Statistics
in Medicine

Special Issue Paper

Received 30 November 2012, Accepted 5 August 2013 Published online 4 September 2013 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5959

Bayesian meta-analysis of diagnostic tests allowing for imperfect reference standards

J. Menten,[a,b*†] M. Boelaert[c] and E. Lesaffre[b,d]

There is an increasing interest in meta-analyses of rapid diagnostic tests (RDTs) for infectious diseases. To avoid spectrum bias, these meta-analyses should focus on phase IV studies performed in the target population. For many infectious diseases, these target populations attend primary health care centers in resource-constrained settings where it is difficult to perform gold standard diagnostic tests. As a consequence, phase IV diagnostic studies often use imperfect reference standards, which may result in biased meta-analyses of the diagnostic accuracy of novel RDTs. We extend the standard bivariate model for the meta-analysis of diagnostic studies to correct for differing and imperfect reference standards in the primary studies and to accommodate data from studies that try to overcome the absence of a true gold standard through the use of latent class analysis. Using Bayesian methods, improved estimates of sensitivity and specificity are possible, especially when prior information is available on the diagnostic accuracy of the reference test. In this analysis, the deviance information criterion can be used to detect conflicts between the prior information and observed data. When applying the model to a dataset of the diagnostic accuracy of an RDT for visceral leishmaniasis, the standard meta-analytic methods appeared to underestimate the specificity of the RDT. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: diagnostic tests; meta-analysis; visceral leishmaniasis

1. Introduction

Accurate diagnosis of infectious diseases is essential in primary health care in developing countries where infections are the most common causes of death and ill health [1]. Rapid and easy-to-use diagnostic tests, which are fit for purpose, have a key role in accurate diagnosis and correct patient management. Management of infectious diseases on the basis of clinical signs and symptoms is often not sufficiently accurate and may lead to giving inappropriate treatment and inducing antibiotic resistance [1]. Detection of infectious organisms in samples through microscopical examination or cultures may be costly, insufficiently sensitive, and difficult or dangerous to perform under field conditions, as for example in the case of the parasitic disease visceral leishmaniasis (VL) [2]. Consequently, there is a need for the development of rapid diagnostic tests (RDTs) for many infectious diseases, which should be evaluated for diagnostic accuracy in field conditions.

In diagnostic accuracy studies, the performance of a diagnostic test or algorithm in correctly identifying diseased and non-diseased subjects is assessed. Most diagnostic accuracy studies report estimates of sensitivity ($S = P(T+|D+)$, where $T+$ is a positive test result and $D+$ indicates subjects with the disease of interest) and specificity ($C = P(T-|D-)$, where $T-$ is a negative test result and $D-$ indicates subjects free of the disease of interest), possibly in combination with other measures of diagnostic accuracy such as the positive and negative predictive values or diagnostic odds ratio [3, 4].

[a]*Clinical Trials Unit, Institute of Tropical Medicine, Antwerp, Belgium*
[b]*L-Biostat, KULeuven, Leuven, Belgium*
[c]*Department of Public Health, Institute of Tropical Medicine, Antwerp, Belgium*
[d]*Department of Biostatistics, Erasmus Medical Centre, Rotterdam, The Netherlands*
**Correspondence to: J. Menten, Clinical Trials Unit, Institute of Tropical Medicine, Antwerp, Belgium.*
[†]*E-mail: jmenten@itg.be*

These measures are usually estimated by comparing the results of the test under evaluation (index test) with that of a reference standard [5], which is the best available approximation of the true disease status [6]. An alternative approach is the use of latent class analysis (LCA), which models the dependence between several diagnostic tests to estimate the diagnostic accuracy of all tests under consideration without explicitly using any of the tests as reference standard [3,4].

Diagnostic accuracy studies tend to be small and give imprecise estimates of S and C [7]. Consequently, there is a need for statistical methods to summarize these studies through meta-analytic techniques. Meta-analyses of diagnostic studies focus mainly on pooling the S and C pairs from each study as these are the most commonly reported accuracy measures. Generally, only methods are advised, which combine studies allowing for correlations among S and C, as on theoretical grounds; these measures should be negatively correlated because of the presence of threshold effects [5], even though empirical studies suggest that methods ignoring possible correlations may perform equally well [8].

Two approaches are described in the literature for the combination S-C pairs. The summary ROC approach models the diagnostic odds-ratio ($DOR = [S \times C]/[(1-S) \times (1-C)]$) [9]. This method has the disadvantage that it limits the analysis to a single measure and does not allow to discriminate between the ability to detect diseased subjects (as described by S) and identify those who do not have the disease in question (as described by C). Consequently, ROC-based approaches may not be appropriate for clinical studies, which assess the use of a diagnostic test in practice [3]. In addition, this approach assumes that the correlation between S and C is negative, which may not always be the case. In contrast to the summary ROC analysis, the bivariate model [5,10] models the study-specific sensitivity and specificity pairs $\{S_i, C_i\}$ for each study i jointly to produce estimates of S and C while allowing for possible correlation between these two measures.

Both approaches have in common that they are applicable only to studies, which use a reference standard and that they produce only valid results if this reference standard can classify all individuals correctly as diseased or non-diseased, that is, a true gold standard. It is well known that if an imperfect reference standard is used as a gold standard, the estimates of the S and C of the index test will be biased. Generally, this imperfect gold standard bias will result in an underestimation of the accuracy of the index test. Only if the errors of the index and reference standard are highly correlated, the opposite will occur, and the S and/or C of the index test will be overestimated [3].

For a correct estimation of S and C, the index and reference tests should be applied to all subjects, and all subjects should be taken into account in the calculation of the diagnostic accuracy of the index test. Investigators tend to believe that only rigorous verification of diseased cases and non-diseased controls, and discarding the data from those subjects for which no definite diagnosis can be made, results in unbiased estimates. However, the opposite is true, and studies requiring the most strict verification of disease status may report the most biased estimates of accuracy [11]. Bias may also be induced when it is impossible to perform the reference standard in the primary health care centers for which the diagnostic test under evaluation is intended. When the diagnostic test is evaluated in a referral center, the diagnostic accuracy of the test may be different because of a different spectrum of subjects being evaluated (spectrum bias) or because of difference in training of staff or test equipment.

Latent class analysis is an alternative approach to estimating S and C. In LCA, the disease status D is an unobserved, or latent, variable, and a probabilistic model is assumed for the relationship between results of several imperfect diagnostic tests results and the latent disease status [12]. Estimation of the model, either through maximum likelihood [13,14] or Bayesian methods [15–17], results in estimates of S and C of each of the diagnostic tests. Even though this approach is not necessarily without bias [18] or problems with respect to interpretation [19–21], it can be a valid approach for diagnostic studies where no gold standard exist or is impossible to perform in field conditions and is increasingly used in the analysis of diagnostic accuracy studies. As a consequence, meta-analysis methods should allow for the combination S and C estimates from studies using reference standard and studies using LCA. In addition, for studies that use reference standards, allowance should be made for the fact that this reference standard may not be perfect.

In this paper, we extend the bivariate model for the meta-analysis of diagnostic studies to accommodate data from both studies using reference standards and studies using LCA and to correct for differing and imperfect reference standards in the primary studies. We do this using a Bayesian analysis with informative priors based on expert opinion. We argue that this allows for a more correct meta-analysis of diagnostic studies when a true gold standard is lacking or difficult to apply in field conditions.

In recent years, a number of papers have addressed the same problem of diagnostic meta-analysis in the absence of a perfect reference standard [22–26]. Our approach differs: (i) by the use of prior infor-

mation and expert opinion on the diagnostic performance of the reference test; and (ii) by allowing for different methods to estimate the index test diagnostic accuracy across studies. Some primary studies may use LCA, in which case we use the resulting estimates of S and C and their confidence intervals. Other primary studies may use a reference test, in which case we use the 2×2 contingency table of index versus reference test results in our meta-analysis.

In Section 2, we describe the dataset that motivated us to extend existing meta-analyses methods for diagnostic studies. We describe the standard bivariate model in Section 3.1 and extend this model to correct for bias due to imperfect reference standards in Section 3.2 and incorporate studies that use LCA in Section 3.3. We describe three variations of the model and study the performance of these model formulations in a simulation study, described in Section 4. We apply the model to motivating example in Section 5 and discuss the use of the extended bivariate model in Section 6.

2. Motivating example

2.1. Introduction

Visceral leishmaniasis, also known as kala-azar, is a deadly protozoal disease transmitted by sandflies. The disease occurs mainly in Eastern Africa, the Indian subcontinent, and Latin America and causes an estimated 200,000 to 400,000 new cases and 20,000 to 40,000 deaths each year [27]. It occurs mainly in rural communities where it affects the poorest of the poor who have access only to the most basic primary health care [28, 29]. Early and accurate diagnosis and treatment are key components of VL control. Diagnostic tests for VL should be highly sensitive, as it is a fatal disease but should also be highly specific as available drugs tend to be toxic. Moreover, these tests should be fit for use in primary health centers in poor and remote rural areas.

Detection of parasites by microscopic examination of aspirates from lymph nodes, bone marrow, or spleen is the classical confirmatory test for VL. The specificity of this procedure is high, but the sensitivity of microscopy varies depending on the type of tissue aspirate. The sensitivity is higher for spleen (93–99%) than for bone marrow (53–86%) or lymph node (53–65%) aspirates [28, 30]. In addition, these techniques are costly and may be difficult or dangerous to perform under field conditions [2]. Bone marrow aspiration is painful and requires sterilization of materials, whereas spleen aspiration can be complicated by life-threatening hemorrhages in $\sim$0.1% of individuals and requires considerable technical expertise and health care facilities [28].

Antibody-based tests in combination with a standardized clinical case definition for VL diagnosis have been shown to be a viable alternative to classical microscopical diagnosis of VL. rK39-based RDTs are considered to be currently the best available diagnostic tool for VL for use in remote areas [28]. The performance of this test was studied in a previous meta-analysis [31] of 13 validation studies in which the rK39-based RDTs showed a sensitivity and specificity of 93.9% (95% CI, 87.7–97.1) and 95.3% (95% CI, 88.8–98.1), respectively [31]. This meta-analysis however combined studies from different clinical stages of the development of the diagnostic test of which only four phase IV studies, that is, studies recruiting clinically suspect patients consecutively in a representative clinical setting. 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 as it avoids spectrum bias of the patient population and is performed by the health care workers that ultimately will use the test in clinical practice [4]. There was consequently a need to update this meta-analysis focussing only on phase IV studies.

2.2. Study data

On the basis of predefined selection rules [32], we included 16 studies in the meta-analysis: six from the Indian subcontinent, eight from Eastern Africa, and one each from Latin America and the Mediterranean. As geographic region has been shown to be an important determinant, we incorporate the effects of geographic region in the analysis using meta-regression techniques [33] and limit the current analysis to the regions from where more than a single study was available, that is, the Indian subcontinent and Eastern Africa.

A forest plot of all S and C estimates of the 14 included studies is shown in Figure 1, and the data are summarized in Supplementary Material 3[‡]. There was wide variation in the reference standards used

[‡]*Supporting information may be found in the online version of this article.*

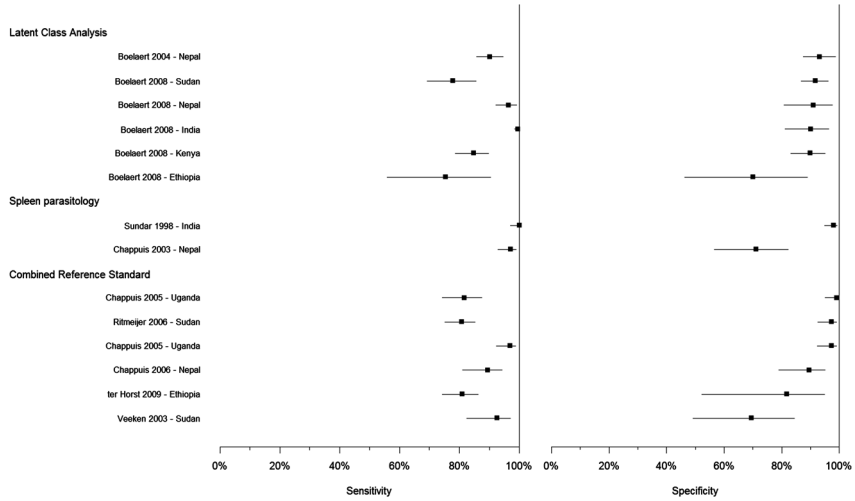


Figure 1. Forest plot of visceral leishmaniasis data: sensitivity and specificity estimates of the rK39 test for visceral leishmaniasis of 16 individual studies, classified by reference standard used.

in the calculation of the diagnostic accuracy of the rK39-based RDTs, even though three studies with insufficiently accurate reference test were a priori excluded. Six studies used LCA: five performed LCA using four diagnostic tests [20, 34]and one study used six diagnostic tests [35].

Two studies used microscopical examination of spleen aspirates as basis for the reference standard [36]. However, spleen aspirates were not performed for all subjects in these two studies. In one study [36], spleen aspiration was preceded by a bone marrow aspirate. If this test was negative, spleen aspiration was performed unless there was clear initial response to treatment of an alternative diagnosis, a refusal of the patient, or his physician; the patients spleen was too small to be punctured; or the patient had a coagulation disorder. In fact, spleen aspiration was only performed in six out of the 45 non-VL controls included in the study. In the second study, out of the 192 non-VL controls, 73 did not have a spleen aspirate performed for patients whose initial blood or sputum smears or chest radiographs indicated a diagnosis of malaria ($n = 52$) or active tuberculosis ($n = 21$) [37]. Consequently, the actual S of the reference test may be less in clinical practice than the 93–99% reported in the literature [28].

The remaining six studies used a combined reference standard of microscopical examination of tissue samples with another serological, diagnostic tests. In the studies included in the meta-analysis, all VL cases were either parasitologically positive or showed high titers to the direct agglutination test (DAT). Controls either had a low DAT titer when spleen or bone marrow aspiration was not required or a borderline DAT titer and a negative spleen or bone marrow aspiration. The studies [38–41] varied with respect to the cutoffs used. Chappuis *et al*. [42] combined information from bone marrow aspirates, DAT results, and response to treatment into a more elaborate reference test.

This example shows how individual studies may vary strongly with respect to the reference test used. Indeed, each investigator attempts to use the best available reference test within the constraints of the study setting or avoids the choice of a single reference test through LCA.

3. The bivariate model

The bivariate model for the meta-analysis of diagnostic studies is a hierarchical model where for each study i, there is a true underlying but unobserved sensitivity S_i and specificity C_i of the index test, that is, the diagnostic test of interest [10]. It allows modeling of the variation of the underlying S_i and C_i among studies through meta-regression [33] and uses a random effects approach to account for additional unexplained variation between studies. This variation may be due to differences in study population, implicit thresholds to qualify subjects as test positive or test negative or variations in test protocol [5].

More specifically, it is assumed that $g(S_i) = \theta_{Si}$ and $g(C_i) = \theta_{Ci}$ follow a bivariate normal distribution:

$$\begin{pmatrix}\theta_{Si}\\ \theta_{Ci}\end{pmatrix} \sim N\left(\begin{bmatrix}\mu_S + \boldsymbol{v_S}\mathbf{Z}_i\\ \mu_C + \boldsymbol{v_C}\mathbf{Z}_i\end{bmatrix}, \Sigma\right) \text{ with } \Sigma = \begin{pmatrix}\sigma_S^2 & \sigma_{SC}\\ \sigma_{SC} & \sigma_C^2\end{pmatrix} \tag{3.1}$$

with $g(.)$ a link function and $\boldsymbol{\nu_S}$ and $\boldsymbol{\nu_C}$ coefficient vectors describing the influence of covariates $\mathbf{Z}_i$ on the mean structure of S and C, respectively [10]. The model allows the study-specific S_i and C_i to be correlated with $\rho_{SC} = \sigma_{SC}/(\sigma_S \times \sigma_C)$ the correlation between θ_{Si} and θ_{Ci}. It is often assumed that this correlation is negative through a test positivity threshold, but the model allows also for a positive correlation. For the link function $g(.)$, usually the logit link function ($g(x) = log[x/(1-x)]$) is proposed [5, 10], but alternative links, such as the complementary loglog link function $g(x) = log[-log(1-x)]$, are possible [24].

The sources of the individual level data can come from studies using a perfect standard, from studies using an imperfect reference standard, and from studies using LCA to estimate the study specific S_i and C_i.

3.1. *Using data from studies using a perfect reference standard*

Results from each study i that estimates the diagnostic accuracy of an index test based on a reference test can be summarized by a contingency table of the cross-classification of the n_i subjects according to index and reference test results (Table I). If the reference test is perfect, the number of diseased and non-diseased subjects is known and is equal to $y_{i.1}$ and $y_{i.0}$, respectively. The numbers of true positives and true negatives are then y_{i11} and y_{i00}. In the standard bivariate model [10], the observed numbers of true positives and true negatives are then assumed to be drawn from two independent binomial distributions $y_{i11} \sim Bin(y_{i.1}, S_i)$ and $y_{i00} \sim Bin(y_{i.0}, C_i)$.

3.2. *Latent class analysis with informative prior distributions for reference test accuracy*

The model in Section 3.1 assumes that the reference standard can perfectly classify patients as diseased or not, that is, that the S and C of the reference tests are both 100%. For many diseases and studies, this perfect reference standard is however unavailable. In many cases, the analysis is then performed using an imperfect reference standard, but presuming it is perfect. This can lead to strongly biased estimates of S and C resulting in a meta-analysis with a highly precise but biased estimate of the diagnostic accuracy of the index test. Instead of assuming all y_{i11} to be true positives and all y_{i00} to be true negatives, we can model directly the counts in contingency Table I using a multinomial distribution.

If y_{ijl} equals the number of subjects in study i with result j (0=negative, 1=positive) to the index test and result l to the reference test, then

$$y_{ijl} \sim \text{Mult}(n_i, p_{ijl}),$$

with

$$\begin{aligned} p_{ijl} = \pi_i &\left[S_i^j \, (1-S_i)^{1-j} \, S_{Ri}^l \, (1-S_{Ri})^{1-l} + (-1)^{j-l} \, cov_{i|D=1} \right] \\ &+ (1-\pi_i) \left[C_i^{1-j} \, (1-C_i)^j \, C_{Ri}^{1-l} \, (1-C_{Ri})^l + (-1)^{j-l} \, cov_{i|D=0} \right], \end{aligned}$$

and

$$\begin{aligned} cov_{i|D=1} &= \rho_{i|D=1} \sqrt{S_i \, (1-S_i) \, S_{Ri} \, (1-S_{Ri})} \\ cov_{i|D=0} &= \rho_{i|D=0} \sqrt{C_i \, (1-C_i) \, C_{Ri} \, (1-C_{Ri})} \end{aligned} \tag{3.2}$$

where π_i is the prevalence in study i, S_{Ri} and C_{Ri} the S and C of the reference test, and $cov_{i|D=1}$ ($\rho_{i|D=1}$) and $cov_{i|D=0}$ ($\rho_{i|D=0}$) the covariances (correlations) between index and reference test results in diseased and non-diseased subjects, respectively.

Table I. Typical data display for a diagnostic accuracy study i, presenting a contingency table of index test and reference test results.

		Reference test result	
		Negative	Positive
Index test result	Negative	y_{i00}	y_{i01}
	Positive	y_{i10}	y_{i11}
	Total	$y_{i.0}$	$y_{i.1}$

For model identifiability, we need deterministic or probabilistic constraints on this model. One possible simplifying deterministic constraint is to assume that index and reference test results are independent conditionally on the disease status, that is, $cov_{i|D=1} \equiv cov_{i|D=0} \equiv 0$. In addition, in most cases, some information is available on the diagnostic performance of the reference test, which can be used as probabilistic constraints [15] in a Bayesian setting. As the different studies in a meta-analysis may employ different reference tests with each of their own S and C, we can categorize studies according to the reference standard R_i in K classes. For example, in our application, we identified two reference standards: 'spleen parasitology' and 'combined reference standard'. For each of these K different reference tests, we can subsequently obtain prior information on the diagnostic accuracy S_{Rk} and C_{Rk}, with $k = 1, 2, \ldots, K$. This information can be obtained from the literature or through elicitation of the opinion of experts in the field. To incorporate this information in the meta-analysis, we can then use one of the following models for each of the K imperfect reference standards identified.

(1) We can model the $g(S_{Ri}) = \theta_{SRi}$ and $g(C_{Ri}) = \theta_{CRi}$ using a bivariate normal, similar to the model used for the index test:

$$\begin{pmatrix}\theta_{SRi}\\ \theta_{CRi}\end{pmatrix} \sim N\left(\begin{bmatrix}\mu_{S_{Rk(i)}}\\ \mu_{C_{Rk(i)}}\end{bmatrix}, \Sigma_{k(i)}\right) \text{ with } \Sigma_{k(i)} = \begin{pmatrix}\sigma^2_{S_{Rk(i)}} & \sigma_{S_{Rk(i)}C_{Rk(i)}}\\ \sigma_{S_{Rk(i)}C_{Rk(i)}} & \sigma^2_{C_{Rk(i)}}\end{pmatrix} \tag{3.3}$$

where $k(i)$ indicates the type of reference test used in study i. We will use informative priors for the hyperparameters $\mu_{S_{l(i)}}$, $\mu_{C_{l(i)}}$ and $\Sigma_{l(i)}$ to ensure identifiability of the model and to incorporate knowledge about the performance of the reference test in our analysis. We will label this model in the remainder of the manuscript as the 'partial pooling' model following the terminology of Gelman and Hill [43].

(2) Alternatively, we can assume that the S and C of the reference tests are constant across studies using the same reference standard, that is, $\theta_{SRi} \equiv \mu_{S_{k(i)}}$ and $\theta_{CRi} \equiv \mu_{C_{k(i)}}$, and performing 'complete pooling' of the study specific $S_{Rk(i)}$ and $C_{Rk(i)}$ estimates. This can be seen as a special case of model 3.3, where $\sigma_{S_{Rk(i)}} \to 0$ and $\sigma_{C_{Rk(i)}} \to 0$ results in the 'complete-pooling' model [43].

(3) In contrast, a 'no-pooling' approach is equally possible by leaving θ_{SRi} and θ_{CRi} unmodeled. This corresponds to $\sigma_{S_{Rk(i)}} \to \infty$ and $\sigma_{C_{Rk(i)}} \to \infty$ in model 3.3 [43].

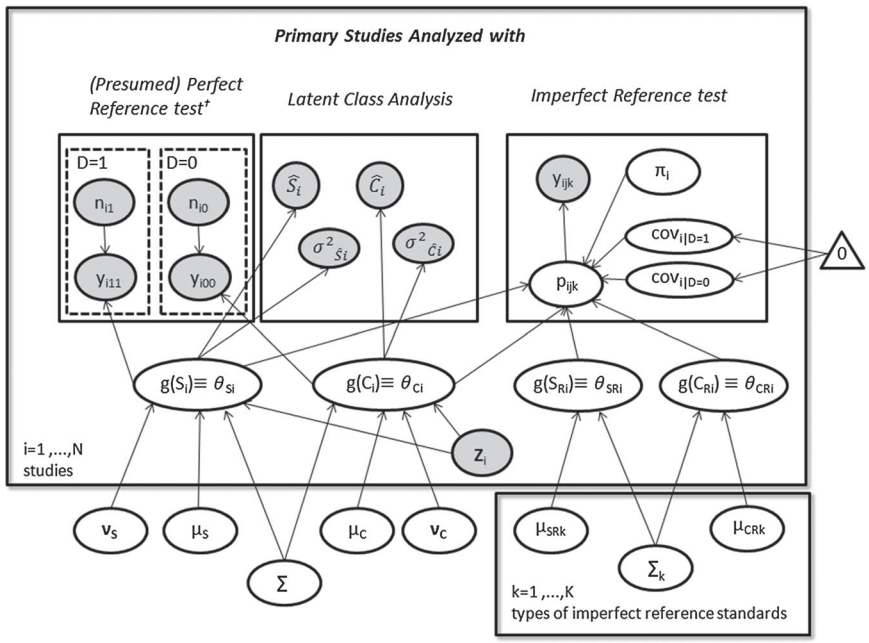
The prevalences π_i at the study level are left unmodeled by providing a Beta(1,1) distribution, equivalent to uniform distributions over the interval [0,1], as prior for π_i.

3.3. Using plug-in estimates from primary studies that are based on latent class analysis

The aforementioned approaches assume at the individual study level that each study uses a single reference standard to classify patients as diseased or not. However, in the last decennia, studies using LCA to calculate the accuracy of novel diagnostic tests are increasingly common. In these studies, several imperfect tests are performed on each subject, and a model-based analysis is performed that allows the joint estimation of the disease prevalence π and the S and C of each test. A wide variety of approaches have been proposed, using maximum likelihood [44] or Bayesian methods [45], allowing for conditional dependence through random or direct effects [16, 20, 46], and correcting for other biases as differential verification bias [47]. The likelihood of the observed data will depend on the number of tests and the presumed dependence structure of the data and may differ between the different studies included in a meta-analysis. These studies will typically report $\hat{S}_i$ and $\hat{C}_i$ together with 95% confidence or credible intervals. From these, we can obtain $g(\hat{S}_i) = \hat{\theta}_{Si}$, $g(\hat{C}_i) = \hat{\theta}_{Ci}$, $\hat{\sigma}_{\hat{\theta}_{Si}}$, and $\hat{\sigma}_{\hat{\theta}_{Ci}}$. Rather than re-analyzing the observed data from these studies, we use these estimates as plug-in estimators, and we assume that the reported $\hat{\theta}_{Si}$ and $\hat{\theta}_{Ci}$ are drawn from two independent normal distributions:

$$\begin{aligned}\hat{\theta}_{Si} &\sim N\left(\theta_{Si}, \sigma^2_{\hat{S}i}\right),\\ \hat{\theta}_{Ci} &\sim N\left(\theta_{Ci}, \sigma^2_{\hat{C}i}\right),\end{aligned} \tag{3.4}$$

with θ_{Si} and θ_{Ci} defined as before and using $\hat{\sigma}^2_{\hat{S}i}$ and $\hat{\sigma}^2_{\hat{C}i}$ as plug-in estimators of σ^2_{Si} and σ^2_{Ci}. Subsequently, the $\{\theta_{Si}, \theta_{Ci}\}$ pairs are assumed to be drawn from the bivariate normal defined in equation (3.1).



Notes: Shaded ellipses show observed data, open ellipses stochastical nodes, and triangles show constants. Parameters are explained in Section 3. † In general, it may not be appropriate to assume that any reference standard is perfect.

Figure 2. Directed acyclic graph of the extended bivariate model.

3.4. Combining data from different sources in the bivariate model

The data from studies using true perfect reference standards, imperfect reference standards, and studies that use LCA can be combined in the basic bivariate model using the methods described in Sections 3.1, 3.2, and 3.3, respectively. A directed acyclic graph of the full model is in Figure 2. In general, it may not be appropriate to assume that any reference standard is perfect and use the standard model described in Section 3.1, as the accuracy of any diagnostic test depends on the availability of well-trained and experienced laboratory technologists, and operator errors can always occur. For example, identification of infectious agents through microscopical examination of tissue samples are known to have a limited sensitivity but are routinely assumed to be 100% specific. However, microscopical misidentification of infectious organisms can occur because of the subjective nature of differentiating similar-appearing organisms on a microscopical slide [48]. In addition, if a test is known to have perfect S or, more commonly, C, this information can be used as deterministic or probabilistic priors for model 3.2 [15].

The hierarchical model will weigh studies according to the precision of the estimation of S and C [49]. Informative priors for the reference test, if they are not in conflict with the data, will—all else being equal—result in a higher precision in estimating the index test S and C and consequently impart a higher weight of the studies in the meta-analysis. Assuming the reference test is perfect, it can be seen as a highly informative prior and will consequently result in higher weights to the studies that use the assumption of perfect reference tests. The precision of the plug-in estimates from primary studies that use latent class models, as described in Section 3.3, will depend on a variety of factors: the sample size and prevalence of the disease, the number of reference tests used, the presumed dependence structure of the data, and the use of informative priors in a Bayesian setting.

4. Simulation study

4.1. Setup

To assess the performance of our approach, we performed a limited simulation study. We simulated a meta-analysis of diagnostic studies of an index test under evaluation with $S = 90\%$ and $C = 90\%$. The simulated meta-analysis combines data from 20 studies. Of these, five studies use LCA to estimate S and C of the index test, and 15 studies use one of three imperfect reference standards: (i) five studies use a reference standard with low S ($S_{R1} = 85\%$) and perfect C ($C_{R1} = 100\%$); (ii) five studies use

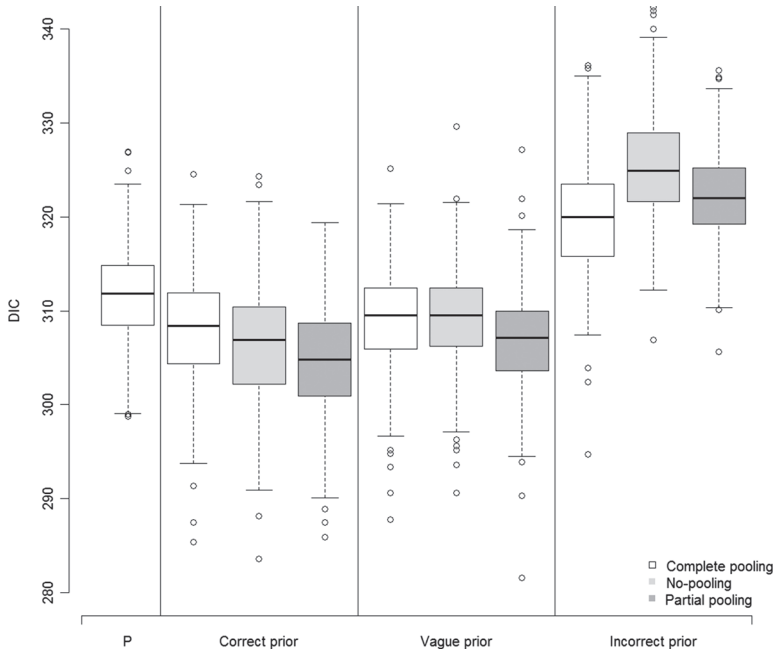
a reference standard with perfect S ($S_{R2} = 100\%$) and low C ($C_{R2} = 85\%$); and (iii) five studies use a reference standard with moderate S and C ($S_{R3} = C_{R3} = 93\%$, as the nearest integer value to the average of 85% and 100%). The results of the index and reference test were simulated from independent distributions, conditional on the disease status of each study subject. The simulated studies had moderate sample sizes (100–300 subjects) and a disease prevalence of 50%. The S_i and C_i of the 20 studies for one of the 500 simulated data sets are shown in Supplementary Material 2. As can be expected, studies with a reference test with a low S tend to underestimate the C of the index test, whereas studies with a reference test with a low C tend to underestimate the S of the index test.

We analyzed each simulated data set using the extended bivariate model described previously using the logit for the link function $g(.)$. Uninformative priors were used for hyperparameters related to the index test (μ_S, μ_C, Σ). Specifically, we used normal priors with mean μ equal to zero and standard deviation σ equal to 1.69 for the logits of index test S and C (μ_S and μ_C). This prior matches a uniform prior over the interval [0,1] in the first two moments on the probability scale [50]. For the variance–covariance matrix Σ, we can use a Wishart prior with 2 degrees of freedom: Wishart $\left[\begin{pmatrix} 0.001 & 0 \\ 0 & 0.001 \end{pmatrix}, 2\right]$ [51] or alternatively uniform priors for σ_S, σ_S, and ρ_{SC}. Prevalences π_i at the study level were left unmodeled by providing a Beta(1,1) distribution. For the reference test, we fitted 'no-pooling', 'complete-pooling', and 'partial-pooling' models as defined in Section 3.2. To assess the influence of prior information, we used four different priors for the accuracy of the reference tests in the analysis (Supplementary Material 4). 'Correct priors' provide information consistent with the true S and C of the reference test; 'vague priors' only indicate that the reference tests are informative of the correct diagnosis (S and C between 50 and 100%); 'incorrect priors' are inconsistent with the simulated S and C.

Results are provided for 10 analyses: one analysis assuming the reference standards are perfect and the three different models for reference test accuracy (no/complete/partial-pooling) with each of the three priors.

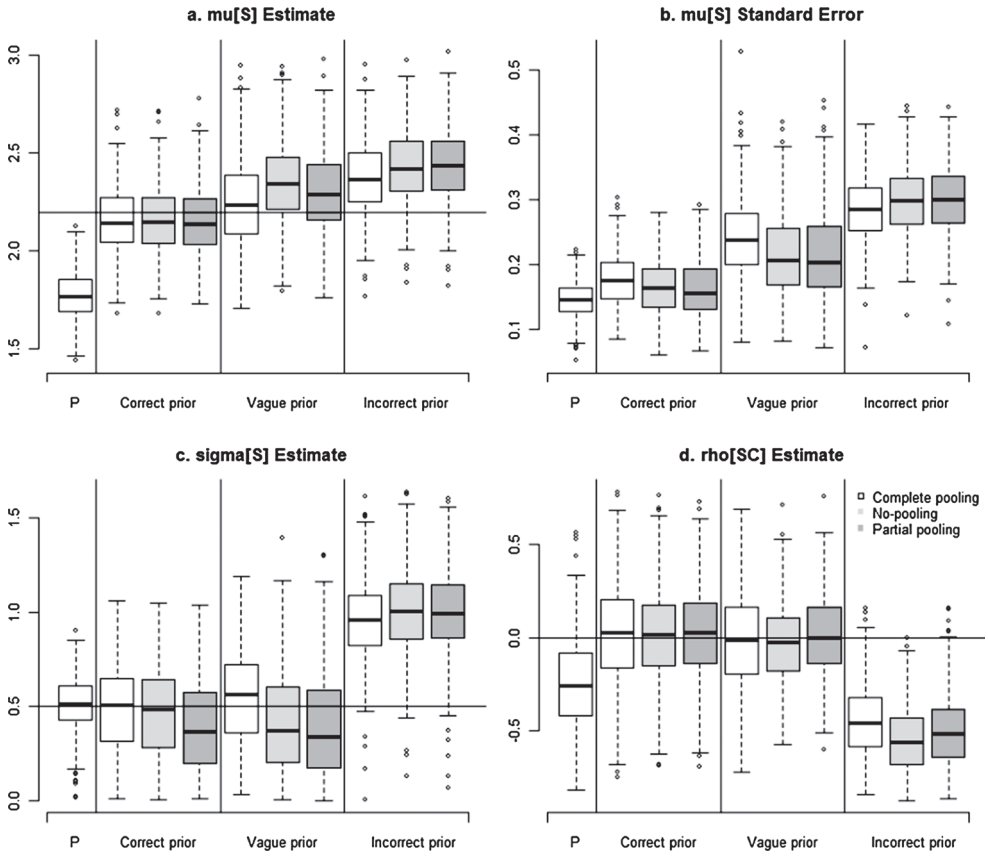
4.2. *Results*

Figures 3 and 4 show the distribution of, respectively, the deviance information criterion (DIC) [52] and $\hat{\mu}_S$, SE($\hat{\mu}_S$), $\hat{\sigma}_S$, and $\hat{\rho}_{SC}$ from the 10 models applied to the 500 simulated data sets. Assessing the fit of



Notes: P indicates the results obtained presuming the reference test is perfect. A full description of the simulation study setup is in Section 4.1. Data were analyzed with the extended bivariate model described in Section 3. DIC=Deviance Information Criterion

Figure 3. Distribution (box plots) of the deviance information criterion: the extended bivariate model applied to 500 simulated datasets.



Notes: P indicates the results obtained presuming the reference test is perfect. A full description of the simulation study setup is in Section 4.1. Data were analyzed with the extended bivariate model described in Section 3. Parameters are explained in Section 3.

Figure 4. Distribution (box plots) of $\hat{\mu}_S$, SE($\hat{\mu}_S$), $\hat{\sigma}_S$, and $\hat{\rho}_{SC}$ from the extended bivariate model applied to 500 simulated datasets.

the model to the data, using the DIC, revealed that both the analysis assuming perfect reference tests and the analysis using incorrect priors showed a worse fit than the analyses using priors consistent with the data (Figure 3). There was however considerable overlap in the distribution of the DIC of the different models over the 500 simulated datasets. The analyses assuming perfect reference tests or using incorrect priors showed also important bias in estimating the index test S (Figure 4a) and C. The analysis assuming perfect reference tests underestimated μ_S (Figure 4a) and μ_C and showed a small standard error for $\hat{\mu}_S$ (Figure 4b) and $\hat{\mu}_C$, resulting in a low coverage (Table II). The analysis with incorrect informative priors resulted in this simulation in an overestimation of μ_S (Figure 4a) and μ_C with low precision (Figure 4b). Both the analyses using correct and vague priors showed acceptable coverages. Vague priors resulted in a small overestimation of μ_S (Figure 4a) and μ_C and reduced precision compared with the analysis using correct informative priors (Figure 4.b) but provided a noticeable improvement compared with the standard analysis assuming perfect reference standards.

With respect to the random effects estimates, the analysis using incorrect priors tended to overestimate the random effects variation (Figure 4c). Both analyses assuming perfect reference tests or using incorrect priors, induced a negative correlation between S_i (Figure 4d) and C_i, while the data were simulated assuming independence. This bias in estimation of ρ_{SC} would be interpreted as evidence of a threshold effect in a standard analysis. However, further study is needed to determine if this negative correlation is particular to our simulation or can be generally seen when incorrectly assuming that the reference tests used were perfect.

Little difference was observed in the results of the three different model formulations for the reference tests (no/complete/partial-pooling). The partial-pooling approach tended to result in the best fitting

Table II. Coverages of the 95% credible intervals for the parameter estimates from the extended bivariate model applied to 500 simulated datasets.

Prior/model for diagnostic accuracy of the reference test	Coverages for parameter estimates of diagnostic accuracy of the index test				
	μ_S	μ_C	σ_S	σ_C	σ_{SC}
Perfect reference test	21.1	23.7	94.7	93.0	90.9
Correct: complete-pooling	94.0	94.7	87.8	87.1	98.1
Correct: no-pooling	92.1	93.0	85.4	84.7	99.0
Correct: partial-pooling	91.1	92.6	81.3	88.0	99.3
Vague: complete-pooling	96.2	95.9	89.2	87.3	99.3
Vague: no-pooling	91.1	87.1	88.5	90.2	99.8
Vague: partial-pooling	92.6	89.4	86.8	87.1	99.8
Incorrect: complete-pooling	98.6	98.3	45.8	35.3	72.9
Incorrect: no-pooling	98.8	93.3	36.5	21.8	50.6
Incorrect: partial-pooling	98.3	96.2	39.6	30.9	65.0

Notes: Each simulation consists of 20 simulated primary study datasets: 15 studies using imperfect reference standards and five studies using latent class analysis. Simulated values were $\mu_S = \mu_C = 2.2$, $\sigma_S = \sigma_C = 0.5$, and $\sigma_{SC} = 0$. Data were analyzed with the extended bivariate model described in Section 3 for index test diagnostic accuracy. Uninformative priors were used for hyperparameters related to the index test. For the reference test, we fitted a model assuming that the reference test is perfect as well as 'no-pooling', 'complete-pooling', and 'partial-pooling' models as defined in Section 3.2 and used three different priors in the analysis. A full description of the simulation study setup is in Section 4.1.

models as assessed by the DIC. Standard errors for $\hat{\mu}_S$ and $\hat{\mu}_C$ were largest for the complete-pooling approach (Figure 4b).

4.3. Conclusions

This simulation study indicates that the proposed methodology is a valid analysis approach to correct for imperfect reference tests in a meta-analysis. The use of both correct and vague priors for the reference test S and C resulted in estimates of the index test S and C, which were less biased compared with those assuming the reference test was perfect. As illustrated, the DIC can be used to identify conflicts between priors and observed data. However, the overlap in DIC distributions between the different models indicates that model choice should not be based on DIC alone. The use of incorrect informative priors resulted in large increases in the standard errors of the parameter estimates $\hat{\mu}_S$ and $\hat{\mu}_C$, which may be preferable to the precise, but biased estimates obtained using incorrectly assuming that the reference test is perfect.

The three modeling approaches for the reference test S and C (no/complete/partial-pooling models) resulted in broadly similar results. The partial pooling model may be most appropriate and showed the best fit to the data but may result in unidentified models in some datasets. In those cases, the complete-pooling model may provide an alternative analysis approach at the expense of a slight loss in precision in estimating $\hat{\mu}_S$ and $\hat{\mu}_C$.

5. Application

5.1. Study Description

We applied our meta-analysis approach to the data from the motivating example described in Section 2 and Supplementary Material 3. We obtained informative priors on the diagnostic accuracy of the two reference tests from seven international *Leishmania* experts. The experts were selected among VL researchers from the different endemic regions of $Leishmania$. The experts were asked to provide the most likely value for S and C of each reference test, together with a 95% prediction interval for the study-specific S_{Ri} and C_{Ri} (Supplementary Material 5). The estimates and prediction intervals were transformed using the selected link function $g(.)$, and a linear pool was constructed by obtaining the average diagnostic and pooled variance over the experts. These estimates were used to construct normal informative priors for the hyperparameters $\mu_{S_{l(i)}}$, $\mu_{C_{l(i)}}$ and $\Sigma_{l(i)}$ in the extended bivariate model

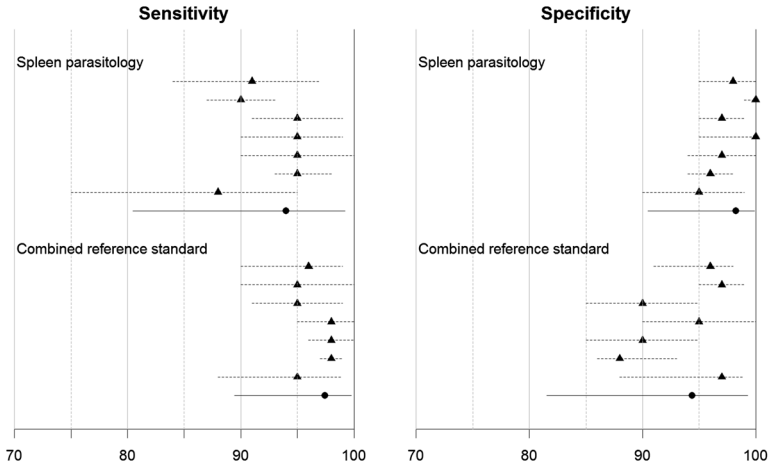


Figure 5. Expert opinion of seven experts (triangles and dotted line) and linear pooled expert opinion (filled circle and full line) on the diagnostic accuracy of the two reference standards used in the visceral leishmaniasis study.

(Figure 5). The resulting priors are given in Supplementary Material 4. In addition, we applied a model using vague priors, assuming with 95% certainty that S_R and S_C were in the interval 50–100%. Again, we used no, complete, and full pooling of the reference test diagnostic accuracy across studies using the same reference test. We used the complementary log-log function as link function, as this provided a better fit to the data and added an effect for geographic region in the model, as an earlier meta-analysis indicated this to be an important predictor of S of serological tests for VL.

5.2. Results

We estimated all models using Markov chain Monte-Carlo methods through Gibbs sampling using OpenBUGS version 3.0.3 called from within R 2.14.1 using the BRugs library. The OpenBUGS code for the models is in Supplementary Material 1. We checked the convergence using visual inspection of trace plots of the Markov chains and the Gelman–Rubin diagnostic statistic [53]. Final results were from 3000 samples obtained out of three chains of each 10,000 Markov chain Monte-Carlo iterations, retaining every 10th draw to reduce autocorrelation, after a burn-in of 40,000 iterations. Gelman–Rubin statistics are $\leqslant 1.01$ for all parameters.

Parameter estimates from applying the extended bivariate model to the data are in Table III. The best fitting model was the partial-pooling model using vague priors, although differences in fit between the different models were modest. The models using expert opinion showed a higher DIC than the models using vague priors, indicating some discordance between the expert opinion of the reference test diagnostic accuracy and the observed data. There was little indication of lack of diagnostic accuracy of the reference tests used in the studies selected for the meta-analysis, and estimates of the extended bivariate model were in line with those from the standard bivariate model assuming perfect reference standards. There may be some underestimation of the C of the index test in the analysis assuming perfect reference standards: 90.1% and 91.1% in the Indian subcontinent and Eastern Africa, respectively, compared with 92.4% and 93.1% in the best fitting model. Credible regions for the means S and C and prediction regions for study specific S_i and C_i are in Figure 6 for the model assuming perfect reference standards and partial-pooling model using vague priors. Credible and prediction regions were similar apart from the slightly higher estimates of C allowing for imperfect reference standards.

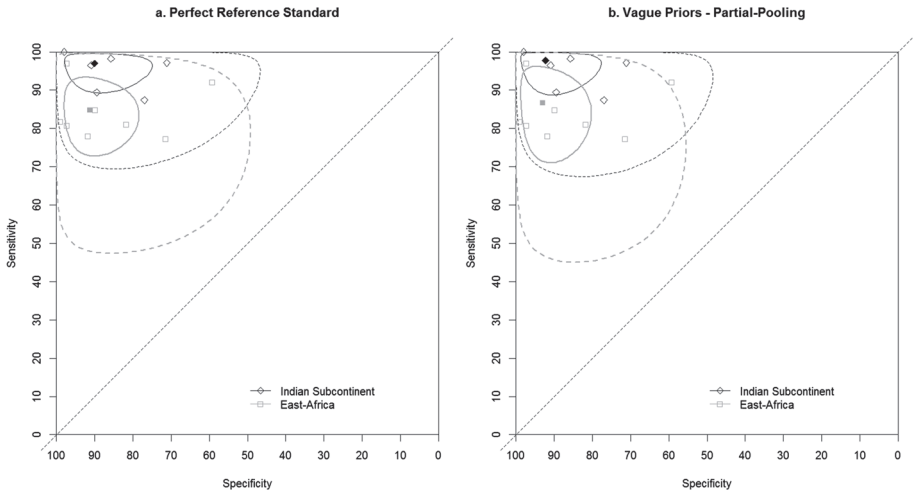
We performed these analyses assuming conditional independence between index and reference test results. To assess the robustness of the results to this assumption, we varied the correlation between index and reference test in model 3.2 from -0.9 to 0.9, in both diseased ($\rho_{i|D=1}$) and non-diseased subjects ($\rho_{i|D=0}$). Results were similar for negative correlations and moderate positive correlations (Supplementary Material 2). With higher positive correlations (0.5 or higher), the estimated S and C were lower compared with the estimates obtained assuming conditional independence. This can be expected as the association between index and reference test results, which under the conditional independence, assumption is explained through the disease status, is then in part explained through the direct correlation between test results.

Statistics in Medicine

Table III. Parameter estimates obtained from seven different models applied to the visceral leishmaniasis data.

Parameter	Perfect reference standard assumed	Model for the reference test diagnostic accuracy					
		Expert priors			Vague priors		
		Complete pooling	Partial-pooling	No-pooling	Complete-pooling	Partial-pooling	No-pooling
DIC	131.3	133.9	132.4	131.9	130.6	130.1	131.4
$\hat{\mu}_{S_{IS}}\left[\text{cloglog}^{-1}\left(\hat{\mu}_{S_{IS}}\right)\right]$	1.25 [96.9]	1.30 [97.4]	1.34 [97.8]	1.37 [98.1]	1.27 [97.2]	1.33 [97.7]	1.38 [98.2]
$\sigma\left(\hat{\mu}_{S_{IS}}\right)$	0.19	0.23	0.24	0.24	0.20	0.22	0.27
$\hat{\mu}_{C_{IS}}\left[\text{cloglog}^{-1}\left(\hat{\mu}_{C_{IS}}\right)\right]$	0.84 [90.1]	1.00 [93.4]	1.05 [94.3]	1.10 [95.0]	0.91 [91.7]	0.95 [92.4]	1.00 [93.4]
$\sigma\left(\hat{\mu}_{C_{IS}}\right)$	0.22	0.34	0.32	0.36	0.28	0.23	0.24
$\hat{\mu}_{S_{EA}}\left[\text{cloglog}^{-1}\left(\hat{\mu}_{S_{EA}}\right)\right]$	0.63 [84.7]	0.69 [86.3]	0.72 [87.2]	0.72 [87.3]	0.65 [85.3]	0.71 [86.9]	0.79 [88.9]
$\sigma\left(\hat{\mu}_{S_{EA}}\right)$	0.16	0.19	0.20	0.20	0.17	0.20	0.24
$\hat{\mu}_{C_{EA}}\left[\text{cloglog}^{-1}\left(\hat{\mu}_{C_{EA}}\right)\right]$	0.88 [91.1]	1.08 [94.7]	1.14 [94.9]	1.12 [95.6]	0.97 [92.9]	0.98 [93.1]	1.06 [94.4]
$\sigma\left(\hat{\mu}_{C_{EA}}\right)$	0.19	0.31	0.31	0.33	0.26	0.20	0.21
$\hat{\sigma}_S$	0.41	0.47	0.48	0.49	0.42	0.45	0.50
$\hat{\sigma}_C$	0.48	0.71	0.66	0.68	0.57	0.45	0.44
$\hat{\rho}_{SC}$	0.16	0.27	0.35	0.44	0.20	0.33	0.48
$\hat{\mu}_{S_{R1}}\left[\text{cloglog}^{-1}\left(\hat{\mu}_{C_{R1}}\right)\right]$	–	1.21 [96.4]	1.14 [95.6]	–	1.75 [99.7]	1.65 [99.4]	–
$\hat{\mu}_{C_{R1}}\left[\text{cloglog}^{-1}\left(\hat{\mu}_{C_{R1}}\right)\right]$	–	1.65 [99.4]	1.42 [98.4]	–	2.47 [100]	2.05 [100]	–
$\hat{\mu}_{S_{R2}}\left[\text{cloglog}^{-1}\left(\hat{\mu}_{S_{R2}}\right)\right]$	–	1.51 [98.9]	1.49 [98.8]	–	1.98 [99.9]	2.33 [100]	–
$\hat{\mu}_{C_{R2}}\left[\text{cloglog}^{-1}\left(\hat{\mu}_{C_{R2}}\right)\right]$	–	1.21 [96.5]	1.10 [95.0]	–	1.89 [99.9]	1.71 [99.6]	–

Notes: IS indicates data in the Indian Subcontinent; EA indicates data from Eastern Africa. Data were analyzed with the extended bivariate model described in Section 3. Uninformative priors were used for hyperparameters related to the index test. For the reference test, we fitted 'no-pooling', 'complete-pooling', and 'partial-pooling' models as defined in Section 3.2 and priors as described in Supplementary Material 4.



Notes: Open symbols show sensitivity-specificity pairs of individual studies. Filled symbols averages by geographic region. The ellipses on the complementary log-log scale show 95% credible regions for the average sensitivity and specificity and the dotted ellipses show a 95% prediction regions for a sensitivity-specificity pair from new study.

Figure 6. Graphical presentation of the results of the analysis of the visceral leishmaniasis data.

We assessed if results from the studies that used LCA as primary analysis method may have been biased by including study type (reference test versus LCA-based primary analysis) as a covariate in model 3.1. There were no significant differences between the two categories of studies. The difference in $\hat{\mu}_S$ between reference test and LCA-based studies was 0.15 (95% credible interval: $-0.29, 0.65$), and the difference in $\hat{\mu}_C$ was 0.13 ($-0.48, 0.79$) (data not shown).

6. Discussion

Diagnostic accuracy studies are still often analyzed under the assumption that perfect reference standards are used. Although meta-analyses of diagnostic studies tend to exclude studies with clearly inadequate reference standards, there remains some variation in the quality of reference standards used. The use of an imperfect reference standard may result in biased estimates of the index test S and C and in additional heterogeneity in S and C estimates across studies.

In this manuscript, we described an extension of the standard bivariate model for the meta-analysis of sensitivity–specificity pairs, correcting for imperfect reference standards and allowing the incorporation of results from studies that use LCA. As usually some information is available on the performance of the reference test, our Bayesian approach allows the use of this information through informative priors. We can obtain these priors from the literature or, as in our example, from expert opinion.

The simulation study shows that even when relatively little information is available on the diagnostic accuracy of the reference test, marked improvement of the estimation of the S and R of the index test is possible compared with the biased estimates obtained when incorrectly assuming that the reference test is perfect. Informative priors, which are in conflict with the data, can be detected through the use of model fit diagnostics as the DIC. However, because of the overlap in DIC distributions between different competing models, model choice should not be based on the DIC alone.

When applied to the motivating example, the best fitting model indicated a somewhat higher C of the index test compared with the analysis assuming a perfect reference test. Compared with the expert opinion, the reference tests appeared to perform better than expected, with S and C close to 100%. A possible explanation is that cases that could not be unequivocally classified with the reference test as diseased or not were removed from the diagnostic studies. Some studies reported the number of excluded patients explicitly [36, 38, 42], but for other studies, we could not determine if or how many clinical suspects who could not unequivocally classified as diseased or not were excluded. Another explanation is the strict criteria for selecting studies in the meta-analysis following the predefined protocol [32].

Limitations of our approach include that currently, we assume conditional independence between index and reference tests in equation (3.2) to ensure identifiability and that we do not correct for the exclusion of subjects who could not be unequivocally diagnosed from the analysis in primary studies.

We assessed the influence of the conditional independence assumption in a sensitivity analysis and concluded that allowing for moderate conditional dependence did not change the study conclusions. Further exploration of these effects are needed but are hampered by the lack of reliable data on correlations between index and reference tests and on the number of exclusions in diagnostic study publications [54]. Individual level meta-analysis of diagnostic studies, which fully report results for two or more diagnostic tests, using network meta-analytic techniques, may help to clarify these issues.

In our modeling, we did not take into account the possibility of differential verification bias where diagnostic work-up depends on the result from the index test, as these types of studies where specifically excluded. In some of the studies included in our meta-analysis, there were differences in the extent of verification of the disease status among subjects, but none depended on information provided by the index text. For example, no spleen aspirate may have been performed on subjects with a diagnosis of tuberculosis or malaria or with coagulation disorders. This may lead to a reduced sensitivity of the reference test, for which we correct in our meta-analysis. However, in none of the included studies, the extent of verification depended on the index test results. If the extent of verification depends on the index test, this may result in additional bias of S and C estimates that induced by the use of an imperfect reference test [55]. In such cases, Bayesian methods, which correct for verification bias [47], can be used to reanalyze the data. The resulting estimates can then be used as plug-in estimators in equation (3.4) for the meta-analysis.

References

1. Peeling RW, Smith PG, Bossuyt PMM. A guide for diagnostic evaluations. *Nature Reviews Microbiology* 2007; **5**(11):S2–S6. DOI: 10.1038/nrmicro1522.
2. Boelaert M, Bhattacharya S, Chappuis F, el Safi SH, Hailu A, Mondal D, Rijal S, Sundar S, Wasunna M, Peeling RW. Evaluation of rapid diagnostic tests: visceral leishmaniasis. *Nature Reviews Microbiology* 2007; **5**(11):S30–S39. DOI: 10.1038/nrmicro1766.
3. Zhou XH, Obuchowski NA, McClish DK. *Statistical Methods in Diagnostic Medicine*. Wiley-Interscience: New-York (US), 2002.
4. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press: Oxford (UK), 2003.
5. 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. *Journal of Clinical Epidemiology* 2005; **58**(10):982–990. DOI: 10.1016/j.jclinepi.2005.02.022.
6. TDR Diagnostics Evaluation Expert Panel. Evaluation of diagnostic tests for infectious diseases: general principles. *Nature Reviews Microbiology* 2007; **5**(11):S17–27. DOI: 10.1038/nrmicro1570.
7. Bachmann LM, Puhan MA, ter Riet G, Bossuyt PM. Sample sizes of studies on diagnostic accuracy: literature survey. *British Medical Journal* 2006; **332**:1127–1129. DOI: 10.1136/bmj.38793.637789.2F.
8. Harbord RM, Whiting P, Sterne JA, Egger M, Deeks JJ, Shang A, Bachmann LM. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. *Journal of Clinical Epidemiology* 2008; **61**(11):1095–1103. DOI: 10.1016/j.jclinepi.2007.09.013.
9. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. *Journal of Clinical Epidemiology* 2003; **56**:1129–1135. DOI: 10.1016/S0895-4356(03)00177-X.
10. 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 *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0.*, Deeks J, Bossuyt P, Gatsonis C (eds). The Cochrane Collaboration: Oxford (UK), 2010; 1–61.
11. Begg CB. Biases in the assessment of diagnostic tests. *Statistics in Medicine* 1987; **6**(4):411–423. DOI: 10.1002/sim.4780060402.
12. Pepe M. Insights into latent class analysis of diagnostic test performance. *Biostatistics* 2007; **8**(2):474–484. DOI: 10.1093/biostatistics/kxl038.
13. Black MA, Craig BA. Estimating disease prevalence in the absence of a gold standard. *Statistics in Medicine* 2002; **21**(18):2653–2669. DOI: 10.1002/sim.1178.
14. Goetghebeur E, Liinev J, Boelaert M, der Stuyft PV. Diagnostic test analyses in search of their gold standard: latent class analyses with random effects. *Statistical Methods in Medical Research* 2000; **9**:231–248. DOI: 10.1177/096228020000900304.
15. Berkvens D, Speybroeck N, Praet N, Adel A, Lesaffre E. Estimating disease prevalence in a Bayesian framework using probabilistic constraints. *Epidemiology* 2006; **17**(2):145–153. DOI: 10.1097/01.ede.0000198422.64801.8d.
16. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple tests. *Biometrics* 2001; **57**:158–167. DOI: 10.1111/j.0006-341X.2001.00158.x.
17. Bernatsky A, Joseph L, Belisle P, Boivin JF, Raja R, Moore A, Clarke A. Bayesian modelling of imperfect ascertainment methods in cancer studies. *Statistics in Medicine* 2005; **24**:2365–2379. DOI: 10.1002/sim.2116.
18. Spencer B. When do latent class models overstate accuracy for diagnostic and other classifiers in the absence of a gold standard? *Biometrics* 2012; **68**:559–566. DOI: 10.1111/j.1541-0420.2011.01694.x.

19. Albert PS, Dodd LE. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics* 2004; **60**(2):427–435. DOI: 10.1111/j.0006-341X.2004.00187.x.
20. Menten J, Boelaert M, Lesaffre E. Bayesian latent class models with conditionally dependent diagnostic tests: a case study. *Statistics in Medicine* 2008; **22**:4469–4488. DOI: 10.1002/sim.3317.
21. Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Statistics in Medicine* 1999; **18**(22):2987–3003. DOI: 10.1002/(SICI)1097-0258(19991130)18:22<2987::AID-SIM205>3.0.CO;2-B.
22. Chu H, Chen S, Louis TA. Random effects models in a meta-analysis of the accuracy of two diagnostic tests without a gold standard. *Journal of the American Statistical Association* 2009; **104**(486):512–523. DOI: 10.1198/jasa.2009.0017.
23. Sadatsafavi M, Shahidi N, Marra F, FitzGerald MH, Elwood KR, Guo N, Marra CA. A statistical method was used for the meta-analyis of tests for latent tb in the absence of a gold standard, combining random-effect and latent-class methods to estimate test accuracy. *Journal of Clinical Epidemiology* 2010; **63**(3):257–269. DOI: 10.1016/j.jclinepi.2009.04.008.
24. Verde PE. Meta-analysis of diagnostic test data: a bivariate Bayesian modeling approach. *Statistics in Medicine* 2010; **29**:3088–3102. DOI: 10.1002/sim.4055.
25. Dendukuri N, Schiller I, Joseph L, Pai M. Bayesian meta-analysis of the accuracy of a test for tuberculous pleuritis in the absence of a gold standard reference. *Biometrics* 2012; **68**:1285–1293. DOI: 10.1111/j.1541-0420.2012.01773.x.
26. Walter SD. Estimation of test sensitivity and specificity when disease confirmation is limited to positive results. *Epidemiology* 1999; **10**:67–72. DOI: 10.1097/00001648-199901000-00009.
27. Alvar J, Velez ID, Bern C, Herrer M, Desjeux P, Cano J, Jannin J, den Boer M, the WHO Leishmaniasis Control Team. Leishmaniasis worldwide and global estimates of its incidence. *PLOS One* 2012; **7**(5):1–12. DOI: 10.1371/journal.pone.0035671.
28. Chappuis F, Sundar S, Hailu A, Ghalib H, Rijal S, Peeling RW, Alvar J, Boelaert M. Visceral leishmaniasis: what are the needs for diagnosis, treatment and control? *Nature Reviews Microbiology* 2007; **5**(11):S7–S16. DOI: 10.1038/nrmicro1748.
29. Boelaert M, Meheus F, Sanchez A, Singh SP, Vanlerberghe V, Picado A, Meessen B, Sundar S. The poorest of the poor: a poverty appraisal of households affected by visceral leishmaniasis in Bihar, India. *Tropical Medicine and International Health* 2009; **14**(6):639–644. DOI: 10.1111/j.1365-3156.2009.02279.x.
30. Zijlstra EE, Ali MS, el Hassan AM, el Toum IA, Satti M, Ghalib HW, Kager PA. Kala-azar: a comparative study of parasitological methods and the direct agglutination test in diagnosis. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 1992; **86**(5):505–507.
31. 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. *British Medical Journal* 2006; **333**(7571):723–7261. DOI: 10.1136/bmj.38917.503056.7C.
32. Boelaert M, Chappuis F, Menten J, van Griensven J, Sunyoto T, Rijal S. Rapid diagnostic tests for visceral leishmaniasis. *Cochrane Database of Systematic Reviews* 2011; **6**. DOI: 10.1002/14651858.CD009135.
33. van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine* 2002; **21**(4):589–624. DOI: 10.1002/sim.1040.
34. Boelaert M, el Safi S, Hailu A, Mukhtar M, Rijal S, Sundar S. Diagnostic tests for kala-azar: a multi-centre study of the freeze-dried DAT, rK39 strip test and katex in East Africa and the Indian subcontinent. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 2008; **102**(1):32–40.
35. Boelaert M, Rijal S, Regmi S, Singh R, Karki B, Jacquet D. A comparative study of the effectiveness of diagnostic tests for visceral leishmaniasis. *The American Journal of Tropical Medicine and Hygiene* 2004; **70**(1):72–77.
36. Chappuis F, Rijal S, Singh R, Acharya P, Karki BM, Das. ML. Prospective evaluation and comparison of the direct agglutination test and an rK39-antigen-based dipstick test for the diagnosis of suspected kala-azar in Nepal. *Tropical Medicine and International Health* 2003; **8**(3):277–285.
37. Sundar S, Reed SG, Singh VP, Kumar PC, Murray HW. Rapid accurate field diagnosis of Indian visceral leishmaniasis. *Lancet* 1998; **351**(9102):563–565. DOI: 10.1016/S0140-6736(97)04350-X.
38. Chappuis F, Mueller Y, Nguimfack A, Rwakimari JB, Couffignal S, Boelaert M. Diagnostic accuracy of two rk39 antigen-based dipsticks and the formol gel test for rapid diagnosis of visceral leishmaniasis in northeastern Uganda. *Journal of Clinical Microbiology* 2005; **43**(12):5973–5977. DOI: 10.1128/JCM.43.12.5973-5977.2005.
39. Ritmeijer K, Melaku Y, Mueller M, Kipngetich S, O'keeffe C, Davidson RN. Evaluation of a new recombinant k39 rapid diagnostic test for Sudanese visceral leishmaniasis. *The American Journal of Tropical Medicine and Hygiene* 2006; **74**(1):76–80.
40. ter Horst R, Tefera T, Assefa G, Ebrahim AZ, Davidson RN, Ritmeijer K. Field evaluation of rk39 test and direct agglutination test for diagnosis of visceral leishmaniasis in a population with high prevalence of human immunodeficiency virus in Ethiopia. *The American Journal of Tropical Medicine and Hygiene* 2009; **80**(6):929–934.
41. Veeken H, Ritmeijer K, Seaman J, Davidson R. Comparison of an rK39 dipstick rapid test with direct agglutination test and splenic aspiration for the diagnosis of kala-azar in Sudan. *Tropical Medicine and International Health* 2003; **8**(2):164–167.
42. Chappuis F, Rijal S, Jha UK, Desjeux P, Karki BM, Koirala S. Field validity, reproducibility and feasibility of diagnostic tests for visceral leishmaniasis in rural Nepal. *Tropical Medicine and International Health* 2006; **11**(1):31–40.
43. Gelman A, Hill J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press: New York (NY, US), 2007.
44. McCutcheon AL. *Latent Class Analysis*, Quantitative Applications in the Social Sciences Series No. 64. Sage Publications: Thousand Oaks, US, 1987.
45. Garrett ES, Eaton WW, Zeger SL. Methods for evaluating the performance of diagnostic tests in the absence of a gold standard: a latent class model approach. *Statistics in Medicine* 2002; **21**:1289–1307. DOI: 10.1002/sim.1105.
46. Hadgu A, Qu Y. A biomedical application of latent class models with random effects. *Journal of the Royal Statistical Society Series C-Applied Statistics* 1998; **47**:603–616. DOI: 10.1111/1467-9876.00131.

47. de Groot JAH, Dendukuri N, Janssen KJM, Reitsma JB, Bossuyt PMM, Moons KGM. Adjusting for differential-verification bias in diagnostic-accuracy studies - a Bayesian approach. *Epidemiology* 2011; **22**(2):234–241.
48. Rosenblatt JE. Laboratory diagnosis of infections due to blood and tissue parasites. *Clinical Infectious Diseases* 2009; **49**(7):1103–1108. DOI: 10.1086/605574.
49. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis (Second Edition)*. Chapman and Hall/CRC: Boca Raton (Florida, US), 2004.
50. Agresti A, Hitchcock DB. Bayesian inference for categorical data analysis: a survey. *Technical report*, Florida (US), University of Florida, 2005.
51. Lesaffre E, Lawson AB. *Bayesian Biostatistics (Statistics in Practice)*. Wiley: New-York (US), 2012.
52. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit (with discussion and rejoinder). *Journal of the Royal Statistical Society, Series B* 2002; **64**:583–639. DOI: 10.1111/1467-9868.00353.
53. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Statistical Science* 1992; **7**:457–472. DOI: 10.1214/ss/1177011136.
54. Westwood ME, Whiting PF, Kleijnen J. How does study quality affect the results of a diagnostic meta-analysis? *BMC Medical Research Methodology* 2005; **5**(20). DOI: 10.1186/1471-2288-5-20.
55. Whiting P, Rutjes AWS, Reitsma JB, Glas AS, Bossuyt PMM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy. *Annals of Internal Medicine* 2004; **140**(3):189–202.