*, **27**: 2540–2553.
33. R Core Team (2022). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria, https://www.R-project.org/.
34. Rai, D., Garg, R., Malhotra, H., Verma, R., Jain, A., Tiwari, S., and Signh, M. (2014). Acute confusional state/delirium: An etiological and prognostic evaluation. *Ann Indian Acad Neurol*, **17**: 30–34.
35. Reitsma, J., Glas, A., Rutjes, A., Scholten, R.J.P.M., Bossuyt, P.M., and Zwinderman, A. (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol*, **58**: 982–990.
36. Riley, R.D., Takwoingi, Y., Trikalinos, T. et al. (2014). Meta-analysis of test accuracy studies with multiple and missing thresholds: a multivariate-normal model. *J Biomet Biostat*, **5**: 196.
37. Riefer, D. and Batchelder, W. (1998). Multinomial modeling and the measurement of cognitive processes. *Psychol Rev*, **95**: 318–339.
38. Rutter, C. and Gatsonis, C. (2001). A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med*, **20**: 2865–2884.
39. Shi, Q., Warren, L., Saposnik, G., and MacDermid, J. (2013). Confusion assessment method: a systematic review and meta-analysis of diagnostic accuracy. *Neuropsych Dis Treat*, **9**: 1359–1370.
40. Stahl, C. and Klauer, K. (2007). Hmmtree: A computer program for latent-class hierarchical multinomial processing tree models. *Behav Res Methods*, **39**: 267–273.
41. Takwoingi, Y., Dendukuri, N., Schiller, I., Rücker, G., Jones, H.E., Partlett, C., and Macaskill, P. (2023). Undertaking meta-analysis. In Deeks, J.J., Bossuyt, P.M., Leeflang, M.M., Takwoingi, Y., Editors. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*, Version 2.0 (updated July 2023). Cochrane. Chapter 10.
42. Takwoingi, Y., Guo, B., Riley, R., and Deeks, J. (2017). Performance of methods for meta-analysis of diagnostic test accuracy with few studies or sparse data. *Stat Methods Med Res*, **26**: 1896–1911.
43. To, D.-K. and Guolo, A. (2021). A pseudo-likelihood approach for multivariate meta-analysis of test accuracy

studies with multiple thresholds. *Stat Methods Med Res*, **30**, 204–220.

44. van Smeden, M., Naaktgeboren, C.A., Reitsma, J.B., Moons, K.G., and de Groot, J.A. (2014). Latent class models in diagnostic studies when there is no reference standard – a systematic review. *Am J Epidemiol*, **179**: 423–431.
45. Zapf, A., Hoyer, A., Kramer, K., and Kuss. O. (2014). Nonparametric meta-analysis for diagnostic accuracy studies. *Stat Med*, **34**, 3831–3841.