Evaluating Diagnostic Tests and Quantifying Prevalence for Tropical Infectious Diseases: a Paradigm of Latent Class Modelling Approaches with and without a Gold Standard for Schistosomiasis Diagnosis

Artemis Koukounari [‡]
Department of Infectious Disease Epidemiology,
London School of Hygiene & Tropical Medicine
and
Haziq Jamil
Department of Statistics, London School of Economics
and
Elena Erosheva
Department of Statistics, School of Social Work,
the Center for Statistics and the Social Sciences
University of Washington, Seattle
Laboratoire J. A. Dieudonné, Université Cotê d'Azur,
CNRS, Nice, France
Irini Moustaki
Department of Statistics, London School of Economics

March 19, 2019

[*]The authors deeply acknowledge the support of Professors Clive Shiff and Russel Stothard for provision of the two real datasets and useful scientific comments and advice throughout this article.
[†]Corresponding Author: Artemis Koukounari, email: Artemis.Koukounari@lshtm.ac.uk

Abstract

Various global health initiatives are currently advocating the elimination of schistosomiasis within the next decade. Schistosomiasis is a highly debilitating tropical infectious disease with severe burden of morbidity and thus operational research accurately evaluating diagnostics that quantify the epidemic status for guiding effective strategies, is essential. Latent class models (LCMs) have been generally considered in epidemiology and in particular in recent schistosomiasis diagnostic studies as a flexible tool for evaluating diagnostics because assessing the true disease status (-via a gold standard-) is not necessary. However, within the biostatistics literature, classical LCM have already been criticized for real-life problems under violation of the conditional independence (CI) assumption and when applied to a small number of diagnostics. Solutions of extensions of this model relaxing the CI assumption and accounting for zero-inflation, as well as collecting partial gold standard information, have been proposed, offering the potential for more robust model estimates. In the current article, we adjusted such approaches to the context of tropical infectious diseases via analysis of two real schistosomiasis datasets and extensive simulation studies. Our main conclusions included the poor model fit in low prevalence settings and that partial gold standard information improved the accuracy and bias of parameter estimates.

Keywords: local independence, sensitivity, specificity, medical diagnosis, lower and middle income countries, communicable diseases

1 Introduction

Despite the successes of the Millennium Development Goals era, infectious diseases still remain a major threat to humankind with severe burden of morbidity and mortality especially in low and middle income countries (LMICs) (Dye, 2014). Availability of accurate diagnosis constitutes an essential component in activities to combat these diseases such as within programs evaluating the effectiveness of interventions including verification of pathogens, elimination and detection of infections with markers of drug resistance (Banoo et al., 2006). Molecular assays constitute the gold standard for the diagnosis of several infectious diseases but the lack of sufficient funds, suitably trained staff and laboratory supplies still hinders their use in LMICs (Mabey et al., 2004). Schistosomiasis, is a long-lived, chronic, and highly debilitating tropical infectious disease caused by schistosome parasites, encountering some of these same challenges.

More precisely, schistosomiasis affects over 250 million people in rural and some urban populations across much of Africa and regions in South America, the Caribbean, People's Republic of China, and Southeast Asia (Fenwick et al., 2003; Steinmann et al., 2006). One of its intestinal forms, *Schistosoma mansoni* being the most widespread of the human-infecting schistosomes, can lead to severe liver and spleen damage if left untreated (McManus et al., 2018; Warren, 1978). Its urogenital form, *Schistosoma haematobium* is associated with squamous cell carcinoma of the urinary bladder, and *S. haematobium* is now classified by the International Agency for Research on Cancer as a carcinogen (IARC Working Group on the Evaluation of Carcinogenic Risks to Humans 2012). Furthermore, manifestations of *S. haematobium* infection may play a part in HIV transmission by increasing the number of inflammatory cells and possibly viral load in semen in women (Midzi et al., 2017). Both of these schistosomiasis forms are also associated with physical growth retardation, cognition and memory problems, exercise intolerance, as well as anaemia (Bustinduy et al., 2011; Jukes et al., 2002; King et al., 2005; Webster et al., 2009). Since the new millennium, schistosomia-

sis interventions are escalating (Fenwick and Webster, 2006) with eliminationand interruption of transmission in selected areas to be set as World Health Organization key goals for 2025. Accurate diagnosis for schistosomiasis is essential to assess the impact of large scale and repeated mass drug administration to control or even eliminate this disease (Shiff, 2012) with some specific molecular technologies to be considered very highly sensitive and specific (Shiff, 2015; van Lieshout and Roestenberg, 2015) but also expensive in endemic countries (Utzinger et al., 2015). Nevertheless, renewed emphasis upon research, and evaluation of schistosomiasis diagnostic tools has been generated (Shiff, 2015; Utzinger et al., 2015) with a notable increase of diagnostic studies particularly during the last five years (Al-Shehri et al., 2018; Beltrame et al., 2017; Booth et al., 2003; Carabin et al., 2005; Clements et al., 2018, 2017; Colley et al., 2013; Coulibaly et al., 2016; DuVall et al., 2013; Ferreira et al., 2017; Foo et al., 2015; Galappaththi-Arachchige et al., 2018; Holmen et al., 2015; Knopp et al., 2014; Koukounari et al., 2009; Lindholz et al., 2018; Shane et al., 2011; Sheele et al., 2013), using a 2-latent class model without a gold standard (Hui and Walter, 1980; Rindskopf and Rindskopf, 1986) to estimate and evaluate schistosomiasis diagnostic accuracy.

Latent class models (LCMs) have been increasingly utilized in medical research for the purpose of describing the relationship among diagnostic tests of schistosomiasis and other infectious diseases as being indicators of latent classes representing usually two states of a disease that of infected and non-infected. Latent class analysis explains the associations among a number of diagnostic tests or observed items in general using a small number of unobserved classes. The analysis aims to create homogeneous groups of respondents based on their results/responses on the observed diagnostic tests. LCM assumes that diagnostic tests are independent given the class. This is known as conditional or local independence assumption. However, the assumption of local independence may be violated and that will have implications in the properties of parameter estimates and statistical inference. LCMs have been considered in general in epidemiology as a flexible tool for evaluation of diagnostics

because assessment of the true disease status (-via a gold standard-) is not necessary. Relying on a probabilistic relationship between the examined diagnostic tests and the unobserved, or latent, disease status, LCM can provide estimates of disease prevalence and diagnostic test accuracy.

The aims of the current article are threefold. First, to study the effect that violations of model assumptions have on estimated parameters such as the prevalence of a disease, specificity and sensitivity of diagnostic tests. Second, to discuss extensions of the classical LCM to account for local dependencies and a large non-pathological group (i.e., zero-inflation). Third, to provide methodological guidance on the available modelling approaches for estimating test accuracy and disease prevalence in the absolute or partial absence of a gold standard for tropical infectious diseases using the paradigm of schistosomiasis by cautioning the practitioner not to blindly apply methods for estimating diagnostic error without a gold standard.

The remainder of the paper is organized as follows. In Section 1 we introduce the conditional independence (CI) assumption in the 2-LCM and describe situations where this could be violated in general and within the context of schistosomiasis diagnosis in particular. In section 1.2 we explain finite mixtures with deterministic all-zero and all-one responses. We subsequently justify the importance of checking model assumptions by highlighting associated inference problems and what potential solutions might exist from the perspective of accurately evaluating diagnostic tests with LCMs. We then introduce the idea of including partially the gold standard for the improvement of bias and precision of parameter estimates.

In Section 2 we provide an overview of the examined statistical models' including goodness-of-fit test statistics and measures of fit. We use examples to explain how the model might be relevant in schistosomiasis diagnostics evaluation research. In Section 3, we fit three different latent variable models and discuss results on real data for *S. haematobium* infection from Ghana (Koukounari et al., 2009) and *S. mansoni* infection from Uganda (Al-Shehri et al., 2018). Both of these datasets contain 5 observed binary items, have similar sample sizes and

lack gold standard evaluations.

In Section 4, we perform simulations on relevant scenarios for *S. mansoni* infection and study the performance of the classical LCM when local independence is violated. Within these simulations we also examine whether partial gold standard information improves the accuracy of parameter estimates (prevalence, sensitivities and specificities for the examined diagnostic tests) in the presence of model violations. Such a process allows us to examine how would our results (bias and precision) change if gold standard information was available for some individuals when three variations on latent class models are fitted. Finally, in Section 5, we propose guidelines for design of future infectious diseases diagnostic accuracy studies and make recommendations for further research. We also provide the JAGS code used to fit the LC, LCRE and FM models in this study in the supplementary material.

1.1 Assumption of Conditional Independence in the 2-latent Class Model

Latent class models typically assume that, conditional on the binary target disease latent status, test results are independent. It also implies that if the true disease status is misclassified by one test, the probability that it will be misclassified by another test will not be affected.

The CI assumption might be plausible when different tests are based on different scientific /technological grounds or when they measure different characteristics of the disease. However, this assumption often fails in practice. For instance, when some individuals without the disease of interest have another condition in common that increases the likelihood of two (or more) tests to render false positives because they are based on a comparable biological principle (Dendukuri, Hadgu, and Wang, 2009)-this can induce correlation between tests beyond the one explained by the disease status of interest. This could be the case in patients with alternative causes of microhaematuria and not infected with urogenital *S. haematobium* infection and perhaps with diagnostic tests such as reagent strips and urine filtration of 10

ml or whole urine samples from single urine specimens (Krauth et al., 2015). The existence of some other disease that has two or more diagnostic tests in common with the disease of primary interest can violate the CI of the examined variables even within latent classes of the disease of primary interest (Rindskopf and Rindskopf, 1986). More precisely, again within the context of diagnosis of urogenital *S. haematobium* infection, a series of self-reported urogenital symptoms (i.e. abnormal discharge colour, abnormal discharge smell, burning sensation in the genitals, bloody discharge, genital ulcer, red urine, pain on urination, stress incontinence and urge incontinence) might be also caused by sexually transmitted diseases such as *Chlamydia trachomatis* (Galappaththi-Arachchige et al., 2018). Individuals in low middle income countries can be co-infected with *Chlamydia trachomatis* and urogenital *S. haematobium* at the same time and such symptoms might be also used for diagnosis of*Chlamydia trachomatis* in schistosomiasis endemic countries. Another cause for conditional dependence among test results could arise if there is a subgroup of individuals with an early or less severe stage of the disease of interest and if these individuals are more likely to be missed by some tests (Brenner, 1996). The latter can be the case in *S. mansoni* infection where the parasitological Kato-Katz (KK) method (Katz, Chaves, and Pellegrino, 1972) to detect Schistosoma eggs in stool exhibits day-to-day variation in faecal egg output, and has low sensitivity in detecting light-intensity infections (Booth et al., 2003). In addition, serological tests, such as screening for antischistosomal antibodies, are of limited use for the diagnosis of active infection, as large parts of the population may carry antibodies due to past cured infections(Shiff, 2015; Utzinger et al., 2015). The need to understand the mechanisms of the tests, and in particular their mutual dependence in diseased and non-diseased subjects, as well as a clear clinical definition for disease has been already previously highlighted (Pepe and Janes, 2007). Finally, if some of the tests are administered by the same nurse or doctor and some are self-reported, there might be a method-effect and the outcomes might not be solely explained by the disease status.

1.2 Finite Mixtures with Deterministic All-Zero and All-One Responses

In epidemiological studies of general populations or pathological populations it is often the case that a large proportion of the sample has none (all-zero) or all (all-one) of the symptoms respectively. That all-zero proportion may include a non-pathological group from whom the diagnostic tests are all expected to be negative (zero). Data that contain an excess number of zeros, relatively to what the model allows/predicts, are called zero-inflated data (see e.g. Hall, 2000; Min and Agresti, 2005; Wang, 2010; Wall, Park, and Moustaki, 2015). In the case of schistosomiasis, it is well known that schistosomes are overdispersed within certain populations and age-groups; a small number of individuals carry most of the parasites and thus the all-zero effect is most relevant, (Gryseels et al., 2006).

The latent class model can be adjusted to include excess numbers of all-zero and all-one respondents in two separate classes which are deterministic (without error). That implies that some individuals in an all-zero or an all-one class will have probability 1 of responding zero or one to all items respectively. In some cases, there will only be a presence of all-zero or all-one effect rather than both.

1.3 LCM without a Gold Standard and an Alternative Solution with Partial Gold Standard

A relatively recent systematic review of the methodology and reporting of LCMs in diagnostic accuracy studies without a gold standard also highlighted that use of these models increased sharply during the last decade, notably in the domain of infectious diseases (van Smeden et al., 2014). However, this review revealed several problematic issues related to the methodology and reporting of studies using LCMs without a gold standard that deserve further attention. More precisely, the majority of the reviewed published studies used only a 2-class independence

LCM to estimate test sensitivity, specificity, and target disease prevalence in the absence of a gold standard without performing and reporting checks of model assumptions. As with any statistical technique or model, the validity of modeling results is jeopardized, leading to systematically overestimated classification accuracy rates (Spencer, 2012) when assumptions are not met. Hence, checking model assumptions through overall goodness-of-fit tests such as chi-square or likelihood ratio tests and measures of fit such as residuals constitute an essential part of the data analysis. For example, the CI assumption could be implicitly checked using goodness-of-fit tests. A model that does not fit the data would imply that more classes might be needed to explain the dependencies and therefore lack of CI of the hypothesized model. In addition, chi-square type measures (bivariate residuals) for pair of diagnostic tests can also identify those items that are not fitted well by the hypothesized model and violate the local independence assumption (Bartholomew, Steele, Moustaki, and Galbraith, 2008). Checking of model assumptions is important to appraise the validity of the reported results when LCMs without a gold standard are used (for further explanation see section 2.4). Such methodological flaws are unfortunately pertinent in some schistosomiasis diagnostic studies too, using LCMs without a gold standard.

Within the biostatistics literature, the latent class approach has already been criticized for real-life problems not only on the ground of violation of the CI assumption but also when applied to a small number of diagnostic tests (Albert and Dodd, 2004). Albert and Dodd (2004) fit two different models for conditional dependence (Albert et al., 2001; Qu et al., 1996) to real and simulated data without a gold standard to a set of 5 items. They found that estimates of sensitivity, specificity, and prevalence were substantially different under the two types of models. They also found that when the number of diagnostic tests is less or equal to 10, it is typically very difficult to discern statistically between the two forms of conditional dependence. They comment that it may be easier to distinguish between these models in a larger number of tests and sample sizes. However, they do also recognize that in most cases the

inclusion of 10 diagnostic tests is unrealistic which is definitively the case of schistosomiasis. Erosheva and Joutard (2014) introduced new models – Grade of Membership and Extended Mixture Grade of Membership – for capturing CI violations in LCM and performed simulation studies to examine recovery of sensitivity and specificity estimates. Erosheva and Joutard (2014) confirmed the findings of Albert and Dodd (2004) about difficulties in distinguishing the latent dependencies structure, now with the expanded class of models, and found that all models tend to underestimate sensitivity and overestimate specificity parameters when deterministic mixture components are present for the all-zero and all-one response patterns.

While these results would caution against using these LCMs, the difficulties of obtaining gold standard verification especially for infectious diseases in LMICs remain a practical reality. Albert and Dodd (2008) proposed solution that collects gold standard information on a subset of subjects but incorporates information from both the verified and nonverified subjects during LCMs estimation, offering the potential for more robust LCM estimates; they conducted simulations assuming common sensitivity and specificity across six diagnostic tests. In this article, we focus on the evaluation of imperfect schistosomiasis diagnostics. Similarly to Albert and Dodd (2008), we conduct simulation studies by considering three latent structure models and different proportions of gold standard. However, unlike Albert and Dodd (2008), we consider different sensitivities and specificities for different diagnostic tests and a range of prevalence settings, appropriate in the context of schistosomiasis. We hope that current findings would be also considered and adapted more broadly to other tropical infectious diseases such as malaria, tuberculosis, dengue and soil transmitted helminths where similar challenges in evaluation of diagnostics research and subsequently accurate quantification of prevalence, are pertinent.

2 Overview of Statistical Models

We discuss three variants of latent class models that are commonly used in epidemiological studies to define disease status in the absence of a gold standard: the latent class model, the latent class random effect model, and the finite mixture latent class model. Let $\mathbf{y} = (y_1, \ldots, y_p)'$ be a vector of p binary observed variables (also referred to as indicators, items, symptoms, diagnostic tests) taking values 0 and 1 indicating a negative and a positive test result respectively. There are in total $R = 2^p$ possible outcomes for vector $\mathbf{y}$. In the absence of gold standard the observed variables $\mathbf{y}$ are considered to be indicators of a discrete latent variable ξ that defines the state of the disease. We assume here two states of the latent variable ξ that of healthy status when $\xi = 0$ and sick status when $\xi = 1$.

2.1 Latent Class Model with Conditional Independence

We first discuss the standard LCM with CI. The marginal probability of a response pattern r is given by:

$$Pr(\mathbf{y}_r) = \sum_{j=0}^{j=1} P(\xi = j) P(\mathbf{y}_r \mid \xi = j), \quad r = 1, \ldots, R, \tag{1}$$

where $\tau = P(\xi = 1)$ defines the disease prevalence parameter for the population of interest. Furthermore, the tests are assumed to be conditionally independent given the true disease status which implies: $P(\mathbf{y}_r \mid \xi = j) = \prod_{i=1}^{p} P(y_i \mid \xi = j)$. Each y_i in turn is modelled with the Bernoulli distribution with class conditional probabilities $\lambda_{1i} = P(y_i = 1 \mid \xi = 1)$ and $\lambda_{2i} = P(y_i = 1 \mid \xi = 0)$. The model parameters are the $\lambda_{1i}, \lambda_{2i}, \tau$, $i = 1, \ldots, p$. Furthermore, for the ith diagnostic test, its sensitivity is the probability of the positive test result given that the true diagnosis is positive, $P(y_i = 1 \mid \xi = 1)$, and its specificity is the probability of a negative response given that the true diagnosis is negative, $P(y_i = 0 \mid \xi = 0) = 1 - P(y_i = 1 \mid \xi = 0)$. The sensitivity and specificity of test i implied by the latent class model are then simply $\mathrm{P}(\mathrm{y}_i = 1 \mid \xi = 1) = \lambda_{1i}$ and $\mathrm{P}(\mathrm{y}_i = 0 \mid \xi = 0) = 1 - \lambda_{2i}$.

2.2 Latent Class Gaussian Random Effects Model

However, in some applications the CI assumption does not hold for all or some of the indicators. To accommodate conditional dependencies among variables, an individual-specific random effect is introduced to capture heterogeneities that cannot be explained by the two classes (Albert et al., 2001; Qu et al., 1996). The random effect is a continuous normally distributed unobserved variable that serves as a summary of individual characteristics that explain together with the disease status the outcome of a test. This is known as a latent class Gaussian random effects model (LCRE). Under this model the y_i is modelled with the Bernoulli distribution with $P(y_i = 1 \mid \xi = j, u) = \Phi(\beta_{ij} + \sigma_j u)$, where Φ is the cumulative distribution function of the normal and u is the individual-specific random effect that follows a standard normal distribution. The parameter σ_j allows the random effect to have a different variance in each latent class defined by ξ. The marginal probability of a response pattern r under CI is given by:

$$Pr(\mathbf{y}_r) = \sum_{j=0}^{j=1} P(\xi = j) \int \prod_{i=1}^{p} P(y_i \mid \xi = j, u)\phi(u)du, \quad r = 1, \ldots, R. \tag{2}$$

The sensitivities and specificities for the latent class Gausian random effects model for each test j are given in closed form:

$$P(y_i = 1 \mid \xi = 1) = \Phi\left(\frac{\beta_{i1}}{(1 + \sigma_1^2)^{1/2}}\right)$$

and

$$P(y_i = 0 \mid \xi = 0) = 1 - \Phi\left(\frac{\beta_{i0}}{(1 + \sigma_0^2)^{1/2}}\right)$$

respectively. When $\sigma_0^2 = \sigma_1^2 = 0$, the model reduces to the latent class model with local independence.

2.3 Finite Mixture Latent Class Model

The above two models do not account for subjects who either have the disease or are healthy with certainty. As already discussed, in the data this will be seen as an excess number of all 0 or all 1 test results. The finite mixture latent class model allows for the modelling of excess zeros (all-zero) and excess ones (all-one) in the data as separate components of a latent class model. To model all-zero and all-one effects, the finite mixture model (FM) (Albert and Dodd, 2004) is employed. The model uses the two-class structure as its basis and adds two point masses for the combinations of all-zero and all-one responses. These point masses correspond to the healthiest and the most severely diseased patients that are always classified correctly. This model can be also considered as a latent class model with four classes of which two classes are fitted as having a point mass. In the truly diseased class the probability of a positive outcome test is 1 and in the truly healthy class the probability of a positive outcome is 0. Let t be an indicator that denotes correct classification. Specifically, let $t = 0$ if a healthy subject is always classified correctly (i.e., has the all-zero response pattern with p tests), $t = 1$ if a diseased subject is always classified correctly, and let $t = 2$ otherwise. Thus, subjects are either always classified correctly, when either $t = 0$ or $t = 1$, or a diagnostic error is possible when $t = 2$. Denote the probabilities for correctly classifying diseased and healthy subjects by $\eta_1 = P(t = 1)$ and $\eta_0 = P(t = 0)$, respectively. Let also $w_i(\xi)$ denote the probability of the ith test making a correct diagnosis when $t = 2$. The finite mixture model of Albert and Dodd (2004) assumes that the test results y_i are independent Bernoulli random variables, conditional on the true disease status and the classification indicator. The model becomes:

$$P(y_i = 1 \mid \xi, t) = \begin{cases} w_i(1), & \text{if } \xi = 1 \text{ and } t = 2 \\ 1, & \text{if } \xi = 1 \text{ and } t = 1 \\ 1 - w_i(0), & \text{if } \xi = 0 \text{ and } t = 2 \\ 0, & \text{if } \xi = 0 \text{ and } t = 0. \end{cases} \tag{3}$$

Note that $P(y_i = 1 \mid \xi = 1, t = 0) = P(y_i = 1 \mid \xi = 0, t = 1) = 0$. The specificity and sensitivity of the ith test under the finite mixture model are then $\mathrm{P}(\mathrm{y}_i = 1 \mid \xi = 1) = \eta_1) + (1 - \eta_1)w_i(1)$ and $\mathrm{P}(\mathrm{y}_i = 1 \mid \xi = 1) = \eta_0 + (1 - \eta_0)w_i(0)$, respectively.

The marginal probability of a response pattern r is given by:

$$Pr(\mathbf{y}_r) = \sum_{t=0}^{2} \sum_{j=0}^{j=1} P(\xi = j \mid t)P(t)P(\mathbf{y}_r \mid \xi = j, t), \quad r = 1, \ldots, R, \tag{4}$$

Note that $P(\xi = 0 \mid t = 0) = 1$ and $P(\xi = 1 \mid t = 1) = 1$.

2.4 Goodness-of-Fit Test Statistics and Measures of Fit

In this section we address the following questions: 1) does the model (LC, LCRE or FM) fit the data? 2) which of the three models (LC, LCRE. FM) provide a better fit to the data and when? and 3) which diagnostic tests are not fitted well by the hypothesized model?

We discuss here two ways for checking the fit of the models. The first way is to use global goodness-of-fit tests that compare the observed and expected (under the model) frequencies across the response patterns such as the likelihood ratio or Pearson chi-squared goodness-of-fit tests. The Pearson chi-squared goodness-of-fit test statistic, X^2, is given by:

$$\mathrm{X}^2 = \sum_{r=1}^{2^p} N \frac{(p_r - \hat{p}_r)^2}{p_r}, \tag{5}$$

where r represents a response pattern, N denotes the sample size, and p_r and $\hat{p}_r$ represent the observed and expected probabilities, respectively, of response pattern r. By multiplying p_r and $\hat{p}_r$ by N we obtain the observed and expected frequencies of pattern r. $\hat{p}_r$ for LC, LCRE and FM is estimated from $Pr(\mathbf{y}_r)$ given in equations (1), (2) and (4) respectively.

If the model holds, (5) is distributed approximately as χ^2 with degrees of freedom equal to the number of different response patterns (2^p) minus the number of independent parameters (prevalence, sensitivities, specificities, variances of random effects in the LCRE model) minus one. For the LC and LCRE with two latent classes, the degrees of freedom are $2^p - 2p - 2$

and $2^p - 2p - 4$ respectively. For the FM model (with four classes) the degrees of freedom are $2^p - 2p - 3$. The data can be considered as a 2^p contingency table. For instance, in our two real datasets from Ghana and Uganda, there are five diagnostic tests in each of them that give a $2 \times 2 \times 2 \times 2 \times 2$ contingency table. The sample sizes for the two datasets are 220 and 258 respectively and only 25 and 17 out of the total 32 response patterns appear in the sample and many patterns that occur in the sample have low (less than 5) frequencies. Theory tells us that small expected cell frequencies (less than 5) (known also as sparseness) have a distorting effect on the chi-square tests. Under sparseness, the test statistic will no longer have the chi-square distribution and so from the practical point of view these tests cannot be used see e.g. Agresti and Yang (1987).

Rather than looking at the whole set of response patterns, a second way could involve instead computing likelihood ratio and chi-square values for the two-way cross-tabulation of the diagnostic tests. That is, we can construct the 2×2 contingency tables obtained by taking two diagnostic tests at a time. For each cell of these bivariate contingency tables, we define a GF-Fit for which we compare the observed frequency with the expected frequency estimated under the corresponding latent variable model (LC, LCRE and FM). We use these terms to emphasize that there are no χ^2-distribution associated with them. For category a of variable i and category b of variable j the GF-Fits are defined as follows

$$GF - Fit_{ab}^{(ij)} = N(p_{ab}^{(ij)} - \hat{p}_{ab}^{(ij)})^2 / \hat{p}_{ab}^{(ij)} \ . \tag{6}$$

As a rule of thumb, if we consider the $GF - Fit_{ab}^{(ij)}$ as having a χ^2 distribution with one degree of freedom, then a value of those fit measures greater than 4 is indicative of poor fit at the 5% significance level (Bartholomew et al., 2008; Jöreskog and Moustaki, 2001). Similarly, summing these measures over a and b give the bivariate GF-Fits for variable i and j and therefore a value greater than sixteen will be then indicative of poor fit.The probabilities $\hat{p}$ can be evaluated under the LC, LCRE and FM models. The chi-squared residuals provide a measure of the discrepancy between the observed and the predicted frequency. A study of the

bivariate chi-squared residuals provides information about where the model does not fit or in other words pair of variables for which the CI assumption is violated (local dependencies). We will illustrate those in one of our real datasets.

3 Results from Real Datasets

We first used data from a study conducted in three villages northwest of Accra in Ghana which examined 220 adults using five *S.haematobium* diagnostic measures: microscopic examination of urine for detection of *S. haematobium* eggs, dipsticks for detection of haematuria, tests for circulating antigens, serological antibody tests (ELISA) and ultrasound scans of the urinary system (Koukounari et al., 2009). We also used data from a most recent study from 258 children near Lake Albert in Uganda using four *S. mansoni* diagnostic measures: microscopy of duplicate Kato-Katz smears from two consecutive stools (these are counting as two observed tests), urine-circulating cathodic antigen (CCA) dipstick, DNA-TaqMan and soluble egg antigen enzyme-linked immunosorbent assay (SEA-ELISA) (Al-Shehri et al., 2018). As mentioned in Section 1, both datasets have five observed binary variables and do not have any observation with known gold standard.

For the dataset from Ghana, all three models perform similarly in terms of model fit (the overall X^2 statistic is 12.79, d.f.=20 for the LC model, 12.42, d.f.=19 for the FM model and 13.23, d.f.=18 for the LCRE model) and provide us with similar estimates of prevalence, sensitivity and specificity. All the bivariate chi-squared residuals have values smaller than 4 indicating also a good fit. The estimated parameters under the LC, LCRE and FM models as well as goodness-of-fit statistics and measures of fit are provided in the shiny application under the Ghana example tab (https://haziqj.shinyapps.io/schis/).

We note that the prevalence level as estimated is fairly small, around 15%. We also note that sensitivity estimates for all diagnostic tests, except for ELISA, have very high associated uncertainty (estimated standard errors). Based on acceptable overall fit from all three models

(LC, LCRE, and FM), in this case, researchers might conclude that the latent class model is appropriate but that no reliable inference is available for item sensitivities.

For the dataset from Uganda, the estimated parameters under the LC, LCRE and FM models as well as goodness-of-fit statistics and measures of fit are provided in the shiny application under the Uganda example tab (https://haziqj.shinyapps.io/schis/). The models again perform similarly in terms of parameter estimates, however, the overall fit for this data set is poor across all three models (the overall X^2 ranges from 36.93, d.f.=20 for LC to 40.25, d.f.=18 for LCRE, to 44.42, d.f.=19 for FM). However, all the bivariate GF-Fit values are very small. Note that the overall test statistics cannot be trusted here due to sparseness (i.e. response patterns with expected frequencies under the model less than 5). More specifically, the GF-Fit values are given in Table 1 in the supplementary material for the pair POC-CCA and SEA-ELISA diagnostic tests and for the LC, LCRE and FM models. The LC model gives a total GF-Fit equal to 10.37. This is the total of the GF-fit values from the four cells. If we apply the rule of thumb given above, then any cell with a value greater than 16 indicates a bad fit or in other words the average across the four cells should not be greater than 4. According to the rule of thumb, all three models show a good fit on the bivariate tables (the GF-fit values are 9.54 and 4.00 for the LCRE and FM models respectively).

On the evidence from the margins, we have no reason to reject any of the three models. The overall significant (not an adequate model fit) result we obtained from the global goodness-of-fit tests cannot therefore be attributed to the relationships between the pairs of variables and in addition cannot be trusted due to sparseness (i.e. there are response patterns with expected frequencies under the model less than 5). The univariate and bivariate GF-Fits for the LC, LCRE and FM models are given in Table 2 in the supplementary material for all pairs of diagnostic tests. Among the three models the LCRE provides the smallest bivariate GF-Fits but all models show adequate fit. Each value that appears in the table is smaller than 16.

In the absence of gold standard and solid prior knowledge on latent dependencies, can researchers make conclusions by using these results? In this paper, we attempt to address this broad question with the simulation study below.

4 Design of the Simulations

To resemble real world problems in the diagnosis of tropical infectious diseases, we consider five imperfect test items and the gold standard which we assume has 100% sensitivity and 100% specificity. In the simulation study, we consider changing four settings: the sample size, the disease prevalence, the availability of gold standard, and the data generating model. Thus, we specify two sample sizes: 250, which is similar to a typical sample size in practice, and 1000, which we use to demonstrate potential improvements solely due to a larger sample. We specify a high, 40%, and a low, 8%, disease prevalence settings to examine how latent variable modeling results might be affected by different prevalence levels. We consider cases when gold standard is not collected (100% of individuals are missing the gold standard), when gold standard is available for 20% (80% missing) and when gold standard is available for 50% (50% missing). Finally, we consider three latent variable models for generating the data: latent class (LC), latent class with individual-specific random effects (LCRE) and the finite mixture model (FM). These settings give us $3 \times 2 \times 2 \times 3 = 36$ simulation scenarios, and we use 128 replications for each one.

Because, in real world problems, we do not know the data generating process, we fit all three models for every simulated data set and study bias and mean squared error for the estimates of prevalence, specificities and sensitivities under the three hypothesized models. The aims of the simulation study are two-fold. First, we explore the idea of sensitivity analysis by using latent structure models with different specifications of latent dependencies, similarly to Albert and Dodd (2008) and Erosheva and Joutard (2014) (but excluding the Grade of Membership and Extended Mixture Grade of Membership models in the current study), now

Table 1: Simulated parameter values for sensitivity and specificity for the diagnosis of *S. mansoni* infection

	Sensitivity	Specificity
Microscopy	0.60	0.99
Dipsticks	0.73	0.45
CAA	0.90	0.87
Antibody	0.90	0.50
LAMP	0.95	0.90
Gold std.	1.00	1.00

with different settings of prevalence, sensitivity and specificity parameters that are particularly relevant for *S. mansoni*. Given that we do not know the data generating model, we examine whether one can make reliable conclusions about prevalence, sensitivity, and specificity for the given tests by relying on results from the various models and goodness of fit test. Second, because many empirical studies only use the LC, we focus on its performance in our simulations. We investigate how the standard LC model performs under CI violation in small and medium sample sizes, two differing proportions of prevalence, and three levels of available gold standard.

The simulated data have been generated using the sensitivities and specificities given in Table 1. Note that, for test items other than gold standard, the true sensitivity values range from 0.6 to 0.95 and the true specificity values range from 0.45 to 0.99. These scenarios were assumed to represent a real world problem for the diagnosis of *S. mansoni* infection.

Following Erosheva and Joutard (2014), we set $\sigma_0 = \sigma_1 = 1.5$ for the random effects variance in the LCRE model, and $\eta_0 = P(t = 0) = 0.5$ and $\eta_1 = P(t = 1) = 0.2$ for generating the zero- and one-excess in the FM model.

We use as performance criteria the bias and mean squared error (MSE) for each parameter given by: $Bias = \frac{1}{R}\sum_{i=1}^{R}\left(\hat{\theta}_i - \theta\right)$, and $MSE = \frac{1}{R}\sum_{i=1}^{R}\left(\hat{\theta}_i - \theta\right)^2$, where R here is the

number of valid replicates, $\hat{\theta}_i$ is the estimate of a parameter or of its asymptotic standard error at the i^{th} valid replication, and θ is the corresponding true value.

4.1 Simulation Results

Simulation results show that the low prevalence setting could result in very unreliable estimates for prevalence and sensitivities for low sample sizes and in the absence of full gold standard. Moreover, even for larger sample size and particularly for low prevalence levels, not having any gold standard observations in the data could prohibit parameter recovery. We note that specificity parameters are generally estimated more reliably than sensitivity parameters in all scenarios (see Figure 1), consistent with findings from prior research (Erosheva and Joutard, 2014). Also consistent with prior research (Erosheva and Joutard, 2014), we find when specificity parameters are biased, they are biased upwards. However, we do not observe a general trend in the direction of biases for sensitivity parameters. The sensitivity parameter estimates acquire large expected biases when the data generating model is LCRE (see Figure 1C); the biases are smaller but substantial when FM is the data generating model (see Figure 1E), and the biases are the smallest when LC is the data generating model (see Figure 1A). A similar pattern arises for the MSE of parameter estimates (see Figure 2). We also observe that it is difficult to distinguish between various forms of latent dependencies solely based on the goodness of fit results (see Model Fit tab in the shiny application provided), which is consistent with prior findings (Erosheva and Joutard, 2014; Albert and Dodd, 2008).

Within each generating model scenario, having a large quantity of gold standard observations is improving the bias in estimates substantially. Thus, when gold standard is available for 50% of observations, we can conclude that specificity parameter estimates acquire little bias for either the large or the small sample size, irrespective of the data generating model and the two different assumed prevalence levels, and, most importantly, irrespective of the fitted model. Likewise, having 20% of gold standard improves with expected biases in sensitivity

estimates, however, we still observe bias up to 0.1 for sensitivity estimates when the sample size is small and LCRE is the data-generating model. Notably, these biases are observed when a simpler LC model is fitted to LCRE-generated data, however, these biases are comparable to expected biases when LCRE model is fit to LCRE-generated data. Thus, sensitivity estimates from both models – the true LCRE and the simpler LCM – acquire expected biases in this case. As expected, these biases improve and become potentially negligible for the larger sample size of 1000. The simulated results under the LC, LCRE and FM models data generating mechanisms as well as for the two different assumed sample sizes and prevalences are provided in the shiny application under the Simulated results tab and by selecting corresponding different simulation scenarios within this tab. Interestingly, the expected impact on parameter estimates due to using an incorrect fitting model is quite negligible as compared to the impact on parameter estimates due to the absence of gold standard particularly for the sensitivity estimates.

We observe that parameter recovery is less reliable when latent dependency structures are more complex than under the CI-LCM, with LCRE data-generating scenario being the most problematic. Thus, for the small sample size of 250, and 8% prevalence, when no gold standard is available and data are generated from the LCRE model, we find that the expected bias in sensitivity estimates ranges from -0.4 to 0.2, with high specificity items experiencing negative biases that are largest in magnitude. Results improve in the case of 40% prevalence and sample size of 250 (see Figures 1 and 2 in the Supplementary file). Under the LCRE generating model, inspecting the fit of the three models, we find that this cannot be considered adequate (even when taking into account that the data are quite sparse); however, we find that estimates from different models are in agreement. In this setting, using the LCRE model for prevalence estimation could result in erroneously estimating the prevalence to be much higher than it actually is (e.g., this could range from 12.3% at the 80th replication to 37.5 % at the 100th replication versus the assumed 8%). For one to obtain the different model estimates and fit

across the 128 replications, in the shiny app go to tab 'Modelfit' and type successively 1-128 in the box 'Replication number'). When LC model was the data generating model, all three models showed good model fit and reasonable estimates of prevalence. When FM was the data generating model, the prevalence estimates were also reasonable for the bigger sample size.

To conclude our observations from the simulation studies, we recommend that in the absence of strong scientific knowledge about the data generating mechanisms, one needs to be careful when interpreting prevalence and sensitivity estimates from latent structure models in the absence of gold standard especially in small sample sizes and in the presence of low prevalence. Our simulation studies show that, under realistic scenarios of a few test items and a sample size of a couple hundred observations, it is possible to obtain severely biased estimates, especially for prevalence and sensitivity parameters, when the true data generating mechanism is more complex than standard LCM with CI. The biases should be of utmost concern when true disease prevalence is low. Based on the simulation studies, we recommend to carefully examine model fit before drawing any conclusions. In our simulations, the fit of LC, LCRE and FM models was poor across the board for small sample sizes when data generating models were complex. It is the case that in real applications the data generating model is likely to be more complex than the simple LCM with CI. If the fit was found to be poor, in the case of small prevalence, sensitivity and prevalence estimates were found to be not reliable even if there is an agreement among different latent structure models. In such cases, we recommend collecting some gold standard. Our simulation results indicate that even 20% of gold standard can drastically improve estimation. Another alternative, which we did not explore in this paper, is to consider covariates such as gender and age on the prevalence and item response probabilities (i.e. the sensitivities and specificities of the diagnostic tests).

Figure 1: Bias of parameter estimates (sensitivities and specificities), as estimated by the LC, LCRE and FM models under differing proportions of missing gold standard, and under differing data generating mechanisms (sample size = 250, prevalence = 0.08).

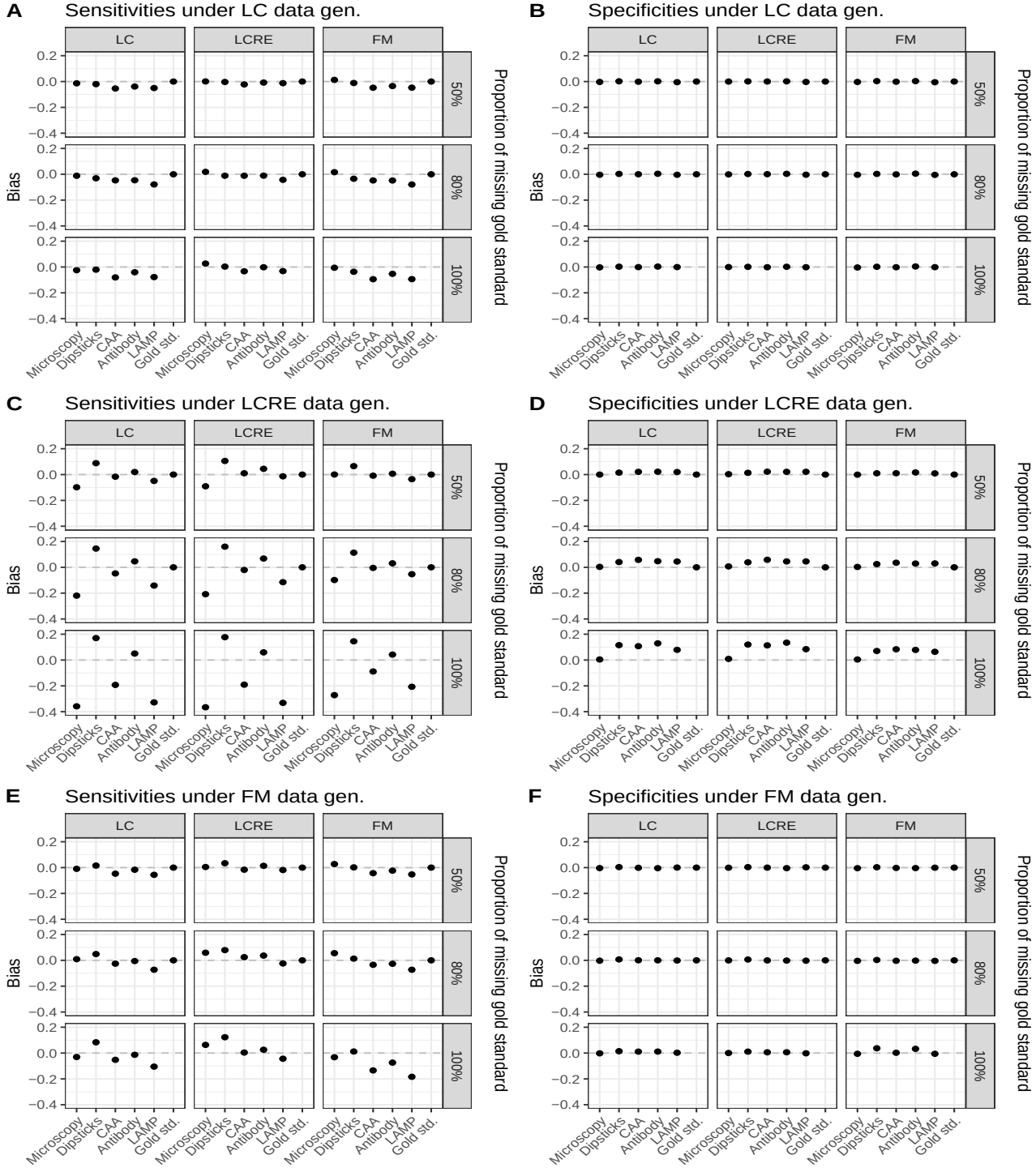
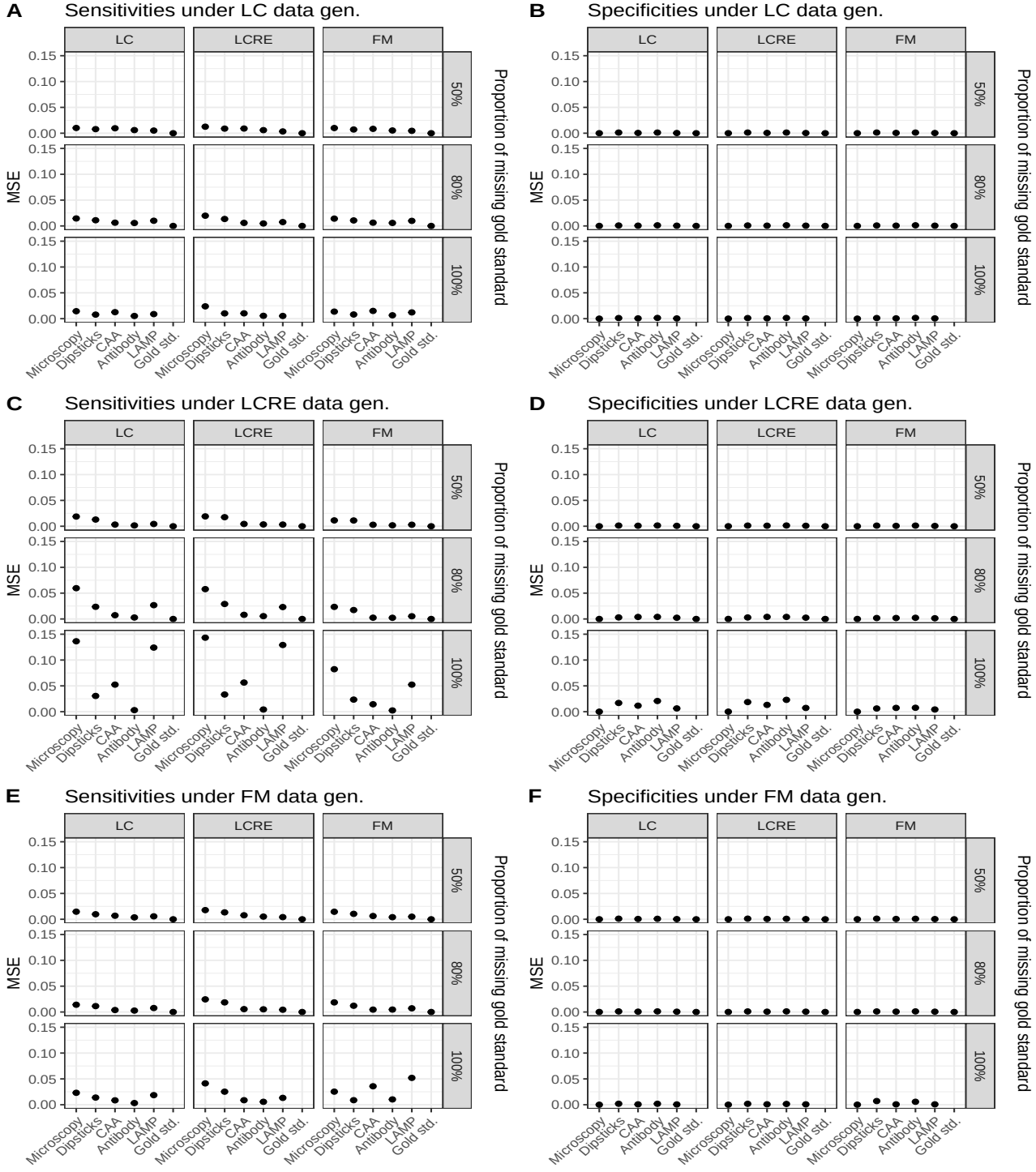


Figure 2: MSE of parameter estimates (sensitivities and specificities) as estimated by the LC, LCRE and FM models under differing proportions of missing gold standard, and under differing data generating mechanisms (sample size = 250, prevalence = 0.08).



5 Discussion

This article has presented the classical LCM as well as two extensions of this model (i.e. LCRE and FM) accounting for local dependencies of diagnostic tests and a large non-pathological group (i.e., zero-inflation), in order to quantify prevalence and evaluate diagnostic tests for tropical infectious diseases. The example of schistosomiasis (i.e. a parasitic disease) was used as a motivation for the presented methodological application. In schistosomiasis diagnostic studies, several inherent study design issues might compromise the accurate estimation of parameters from the presented latent variable models. More specifically, those study design issues are: the inclusion of small number of diagnostic tests (i.e. most often ≤ 5) with different mutual dependencies in diseased and non-diseased subjects; the lack of a gold standard due to mainly lack of appropriate equipment and training of technicians across the endemic countries because of scarce financial resources; the current use of relatively small sample sizes in relevant studies due to lack of research funding, and finally the inherent differing levels of prevalence of the studied disease even within the same area of one endemic country. In general, some of these issues can also be pertinent in the diagnosis of other major tropical infectious diseases such as malaria, tuberculosis, dengue and soil transmitted helminths and thus we hope that current findings would be also considered and adapted more broadly.

We have initially fitted these three different latent variable models in two real schistosomiasis datasets from sub-Saharan Africa in order to caution the practitioners not to blindly apply such methods for estimating diagnostic error without a gold standard. We have subsequently explored solutions from the biostatistical literature (Albert and Dodd, 2008) of whether including partially a gold standard in these models improves bias and precision of parameter estimates for the problem of schistosomiasis diagnosis via extensive simulation studies (assuming different sensitivities and specificities for different diagnostic tests and a low- and a high-prevalence settings). For the real dataset from Ghana, all three models provided acceptable overall model fit, with all the bivariate chi-squared residuals also indicating a good fit.

Estimates of prevalence, sensitivity and specificity were all similar across the 3 fitting models and thus in this case, researchers might conclude that the latent class model is appropriate but that no reliable inference is available for item sensitivities (due to large standard errors) in the absence of partial gold standard information. However, for the real dataset from Uganda although the models performed similarly in terms of parameter estimates, the overall fit was poor across all three models but with all the bivariate chi-squared residuals indicating a good fit. For this specific dataset from Uganda and based on our simulation results, we conclude that inclusion of partial gold standard information would have added more confidence in the parameter estimates and derived conclusions. It should be also noted that if the inclusion of partial gold standard is not feasible due to mainly financial constraints, then alternatively the inclusion of covariates might have improved model fit.

Furthermore, our simulation studies indicated that the sensitivity parameter estimates acquired some degree of expected biases in small sample sizes. We explain the reasons for these findings below. For instance, in our simulation scenaria, when we generate data from the FM model with an excess for all-one and all-zero responses and a sample size of 250 we are left with 75 observations to be fitted by the hypothesized model. When prevalence is also low, we then get very unstable results. On the other hand, the LCRE model fits a continuous latent variable (random effect) in addition to the two latent classes to explain the dependencies among the five (in the full absence of gold standard) or six items (when there is partial gold standard information), respectively, beyond the ones explained by the two latent classes. With an LCRE model we do not estimate loadings for the random effects but again the model is probably overfitting and could also result in unstable estimation. Another inherent problem with the LCRE fitting for the scenario of the small sample size, is the difficulty of estimating random effects variance from small samples. Those two might be the reasons that even when we estimate from LCRE or FM models to their respective data generating models, the number of items is small or the sample size is small after we eliminate the 1 and 0 response patterns.

More generally, the simulation studies showed that the fit of all three examined models was poor in the case of low prevalence and that partial gold standard information improved the accuracy and bias of parameter estimates (prevalence, sensitivities and specificities for the examined diagnostic tests) in the presence of model violations. In particular, when a future study could afford recruiting only around 250 participants, we recommend the inclusion of at least 20% gold standard information (i.e., collect gold standard on 50 randomly selected individuals of the initial sample) and, most ideally, of 50% gold standard (i.e., collect gold standard on 125 randomly selected individuals of the initial sample) for both low and higher prevalence scenaria. However, even for sample size of 1000, particularly for low prevalence levels, the full absence of gold standard could lead to erroneous parameter estimates but inclusion of 20% gold standard (i.e., gold standard collected on 200 randomly selected individuals of the initial sample) can also notably improve the obtained parameter estimates. If a high prevalence is expected, and the sample size can be as large as 1000, the inclusion of partial gold standard for the improvement of bias and precision of model estimates becomes less important. For all different prevalence levels and sample sizes, we definitely recommend to carefully examine different models fit of real world data before drawing any conclusions. It should be noted that the more complex models presented here are not possible to be fitted with less than 5 diagnostic items. With a higher number of items, it could also be warranted to explore other formulations for latent dependencies such as mixed membership (Airoldi, Blei, Erosheva, and Fienberg, 2014), which may not necessarily be helpful for only 5 items as in our case because of known difficulties in determining true latent dependency structures (Albert and Dodd, 2008; Erosheva and Joutard, 2014). In applied studies where only three diagnostic tests would be available and no gold standard would be feasible to be collected, we would also recommend Bayesian inference and the inclusion of informative priors on latent class modelling parameters. Such an approach would represent our (un)certainty in the model parameters and this could also improve the model estimation accuracy (Krolewiecki,

Koukounari, Romano, Caro, Scott, Fleitas, Cimino, and Shiff, 2018).

Finally, future methodological research should also explore the more precise merits of incorporating covariates (such as gender and age) on the prevalence and item response probabilities and how those could improve the fit of these models. Overall, the inclusion of covariates can strengthen the prediction power of a model. With latent variable models such as LC, LCRE and FM, the inclusion of individual level characteristics on the sensitivities and specificities can identify individual-item effects that might be of medical importance and interest (e.g., diagnostic tests behaving differently between men and women or different age groups) but also covariates can be seen as variables that together with the latent structure explain local dependencies and reduce the variance of the random effects. The inclusion of individual level covariates on the prevalences can identify groups of individuals that are more or less prone to a disease. For instance, in schistosomiasis, chronic infections occur in adults of increasing age, but in these groups it has been difficult to detect infection based on egg detection while serological and immunological tests can be more appropriate to detect duration of infection (Shiff, 2014) so assessing diagnostic accuracy for different age groups in schistosomiasis studies, is highly relevant. In addition, gender may influence the accuracy of estimates through factors such as menstruation and genitourinary tract infection in females (French, Rollinson, Basáñez, Mgeni, Khamis, and Stothard, 2007) and thus gender is another highly relevant covariate to be considered. Such approaches have been also empirically explored and applied in schistosomiasis diagnostic accuracy studies (Ibironke, Koukounari, Asaolu, Moustaki, and Shiff, 2012). Other covariates relevant to diagnostic accuracy could be multiple labs or technicians, different locations, schools or different countries. Such scenaria could be possible if not identical training or equipment is used across the different locations or if multi-country diagnostic studies were in place and the researcher analysing the data would be keen to check and estimate such differences. There are currently various debates and initiatives within the global health arena towards the elimination of schistosomiasis within the next decade (Fenwick

and Jourdan, 2016). Thus, operational research accurately evaluating existing and new diagnostic tools as well as quantifying the epidemic status for guiding effective and well-focused strategies, is essential (Secor and Colley, 2018). Our article apart from outlining the mathematical details of these models and their optimal usage for modelling diagnostic errors in the context of tropical infectious diseases, also provides the JAGS code so that readers can fit the discussed models to other relevant datasets and perform their own sensitivity analysis. In this way, we strongly believe that the current article could contribute notably valuable tools to the operational research agenda mentioned above.

SUPPLEMENTARY MATERIAL

Fit Measures: Univariate and bivariate GF-Fits for the LC, LCRE and FM models for the Uganda Data set

Figures: Figures 1 and 2 for Bias and MSE respectively of estimated sensitivities and specificities for the simulation scenario of sample size=250 and prevalence=0.4.

JAGS code: Code for fitting the latent class model under conditional independence, the latent class model with Gaussian random effects and the finite mixture latent class model.

References

Agresti, A. and M. C. Yang (1987). An empirical investigation of some effects of sparseness in contingency tables. *Computational Statistics and Data Analysis 17*, 9–21.

Airoldi, E. M., D. Blei, E. A. Erosheva, and S. E. Fienberg (2014). *Handbook of Mixed Membership Models and their Applications*. CRC press.

Al-Shehri, H., A. Koukounari, M. C. Stanton, M. Adriko, M. Arinaitwe, A. Atuhaire, N. B. Kabatereine, and J. R. Stothard (2018). Surveillance of intestinal schistosomiasis during control: a comparison of four diagnostic tests across five Ugandan primary schools in the Lake Albert region. *Parasitology 145*(13), 1715–1722.

Albert, P. S. and L. E. Dodd (2004). A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics 60*(2), 427–435.

Albert, P. S. and L. E. Dodd (2008). On estimating diagnostic accuracy from studies with multiple raters and partial gold standard evaluation. *Journal of the American Statistical Association 103*, 61–73.

Albert, P. S., L. M. McShane, J. H. Shih, and U.S. National Cancer Institute Bladder Tumor Marker Network (2001). Latent class modeling approaches for assessing diagnostic error without a gold standard: with applications to p53 immunohistochemical assays in bladder tumors. *Biometrics 57*, 610–619.

Banoo, S., D. Bell, P. Bossuyt, A. Herring, D. Mabey, F. Poole, P. G. Smith, N. Sriram, C. Wongsrichanalai, R. Linke, R. O'Brien, M. Perkins, J. Cunningham, P. Matsoso, C. M. Nathanson, P. Olliaro, R. W. Peeling, and A. Ramsay (2006). Evaluation of diagnostic tests for infectious diseases: general principles. *Nature Reviews Microbiology 4*, S21–31. 9 Suppl.

Bartholomew, D., F. Steele, I. Moustaki, and J. Galbraith (2008). *Analysis of Multivariate Social Science Data* (2nd ed.). Chapman and Hall/CRC.

Beltrame, A., M. Guerriero, A. Angheben, F. Gobbi, A. Requena-Mendez, L. Zammarchi, F. Formenti, F. Perandin, D. Buonfrate, and Z. Bisoffi (2017). Accuracy of parasitological and immunological tests for the screening of human schistosomiasis in immigrants and refugees from African countries: An approach with latent class analysis. *PLoS Negl Trop Dis 11*(6).

Booth, M., P. Vounatsou, E. K. N'goran, M. Tanner, and J. Utzinger (2003). The influence of sampling effort and the performance of the Kato-Katz technique in diagnosing Schistosoma mansoni and hookworm co-infections in rural Cte d'ivoire. *Parasitology 127*, 525–531.

Brenner, H. (1996). How independent are multiple independent diagnostic classifications? *Stat Med 15*, 1377–1386.

Bustinduy, A. L., C. L. Thomas, J. J. Fiutem, I. M. Parraga, P. L. Mungai, E. M. Muchiri, F. Mutuku, U. Kitron, and C. H. King (2011). Measuring fitness of Kenyan children with polyparasitic infections using the 20-meter shuttle run test as a morbidity metric. *PLoS Negl Trop Dis 5*(7).

Carabin, H., E. Balolong, L. Joseph, S. T. McGarvey, M. V. Johansen, T. Fernandez, A. L. Willingham, and R. Olveda (2005). Estimating sensitivity and specificity of a faecal examination method for Schistosoma japonicum infection in cats, dogs, water buffaloes, pigs, and rats in Western Samar and Sorsogon Provinces. *Int J Parasitol 35*(14), 1517–1524.

Clements, M. N., P. L. A. M. Corstjens, S. Binder, C. H. Jr, C. J. Campbell, A. de Dood, W. Fenwick, D. Harrison, C. H. Kayugi, D. King, O. Kornelis, G. Ndayishimiye, M. S. Ortu, A. Lamine, D. G. Zivieri, G. J. Colley, and J. van Dam (2018). Latent class analysis to evaluate performance of point-of-care CCA for low-intensity Schistosoma mansoni infections in Burundi. *Parasit Vectors 11*.

Clements, M. N., C. A. Donnelly, A. Fenwick, N. B. Kabatereine, S. C. L. Knowles, A. Meit, E. K. N'Goran, Y. Nalule, S. Nogaro, A. E. Phillips, E. M. Tukahebwa, and F. M. Fleming (2017). Interpreting ambiguous 'trace' results in Schistosoma mansoni CCA

Tests: Estimating sensitivity and specificity of ambiguous results with no gold standard. *PLoS Negl Trop Dis 11*.

Colley, D. G., S. Binder, C. Campbell, C. H. King, L. A. T. Tchuent, E. K. N'Goran, B. Erko, D. M. Karanja, N. B. Kabatereine, L. van Lieshout, and S. Rathbun (2013). A five-country evaluation of a point-of-care circulating cathodic antigen urine assay for the prevalence of Schistosoma mansoni. *Trop Med Int Health 18*(4), 477–484.

Coulibaly, J. T., M. Ouattara, S. L. Becker, N. C. Lo, J. Keiser, E. K. N'Goran, D. Ianniello, L. Rinaldi, G. Cringoli, and J. Utzinger (2016). Comparison of sensitivity and faecal egg counts of Mini-FLOTAC using fixed stool samples and Kato-Katz technique for the diagnosis of Schistosoma mansoni and soil-transmitted helminths. *Acta Trop 164*, 107–116.

Dendukuri, N., A. Hadgu, and L. Wang (2009). Modeling conditional dependence between diagnostic tests: a multiple latent variable model. *Stat Med 28*, 441–461.

DuVall, A. S., J. K. Fairley, L. Sutherland, A. L. Bustinduy, P. L. Mungai, E. M. Muchiri, I. Malhotra, U. Kitron, and C. H. King (2013). Development of a specimen-sparing multichannel bead assay to detect antiparasite IgG4 for the diagnosis of Schistosoma and Wuchereria infections on the coast of Kenya. *Am J Trop Med Hyg 88*(3), 426–432.

Dye, C. (2014). After 2015: infectious diseases in a new era of health and development. *Philos Trans R Soc Lond B Biol Sci 369*.

Erosheva, E. A. and C. Joutard (2014). Estimating diagnostic error without gold standard: A mixed membership approach. In E. M. Airoldi, D. M. Blei, E. A. Erosheva, and S. E. Fienberg (Eds.), *Handbook on Mixed-Membership Models*. Chapman Hall/CRC.

Fenwick, A. and P. Jourdan (2016). Schistosomiasis elimination by 2020 or 2030? *Int J Parasitol. 46*(7), 385–388.

Fenwick, A., L. Savioli, D. Engels, N. R. Bergquist, and M. H. Todd (2003). Drugs for the control of parasitic diseases: current status and development in schistosomiasis. *Trends Parasitol 19*, 509–515.

Fenwick, A. and J. P. Webster (2006). Schistosomiasis: challenges for control, treatment and drug resistance. *Curr Opin Infect Dis 19*, 577–582.

Ferreira, F. T., T. A. Fidelis, T. A. Pereira, A. Otoni, L. C. Queiroz, F. F. Amncio, C. M. Antunes, and J. R. Lambertucci (2017). Sensitivity and specificity of the circulating cathodic antigen rapid urine test in the diagnosis of Schistosomiasis mansoni infection and evaluation of morbidity in a low- endemic area in Brazil. *Rev Soc Bras Med Trop 50*, 358–364.

Foo, K. T., A. J. Blackstock, E. A. Ochola, D. O. Matete, P. N. Mwinzi, S. P. Montgomery, D. M. Karanja, and W. E. Secor (2015). Evaluation of point-of-contact circulating cathodic antigen assays for the detection of Schistosoma mansoni infection in low-, moderate-, and high-prevalence schools in western Kenya. *Am J Trop Med Hyg 92*, 1227–1232.

French, M. D., D. Rollinson, M. G. Basáñez, A. F. Mgeni, I. S. Khamis, and J. R. Stothard (2007). School-based control of urinary schistosomiasis on Zanzibar, Tanzania: Monitoring micro-haematuria with reagent strips as a rapid urological assessment. *Journal of pediatric urology 3*(5), 364–368.

Galappaththi-Arachchige, H. N., S. Holmen, A. Koukounari, E. Kleppa, P. Pillay, M. Sebitloane, P. Ndhlovu, L. van Lieshout, B. J. Vennervald, S. G. Gundersen, M. Taylor, and E. F. Kjetland (2018). Evaluating diagnostic indicators of urogenital Schistosoma haematobium infection in young women: A cross sectional study in rural South Africa. *PLoS One 13*(2).

Gryseels, B., K. Polman, J. Clerinx, and K. L. (2006). Human schistosomiasis. *The Lancet 368*(9541), 1106–1118.

Hall, D. B. (2000). Zero-inflated Poisson and Bionomial regression with random effects: a case study. *Biometrics 56*, 1030–1039.

Holmen, S. D., E. Kleppa, K. L. K, P. Pillay, L. van Lieshout, M. Taylor, F. Albregtsen, B. J. V. M., Onsrud, and E. F. Kjetland (2015). The first step toward diagnosing female genital schistosomiasis by computer image analysis. *Am J Trop Med Hyg 93*, 80–86.

Hui, S. L. and S. D. Walter (1980). Estimating the error rates of diagnostic tests. *Biometrics 36*, 167–171.

Ibironke, O., A. Koukounari, S. Asaolu, I. Moustaki, and C. Shiff (2012). Validation of a new test for Schistosoma haematobium based on detection of Dra1 DNA fragments in urine: evaluation through latent class analysis. *PLoS Negl Trop Dis 6*(1).

Jöreskog, K. G. and I. Moustaki (2001). Factor analysis of ordinal variables: a comparison of three approaches. *Multivariate Behavioral Research 36*, 347–387.

Jukes, M. C., C. A. Nokes, K. J. Alcock, J. K. Lambo, C. Kihamia, N. Ngorosho, A. Mbise, W. Lorri, E. Yona, L. Mwanri, A. D. Baddeley, A. Hall, D. A. Bundy, and P. for Child Development (2002). Heavy schistosomiasis associated with poor short-term memory and slower reaction times in Tanzanian schoolchildren. *Trop Med Int Health 7*, 104–117.

Katz, N., A. Chaves, and J. Pellegrino (1972). A simple device for quantitative stool thick-smear technique in schistosomiasis mansoni. Technical Report 14, Rev Inst Med Trop So Paulo.

King, C. H., K. Dickman, and D. J. Tisch (2005). *Reassessment of the cost of chronic helmintic infection: a meta-analysis of disability-related outcomes in endemic schistosomiasis*, Volume 365. Lancet Infect Dis.

Knopp, S., P. L. Corstjens, A. Koukounari, C. I. Cercamondi, S. M. Ame, S. M. Ali, C. J. de Dood, K. A. Mohammed, J. Utzinger, D. Rollinson, and G. J. van Dam (2014). Sensitivity and Specificity of a Urine Circulating Anodic Antigen Test for the Diagnosis of Schistosoma haematobium in Low Endemic Settings. *Am J Trop Med Hyg 90*(4), 638–645.

Koukounari, A., J. P. Webster, C. A. Donnelly, B. C. Bray, J. Naples, K. Bosompem, and

C. Shiff (2009). Sensitivities and specificities of diagnostic tests and infection prevalence of Schistosoma haematobium estimated from data on adults in villages northwest of Accra, Ghana. *Am J Trop Med Hyg 80*(3), 435–441.

Krauth, S. J., H. Greter, K. Stete, J. T. Coulibaly, S. I. Traor, B. N. Ngandolo, L. Y. Achi, J. Zinsstag, E. K. N'Goran, and J. Utzinger (2015). All that is blood is not schistosomiasis: experiences with reagent strip testing for urogenital schistosomiasis with special consideration to very-low prevalence settings. *Parasit Vectors 8*.

Krolewiecki, A. J., A. Koukounari, M. Romano, R. N. Caro, A. L. Scott, P. Fleitas, R. Cimino, and C. J. Shiff (2018). Transrenal DNA-based diagnosis of *strongyloides stercoralis* (Grassi, 1879) infection: Bayesian latent class modeling of test accuracy. *PLoS Negl Trop Dis*.

Lindholz, C. G., V. Favero, C. M. Verissimo, R. R. F. Candido, R. P. de Souza, R. R. D. Santos, A. L. Morassutti, H. R. Bittencourt, M. K. Jones, T. G. S. Pierre, and C. Graeff-Teixeira (2018). Study of diagnostic accuracy of Helmintex, Kato-Katz, and POC-CCA methods for diagnosing intestinal schistosomiasis in Candeal a low intensity transmission area in northeastern Brazil. *PLoS Negl Trop Dis 12*(3).

Mabey, D., R. W. Peeling, A. Ustianowski, and M. D. Perkins (2004). Diagnostics for the developing world. *Nature Reviews Microbiology 2*.

McManus, D. P., D. W. Dunne, M. Sacko, J. Utzinger, B. J. Vennervald, and X. N. Zhou (2018). Schistosomiasis. *Nat Rev Dis Primers 4*(1).

Midzi, N., T. Mduluza, B. Mudenge, L. Foldager, and P. D. C. Leutscher (2017). Decrease in seminal HIV-1 RNA load after praziquantel treatment of urogenital schistosomiasis coinfection in HIV- Positive men- an observational study. *Open Forum Infect. Dis. 4*.

Min, Y. and A. Agresti (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling 5*, 1–19.

Pepe, M. S. and H. Janes (2007). Insights into latent class analysis of diagnostic test performance. *Biostatistics 8*, 474–484.

Qu, Y., M. Tan, and M. H. Kutner (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics 52*, 797–810.

Rindskopf, D. and W. Rindskopf (1986). The value of latent class in medical diagnosis. *Statistics in Medicine 5*, 21–27.

Secor, W. and D. Colley (2018). When should the emphasis on schistosomiasis control move to elimination? *Trop Med Infect Dis. 3*(3).

Shane, H. L., J. R. Verani, B. Abudho, S. P. Montgomery, A. J. Blackstock, P. N. Mwinzi, S. E. Butler, D. M. Karanja, and W. E. Secor (2011). Evaluation of urine CCA assays for detection of Schistosoma mansoni infection in Western Kenya. *PLoS Negl Trop Dis 5*.

Sheele, J. M., J. H. Kihara, S. Baddorf, J. Byrne, and B. Ravi (2013). Evaluation of a novel rapid diagnostic test for Schistosoma haematobium based on the detection of human immunoglobulins bound to filtered Schistosoma haematobium eggs. *Trop Med*

Int Health 18, 477–484.

Shiff, C. (2012). The importance of definitive diagnosis in chronic schistosomiasis, with reference to Schistosoma haematobium. *J Parasitol Res 2*.

Shiff, C. (2014). New diagnostics reform infectious parasite epidemiology. *Lancet Infect Dis. 14*, 446–448.

Shiff, C. (2015). Accurate diagnostics for schistosomiasis: a new role for PCR? *Reports in Parasitology 4*, 23–29.

Spencer, B. (2012). When do latent class models overstate accuracy for diagnostic and other classifiers in the absence of a gold standard? *Biometrics 68*, 559566.

Steinmann, P., J. Keiser, R. Bos, M. Tanner, and J. Utzinger (2006). Schistosomiasis and water resources development: systematic review, meta-analysis, and estimates of people at risk. *Lancet Infect Dis 6*, 411–425.

Utzinger, J., S. L. Becker, L. van Lieshout, G. J. van Dam, and S. Knopp (2015). New diagnostic tools in schistosomiasis. *Clin Microbiol Infect 21*, 529–542.

van Lieshout, L. and M. Roestenberg (2015). Clinical consequences of new diagnostic tools for intestinal parasites. *Clin Microbiol Infect 21*, 520–528.

van Smeden, M., C. A. Naaktgeboren, J. B. Reitsma, K. G. Moons, and J. A. de Groot (2014). Latent class models in diagnostic studies when there is no reference standard–a systematic review. *Am J Epidemiol 179*, 423–431.

Wall, M. M., J. Y. Park, and I. Moustaki (2015). IRT modeling in the presence of zero-inflation with application to psychiatric disorder severity. *Applied Psychological Measurement 39*, 583–597.

Wang, L. (2010). IRTZIP modeling for multivariate zero-inflated count data. *Journal of Educational and Behavioral Statistics 35*, 671–692.

Warren, K. S. (1978). The pathology, pathobiology and pathogenesis of schistosomiasis. *Nature 273*, 609–612.

Webster, J. P., A. Koukounari, P. H. Lamberton, J. R. Stothard, and A. Fenwick (2009). Evaluation and application of potential schistosome-associated morbidity markers within large-scale mass chemotherapy programmes. *Parasitology 136*, 1789–1799.