Menten and Lesaffre *BMC Medical Research Methodology* (2015) 15:70
DOI 10.1186/s12874-015-0061-7

BMC
Medical Research Methodology

TECHNICAL ADVANCE Open Access

A general framework for comparative Bayesian meta-analysis of diagnostic studies

Joris Menten[1,2*] and Emmanuel Lesaffre[2]

Abstract

Background: Selecting the most effective diagnostic method is essential for patient management and public health interventions. This requires evidence of the relative performance of alternative tests or diagnostic algorithms. Consequently, there is a need for diagnostic test accuracy meta-analyses allowing the comparison of the accuracy of two or more competing tests. The meta-analyses are however complicated by the paucity of studies that directly compare the performance of diagnostic tests. A second complication is that the diagnostic accuracy of the tests is usually determined through the comparison of the index test results with those of a reference standard. These reference standards are presumed to be perfect, i.e. allowing the classification of diseased and non-diseased subjects without error. In practice, this assumption is however rarely valid and most reference standards show false positive or false negative results. When an imperfect reference standard is used, the estimated accuracy of the tests of interest may be biased, as well as the comparisons between these tests.

Methods: We propose a model that allows for the comparison of the accuracy of two diagnostic tests using direct (head-to-head) comparisons as well as indirect comparisons through a third test. In addition, the model allows and corrects for imperfect reference tests. The model is inspired by mixed-treatment comparison meta-analyses that have been developed for the meta-analysis of randomized controlled trials. As the model is estimated using Bayesian methods, it can incorporate prior knowledge on the diagnostic accuracy of the reference tests used.

Results: We show the bias that can result from using inappropriate methods in the meta-analysis of diagnostic tests and how our method provides more correct estimates of the difference in diagnostic accuracy between two tests. As an illustration, we apply this model to a dataset on visceral leishmaniasis diagnostic tests, comparing the accuracy of the RK39 dipstick with that of the direct agglutination test.

Conclusions: Our proposed meta-analytic model can improve the comparison of the diagnostic accuracy of competing tests in a systematic review. This is however only true if the studies and especially information on the reference tests used are sufficiently detailed. More specifically, the type and exact procedures used as reference tests are needed, including any cut-offs used and the number of subjects excluded from full reference test assessment. If this information is lacking, it may be better to limit the meta-analysis to direct comparisons.

Keywords: Meta-analyses, Diagnostic test accuracy, Bayesian statistics, Latent class model

*Correspondence: jmenten@itg.be
[1]Clinical Trials Unit, Institute of Tropical Medicine, Nationalestraat 155, B-2000 Antwerp, Belgium
[2]L-Biostat, KULeuven University of Leuven, Kapucijnenvoer 35, B-3000 Leuven, Belgium

Background

There is a growing interest in diagnostic test accuracy (DTA) reviews to select the best diagnostic test procedure [1] for a given setting. Most meta-analyses of diagnostic tests, however, estimate the diagnostic accuracy of a single test [2, 3]. Selection of the best test is usually done by undertaking separate meta-analyses for each test and then comparing the results [3]. Even when formally comparing diagnostic tests in a single systematic review, the analysis may ignore study effects. Such an approach can lead to biased comparisons due to confounding by study effects, as shown in a recent review [3]. Takwoingi and colleagues showed that results from comparative studies, where two tests were directly compared and which provide the most robust comparisons, differed from those of non-comparative studies. However, only 31 % of available studies were comparative. This indicates that there is a need for meta-analytical methods via direct and indirect comparisons. Data from direct comparisons may be inconclusive while a combined analysis of direct and indirect comparisons may be conclusive and can result in more accurate estimates [4, 5].

A particular aspect of comparisons between diagnostic tests is that the diagnostic performance of the index test is nearly always determined by comparison with a second test, the reference standard. Such a reference standard is presumed to 100 % correctly classify subjects as diseased or not. However, for many diseases it is impossible to determine the true disease status with certainty [6] and reference standards are imperfect. It is well known that the use of imperfect reference standards may bias estimates of the accuracy of the index test [7]. This consideration leads to a second requirement for comparative meta-analyses of diagnostic studies: the meta-analytic methods should adjust for the use of imperfect reference standards.

The aim of this manuscript is to develop a model that can be used for the comparative meta-analysis of two diagnostic tests that conforms to the two requirements sketched above. First, we assess possible biases in the estimation of the relative accuracy of two index tests due to the use of imperfect reference tests. We describe the different parameters that can be used to estimate the relative accuracy of two tests and assess the bias resulting from the use of imperfect reference standards. This allows us to select the most appropriate summary measure to use in the comparative meta-analysis of two diagnostic tests. Subsequently, we describe and develop models that can be used in the meta-analysis of diagnostic studies to compare the relative accuracy of two tests. We start with models that presume a perfect reference test is used in each primary study and extend these models allowing for imperfect reference tests. We estimate these models using Bayesian methods, specifically using Markov-Chain Monte-Carlo (MCMC) methods through Gibbs sampling [8]. For each model we provide the model specification and offer suggestions for appropriate informative or vague priors. In addition, we assess in a simulation study the value of these newly developed models but also the bias induced by the use of incorrect methods. Finally, we apply the methods to a real data example in the field of leishmaniasis.

Methods

Our aim is to estimate and test the difference in diagnostic accuracy of two or more index tests in a meta-analysis, combining data across all available studies. The studies included in a DTA review typically test each subject with one or more index tests and with one reference test. This reference test may differ between studies. To set the scene, data from a hypothetical meta-analysis are presented in Table 1. In this example, there are three index tests (T_1, T_2, T_3) and two possible reference tests (T_4, T_5). For example, in Study 1 index tests T_1 and T_2 are performed on all subjects as well as reference test T_4. There are 30 subjects with positive results on all three tests, one subject shows positive results on T_1 and T_2 and a negative result on T_4, etc. Studies 1 and 2 allow direct estimation of the relative accuracy of T_1 and T_2. Studies 3, 4 and 5 allow the estimation of the accuracy of T_1 (studies 3 and 4) or T_2 (study 5), but allow no direct comparison of T_1 and T_2. For these studies, the relative accuracy of T_1 and T_2 can only be estimated by estimating the diagnostic accuracy of each test separately and then comparing these estimates. This is complicated by the fact that the reference test is not the same for each study. Studies 6 and 7 do not allow direct comparison of the accuracy T_1 and T_2, but offer the possibility of an indirect comparison through the third index test T_3. The information from this third test may help to eliminate differences among the studies.

As a first step in a comparative DTA meta-analysis, we have to select an appropriate statistic to compare the two tests. The best statistic would be one which is readily interpretable by users of the meta-analysis and which is least prone to bias. We describe the possible choices below together with the results from a small simulation study. Subsequently, we need to develop a model which allows the incorporation of all available data while ensuring that results are valid and are not biased by differences in study characteristic, such as the selection of the reference standard used. Some possible models are described below. We assessed the value of these models in a simulation study and in a practical application.

Measures of relative value of diagnostic tests

Diagnostic accuracy is characterised by sensitivity S and specificity C. These two quantities are related and comparisons between tests need to take both S and C into account. Comparisons between two tests can be

Table 1 Tabulation of an hypothetical diagnostic test accuracy meta-analysis. Columns T_1, T_2, T_3 indicate results for the 3 possible index tests. Columns T_4, T_5 indicate results for the 2 possible reference tests. + indicates a positive test result, - a negative test result. NA indicates that the test was not performed in that particular study. The observed frequency column report the number of subjects with a specific test result pattern in each study

Study	Index tests			Reference		Observed
Nr.	T_1	T_2	T_3	T_4	T_5	Frequency
1	+	+	NA	+	NA	30
1	+	+	NA	-	NA	1
1	+	-	NA	+	NA	3
1	+	-	NA	-	NA	6
1	-	+	NA	+	NA	0
1	-	+	NA	-	NA	3
1	-	-	NA	+	NA	8
1	-	-	NA	-	NA	160
2	+	+	+	NA	+	49
…	…	…	…	…	…	…
2	-	-	-	NA	-	99
3	+	NA	NA	+	NA	115
…	…	…	…	…	…	…
3	-	NA	NA	-	NA	244
4	+	NA	NA	NA	+	11
…	…	…	…	…	…	…
4	-	NA	NA	NA	-	19
5	NA	+	NA	+	NA	66
…	…	…	…	…	…	…
5	NA	-	NA	-	NA	29
6	+	NA	+	NA	+	27
…	…	…	…	…	…	…
6	-	NA	-	NA	-	56
7	NA	+	+	NA	+	77
…	…	…	…	…	…	…
7	NA	-	-	NA	-	13
8	+	+	+	NA	NA	143
…	…	…	…	…	…	…
8	-	-	-	NA	NA	85

summarized using the difference or relative risk in S and C for the two tests. An alternative parameterization uses the diagnostic odds ratio $DOR = (S \times C)/[(1-S) \times (1-C)]$ which summarizes the accuracy of a test in a single number [9]. This parameter could be used as a summary in a meta-analysis, for example by calculating the relative DOR of two tests [3]. However, the use of imperfect reference standards can bias all above measures of relative accuracy of two index tests. We assessed the direction and magnitude of bias on these measures in a small simulation study. A description of the simulation study setup is given in Additional file 1.

Models for the comparative meta-analysis of diagnostic tests

In this section we develop models to compare J tests, by combining data across I studies in a comparative meta-analysis. All these models are hierarchical in nature. At the first level of the hierarchy, the models describe the observed data of the individual studies. The observed test outcomes depend on the disease prevalence and the accuracy of the tests in each study, and possible covariation among the test results. We describe the accuracy of the tests in terms of the study-specific sensitivity S_{ij} and specificity C_{ij} of test j in study i. At the second level, we specify

a model for these study-specific sensitivity-specificity pair $\{S_{ij}, C_{ij}\}$. Five possible models are described; they are listed in Table 2.

Meta-analytic models when a perfect reference standard is available

If a perfect reference test is available, the number of diseased N_{Di} and non-diseased N_{NDi} subjects in study i is known, as are the numbers of true positives N_{TPij} and true negatives N_{TNij} for each test j. In the standard bivariate model for the meta-analysis of a diagnostic test [9], the observed numbers of true positives and true negatives for each index test are assumed to be drawn from two independent binomial distributions $N_{TPij} \sim Bin(N_{Di}, S_{ij})$ and $N_{TNij} \sim Bin(N_{NDi}, C_{ij})$. The transformed values $g(S_{ij}) = \theta_{Sij}$ and $g(C_{ij}) = \theta_{Cij}$ are modeled at the next level, where $g(.)$ is a link function to allow the use of the normal distribution. Common choices for $g(.)$ are the logit, complementary log-log or probit functions. Several models are possible to incorporate comparisons of the diagnostic accuracy of different tests in this framework. We discuss three models below. These models can be further expanded to allow for covariates, other dependence structures or alternative parameterizations.

Model 1: Standard bivariate model for the meta-analysis of diagnostic tests A basic approach is to estimate the average diagnostic accuracy of each test separately and subsequently compare the estimates of the average S_j and C_j across the different studies. In this approach, the standard bivariate model for the meta-analysis of diagnostic tests [2] can be used for each test separately. All $g(S_{ij}) = \theta_{Sij}$ and $g(C_{ij}) = \theta_{Cij}$ pairs are assumed to follow independent bivariate normal distributions:

$$\begin{pmatrix} \theta_{Sij} \\ \theta_{Cij} \end{pmatrix} \sim N\left(\begin{bmatrix} \mu_{S_j} \\ \mu_{C_j} \end{bmatrix}, \Sigma_j \right),$$

$$\text{with } \Sigma_j = \begin{pmatrix} \sigma^2_{S_j} & \sigma_{S_jC_j} \\ \sigma_{S_jC_j} & \sigma^2_{C_j} \end{pmatrix}, \tag{1}$$

where $\rho_{S_jC_j} = \sigma_{S_jC_j}/(\sigma_{S_j} \times \sigma_{C_j})$ is the correlation between $\theta_{S_{ij}}$ and $\theta_{C_{ij}}$. Estimates of the relative accuracy of the tests are obtained from the estimated $g^{-1}(\mu_{S_j})$ and $g^{-1}(\mu_{C_j})$. For example, the average difference in S between T_1 and T_2 is estimated as $\hat{S}_{D21} = g^{-1}(\hat{\mu}_{S_2}) - g^{-1}(\hat{\mu}_{S_1})$. The advantage of the standard bivariate model is that it is relatively easy to fit using both Bayesian or frequentist techniques, with SAS [9] and WinBUGS [10, 11] example code available. However, as this model is not based on the comparisons between the index tests, but on the pooling of results for each test across all available studies, the results may be biased by study characteristics. This is equivalent to pooling findings from the active treatment arms of RCTs and comparing these estimates, an approach which is considered not to be appropriate for the meta-analysis of RCTs [12].

Model 2: Meta-Analysis Based on Direct Comparisons

To take study effects into account, the overall probability of testing positive in diseased subjects μ_{Si} or in non-diseased subjects μ_{Ci} for each study i could be modeled and S_{ij} and C_{ij} of the individual tests described as contrasts from this overall probability.

If we limit the data to studies which compare the two tests directly, we can write the study specific, transformed sensitivities $g(S_{ij}) = \theta_{Sij}$ and specificities $g(C_{ij}) = \theta_{Cij}$ as follows:

$$\begin{aligned} \theta_{Si1} &= \mu_{Si} + \delta_{Si}/2, \\ \theta_{Si2} &= \mu_{Si} - \delta_{Si}/2, \\ \theta_{Ci1} &= \mu_{Ci} - \delta_{Ci}/2, \\ \theta_{Ci2} &= \mu_{Ci} + \delta_{Ci}/2. \end{aligned} \tag{2}$$

In case g is the logit function, $\delta_{Si} = \log(S_{OR12})$ and $\delta_{Ci} = \log(C_{OR12})$, i.e. the log of the ORs of testing positive in diseased subjects for T_1 compared to T_2 and the log of the ORs of testing negative in non-diseased subjects in study i, respectively. To obtain average estimates of the difference in diagnostic accuracy between the two tests, δ_{Si} and δ_{Ci} are modeled using a bivariate normal distribution:

$$\begin{pmatrix} \delta_{Si} \\ \delta_{Ci} \end{pmatrix} \sim N\left(\begin{bmatrix} \nu_{\delta_S} \\ \nu_{\delta_C} \end{bmatrix}, \Sigma \right)$$

$$\text{with } \Sigma = \begin{pmatrix} \sigma^2_{\delta_S} & \sigma_{\delta_S\delta_C} \\ \sigma_{\delta_S\delta_C} & \sigma^2_{\delta_C} \end{pmatrix}. \tag{3}$$

Table 2 Description of the different models. Example code for the models is given in Additional file 2. S_{ij} and C_{ij} represent the sensitivity and specificity of test j in study i

Model	Reference standard	Model estimation
1	Assumed to be perfect	Independent estimation of S_{ij} and C_{ij}
2	Assumed to be perfect	Direct comparisons only
3	Assumed to be perfect	Direct and indirect comparisons
4	Allowing for imperfect reference standards	Hierarchical latent class model
5	Allowing for imperfect reference standards	Network-based latent class model

The ν_{δ_S} and ν_{δ_C} are the average log OR of the S and C between tests T_1 and T_2, respectively. The μ_{Si} and μ_{Ci} account for the dependence of test results obtained from the same study and can be estimated as fixed effects of in their turn modeled using bivariate normal distributions. This model is equivalent to the Smith−*Spiegelhalter*−Thomas model for two-treatment comparisons of RCTs [13, 14]. A similar model, but assuming a fixed, rather than random, relative accuracy between the different index tests is described in the Cochrane Handbook for Systematic Reviews of DTA studies [9].

Model 3: Meta-Analysis Based on Direct and Indirect Comparisons As shown in Lu et al. [14] in the case of meta-analysis of RCTs, the Smith−*Spiegelhalter*−Thomas model can be expanded to a mixed treatment-comparison meta-analysis of more than two treatments. Similarly, we can expand Model 2 to J diagnostic tests. By taking diagnostic test T_J as baseline, we can rewrite eqs. 2 and 3 as:

$$\begin{aligned}
\theta_{Si1} &= \mu_{Si} + (J-1)\times\delta_{Si1}/J - \delta_{Si2}/J - \ldots - \delta_{Si(J-1)}/J,\\
\theta_{Si2} &= \mu_{Si} - \delta_{Si1}/J + (J-1)\times\delta_{Si2}/J - \ldots - \delta_{Si(J-1)}/J,\\
&\vdots\\
\theta_{SiJ} &= \mu_{Si} - \delta_{Si1}/J - \delta_{Si2}/J - \ldots - \delta_{Si(J-1)}/J,\\
\\
\theta_{Ci1} &= \mu_{Ci} + (J-1)\times\delta_{Ci1}/J - \delta_{Ci2}/J - \ldots - \delta_{Ci(J-1)}/J,\\
\theta_{Ci2} &= \mu_{Ci} - \delta_{Ci1}/J + (J-1)\times\delta_{Ci2}/J - \ldots - \delta_{Ci(J-1)}/J,\\
&\vdots\\
\theta_{CiJ} &= \mu_{Ci} - \delta_{Ci1}/J - \delta_{Ci2}/J - \ldots - \delta_{Ci(J-1)}/J,
\end{aligned}$$

with

$$\left(\delta_{Si1}, \delta_{Si2}, \ldots, \delta_{Si(J-1)}, \delta_{Ci1}, \delta_{Ci2}, \ldots, \delta_{Ci(J-1)}\right) \sim N(\boldsymbol{\nu_\delta}, \Sigma). \tag{4}$$

and $\boldsymbol{\nu_\delta} = (\nu_{\delta_{S1}}, \ldots, \nu_{\delta_{S(J-1)}}, \nu_{\delta_{C1}}, \ldots, \nu_{\delta_{C(J-1)}})$ represents the average log ORs for S and C of the $J-1$ tests compared to the baseline test T_J. The differences in S and C between T_1 and T_2 on the logit scale are estimated by $\nu_{\delta_{S1}} - \nu_{\delta_{S2}}$ and $\nu_{\delta_{C1}} - \nu_{\delta_{C2}}$, respectively. This method allows indirect comparisons of T_1 and T_2 through comparison with a third test, similar to mixed treatment comparisons meta-analysis of RCTs. One complication of this model, is the specification and estimation of the variance-covariance matrix Σ. Specifying a structured variance-covariance matrix is in general complex and difficult to handle in MCMC estimation since each sampled variance-covariance matrix should be positive-definite [15]. In addition, model identification of the model with a general variance-covariance matrix will be difficult, especially when number of tests of interest is large. As an initial exploration of this model we can use a simplified variance-covariance structure, for example a diagonal or block diagonal matrix, and subsequently assess the effects of relaxing the simplifying assumptions. We describe some possible simplified variance-covariance structures in Additional file 2: Section 2.6.

Meta-analytic models when no perfect reference standard is available

Introduction The models described above presume that the disease status of all subjects in all studies is known, and consequently that the N_{Di}, N_{NDi}, N_{TPij} and N_{TNij} for each study i and test j is available. However, if only imperfect reference standards are available, the reported estimates of these quantities may be biased. The models described above can be expanded through latent class analysis (LCA) [16] to allow for the use of imperfect reference standards. In LCA, the true disease status of the participants of the basic studies is an unobserved, or latent, variable with two mutually exclusive categories, "diseased" and "non-diseased". This unobserved variable determines the probability to test positive or negative to a number of diagnostic tests which may include one or more imperfect reference tests. LCA models have been described for a variety of situations ranging from when a single imperfect test is observed in each study to more complex designs involving multiple tests. When multiple tests are involved, they may be treated as independent conditional on the disease status or the conditional dependence between them may be modeled using a variety of approaches [17–20]. An important underlying assumption of the latent class model is that the tests included in the model all correspond to the same underlying disease state [21]. Especially in a meta-analysis, where each study may use a different set of tests, this assumption is critical. If this assumption is not met, the underlying latent variable may differ among studies.

Description of the conditional independence latent class model In this section, we describe the basic latent class model at the level of the individual study i in the meta-analysis. To simplify notation, we temporarily suppress the i-subscript for the study level. For latent class analysis, the basic data is not the number of true positives and true negatives for each test j, but rather the number of subjects that show a certain pattern of outcomes across the J tests performed in a study. The number of subjects with pattern $\mathbf{y} = (y_1, y_2, \ldots, y_j)$ can be denoted as $N_\mathbf{y}$ and is assumed to follow a multinomial distribution $N_\mathbf{y} \sim \text{Mult}\left[N, P(\mathbf{y})\right]$, with y_j the observed binary outcome (0 = negative, 1 = positive) for test T_j, N the total sample size and $P(\mathbf{y})$ the probability that $\mathbf{y}$ occurs.

Denoting the unobserved disease status as D (not diseased $D = 0$, diseased $D = 1$) and under the conditional independence assumption $P(\mathbf{y}|D = k) =$

$\prod_{j=1}^{J} P\left(y_j|D=k\right)$, the class probabilities $P(\mathbf{y})$ can be described in terms of the S_j and C_j of the J tests. That is:

$$P(\mathbf{y}) = \sum_{k=0}^{1} P(D=k)\, P(\mathbf{y}|D=k) =$$

$$\pi \prod_{j=1}^{J} S_j^{y_j} \left(1-S_j\right)^{(1-y_j)} + (1-\pi) \prod_{j=1}^{J} C_j^{(1-y_j)} \left(1-C_j\right)^{y_j}, \quad (5)$$

with π the disease prevalence.

Thus LCA provides estimates for the study specific prevalence of disease π_i and the S_{ij} and C_{ij} of the J_i tests used in study i, which is a subset of the J different tests used across the I studies of the meta-analysis.

Model 4: Hierarchical Latent Class Model In essence, the most basic hierarchical latent class model (Model 4) is constructed through a combination of equations 1 and 5. While previously the reference test was presumed to be 100 % sensitive and specific, in Model 4 all S_{ij} and C_{ij}, including those of the reference tests, are modeled using separate bivariate normal distributions as in Equation 1. The observed data is assumed to come from the multinomial distribution described in Equation 5. Again, like Model 1, this model ignores the correlation among test results from the same study. The prevalences π_i can be assumed to be different for each study or to have a common normal distribution, $\pi_i \sim N\left(\mu_\pi, \sigma_\pi^2\right)$.

Model 5: Network-based Hierarchical Latent Class Model By rewriting the θ_{Sij} and θ_{Cij} in terms of μ_{Si}, μ_{Ci}, δ_{Sij}, and δ_{Cij} as in Eq 4, we can again take into account study level effects. The hierarchical modeling is equal to Model 3, the only difference is at the study level as described in Eq 5. This model thus adjusts the meta-analysis for the use of imperfect reference tests. By using the expanded Smith–*Spiegelhalter*– Thomas model of Lu et al. [14] at the second level of the hierarchy, study level effects are eliminated without the need to limit the analysis to direct comparisons only.

Model estimation and prior specification

Models are estimated in a Bayesian framework using Markov Chain Monte Carlo (MCMC) methods with OpenBUGS 3.0.3 called from within R 3.0.1 using the BRugs library. The Bayesian approach allows the estimation of complex, joint models and the combination of prior information, e.g. on the value of the reference test used, in the meta-analysis of new diagnostic tests. To complete the Bayesian model, priors need to be provided for all model parameters. OpenBUGS code for the models and full specifications of the priors are in Additional file 2. Convergence was checked using visual inspection of trace plots of the Markov chains and the Gelman-Rubin diagnostic statistic [22].

For parameters related to the index tests of interest, we consider it generally most appropriate to use uninformative priors. Specifically, we used normal priors with mean μ equal to zero and standard deviation σ equal to 1.69 for logit-transformed probabilities. This prior matches a uniform prior over the interval [0,1] in the first two moments on the probability scale [23]. When appropriate, these priors were bounded to avoid label switching [20]. Label switching is a problem arising in MCMC estimation of latent class models when two equivalent solutions are possible which give rise to identical observed data [24, 25]. The problem can be avoided by constraining S or C of one or more test to be $\geq$ 0.5. For the contrast in S and C, expressed as log ORs, normal priors with $\mu = 0$ and a large standard deviation, e.g. $\sigma = 10$ can be used. For the variance-covariance matrices, we construct non-informative priors using uniform priors for standard deviations and correlations.

The model was specified using a logit link function and results are estimated on the log-odds scale. The MCMC approach as implemented in OpenBUGS allows to obtain posterior distributions of all functions of the estimated parameters, as the average S and C of the index tests and differences between S and C of the different tests. We illustrate this in the OpenBUGS code in Additional file 2. We used the 2.5 and 97.5 th percentiles of the sampled posterior distribution of the statistics of interest as bounds for the 95 % credible intervals.

If we want to use information from previous phases of the research, we can use informative priors. It may for example be appropriate to use information obtained from a previous meta-analysis of case-control studies when performing a meta-analysis of phase IV studies, i.e. studies recruiting clinically suspect patients consecutively in a representative clinical setting [7]. However, given that the phase IV design ensures the most realistic assessment of the performance of a test when used as a diagnostic tool in the target population [6], we may want to reduce the influence from these prior phases by using a prior which is more diffuse than the actual results from the prior meta-analysis. In the latent class model Model 4, we can use informative priors for the diagnostic accuracy of the reference test. It is likely that some information on the accuracy of the reference tests is available. In fact, standard analysis assumes S and C of the reference test to be 100 %, which can be considered to be very strong deterministic prior from a Bayesian viewpoint [26]. Priors for the accuracy of reference test can be obtained from the literature or expert opinion [10].

Simulation study

To assess the performance of the different models and to

uncover possible bias of combining data without proper control for study specific effect or adjustment for the use of imperfect reference standards, we performed a simulation study using two different scenarios. For each scenario, we generated 250 sets of 20 diagnostic studies. We analyzed each simulated data set using the models described above using the logit for the link function $g(.)$. We evaluated the models using coverage probabilities (the proportion of replications in which the 95 % credible interval contained the true value) and power (the proportion of replications in which a difference in S and C between the two tests of interest was detected). In Scenario 1, we simulated a setting without systematic bias but where a common imperfect reference test is used to assess the diagnostic accuracy of the index tests in all primary studies. In Scenario 2, we simulated the situation of two index tests which are assessed in primary studies that tend to use different reference standards. This situation may rarely occur in practice, but was selected to assess how the model performed in an extreme situation with systematic bias due to imperfect reference tests. A full description of the simulation study setup is in Additional file 3.

Real data example

We applied the models to data obtained in a meta-analysis of rapid diagnostic tests for visceral leishmaniasis, which we described earlier [10, 27]. In the published meta-analysis, the focus was on estimating the diagnostic accuracy of individual tests. We extracted the data relevant for the comparison of one rapid diagnostic test, the RK39 dipstick, with that of the direct agglutination test (DAT) as a test case for the application of the methods developed in the current paper. We limited the data for this test case to primary studies that included the RK39 dipstick or DAT with at least one other index test and microscopical examination as a reference test for which full data was available in the published primary study. We selected all index tests which were used in more than one study. In total, we included 10 primary studies, four index tests (DAT, RK39-dipstick, IFAT, KAtex) and two reference tests (parasitology including spleen aspirate, parasitology not including spleen aspirate) (Table 3). All tests are specific to VL and consequently can be expected to related to the same underlying latent variable. The data are shown in Fig. 1 and Appendix 4. Note that the current study is used as "proof-of-concept" of the statistical modeling approach and not as a complete meta-analytic comparison of the two tests which would require a more extended search strategy.

The aim of this modeling exercise was to estimate the differences in S and C between the RK39-dipstick and DAT. A previous meta-analysis indicated that the diagnostic accuracy of the RK39, and possibly also of the DAT, may be lower in East-Africa compared to other geographic regions [28]. To correct for these differences, we included a fixed region effect (East-Africa vs. rest of the world) for S. We fitted the five models listed in Table 2. In the previous study [10], we obtained expert opinion on the diagnostic accuracy of the two reference tests. Expert opinion on the diagnostic accuracy of parasitology including spleen aspirate varied between 88 and 95 % for S and between 95 and 100 % for C. For parasitology without spleen aspirate, expert opinion varied for S between 70 and 80 % and between 95 and 100 % for C. We used this information to determine the priors in estimation of the models allowing for imperfect reference standards.

Table 3 Overview of the real data example: a comparative meta-analysis of the RK39 dipstick and direct agglutination test (DAT) for the diagnosis of visceral leishmaniasis. The total sample size (N) and availability of test results (X) is given for all 10 studies. Other tests: IFAT=indirect fluorescent antibody test, KAtex=latex agglutination test, spleen=parasitological examination of tissue aspirates including spleen sample, no spleen: parasitological examination of tissue aspirates not including spleen sample

Study information		Index tests				Reference test		
Publication	Country	RK39	DAT	KAtex	IFAT	Spleen	No Spleen	N
Boelaert-1999	Sudan		X		X		X	59
Boelaert-2004	Nepal	X	X		X	X		309
Boelaert-2008	Nepal	X	X	X		X		158
Boelaert-2008	India	X	X	X		X		352
Boelaert-2008	Kenya	X	X	X		X		307
Boelaert-2008	Ethiopia	X	X	X		X		35
Boelaert-2008	Sudan	X	X	X		X		291
de Assis-2012	Brazil	X	X		X		X	407
Toz-2004	Turkey	X			X		X	42
Veeken-2003	Sudan	X	X			X		77

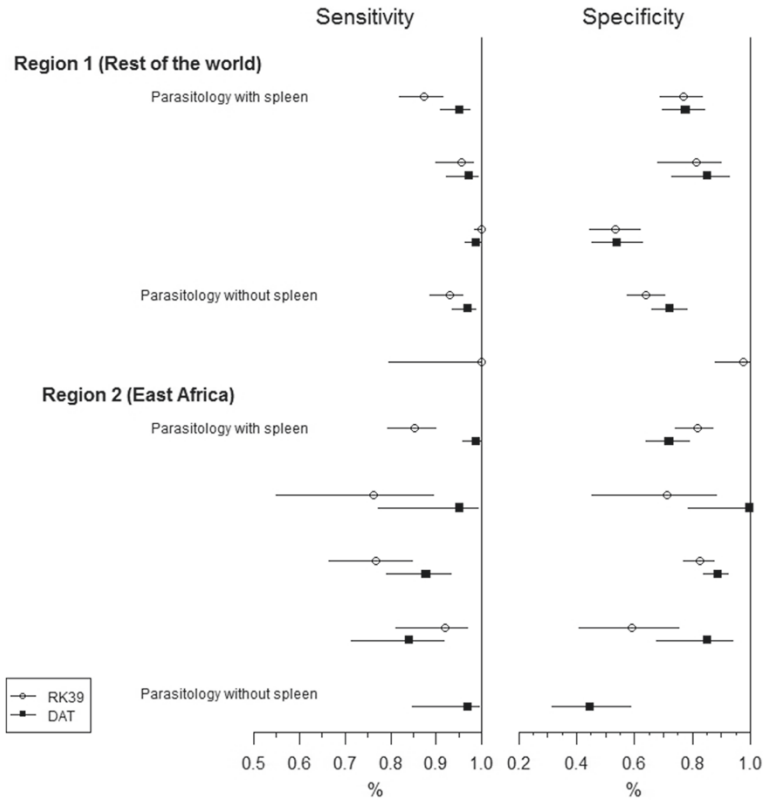

Fig. 1 Forest plot for real data example. Estimated sensitivity and specificity of the RK39 dipstick (open circles) and DAT (closed squares) with 95 % confidence interval, using parasitology as gold standard

Results

Measures of relative value of diagnostic tests

The results of our simulation study indicated that in a realistic setting, bias in estimating the difference in S and C between two index tests due to the use of an imperfect reference standard can be relatively limited (Additional file 1). Strong bias only occurred if the errors of one index test were strongly correlated with those of the reference test while the errors of the second index test were uncorrelated with those of the reference test. Similar observations can be made for the relative S and C. When the comparisons were expressed as odds-ratios or when using the relative Diagnostic Odds Ratio as a summary statistics, bias was more substantial and occurred even with uncorrelated errors. This corresponds to the findings from Zhang et al. who report that also in the meta-analysis of RCTs the odds-ratio is not always a suitable summary statistic [5].

Model performance: simulation study

Results of the simulation study of the model performance are described in detail in Additional file 3. The bias in estimating the contrasts in S and C between T_1 and T_2, expressed as a difference, relative risk or odds-ratio is summarized in Fig. 2.

In Scenario 1, where a common imperfect reference standard with moderate S and high C was used, a naive analysis assuming the reference test was perfect (Models 1–3), resulted in bias in estimating the odds-ratios (Fig. 2c) and to a lesser extent also the relative risks (Fig. 2b). If the contrast of interest was expressed as a difference (Fig. 2a), a true gold standard was available (Fig. 2a–c, Models 1–3), or if a latent class model was used to allow for imperfect reference tests (Fig. 2a–c, Models 4 and 5), no bias was apparent. Allowing for imperfect reference tests resulted in a lower power compared to the situation that a perfect reference test was available (Additional file 3, Table 3).

In Scenario 2, the reference standard of the simulated studies varied according to the index tests studies, which could result in systematic bias. In the analysis of Scenario 2, Models 4.i and 5.i correspond to the situation where the researchers knew of the differences in reference standard used across studies and that the variation in reference standard was thus a known source of bias. Models 4.ii and 5.ii correspond to the situation that researchers are unaware of the differences in reference standards among studies and that consequently the variation in reference standard was an unknown source of bias. This may occur if researchers do not provide sufficient detail in

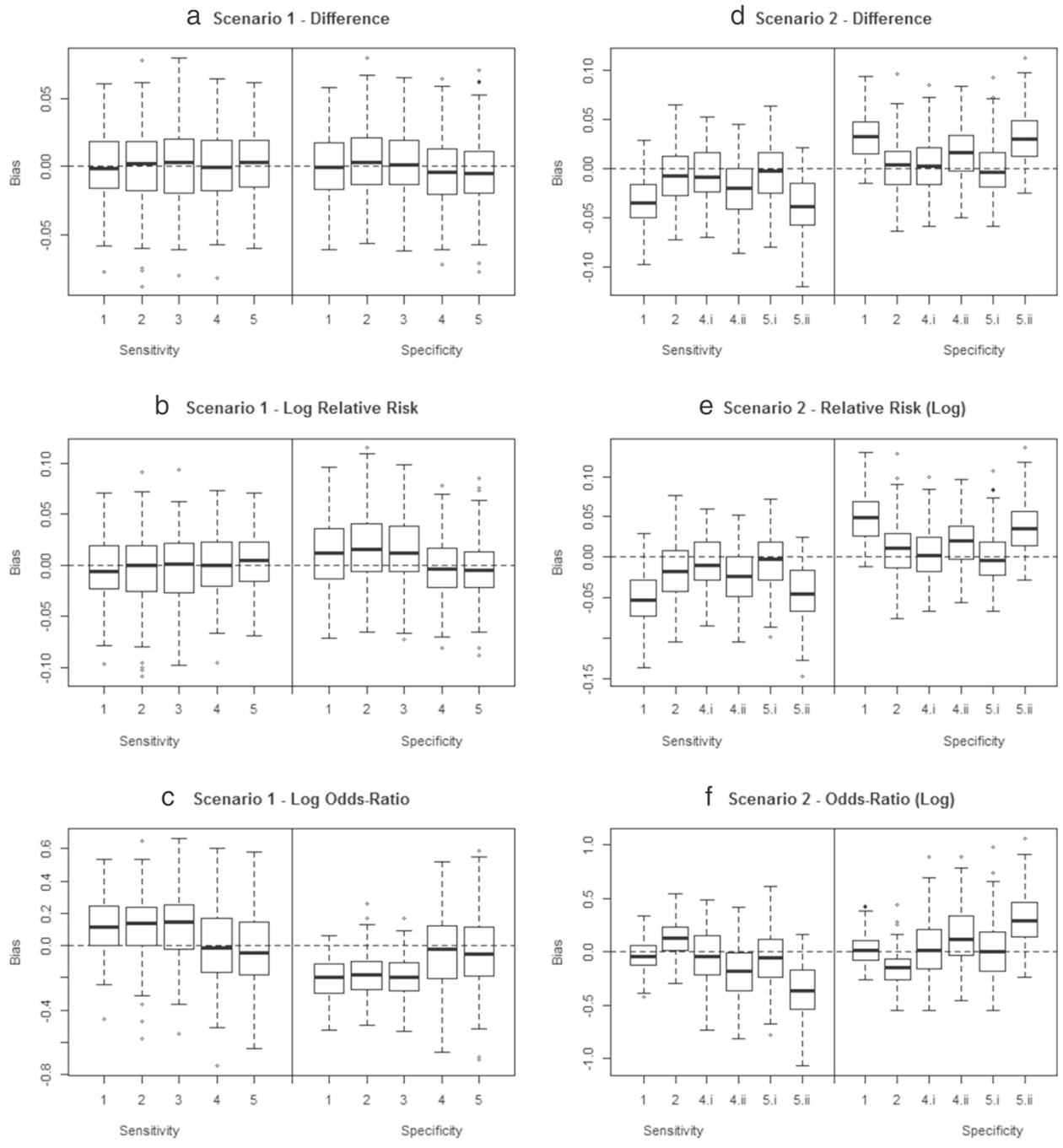

Fig. 2 Summary of simulation results. Bias in estimates of the contrasts in diagnostic accuracy from the proposed meta-analytical models applied in the simulation study. The boxplots present the bias in $\hat{S}_{D12}$ and $\hat{C}_{D12}$ (first row), $\hat{S}_{RR12}$ and $\hat{C}_{RR12}$ (second row), $\hat{S}_{OR12}$ and $\hat{C}_{OR12}$ (third row). The first column presents Scenario 1 where a common imperfect reference standard with moderate *S* and high *C* was used, the second scenario 2 where systematic bias is induced by differing reference standards. Full explanation of the model is in the text; full explanation of the simulation setup and results in Additional file 3. Note for Scenario 1: For models 1 to 3 disease status was estimated from the results of T_4. Note for Scenario 2: Models 4 and 5 were applied both assuming it is known that the reference tests differ across studies (4a and 5a) and ignoring the difference in reference tests (4b and 5b)

the primary publications on the exact modalities of the reference test procedures. For example, in the diagnosis of VL microscopical examination of spleen aspirates is the preferred reference test, while bone marrow aspirates show a limited *S*. Often researchers indicate their reference test to be based on spleen aspiration. However in

closer assessment of these publications, it can become apparent that some researchers perform spleen aspiration on nearly all subjects while others may preform spleen aspiration on only a minority of subjects. Ignoring these difference in reference tests may lead to bias. Incorrectly assuming the reference test were perfect resulted in substantial bias, especially when ignore study level effects (Fig. 2d–f, Model 1). When correcting for the use of imperfect reference tests using LCA (Models 4.i and 5.i), unbiased estimates for the differences in diagnostic accuracy between T_1 and T_2 were obtained (Fig. 2d–f, Models 4.i and 5.i). If the data were however analyzed ignoring the differences between the reference tests, the differences in diagnostic accuracy between T_1 and T_2 were overestimated (Fig. 2d–f, Models 4.ii and 5.ii).

Real data example: diagnostic tests for visceral leishmaniasis

Results of modeling of the VL data are in Table 4. All models indicated that S of DAT (S_2) was 8 to 11 % higher, compared to the S of RK39 (S_1), in East-Africa, but this difference did not reach statistical significance. In the rest of the world, estimates of S_1 and S_2 were similar. Differing modeling strategies or allowing for imperfect reference standards did not impact estimates of S or comparisons of S between the two index tests of interest. This is as expected as the parasitological reference tests show a similar and high C.

In contrast, allowing for imperfect reference tests (models 4 and 5) resulted in considerably higher estimates for C of both the RK39 dipstick and DAT compared to models assuming perfect reference tests were used (models 1–3). False negative results for the reference tests may have resulted in reduced estimates of C_1 (75.5–78.6 %) and of C_2 (80.1–81.5 %). Allowing for imperfect reference standards resulted in considerable higher estimates for C_1 (90.2–91.0 %) and C_2 (93.0–94.1 %).

In the analyses that used parasitology as a, presumed perfect, reference test, a substantial difference between C_1 and C_2 ($\hat{C}_2 - \hat{C}_1$ = 5.3 %) was observed when limiting the analysis to direct comparisons only (Model 2). On the other hand, Model 1, based on independent estimation of C_1 and C_2, showed a much smaller difference ($\hat{C}_2 - \hat{C}_1$ = 1.5 %). The model using direct and indirect comparisons (Model 3) showed intermediate results (3.2 %). This can be explained by the fact that the studies in which no direct comparison was possible between the RK39 dipstick and DAT showed contradictory results to the studies with direct comparisons. These studies also used the least sensitive reference standard which may explain that results of Models 4 and 5, both allowing for imperfect reference standards, were similar.

Discussion

In this paper, we developed a novel model to perform a comparative meta-analysis of the accuracy of two or

Table 4 Results of the meta-analysis of the diagnostic tests for visceral leishmaniasis

	Parasitology as Gold Standard			No Gold Standard	
	Model 1	Model 2	Model 3	Model 4	Model 5
Parameter	Estimate	Estimate	Estimate	Estimate	Estimate
S_1 (R1)	94.2	95.5	94.8	95.9	94.7
S_1 (R2)	85.2	84.0	86.5	84.6	88.1
S_2 (R1)	95.7	97.2	96.4	96.3	96.4
S_2 (R2)	94.4	93.5	95.4	96.1	96.5
C_1	78.6	75.5	78.3	90.2	91.0
C_2	80.1	80.9	81.5	93.0	94.1
S_{D12} (R1)	1.5	1.7	1.6	0.4	1.7
S_{D12} (R2)	9.2	9.6	8.9	11.5	8.4
C_{D12}	1.5	5.3	3.2	2.8	3.1
S_{RR12} (R1)	1.02	1.02	1.02	1.00	1.02
S_{RR12} (R2)	1.11	1.12	1.10	1.14	1.10
C_{RR12}	1.02	1.07	1.04	1.03	1.03
S_{OR12} (R1)	1.3	1.7	1.5	1.1	1.6
S_{OR12} (R2)	2.7	2.8	3.3	4.7	3.9
C_{OR12}	1.1	1.4	1.2	1.5	1.7

S_i and C_i: sensitivity and specificity of Test i; S_{D12} and D_{D12}: difference in sensitivity and specificity between Test 1 and Test 2; S_{RR12} and D_{RR12}: relative sensitivity and specificity of Test 1 compared to Test 2 as a relative risk; S_{OR12} and D_{OR12}: relative sensitivity and specificity of Test 1 compared to Test 2 expressed as an odds-ratio. R1 and R2 indicate estimates obtained for East-Africa and the rest of the world, respectively

more diagnostic tests when a perfect reference standard is unavailable. In a first step, we assessed the bias of comparative measures of the diagnostic accuracy of two tests induced by the use of an imperfect reference test. We observed that the difference in *S* and *C* may be the least subject to bias while at the same time being easily understandable to users of the meta-analysis results. In our modeling approach, we combined LCA with models developed for the mixed treatment comparisons meta-analysis of RCTs. The modeling framework accommodates a broad range of studies, including "Multiple Test Comparison", "Randomized Test Comparison", and "Between-Study Test Comparison" studies according to the terminology of Takwoingi et al., with the first two designs offering the most robust comparative data [3]. In a simulation study, the resulting model showed adequate performance, even if some aspects of the data generating mechanism were ignored. The simulation study also stressed the importance of accurate and complete extraction of the data from the primary studies when performing a DTA review. When differences in reference tests were ignored, biased estimates of the relative accuracy of the competing tests were unavoidable. This highlights the importance of complete and transparent reporting of DTA studies as promoted by the STARD initiative [29]. For a correct analysis of the data, the index and reference tests should be accurately described. Any cut-offs used to classify test results as positive or negative should also be reported and results for all subjects should be given, including subjects with incomplete or equivocal test results. The cross-classification of all test results should be presented in a format similar to that of the motivating example dataset in Additional file 4. The fact that meta-analysis is possible using imperfect reference tests suggests it may be more efficient to design future studies with multiple imperfect tests rather than using a single "as-accurate-as-possible" reference test, as has been shown in the analysis of epidemiological studies with imperfect measures of exposure [30, 31]. When applied to a dataset on visceral leishmaniasis diagnostic tests, the model indicated that *C* of the two tests of interest may have been underestimated due to the use of imperfect reference test. Our novel modeling approach, combining latent class analysis with hierarchical meta-analysis modeling, allowed the estimation of the difference in accuracy of the two index tests without making strong assumptions on the performance of the reference tests used. However, as in all meta-analyses, care should be taken that the studies combined are in fact comparable. While our approach corrects for bias and heterogeneity induced by the use of imperfect reference tests, other types of bias as publication and spectrum bias, can result in incorrect meta-analysis results. The approach can be combined with meta-regression techniques to reduce heterogeneity.

As limitations of our approach the following points can be given, which can indicate future avenues for further progress in this field. As a first limitation, we chose to compare diagnostic tests based on the sensitivity and specificity, and in particular based on the difference in these quantities among competing tests. Focusing on differences in *S* and *C* leads to results which are easily understandable for potential users. However, a test can be superior to another with respect to *S* while inferior with respect to *C*. In this case, selecting the optimal test can be difficult. Using a single summary measure of diagnostic accuracy, as the relative diagnostic odds-ratio (rDOR) can make comparisons among tests easier [32]. Theoretically, the test with the highest DOR may be preferred. However, this may not always be the case as the potential risk of a false positive result may be different from the risk of a false negative result. It may be easier for users to balance an increase in *S* versus decreases in *C*. In addition, the rDOR may be more prone to bias as we have shown for the OR difference in *S* and *C*. In our model formulation, the rDOR can be easily obtained. If the primary parameter of interest is however the rDOR, an alternative model formulation, for example an extension of the hierarchical summary ROC model [9, 33], may be more appropriate.

To allow estimation of the model, we made considerable simplifications to the variance-covariance structure of our parameter space. Not all these simplification may be warranted and a more general variance-covariance structure may refine estimates from this model. Fitting a general variance-covariance matrix however results in important computational difficulties. Our simulation study indicated that these limitations do not necessarily invalidate analysis results, but further research is needed to assess when this may no longer be the case. Modeling the variance-covariance matrix via partial autocorrelations [15] may allow the fitting of more complex model. We accommodated study effects using contrast-based (CB) approaches. However, in RCT arm-based approaches that correctly incorporate correlations have been shown to be superior to CB methods [5]. Further development of the equivalent models for DTA reviews, Models 1 and 4 in our setting, incorporating the correlations induced by study levels, is needed. Model 4 which corrects for imperfect reference tests, but ignores study effects, performed well in our simulation study. However, it is vulnerable to bias from study-specific effects. the model would need extensions to incorporate dependencies between test results from the same study, as for example was done for the meta-analysis of RCTs in Zhang et al. 2014 [5], before it is recommended as a general method for the meta-analysis of DTA studies above Model 5. However, if the accuracy of the different index tests is not strongly correlated across studies, Model 4 may perform equally well as Model 5 and may offer

advantages in terms of identifiability and computational feasibility.

We showed how prior information, e.g. on the diagnostic accuracy of the reference test, can be used to aid model estimation in the case of the hierarchical latent class model. This is in line with the methods we have earlier developed for the meta-analysis of the diagnostic accuracy of a single test when using an imperfect reference standard [10]. In the case of the network based latent class model, it is however much less clear how this information can be used. The diagnostic accuracy of each test is in this model a linear combination of an overall, study-specific, probability of testing positive on all tests and a number of contrasts in diagnostic accuracy among these tests. More research is needed on how priors can be constructed for this model, e.g. using the priors for conditional probabilities rather than for S and C directly [26].

DTA studies can be expected to exhibit considerable heterogeneity and may be more prone to bias and inconsistency between direct and indirect comparisons compared to RCTs. Applications of network meta-analytic model to DTA studies must be performed with care and further development of statistical methods are needed. The literature on network-based meta-analysis of RCTs contains many additional tools, for example to assess consistency of estimates obtained from direct versus indirect comparisons [34–36], assess heterogeneity among studies [37], detect outlying studies [38] and correct for bias [39]. We only performed a limited application of techniques developed in this context. Expanding these techniques to DTA meta-analyses may be a valuable direction of research. In particular, it is important to expand the concept of consistency of comparisons across networks to the context of DTA reviews [40, 41].

Alternative approaches to the comparative meta-analysis of diagnostic tests are proposed. The regression approach of Macaskill et al. [9] can be seen as a variation of our model 2 in which the relative S and C, expressed as an odds-ratio, between tests is constant across studies. For the case all tests are applied to all subjects, Trikalinos at al. [42] describe a model which fully accounts for the within-study correlation between the tests' subject-specific S and C. This approach can be more efficient than the methods proposed in the current manuscript. However, both approaches need further empirical and simulation studies to assess their relative merits. Different models may be most appropriate depending on the application. In case it is suspected that reference tests may show only limited S or C, the analysis method should allow for the use of imperfect reference test. If important study-level effects are expected, proper control for confounding by these effects is needed. If there is important uncertainty on the value of the reference test or the presence of study level effects, it will be preferable to fit several models and assess the robustness of the results to the assumptions. At this stage of research, it is not possible to provide a general recommendation on the optimal modeling approach for the meta-analysis of comparative DTA reviews.

Conclusions

The models developed in this paper are promising and can improve the comparison of the diagnostic accuracy of competing tests in DTA systematic review. This is however only true if the studies and especially information on the reference tests used are described in sufficient detail. If the reporting of the studies does not provide sufficient detail, it may be better to limit the meta-analysis to direct comparisons. Further work refining the modeling approach, especially with respect to the specification of more general covariance structures and the use of measures of consistency of direct versus indirect comparisons, can further improve these methods.

Additional files

Additional file 1: Simulation Study for the Selection of Appropriate Statistics for a Comparative DTA Review. (PDF 1208 KB)

Additional file 2: Software Code, Prior Specification, and Possible Structures for Variance-Covariance Matrices. (PDF 149 KB)

Additional file 3: Simulation Study of the Modeling Approach. (PDF 491 KB)

Additional file 4: Visceral Leishmaniasis Data. (PDF 44 KB)

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
JM conceived research questions, developed study design and methods, carried out statistical analysis, interpreted results and drafted the manuscript. EL advised on study design, methods, statistical analysis and commented on successive drafts. Both authors read and approved the final manuscript.

Acknowledgments
The works was supported by the Department of Economy, Science and Innovation of the Flemish Government. JM thanks Marleen Boelaert for her support and advice.

Received: 4 December 2014 Accepted: 28 July 2015
Published online: 28 August 2015

References

1. Leeflang MMG, Deeks JJ, Takwoingi Y, Macaskill P. Cochrane diagnostic test accuracy reviews. Syst Rev. 2013;2:82.
2. Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Ethics. 2005;58(10):982–90.
3. Takwoingi Y, Leeflang MMG, Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. Ann Intern Med. 2013;158:544–54.
4. Caldwell DM, Ades AE, Higgins JPT. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. Br Med J. 2005;331:897–900.

5. Zhang J, Carlin BP, Neaton JD, GG GGS, Nie L, Kane R, et al. Network meta-analysis of randomized clinical trials: Reporting the proper summaries. Clin Trials. 2014;11(2):246–62.
6. Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford (UK): Oxford University Press; 2003.
7. Zhou XH, Obuchowski NA, McClish DK. Statistical Methods in Diagnostic Medicine. New-York (US): Wiley-Interscience; 2002.
8. Lesaffre E, Lawson AB. Bayesian Biostatistics (Statistics in Practice). New-York (US): Wiley; 2012.
9. Macaskill P, Gatsonis C, Deeks J, Harbord R, Takwoingi Y. Cochrane handbook for systematic reviews of diagnostic test accuracy - Chapter 10: Analysing and presenting results In: Deeks JJ, Bossuyt PM, Gatsonis C, editors. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0. London (UK): The Cochrane Collaboration; 2010. p. 1–61.
10. Menten J, Boelaert M, Lesaffre E. Bayesian meta-analysis of diagnostic tests allowing for imperfect reference standards. Stat Med. 2013;32(30): 5398–413.
11. Verde PE. Meta-analysis of diagnostic test data: a bivariate Bayesian modeling approach. Stat Med. 2010;29:3088–102.
12. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. J Clin Epidemiol. 1997;50(6):683–91.
13. Smith TC, Spiegelhalter DJ, Thomas SL. Bayesian approaches to random-effects meta-analysis: a comparative study. Stat Med. 1995;14: 2685–699.
14. Lu G, Aedes AE. Combination of direct and indirect evidende in mixed treatment comparisons. Stat Med. 2004;23:3105–124.
15. Daniels MJ, Pourahmadi M. Modeling covariance matrices via partial autocorrelations. J Multivar Anal. 2009;100(10):2352–363.
16. McCutcheon AL. Latent Class Analysis. Quantitative Applications in the Social Sciences Series No. 64. Thousand Oaks, US: Sage Publications; 1987.
17. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple tests. Biometrics. 2001;57:158–67.
18. Qu Y, Tan M, Kutner MH. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. Biometrics. 1996;52:797–810.
19. Qu Y, Hadgu A. A model for evaluating sensitivity and specificity for correlated diagnostic tests in efficacy studies with an imperfect reference test. J Am Stat Assoc. 1998;93:920–8.
20. Menten J, Boelaert M, Lesaffre E. Bayesian latent class models with conditionally dependent diagnostic tests: a case study. Stat Med. 2008;27(22):4469–488.
21. Dendukuri N, Hadgu A, Wang L. Modeling conditional dependence between diagnostic tests: A multiple latent variable model. Stat Med. 2009;28:441–61.
22. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Stat Sci. 1992;7:457–72.
23. Agresti A, Hitchcock DB. Bayesian Inference for Categorical Data Analysis: A Survey. 2005.
24. A. Jasra CCH, Stephens DA. Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. Stat Sci. 2005;20(2):50–67.
25. Stephens M. Dealing with label switching in mixture models. J R Stat Soc Ser B Stat Methodol. 2000;62(1):795–809.
26. Berkvens D, Speybroeck N, Praet N, Adel A, Lesaffre E. Estimating disease prevalence in a Bayesian framework using probabilistic constraints. Epidemiology. 2006;17(2):145–53.
27. Boelaert M, Chappuis F, Menten J, van Griensven J, Sunyoto T, Rijal S. Rapid diagnostic tests for visceral leishmaniasis. Cochrane Database Syst Rev. 2011;6:.
28. Chappuis F, Rijal S, Soto A, Menten J, Boelaert M. A meta-analysis of the diagnostic performance of the direct agglutination test and rk39 dipstick for visceral leishmaniasis. Br Med J. 2006;333(7571):723–6.
29. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and,accurate reporting of studies of diagnostic accuracy: the stard initiative. Br Med J. 2003;326(7379):41–4.
30. Chu H, Cole SR, Wei Y, Ibrahim JG. Estimation and inference for case-control studies with multiple non-gold standard exposure assessments: with an occupational health application. Biostatistics. 2009;10:591–602.
31. Zhang J, Cole SR, Richardson DB, Chu H. A bayesian approach to strengthen inference for case-control studies with multiple error-prone exposure assessments. Stat Med. 2013;32(25):4426–437.
32. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol. 2003;56:1129–1135.
33. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med. 2001;20(19):2865–884.
34. Lu G, Aedes AE. Assessing evidence inconsistency in mixed treatment comparisons. J Am Stat Assoc. 2006;101(474):447–59.
35. Caldwell DM, Welton NJ, Ades AE. Mixed treatment comparison analysis provides internally coherent treatment effect estimates based on overviews of reviews and can reveal inconsistency. J Clin Epidemiol. 2010;63:875–82.
36. Salanti G, Higgins JPT, Ades AE, Ioannidis JPA. Evaluation of networks of randomized trials. Stat Methods Med Res. 2008;17:279–301.
37. Cooper NJ, Sutton AJ, Morris D, Ades AE, Welton NJ. Addressing between-study heterogeneity and inconsistency in mixed treatment comparisons: Application to stroke prevention treatments in individuals with non-rheumatic atrial fibrillation. Stat Med. 2009;28:1861–1881.
38. Zhang J, Fu H, Carlin BP. Detecting outlying trials in network meta-analysis. Stat Med. 2015;34(Epub ahead of print):1–3.
39. Dias S, Welton NJ. Estimation and adjustment of bias in randomized evidence by using mixed treatment comparison meta-analysis. J R Stat Soc Ser A. 2010;176(3):613–29.
40. Higgins JPT, Jackson D, Barrett JK, Lu G, Ades AE, White IR. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. Res Synth Meth. 2012;3:98–110.
41. White IR, Barrett JK, Jackson D, Higgins JPT. Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression. Res Synth Meth. 2012;3:111–25.
42. Trikalinos TA, Hoaglin DC, Small KM, Terrin N, Schmid CH. Methods for the joint meta-analysis of multiple tests. Res Synth Meth. 2014;5:294–312.