Published in final edited form as:
*Stat Med*. 2014 October 30; 33(24): 4141–4169. doi:10.1002/sim.6218.

# Estimation of diagnostic test accuracy without full verification: a review of latent class methods

**John Collins**[*,†] and **Minh Huynh**
Rehabilitation Medicine Department, National Institutes of Health, Bethesda MD 20892, U.S.A

## Abstract

The performance of a diagnostic test is best evaluated against a reference test that is without error. For many diseases, this is not possible, and an imperfect reference test must be used. However, diagnostic accuracy estimates may be biased if inaccurately verified status is used as the truth. Statistical models have been developed to handle this situation by treating disease as a latent variable. In this paper, we conduct a systematized review of statistical methods using latent class models for estimating test accuracy and disease prevalence in the absence of complete verification.

### Keywords

no gold standard; latent class model; sensitivity; specificity; diagnostic testing; review

## 1. Introduction

Effective medical treatment often relies upon an ability to synthesize the results of imperfect and subjective assessments of patient status. Definitive or ‘gold standard’ tests are not available for many diseases or may be expensive or unethical to perform without prior determination of exposure to a risk factor. In other cases, a gold standard may, in fact, be highly error prone when used on certain subpopulations or only be accepted as a gold standard in conjunction with a second test. As an example of the former, it is unethical to perform a biopsy on an otherwise healthy individual without first screening for the presence of disease. In the latter case, the Western blot test for HIV infection is understood to be inadequate for detection of recent HIV infections while possessing a very low false positive rate and so is often used to confirm the results of highly sensitive enzyme-linked immunosorbent assay screening tests. In order for imperfect test results to be weighed properly when coming to a diagnosis or course of treatment, the performance of each test must be determined by probabilistic estimates of its accuracy conditional on disease status. This is straightforward when a gold standard is available for verifying disease status, but when it is not, confirmation must somehow be obtained through the use of one or more additional imperfect tests.

[*]Correspondence to: John Collins, 6100 Executive Blvd, Bethesda MD 20892, U.S.A.
[†]collinsjp@mail.nih.gov

In this review, we focus on the use of latent class models for estimating test accuracy and disease prevalence in the absence of a gold standard. The first work published on this topic was that of Hui and Walter in 1980 [1], and over the following decade, most publications on this topic applied their model to data with varying numbers of tests and population groups. During the 1990s, new methods were developed for examining the effects of dependent tests on accuracy estimation, the use of Bayesian methods, and modeling ordinal-valued test accuracy with receiver operating characteristic (ROC) curves. Since then, the field has broadened further to study tests whose accuracy varies between individuals, disease status that changes over the course of multiple tests, imputation techniques for handling nonignorable missing data, and the increasing use of nonidentifiable Bayesian models to avoid imposing questionable assumptions.

As the literature developed, its applications broadened from medical to veterinary testing and then to a variety of nonmedical subjects. Among the medical applications, extensive work has been carried out for HIV [2-8] and cancer screening [9-23] in particular. Since 2006, there has been increasing interest in studying paratuberculosis (Johne's disease) [24-31] and tropical diseases [32-36]. Applications beyond human and veterinary medicine include polygraph validity [37], toxic chemicals testing [38], Social Security disability determinations [39, 40], jury decisions [41], wildlife tracking [42, 43], asbestos fiber detection [44], administrative records quality [45, 46], and concordance between questions in national surveys [47].

While the breadth of applications speaks to the success of these methods, it has resulted in the literature being spread across a large number of journals. At the same time, the models in use have grown increasingly complex and exhibit subtleties that, while known by many practitioners, are not always addressed directly in print. This raises the risk of duplication of effort and slows the adoption of the best modeling practices. There have been efforts to address some of these computational and modeling issues through small and focused reviews using real and simulated data [48-50], but there remains a need for a broader review to bring the disparate branches of the literature together and identify areas ripe for future research. This review aims to fill those gaps. In addition, we present the Bayesian methods in a framework that is intelligible to non-Bayesian audiences and provide guidelines for the use of previously developed model structures.

### 1.1. Objectives

A systematized review of the literature, as defined by Grant and Booth [51], was conducted for latent class models that have been proposed to analyze test accuracy or estimate disease prevalences in the absence of a 'gold' standard or reference test without error. This type of review describes those that nearly meet the criteria for a full systematic review but fall short in one or more areas. In our case, only one author was in charge of the literature search, but all other criteria for a systematic review were met. This review has the following aims:

- To give an overview of the characteristics of latent class models that have been proposed for situations when there is no gold standard;
- To discuss the assumptions, strengths, weaknesses, and tradeoffs of existing models;

- To contextualize the development of the literature in terms of the tradeoffs identified;
- To provide guidance to researchers on the least biased methods of estimating test accuracy and disease prevalence in the absence of a gold standard.

We restrict ourselves to latent class models for estimating test accuracy. The use of discrepant analysis, Cohen's kappa, relative true positive fraction, expert panels, and related methods will not be discussed as each has been covered in a separate review by Rutjes *et al*. [52]. We strongly discourage the use of discrepant analysis in particular, as it has been shown to lead to uncorrectable bias in its estimates [53-58]. Cohen's kappa and other correlation-type measures mistake the agreement of two tests for the truth and penalize a new test that disagrees with the existing test even if the new test is more accurate. Expert panels are effective for bringing together the combined knowledge of many distinguished researchers but lack a formal data-driven methodology. If two expert panels came to different conclusions, how should this be interpreted?

### 1.2. Methods

The existing literature is segmented by subject matter, with resulting differences in terminology and a relatively sparse network of citations between clusters of authors. Consequently, we used a two-stage process for the identification of papers that met our inclusion criteria. The first stage accumulated all papers from two types of searches using Scopus, PubMed, and Google Scholar. The first type of search used combinations of keywords from the list given in the succeeding text such that at most 500 papers were returned. In the second type of search, all papers citing or cited by articles that had already met our inclusion criteria were identified. In the second stage, we removed papers obtained in the first stage, which did not meet the inclusion criteria.

Our inclusion criteria are as follows:

1. Published in English in 1991 or later;
2. Has a latent class model for prevalence of a disease and estimates at least one of sensitivity, specificity, positive predictive value, or negative predictive value for one or more diagnostic tests with imperfect accuracy; or
3. Performs an analysis of the robustness or necessary sample sizes of existing models or develops code implementing an existing model.

All criteria except for the lower bound on publication date were decided in advance. The publication date criterion was chosen to avoid overlap with the thorough 1988 review of Walter and Irwig [59], while allowing duplication with the two small 1998 reviews of Hui and Zhou [60, 61]. This overlap in publication date range was decided after papers published between 1991 and 1998 that had not been included in either review were identified, and the inclusion of many papers cited in either review was deemed necessary to properly contextualize later developments in the literature.

A publication could satisfy these criteria if either it satisfied the first two criteria or the second criterion was not met in full perhaps because no estimation was performed, then the

first and third criteria were satisfied. For many papers returned from keyword searches, it was unambiguously clear from their abstract alone that the inclusion criteria would not be met. These papers were discarded in the first stage of our process. If there was any uncertainty on whether the criteria might be satisfied, a copy of the paper was obtained for the closer examination of the second stage. The second stage of this process involved reading all papers included from the first stage to determine with certainty if they met the criteria.

Keyword searches identified 1773 papers, and searches of bibliographies and works cited identified 3180 papers, with up to 3969 unique publications identified in total. The first stage of the process described earlier resulted in the immediate exclusion of at least 90% of this total. For each of the remaining papers, a PDF copy was obtained and read to determine if it in fact met the inclusion criteria. Only 248 papers remained that met the inclusion criteria, of which a further 13 were excluded for one of two additional reasons. These 13 either directly applied an existing model for which multiple examples were already available among the 235 papers not excluded and which did not contain discussion of any of the model or sample's technical issues, or focused entirely on the well-understood fact that estimates are biased if an imperfect test is treated as if it were a gold standard when estimating prevalence or accuracy.

The list of keywords used for searches was initially very small. As papers were identified that met our inclusion criteria if any of the paper's keywords or words in its title were relevant to the review topic and not on the list already then they were added. The initial list consisted of only the words Hui–Walter, sensitivity, specificity, prevalence, and latent class model. Our final list of keywords is given in the succeeding text:

> Misclassification, pseudogold standard, (no) gold standard, verification bias, imperfect reference, screening test, (diagnostic) test accuracy, sensitivity, specificity, prevalence, ROC curve, receiver operating characteristic curve, Hui–Walter, latent class model, nonidentified, nonidentifiable, evaluation of diagnostic test, predictive values, dependence, dependent test.

### 1.3. Structure of the review

Models of accuracy without gold standard differ from each other because of the objective differences in how testing is performed and the subjective choices of model properties meant to capture latent characteristics of the sample. Objective aspects are those variables fixed once all data are collected and include the following: the number of tests performed, sampling design, availability of a gold standard, and whether each test is scored on a binary, ordinal, or continuous scale. In particular, there is a continuum between all subjects being tested with a gold standard and its complete absence in which a percentage of all subjects have their status verified. Subjective elements include the choice between frequentist or Bayesian statistical methods, inclusion of covariates, the modeling of conditional dependence between tests, and whether test accuracy is assumed to be fixed or allowed to vary between subpopulations.

The review is divided into five sections. Its focus is primarily on the subjective aspects of modeling, as defined earlier, because these defy simple recommendation for best practice. In the introduction, the criteria and methodology of the review are discussed, and the foundational model of Hui and Walter [1] is presented. The second section considers the main extensions of the Hui–Walter model to reduce bias, including modeling dependent tests, test accuracy that varies between populations, and methods to handle issues that arise from repeat testing possibly with time lapse between tests. The third section provides an introduction to Bayesian techniques for a general audience as well as a discussion of issues specific to nonidentifiable Bayesian models. These models often arise from attempts to properly account for the issues discussed in the second section without resorting to unjustifiable assumptions about the data. The fourth section examines specialized topics and model selection techniques, such as the use of partially verified data, defining and estimating the accuracy of a nonbinary test, sample size estimation, and the availability of code for implementing the standard models. The final section discusses weaknesses of the review, summarizes the active areas of research, and analyzes open problems in the field.

### 1.4. Terminology

Let $D$ denote the disease of interest. It is a latent variable, meaning that its existence is not directly measured in all subjects and can only be inferred from the results of imperfect tests and covariates. Throughout, we assume that $D$ is binary, representing diseased ($D = 1$) and not diseased ($D = 0$) status. The prevalence of disease in the population is $\pi = \mathbb{P}(D = 1)$. Diagnostic tests are allowed to be binary, to be ordinal ($N$-valued for $N > 2$), or to take values in a continuum. If the test is binary, its sensitivity is the probability of a positive test given disease is present, denoted $\alpha = \mathbb{P}(T = 1 \mid D = 1)$, and specificity is the probability of a negative test result given disease is absent, denoted $\beta = \mathbb{P}(T = 0 \mid D = 0)$. The sensitivity and specificity of ordinal and continuous-valued tests are defined with respect to a cutoff threshold for positivity. Namely, for threshold $\tau$, the test takes values $T_\tau = 1$ if $T \geq \tau$ and $T_\tau = 0$ if $T < \tau$. Sensitivity and specificity dependent on cutoff $\tau$ are denoted $\alpha_\tau$ and $\beta_\tau$, respectively, although it is customary to suppress the subscripts whenever the cutoff is clear from context. The false positive and false negative rates are, respectively, $1 - \beta$ and $1 - \alpha$. Some authors prefer to use the positive predictive value, $\text{PPV} = \mathbb{P}(D = 1 \mid T = 1)$, and negative predictive value, $\text{NPV} = \mathbb{P}(D = 0 \mid T = 0)$, as these are the probabilities of concern to a clinician when deciding on a course of treatment. Despite this, researchers focus on sensitivity and specificity as PPV and NPV are prevalence dependent and so can give misleading information for very low and high prevalence populations. Furthermore, PPV and NPV can be computed from knowledge of disease prevalence, sensitivity, and specificity, of the test using Bayes' theorem:

$$PPV = \frac{\alpha\pi}{\alpha\pi + (1-\beta)(1-\pi)}$$
$$NPV = \frac{\beta(1-\pi)}{\beta(1-\pi) + (1-\alpha)\pi}$$

The dependence of predictive values on prevalence is easily observed: if $\pi = 0.05$, $\alpha = \beta = 0.95$, then the PPV is 0.5 and the NPV is 0.997. Even though the test is highly accurate, a positive test outcome is at best a coin flip for determining disease status.

A model is said to be nonidentifiable if there exist at least two choices of parameters for which the distributions of observable data are the same, otherwise it is identifiable. Models of accuracy are trivially nonidentifiable in that they can suffer from label switching, wherein positive (negative) test results are interpreted as predictions of disease absence (presence). Label swapping replaces estimation of $(\pi, \alpha, \beta)$ with $(1 - \pi, 1 - \beta, 1 - \alpha)$, and so is easily recognizable in practice and can be prevented by imposing the condition $\alpha + \beta > 1$. For this reason, we will ignore label switching when discussing a model's identifiability, although it must be taken into account when performing maximum likelihood estimation (MLE) or Markov chain Monte Carlo (MCMC). When a model has a number of parameters equal to its degrees of freedom, it is not necessarily identifiable, and some authors have emphasized these situations by calling such models weakly identifiable [62]. An educational example of the relationship between degrees of freedom, model parameters, and identifiability is given in the following section.

### 1.5. Foundational model: Hui–Walter

We introduce the methodology by which latent class models can estimate test accuracy and disease prevalence in the absence of a gold standard through a concrete example and conclude with a description of the foundational model of Hui and Walter. The data in Table I were studied by Hui and Walter [1] and represent the results of two tests for tuberculosis given to a general population group of children in a single school district (pop. 1) and to a high-risk group of individuals at a state sanatorium (pop. 2).

In order to understand the effect of an absence of gold standard, suppose first that test $T_2$ is a gold standard and focus on population 1. The prevalence and properties of $T_1$ are then simple to determine, with prevalence $\hat{\pi} = 23/555 \approx 0.041$, sensitivity $\hat{\alpha_1} = 14/23 \approx 0.609$, and specificity $\hat{\beta_1} = 528/532 \approx 0.992$. The same process could be repeated for the second population or carried out for the combined sample.

Consider the case where neither test is a gold standard. Preliminary knowledge of the accuracy of both tests can still be gleaned from the size of their disagreement relative to the total sample size. If both tests are highly accurate, they will tend to correctly agree with high probability, and so the relative frequency of off-diagonal entries in the table will be very small. If instead the tests have poor accuracy and are independent, disagreement will be frequent, and the relative frequency of the off-diagonal will be high. Similar comparison of the size of the two diagonal entries gives insight into the prevalence.

Difficulties arise once we attempt to derive estimates of both tests' properties as well as disease prevalence without assuming that either test is a gold standard. With two tests, in a single population, knowing three of the four cells uniquely determines the fourth because the sample size is known. Therefore, with two tests, each distinct population provides three degrees of freedom. When we assume that $T_2$ is the gold standard, one population is sufficient as its 3 degrees of freedom are all that is needed for the three parameters being estimated. If $T_2$ is not a gold standard, however, the two tests each have two accuracy parameters to be estimated, for a total of five parameters including prevalence. As there are two more parameters than degrees of freedom, this model is nonidentifiable; there will be many combinations of test accuracy and prevalence that fit the data. The addition of a

second population increases the available degrees of freedom by three while adding only one parameter for the new population's prevalence, resulting in an equal number of degrees of freedom and parameters. The beneficial impact of studying multiple populations is one of the key insights of Hui and Walter [1].

Their model, hereafter the HW model, assumes that each population has distinct disease prevalence, that test sensitivity and specificity do not vary between populations, and that test results are independent conditional on disease status. It has six parameters: two for the disease prevalence in each population and four for the sensitivity and specificity of each test. The likelihood of sampling $n_{sij}$ individuals in population $s = 1, 2$ with test results $T_1 = i$, $T_2 = j$, for $i, j = 0, 1$ under the models assumptions, is

$$L = \prod_{s=1}^{2} [\pi_s \alpha_1 \alpha_2 + (1-\pi_s)(1-\beta_1)(1-\beta_2)]^{n_{s11}} [\pi_s(1-\alpha_1)\alpha_2 + (1-\pi_s)\beta_1(1-\beta_2)]^{n_{s01}} \times [\pi_s \alpha_1(1-\alpha_2) + (1-\pi_s)(1-\beta_1)\beta_2]^{n_{s10}} [\pi_s(1-\alpha_1)(1-\alpha_2) + (1-\pi_s)\beta_1\beta_2]^{n_{s00}} \quad (1)$$

In the likelihood earlier, an individual that is positive on both tests can either be diseased or healthy. In the former, both tests were correctly positive, and the probability of this outcome is $\pi_s \alpha_1 \alpha_2$. Otherwise, both tests represent false positives with the probability of this event given by $(1 - \pi_s)(1 - \beta_1)(1 - \beta_2)$. The other three cases proceed by similar logic. Exact analytic solutions for two tests and two populations [1] and three tests and one population [63] are available. The general case with $n$ tests and $g$ populations has $2n + g$ parameters and $g(2^n - 1)$ degrees of freedom, with $g$ parameters for prevalences in $g$ populations and $2n$ for the sensitivities and specificities of $n$ tests.

## 2. Extending the HW model

Incorrectly specified latent class models may systematically overestimate accuracy rates [64, 65]. Consequently, as the HW model gained in popularity, it became necessary to examine its robustness and develop alternative models that weakened its assumptions. The HW model's lack of robustness for conditionally dependent tests is well known [66, 67] and has also been established for tests with prevalence-dependent accuracy [49]. In this section, extensions of the HW model using conditional test dependence, explanatory covariates, and nonconstant accuracy rates are discussed. Concern for model identifiability places limits upon how far these assumptions can be weakened, eventually leading to a rise in the use of Bayesian methods and model selection techniques. For the remainder of this review, we will omit writing 'conditional' when discussing conditional test dependence as this is the only type of dependence considered.

### 2.1. Dependence: covariance

The covariance model was introduced for a model of two tests and two populations by Vacek [67] but with parameter constraints necessary for identifiability and later by Sinclair and Gastwirth [68] to study survey response reconciliation bias. The model was then generalized to arbitrarily many tests and an analysis of its properties conducted by Torrance-Rynard and Walter [69]. Dependence is introduced through the equations

$$\mathbb{P}(T_i, T_j|D=1)=\mathbb{P}(T_i|D=1)\mathbb{P}(T_j|D=1)+(-1)^{|T_i-T_j|}\delta_{ij} \quad (2)$$

$$\mathbb{P}(T_i, T_j|D=0)=\mathbb{P}(T_i|D=0)\mathbb{P}(T_j|D=0)+(-1)^{|T_i-T_j|}\varepsilon_{ij} \quad (3)$$

for each pair of tests $T_i$, $T_j$. If $\delta_{ij} = \varepsilon_{ij} = 0$, the tests are independent, and the HW model is recovered. The relationship between nonidentity covariance matrices and dependent random variables is well understood, making this dependence structure conceptually simple to understand. The primary disadvantage of this model is that its full dependence structure uses $n^2 + n$ parameters to model accuracy when $n$ tests are used. In particular, identifiability requires at least three populations and two tests, two populations and three tests, or one population and five tests. Despite having degrees of freedom equal to its number of parameters, the first of these three cases is nonidentifiable [70], and the last usually requires repeat testing as few applications have more than three distinct tests. This model also assumes that no higher order dependencies exist and in particular, that dependence between three tests can be modeled in terms of pairwise relations.

It is frequently assumed that the covariance matrix has a block diagonal form, with tests from distinct blocks independent of each other. This reduces the number of parameters needed and is an intuitively plausible assumption when tests measure distinct biological mechanisms [13,30, 71]. The most common version of the block diagonal method is to assume that one test is independent of the others while allowing the full covariance matrix between the remainder [11, 72, 73]. Alternatively, by setting $\varepsilon_{ij} = 0$ for all $i$, $j$, dependence between tests is restricted to diseased subjects [62].

In order to account for higher order interactions that were assumed absent by the covariance model, Bayesian models have considered full characterization of the conditional distributions between tests [74-79]. This requires the use of $2^n$ accuracy parameters with $n$ tests and so tends to result in nonidentifiable models that necessitate the use of informative priors, as discussed in Section 3.3. If at least one pair of tests is independent out of each collection of three tests, this more general dependence structure reduces to the standard covariance model [72, 80].

### 2.2. Dependence: random effects

The Gaussian random effects (GRE) model [7] captures conditional dependence as an unobserved random effect common to each of the tests. This effect is designed to capture both subject-specific characteristics not directly recorded as well as inter-rater dependence resulting from subject-specific effects in multiple testing. In a later paper [81], observed covariates were explicitly included, and we present that more general version here. Let $X_i$ be the vector of covariates associated to the $i^{th}$ test and $t$ a standard Normal random variable. The GRE model uses a probit link

$$\mathbb{P}(T_i=1|D, t, X_i)=\Phi(a_{iD}+b_{iD}t+c_D X_i) \quad (4)$$

where Φ() is the distribution function of the standard Normal variate, and then integrates over $t$ to obtain sensitivity and specificity conditional only on $X_i$. If the values of the covariates are fixed and absorbed into the $a_{iD}$ term, there are formulas for sensitivity and specificity,

$$\mathbb{P}(T_i{=}1|D{=}1){=}\Phi\left(\frac{a_{i1}}{\sqrt{1{+}b_{i1}^2}}\right) \quad and \quad \mathbb{P}(T_i{=}0|D{=}0){=}\Phi\left(\frac{-a_{i0}}{\sqrt{1{+}b_{i0}^2}}\right) \quad (5)$$

In a model with $n$ tests and no covariates, GRE requires the use of $4n$ parameters to describe test accuracy as opposed to the $2n$ parameters of the HW model. A restricted form of GRE common in the literature assumes equal variance of the random effect across all tests, reducing to $2n+2$ parameters for test accuracy [81-85]. A general framework combining covariance and random effects models as special cases was developed by Xu and Craig [86], and a generalization of random and linear mixed effects with covariates by Shih and Albert [87]. It has traditionally been assumed that all random effects are independent, but when subject-specific and rater-specific random effects are present using a single Gaussian random effect will result in biased estimates [88]. This bias can be minimized if a mixture distribution is used, but in practice, it is difficult to identify the correct distribution, as we discuss in Section 2.4. Unbiased choice of a best-fitting random effect structure can be performed using information criteria [89], which are discussed in greater detail in Section 4.1.

The GRE model results are robust against skewness in the random effect's distribution if the Normal random effect is replaced with a $t$-distribution [90]. The random factor $t$ and covariates $X$ allow sensitivity and specificity to vary between populations because of observed and unobserved factors, generalizing HW model's assumption of constant sensitivity and specificity. This assumes, however, that test dependence is the result of outcomes being drawn from the same distribution [50]. Moreover, it assumes that unobserved factors are equally likely to positively or negatively affect the probability of a positive test, while in practice, negative correlations between tests are virtually unknown. Consequently, GRE may be more suitable for comparisons between labs all using a single test than for modeling dependence between distinct diagnostic tests because differences in training or variation in the quality of lab equipment are plausibly explained as random effects. Random effects models using other distributions have been developed: a Dirichlet distribution effect for studying subject-specific accuracy rates [91] and a beta binomial effect for repeat testing dependence using a single test [2, 92].

### 2.3. Dependence: other structures

The first dependence model considered in the literature used a multiplicative parameter as a measure of dependence, $\mathbb{P}(T_1 = i \mid D = d, T_2 = 0) = \theta_d \mathbb{P}(T_1 = i \mid D = d, T_2 = 1)$, for $d, i = 0, 1$ [93]. If $\theta_d = 1$, then the tests are independent conditional on $D = d$. While there exists a change of parameters relating this model to the standard covariance model [94], since its publication, this parametrization has seen limited use [14, 20, 95, 96], as the dependence parameters of the covariance model (Equations 2 and 3) have a more intuitive interpretation.

The finite mixture model (FMM) assumes the existence of 'verified' subpopulations of healthy and diseased individuals whose disease statuses are correctly identified by all tests. This structure was suggested by data from a study of expression levels of a tumor suppressing protein [9] in which about half of subjects had expression levels in the $1^{st}$ or $99^{th}$ percentiles. FMM identifies these ranges of test values as giving verified disease status even though the test may have significant error in an intermediate range. While FMM is a dependence structure, it is better thought of as a model for identifying a breakdown in the assumption of constant test accuracy across populations. Specifically, FMM describes the case where test accuracy is constant after stratifying along verified status, with all tests having perfect accuracy in the verified subpopulation.

As related alternatives to the full covariance model, log-linear [41] or logit functions for the marginal distributions [97] can be used in order to capture higher order interactions. A Bayesian variation of the latter exists as well [98]. Unless many tests and populations are available, however, identifiability of these models requires that strong constraints be imposed on the interactions between some of the tests in order to have free parameters for the remaining higher order interactions. When only second-order effects are used, these models can be viewed as reparametrizations of the full covariance model.

The use of block covariance matrix designs has been generalized by the multiple latent variable model [83]. Tests are organized into types on the basis of the proxy variable for disease used. Dependence structures are then imposed separately upon the collection of tests associated with each proxy. The multiple latent variable model formalizes methods used in earlier work [17,92,99,100] and has since been applied to the diagnosis of tuberculosis in elephants [101].

### 2.4. Testing for dependence

The decision of whether to control for dependence between tests is crucial in choosing a model of accuracy without gold standard. Statistical tests exist to aid researchers in this choice. The first of these are the goodness-of-fit tests, such as $\chi^2$, $G^2$, and $CR$. Graphical methods have been proposed for correlation residual plots [7], biplot graphical displays [102], or through a log-odds ratio check [103]. The effectiveness of these methods in detecting correlations has been analyzed, finding that goodness-of-fit tests are better than both graphical methods [104]. In particular, log-odds ratio check suffers from very low detection rates. A statistical test for nonzero correlation under repeat testing and time-varying disease status has also been considered [19]. MLE of the conditional covariance [105] or kappa statistics [106] and their 95% confidence intervals have been proposed. None of the aforementioned methods has seen much use since the rise of Bayesian methods, as recent models tend to use information criteria to select a dependence structure or fully characterize the conditional dependence between tests.

There is also the informal test of 'practical significance', where a dependence parameter may have statistically significant difference from zero without impacting model estimates [26]. Practical significance is not a viable method in the absence of exhaustive study of dependence structures because model estimates can vary widely between some structures while remaining similar for others, leaving the researcher at a loss as to whether observed

agreement is a sign of accurate estimates. For example, consider Handelman's dentistry data as analyzed by Albert and Dodd in which estimates from independent and FMM are highly similar, while GRE differs from both [82].

It should be noted that dependence testing will determine only for which pairs of tests, if any, there is a significant level of dependence. These tests do not identify the correct dependence structure itself. While significant asymptotic bias in prevalence estimation can occur under a misspecified dependence structure [9, 107], there are conflicting results in the literature showing that close agreement can exist across multiple structures despite varying goodness-of-fit values [84, 85, 97]. Even if the true model is among those under consideration, six tests are not sufficient to distinguish it among a finite collection of alternatives [82]. On the other hand, partial verification models (discussed in Section 4.3) are robust against bias from dependence structure misspecification, with bias decreasing as the proportion of verification increases [107, 108].

Perhaps for this reason, it has become standard practice to fix a particular dependence structure in advance and determine only which tests have dependence that must be included under the model. This is not recommended practice. A better strategy might begin with a careful study of the characteristics of the sample population and tests, followed by tailoring the dependence structure to any unique characteristics identified [109]. If a choice is made without comparison between multiple distinct structures, we recommend the use of a nested sequence of models that allows for higher order interactions, such as those used by Spencer [41] or Berkvens *et al*. [74]. Model selection methods, as discussed in Section 4.1, can then be used to select the appropriate combination of interactions to include.

### 2.5. Inclusion of covariates

Covariates may be included to improve model fit under any of the existing dependence structures by use of a link function, $g$, giving a relationship between an expected value and the model's covariate vector $X$,

$$\mathbb{P}(T_i=1|X, D=d)=g(\gamma X+\sigma_d d) \quad (6)$$

The most frequent choices of link are the logit, inverse Normal, and probit. If a standard Normal random variable is included in the covariate vector, then we recover the random effect model of Equation 4.

The inclusion of covariates may reduce dependence between tests if it arises from characteristics of subpopulations defined by these covariates. In addition, this allows disease prevalence and test sensitivity and specificity to vary between the subpopulations defined by each choice of covariate vector. If one or more tests have accuracy rates that vary between subpopulations defined by known covariates, then the standard assumption of constant test accuracy between populations can be accounted for by the inclusion of these covariates into the model. In Section 2.6, we consider what can be carried out when there is suspected variation in accuracy between populations but no covariates can be identified that suitably account for this variation.

Covariates may be used for partitioning into subpopulations in order to increase the number of degrees of freedom available [110-114] or to better predict outcomes [35, 115-118]. A common stratification method is to use indicator variables for high-risk subpopulation(s) with the general population or a confirmed uninfected subpopulation as the comparison group [119]. If such a population is not identified in advance, grouping can be performed using ranges of values of an ordinal or continuous-valued test [120], through observed covariates in multi-stage studies [121, 122], or, if necessary, using a correlate of disease that is not observed until after data are collected [34].

Stratification is not always possible even when subpopulations are identified because it is assumed that all strata have distinct prevalences. Model estimates tend to be biased when applied to data with modest but nonzero differences in prevalence. This bias persists even as sample size increases. For prevalence differences of less than 20%, an increase in sample size from 200 to 2400 had minimal impact on confidence interval width in the HW model [50]. Experience has shown that the HW model returns implausible parameter estimates or confidence intervals that cover the entire [0, 1] interval when population prevalences differ by less than 15%. If covariates are used only to model finer differences between the subpopulations of a sample defined by each covariate vector, and not for stratifying the sample into distinct populations each with their own unique disease prevalence parameter, then these issues with stratification will not arise. In this case, there is also no increase in the available degrees of freedom, and so the needs of the situation must be considered when deciding how to incorporate covariates into the model.

Choosing appropriate subpopulations is particularly challenging in the study of rare diseases. Identifiable high-risk groups may still have only moderately elevated disease prevalence compared with that of the general population. In addition, it may be that any variables that can be used to define strata with sufficiently distinct prevalence are achieving this by identifying population subgroups in which one or more tests have improved sensitivity. To properly account for this, the model should not assume constant accuracy across all populations. We discuss methods for handling this issue in the following section, but we first present an original example using survey data for how these models can be inappropriate for rare conditions without careful stratification and attention to model assumptions.

For example, the authors attempted to stratify the 1996 panel of the Survey of Income and Program Participation into low-risk and high-risk groups for disability by the presence of activities of daily living restrictions, using self-reported disability and disability beneficiary status as tests of the unobserved true state (Table II). The analytic solution to the HW model for these data claims moderate sensitivity, $\alpha_1 = 0.798$, $\alpha_2 = 0.622$, and that false positives are very rare, $\beta_1 = 1.00$, $\beta_2 = 0.980$, with all confidence interval widths less than 0.025. While the poor sensitivity of beneficiary status can be explained by not having restricted the sample to benefit applicants, the estimated specificities are suspiciously high for both tests. It is likely that the model has broken down because of the similar prevalences and variation in test accuracy between the two groups introduced by stratifying along covariates correlated with disability.

### 2.6. Nonconstant accuracy

The HW model assumes that test accuracy does not vary between populations or equivalently that it is independent of prevalence. In clinical settings when this variation occurs, it is known as spectrum bias. Few accuracy models have considered alternatives to the HW model assumption because of the high parameter cost of relaxing it; in its absence, $2g$ parameters are needed for accuracy of each test when $g$ populations are used. This results in a counterintuitive scenario where sampling from a larger number of populations requires more tests to achieve identifiability. In particular, three independent tests are necessary with one population, but five are required with two populations. In this section, we illustrate the HW model's lack of robustness when this assumption is violated and discuss alternative models from the literature.

Robustness of the HW model against varying accuracy rates has been tested using simulated data [49, 109, 123, 124]. The first of these, whose results presented in Table III, used two independent and identical tests with population-varying sensitivities. Disease prevalence was low, and there was a small difference of 3% between populations. True values near 0 or 1 can cause difficulties for MLE, and we have already raised concerns about the use of population strata with near-identical prevalence. Consequently, the published results are compared with medians computed from MCMC estimation in WinBUGS using noninformative priors. As the true and estimated values were equal for both tests, only results for one test are shown.

The MLE is exhibiting stereotypical 'corner' behavior, where estimates of one or more parameters become pinned at an extreme value of 0 or 1 and estimated confidence intervals, while not reported in the paper, were likely either undefined or covered the entire [0,1] interval. The comparison to make in these situations is not between the true and MLE sensitivity values, but rather how much the true values must change before estimates are strictly within the interval (0,1). Except for modest overestimation of prevalence in the first population of both data sets under MCMC, nonconstant sensitivity is only biasing sensitivity estimates for either method. In the second data set, a best worst-case scenario is reached for MCMC, where estimated sensitivity approximately equals the mean of the population-specific true values.

If expert opinion is available for elucidation of informative priors, as covered in Section 3.2, a more appropriate model when faced with nonconstant accuracy rates may be to use only one of the populations, as in Joseph *et al*. [125]. As an example, we used the first population of the first data set, with a modestly informative Beta(1,19) prior for prevalence, Beta(9.5,0.5) for both test sensitivities, and Beta(8.5,1.5) priors for both specificities. This resulted in significantly better median and narrower 95% credible interval estimates for prevalence $\pi_1 = 0.049$ (0.035, 0.075), sensitivity $\alpha = 0.975$ (0.757, 0.999), and specificity $\beta = 0.850$ (0.831, 0.870), despite the model's nonidentifiability.

Some characteristics of samples that can result in population-varying accuracy rates have been identified in the literature. When the test takes a continuum of values, unobserved confounders unrelated to the disease can influence results, resulting in population-dependent accuracy rates if the confounders are observed [126]. In the standard accuracy models,

disease status is given by a binary variable, but some combinations of population and test can be better fit by a model that includes levels of disease severity or stages of progression [127-130]. Intermediate disease stages, such as from age-dependent testing [111, 131] or subclinical infections and infection before immune response [27], will lead to variation in test accuracy. This effect of intermediate disease stages is pronounced for continuous tests, with the size of the effect depending on the threshold chosen to distinguish positive and negative test outcomes [38]. If the threshold is set low to minimize the effect of intermediate stages, the test's specificity will be reduced. This issue is discussed in greater detail in Section 4.2. In general, when considering the inclusion of a new model characteristic, such as ordinal disease status, it is best practice to construct at least two models for its presence and exclusion and to select the better fitting one using statistical methods for model selection [132].

Variation in test accuracy is possible when multiple raters or labs have similar but not identical levels of training or equipment quality and the researcher wants to estimate their average accuracy. If all samples are tested by each rater, then a repeat testing model can be used with each rater representing an application of a common test. Otherwise, if similarities between raters allow them to be viewed as draws from a single population, then the use of a random effects model (Section 2.2) for differences in rater accuracies is justified [92, 100]. Random effects models can also be used when variation in accuracy is suspected between individuals rather than raters because of observed [5,41] or unobserved [91,133] characteristics. When variation is suspected between individuals and between raters, the use of two independent random effects is warranted [109]. While it was not developed to address this situation, the multiple latent variable model [83] may be of use when tests have a subjective aspect [13, 47], such as with survey questions where certain responses can carry social stigma.

### 2.7. Repeat testing

Repeat testing models are an effective means of increasing the number of degrees of freedom available in the absence of additional distinct tests or populations. These models assume that one or more tests are applied to the same population multiple times. Repeat testing models are sometimes appropriate even if, strictly speaking, distinct tests are performed. Multiple clinicians examining the same X-ray can be modeled as repeat testing if we assume each clinician has a similar level of training. Repeated tests are not independent and require careful study in choosing an appropriate dependence structure.

A wide variety of repeat testing models exist in the literature. Models have been constructed using a distribution moments method [134], through application of Bayes' theorem to estimate odds ratios [15], and as draws from a binomial distribution [135], although none of these accounts for repeat testing dependence. More recently, a mixture model with repeat test dependence was constructed, although it still assumed that distinct tests were independent to preserve identifiability [44].

Time-dependent models consider repeat testing under the condition that disease status has a nontrivial probability of changing between each testing time period. These models were introduced to the accuracy literature through an application of model selection methods,

discussed in Section 4.1, to find best-fit generalized linear or quadratic models for longitudinal data [136]. Other linear models studied include the comparison of a naive moving average model with an autoregressive time-lagged process [137]. Trigonometric and quadratic models were considered by Billiouw *et al*. for modeling seasonal parasitic infections [138]. Both frequentist and Bayesian versions of identifiable Markov transition models were then published [139, 140] although the former notes correctly that the Markov assumption will fail if an observed change in status would lead to differences in treatment. A two-period logit model, with time interaction factor and independent tests, [141] has been extended greatly to allow for dependent tests and time-increasing sensitivity post-infection [142, 143]. Time-varying sensitivity has also been used with explanatory covariates for HIV testing [144]. Norris *et al*. model movement between three disease states using a reversible jump MCMC algorithm [27]. Explanatory and response covariates are chosen to inform a change-point process [31,145], or Bayesian version of the HW model where data are treated at the testing rather than subject level [28]. Lastly, a two-component model with independent tests and covariate-dependent sensitivity and specificity has been proposed [146], with variations that introduce biomarker serial correlation or delayed response after infection.

Modeling time dependence requires the use of many parameters, as repeat testing, between-test, and time-dependent status factors may all be involved. Consequently, identifiability of these models is not easily achieved. If repeat testing is used specifically to achieve identifiability so that studying time-varying status is not an explicit goal, we recommend that multiple raters of each subject be used within a short time window. This will reduce the number of necessary parameters in the model, a technique that has been used with some success [23, 92, 107].

## 3. Bayesian models

Since the publication of the first Bayesian models of accuracy without gold standard [3, 147, 148], Bayesian methods have increasingly dominated this field. Despite their increasing popularity, the terminology and methods used can create barriers for potential users. It therefore appears useful to provide a framework through which Bayesian models can be interpreted for frequentist and nonspecialist audiences.

Bayesian statistics is a theory for interweaving new and existing data by taking both sources into account. Parameters are no longer unobserved constants but instead are given by random variables; our uncertainty about the parameter is captured by the random variable's distribution. This view comes from an interpretation of Bayes' theorem

$$\mathbb{P}(\theta|X)=\frac{\mathbb{P}(X|\theta)\mathbb{P}(\theta)}{\mathbb{P}(X)} \quad (7)$$

in which $X$ represents sample data and $\theta$ the parameters of a model. The prior distribution, $\mathbb{P}(\theta)$, represents our belief in the plausibility of parameter values before conducting the experiment. $\mathbb{P}(X \mid \theta)$ is the model's likelihood, and $\mathbb{P}(\theta \mid X)$ is the posterior distribution, or result of weighting our prior assumptions by the likelihood of their conforming to sampled

data. Under reasonable conditions, as sample sizes increase, the variance of the posterior will decay to zero and results in asymptotic agreement with maximum likelihood estimates.

### 3.1. Sampling from the posterior: MCMC

The likelihood equations of all but the simplest accuracy models result in posterior distributions with a complicated relationship to the prior. Obtaining the posterior requires the evaluation of high-dimensional integrals, which can be very difficult computationally. MCMC attempts to sidestep the curse of dimensionality through the use of probabilistic sampling from the posterior distribution. Monte Carlo integration is a numerical integration method using random numbers from a probability distribution to choose points for evaluating the integrand rather than through a regular grid. When sampling from a complex distribution for Monte Carlo integration, algorithms such as Metropolis–Hastings are used to preferentially sample from regions of high density of this distribution. They compare the density at the previous sample point with a randomly chosen real number. Because only the previous sample is used in this comparison, the result is a Markov chain of samples used to perform the Monte Carlo integration.

The most common MCMC algorithm in Bayesian statistics is the Gibbs sampler. This algorithm samples from the conditional distributions of parameters when their joint distributions are unknown. It is also efficient at converging to the posterior distribution outside of pathological cases such as when the joint distribution is multimodal with regions of very low probability separating modes. In the worst cases, the sampler will still converge, but the number of samples required can be computationally unfeasible. It is also an unfortunate truth that the choice of posterior sample size is more art than science. Knowing that a chain will eventually converge as the sample size increases to infinity does not mean that a theoretically justified necessary posterior sample size exists for a given model. With the computational power of modern computers, obtaining very large posterior sample sizes from Bayesian accuracy models is practical and therefore strongly recommended. The reader is referred to Smith and Roberts [149] for an introduction to MCMC and Gibbs sampling, and Roberts and Smith [150] for convergence properties. Practical issues with the use of priors, identifying autocorrelation and evaluating chain convergence in WinBUGS are discussed in Toft *et al*. [151].

While sufficient levels of convergence cannot be guaranteed in advance, there are additional procedures that can be used to give confidence that the results obtained are valid. Because the sampling process is given by a Markov chain, there is autocorrelation between adjacent samples that can slow the rate of convergence. This can be overcome by thinning the sample, meaning that only every $n^{th}$ sample for some $n$ is preserved when constructing the sampling distribution so that the remainder are approximately independent samples from the posterior. Thinning by a factor of $n$ to achieve a final sample of 50,000 requires $50,000n$ unthinned samples to be computed. It has also been shown that using the full (unthinned) chain can often result in better estimates of the posterior, even for the tails of the distribution and despite the effect of autocorrelation [152]. For small samples, thinning should result in better estimates and makes precision estimation easier as there is less bias in assuming samples to be independent. We strongly recommend using multiple chains in order to assess

convergence to a unique distribution for nonidentifiable models and to guard against label-flipping in particular. This is analogous to running MLE from multiple sets of initial conditions to attempt to determine if the results are a local or a global maximum. If all chains give comparable results, their samples can be pooled. Standard practice recommends that an unspecified number among the first samples taken be dropped from estimation, called the burn-in. This is carried out under the assumption that the Markov chain will have converged to its steady state after the burn-in, which means it has entered a region of high density in the posterior. While starting the chain from a region of high density will improve estimation, there is no justification available for why a particular burn-in length would accomplish this and beginning estimation from a region of low density will only slow convergence rates. Both problems are better solved by choosing 'large' samples rather than attempting to finesse these difficulties. If the user remains suspicious of early samples, there exist diagnostic tests for stationarity of which one of the simplest is Geweke's comparing the mean of the first 10% of samples with the mean of the last 50% [153].

### 3.2. Choice of prior

There are two approaches to prior distributions in the accuracy literature, each appropriate to a different setting. In the first, noninformative priors are used so that posteriors are determined by the data rather than assumptions of the prior, while in the second, highly informative priors are used to effectively increase the number of degrees of freedom by restricting the likely ranges of some or all parameters of the model. A prior is noninformative if it encodes a lack of information about possible parameter values, else it is informative.

Noninformative priors are preferred for smaller sample sizes when there is no source of expert opinion to use for choice of priors [114]. Small sample sizes result in posteriors that are significantly affected by the choice of prior, as the independence of posterior from choice of prior is only guaranteed asymptotically under reasonable conditions by the Bernstein–von Mises theorem. In some cases, expert opinion exists for a related but not identical population, leading researchers to use a downweighted prior to account for doubts of the direct applicability of this knowledge [3].

Informative priors are frequently used when expert opinion is available. This is particularly useful as a means of imposing weak, probabilistic constraints on parameters in order to compensate for a lack of model identifiability [48, 154-158]. Probabilistic constraints are considered preferable to questionable deterministic constraints for achieving identifiability because sharp constraints can bias other model parameters [22,63,159]. Even for identifiable models, mildly informative priors are useful for preventing label flipping.

Beta($a$, $b$) priors are standard in the literature for prevalence, sensitivity, and specificity with $a$ representing a count of successes and $b$ of failures. The Dirichlet distribution generalizes the Beta to the case of joint distributions, giving it appeal for nonparametric modeling of test dependence [8, 160, 161], at the cost of being difficult to elicit properly [162]. If prior information exists for diseased and healthy populations, then a mixture distribution prior can capture this information [163]. Researchers have also used counts from existing studies [4, 164] or estimates from published literature [165, 166], but in these cases, it is recommended

that priors be downweighted so that differences between study populations or estimation errors do not propagate.

As the literature has developed, best practices for elucidation of priors have been identified. Particularly in the case of beta priors, expert opinion is easier to obtain for prior mode and the 5% or 95% quantiles, which are then used to derive the hyperparameters for the prior [27, 48, 167]. When either sensitivity or specificity is known to be very close to 100%, we recommend that a highly informative Beta prior or a uniform prior restricted to a range such as 90–100% be used [45] instead of fixing exact values [33], as fixing values prevents learning if the estimate is incorrect. Finally, uniform priors for one or two parameters on large subintervals of [0, 1], as used in the noninformative case given in Table IV, are effective for preventing label flipping.

### 3.3. Nonidentifiable models

As the accuracy without gold standard literature has grown, researchers have attempted to weaken the restrictive assumptions imposed by the early models. This has led to a tension between the number of parameters necessary to account for test dependence and, possibly, time-varying status, and the availability of sufficient tests and populations with distinct prevalences to obtain an identifiable model. Consequently, the use of nonidentifiable Bayesian models has become increasingly common. Recall that a model is necessarily nonidentifiable if its number of parameters exceeds its degrees of freedom, while if these are equal, it is called weakly identifiable but may not in fact be identifiable. In this section, the merits of using nonidentifiable models and the extent to which informative priors can compensate for nonidentifiability are considered.

Researchers have in the past been wary of nonidentifiable models for several reasons: their posterior distributions will not asymptotically converge [168, 169], they may produce posterior estimates worse than that of the prior [170], or they may have weak estimability because of wide credible intervals [74, 171]. Despite these weaknesses, nonidentifiable models are not without merit: information in the data set leads to partial updating of posteriors [125], a nonidentifiable model may still be valid [103], and stratifying the sample to obtain identifiability may not improve estimates [172]. The posteriors of an identifiable model may also fail to converge as sample size increases [173], and so failure to converge should not by itself discourage us from considering nonidentifiable models. In some cases, partial updating of posteriors results in a subset of parameters converging fully, while the rest remain uninformative [174]. In general, indirect learning may occur for nonidentifiable parameters as a result of learning in the identifiable part of the model when such a change of variables does not exist [172]. Measures of Bayesian learning are available and include the difference in precision, the reciprocal of variance, between posterior and prior distributions for Normal distributions, or more generally the Kullback–Leibler divergence of posterior and prior distributions [175]. Consequently, the question of whether or not a nonidentifiable model is appropriate should be answered on a case by case basis after taking into account the characteristics of the data.

When a nonidentifiable model is deemed appropriate, every effort should be made to obtain informative priors for model parameters as highly informative priors serve to narrow the

parameter ranges searched by MCMC. The strength of the constraint imposed by choice of priors is measured by a summand of the deviance information criterion (DIC) called the effective number of parameters and denoted $p_D$ [176]. The relationship between informative priors and $p_D$ can be observed by their relation to the width of credible intervals for the model's parameters, as illustrated in Table IV. The data for this table were taken from a myocardial infarction data set [113], after collapsing across weight and gender covariates and treating subsequent events as a third imperfect test. A nonidentifiable Bayesian covariance model was used with three tests and one population, with subsequent events assumed independent of the other two tests, for a total of 7 degrees of freedom and nine parameters.

Bayesian inference using Gibbs sampling was performed in WinBUGS [177], and in each case, four chains with a posterior sample size of 100,000 were obtained without burn-in or thinning and with randomly generated initial values. Label flipping was prevented in the uninformative case by restriction of the prevalence prior: $\pi \sim$ Unif[0, 0.4]. In the moderately informative case, Beta($a$, $b$) priors on $\pi$, $\alpha$, $\beta$ are chosen for tests 1 and 2 with means in close agreement with the estimates from the noninformative case, and such that $14 \leq a + b \leq 18$:

$$\begin{array}{cccc} \pi \sim Beta(2.5, 15.5) & \alpha_1 \sim Beta(7, 9) & \beta_1 \sim Beta(14, 1.3) & \\ \alpha_2 \sim Beta(13, 2) & \beta_2 \sim Beta(12, 4.5) & \alpha_3 \sim Unif(0, 1) & \beta_3 \sim Unif(0, 1) \end{array}$$

Informative priors were not used for the third test as subsequent events were not used to predict disease in the paper. Highly informative priors were obtained by multiplying the counts in all beta priors above by 3. Covariance priors were chosen using uniform distributions over the possible range between perfect negative and positive correlations, although similar results are obtained if the standard practice of restricting to nonnegative correlation is followed [178]. Note that all choices of prior are made solely to demonstrate the effect of choice of prior and not to claim that this example illustrates a valid method for elicitation of priors.

The credible intervals for all parameters show consistent narrowing between the uninformative and highly informative cases. Credible interval narrowing is not, however, distributed evenly across all parameters. In particular, with the exception of $\beta_1$ and $\beta_3$, all parameters show a decrease in credible interval width of at least 0.158, while for these two, a reduction of not more than 0.033 was observed. Finally, we note that the property of indirect learning introduced by Gustafson [172] is unambiguously active in the case of $\alpha_3$, whose credible interval narrowed dramatically even though its prior was uninformative and unchanged in all cases.

## 4. Special topics

### 4.1. Finding the best fit: model selection

Researchers may find that there is no a priori best model for their data, and so they need an unbiased method for assessing the tradeoffs between different sets of modeling assumptions. These tradeoffs exist in two types: first, as in statistical modeling generally, a decision of which covariates to include to improve model fit while avoiding overfitting, and second, a

decision of how far to weaken the assumptions of the HW model to reduce bias while avoiding nonidentifiability or limiting its negative effects. In other words, this is a problem of model selection.

The problem of model selection is one of balancing efforts to reduce model variance and model bias. Increasing the number of parameters in a model can reduce modeling bias but at the cost of increased variability in estimates due to sampling variation. Information criteria seek to balance these competing objectives by penalizing a goodness-of-fit measure with a measure of complexity. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) both penalize twice the model's log-likelihood $\ell(M)$ with a multiple of its number of parameters dim($M$),

$$AIC(M) = -2\ell(M) + 2\,dim(M) \quad BIC(M) = -2\ell(M) + log(n)\,dim(M) \quad (8)$$

where $n$ is the sample size. Despite its name, BIC does not require a Bayesian model. The DIC, $DIC(M) = D(\bar{\theta}) + p_D$, however, is truly a 'Bayesian' criterion, where $D(\bar{\theta})$ is the Bayesian deviance at the posterior mean of the parameters [176]. Smaller values of each criterion correspond to models that better fit the data without overfitting, although the reported values have no meaning in themselves as they are not on an absolute scale, and values from different criteria are not comparable.

In practice, it is recommended to compute both AIC and BIC for frequentist models. AIC and BIC have different asymptotic behaviors: AIC will choose the model that minimizes mean squared error, while BIC chooses the model with minimal Kullback–Leibler distance to the true model [179]. While all three criteria can be appropriate for comparing Bayesian models, DIC is recommended if informative priors are used in any of the models so that the number of parameters is penalized properly to allow for an informed decision of the tradeoffs between possible prior misspecification and reduction in $p_D$. DIC has the added advantage of being directly estimable from the results of MCMC, while AIC and BIC require a follow-up computation of the log-likelihood. For additional details on the definition and construction of each criterion and their asymptotic behavior, refer to Claeskens and Hjort [179].

Information criteria have found use in the accuracy literature primarily as tool for choosing between multiple dependence structures. In a typical application, a subset of all possible covariance structures between a collection of tests is decided in advance, the model resulting from each choice is determined, and the criteria then used to select candidates [18,34,41,76,85,132,180,181]. Other applications include analysis of prior information as in Table IV [74, 75], fine-tuning time-dependence effects [139, 146], or choosing between nonignorable missing data models [182].

It can occur that the smallest value of AIC, BIC, or DIC is proportionately nearly identical to that of one or more other models. On Table IV, the noninformative prior model has DIC less than 3% larger than with highly informative priors. In this situation, all models with criterion values approximately equal to the minimum can be accepted and the conclusions obtained from each discussed [83], or a composite constructed using Bayesian model averaging. Bayesian model averaging uses posterior probability-weighted averages of model

estimates to arrive at composite predictions [183]. This method is computationally feasible for large collections of models if the collection is first winnowed using information criteria. For smaller classes of models, Bayesian model averaging can be used as an alternative to information criteria without increasing estimation bias [108, 171, 184].

### 4.2. Methods for nonbinary tests

All of the methods considered thus far assume that test outcomes are binary. In practice, many of the commonly used tests, such as enzyme-linked immunosorbent assays and antigen level tests (e.g., prostate-specific antigen), are nonbinary. There are a variety of methods in use for this situation, which we will consider in order of their complexity.

To obtain a binary outcome, a threshold for positivity can be imposed (we will assume that larger test values are associated with disease). As shown in Figure 1, the choice of threshold determines the false positive and false negative rates, where increasing the threshold decreases the false positive rate at a cost of increasing the rate of false negatives.

If true status is also ordinal, a series of thresholds are imposed to first dichotomize true status and then threshold the test [129, 130]. Choosing an optimal threshold will depend upon the test's use: screening in or out, minimizing total error, or confirmatory testing after screening. For example, screening in requires a highly sensitive test so that few true positives are misclassified while still preserving moderate specificity to reduce the burden on the confirmatory test.

The optimal choice of threshold can be modeled by a cost function [185, 186]. Geisser [185] did not explicitly consider thresholds, but their method can be applied to compare the merits of two distinct choices of threshold for the same test. The methods of both papers require, however, that the costs of true positive, true negative, false positive, and false negative test results be understood. Carefully addressing this issue is often difficult, as a survey of the research on quality-adjusted life years or rationing of medical care will show. An alternative is to characterize the distributions of healthy $G_1(x)$ and diseased $G_2(x)$ test scores from the mixture distribution of the whole sample $F(x) = (1 - \pi)G_1(x) + \pi G_2(x)$ [187]. This approach is effective if the distributions are known to belong to particular families or if very large samples are available.

The ROC curves are a method of comparing sensitivity and specificity across a range of possible thresholds rather than fixing a threshold in advance. To be precise, an ROC curve graphs sensitivity as a function of the false positive rate (1 - specificity) and can be determined by estimating sensitivity and specificity at a large number of thresholds, as illustrated in Figure 2.

The area under the ROC curve (AUC) is then a measure of the average accuracy of the test; a perfect test has an AUC of 1, while a coin flip has AUC of 0.5. While AUC is the average accuracy, a fixed threshold must be set in practice when using a test, so we do not find it to be the most efficient measure of test value. Optimal choice of this threshold will instead depend upon the relative cost, whether monetary or through health outcomes, of false positive and false negative tests. If both error types are treated equally, the optimal threshold

can be determined by the Youden index or minimization of the distance from the ROC curve to the point (0,1), which corresponds to perfect sensitivity and specificity. The Youden index maximizes the vertical distance between the ROC curve and the line $y = x$, the line on which false positive rate equals true positive rate. Equivalently, the Youden index equals $\max(\alpha + \beta) - 1$ over all possible thresholds.

More generally, if we have an estimate of the relative cost of a false positive $1 - \beta$ compared with a false negative $1 - \alpha$, labeled by $c$, we can minimize the resulting cost function $C_c(\beta)$ or weighted distance $D_c(\beta)$

$$C_c(\beta) = (1 - \alpha) + c(1 - \beta) \quad (9)$$

$$D_c(\beta) = \sqrt{(1 - \alpha)^2 + c(1 - \beta)^2} \quad (10)$$

This method gives rise to simple conditions upon the slope $d\alpha/d\beta$ of the ROC curve at the maximum:

$$min\, C_c(\beta) \Rightarrow \frac{d\alpha}{d\beta} = c \quad (11)$$

$$min\, D_c(\beta) \Rightarrow \frac{d\alpha}{d\beta} = c\frac{1 - \beta}{1 - \alpha} \quad (12)$$

Concavity of ROC curves implies that, with either condition, as $c$ increases, $\alpha$ must decrease and $\beta$ increase. This is to be expected as larger values of $c$ correspond to higher costs from false positives and therefore a need for greater specificity. Utility of medical procedures has been considered as a measure of cost in the context of ROC curve analysis [188].

An early application in this literature used summary ROC (SROC) curves to perform a meta-analysis of cervical cancer data with a frequentist parametric model [189]. Since then, development has been rapid as new dependence structures, semiparametric and nonparametric models of test scores, and Bayesian techniques were applied to ROC curve analysis. In particular, ROC curves have been determined using covariates and a parametric model of test outcomes for diseased individuals with covariates [29], semi-parametric models of test outcomes both with [190] and without covariates [191], and nonparametric models of test outcomes with independent [30] or correlated tests [23, 24]. The effect of nonignorable verification bias resulting from partial verification designs, covered in Section 4.3, has been well studied [192-196], and AUC estimated in a frequentist model under verification bias [197]. Bayesian meta-analysis of tuberculosis testing using summary ROC curves in a model with dependent tests has also been performed [198]. When a test has upper or lower detection limits, such as polymerase chain reaction for detecting viral loads, its extreme results must be treated as censored data to avoid bias and the likelihood function adjusted to compensate during estimation of the ROC curve [199]. These models are efficient in that they estimate a large, finite number of points of an ROC curve at once. It is possible, however, to create an ROC curve by applying any of the accuracy models

discussed previously in this review at each of a number of test thresholds, one at a time [120, 181].

In some testing designs, a combination of tests is used to make a single prediction, often by defining their combination to be either the minimum or maximum observed after thresholding. These extremes of test combination may not be ideal, and so there is interest in finding an optimal combination. Under the assumption that all tests have normally distributed values in each of the healthy and diseased populations, there exist optimal combinations of the tests to maximize AUC [200] or AUC restricted to a range of specificity [201]. A more general method that allows for inclusion of correlated tests through observed covariates exists, although optimality of its combinations was not guaranteed [202]. Finally, Jin and Lu have proposed the use of noninferiority testing to determine if a given combination of tests has AUC no worse than a theoretical combination optimizing the likelihood ratio [203].

The inclusion of multiple latent classes for intermediate stages of disease, or levels of disease severity, was recommended early in the literature to sharpen the distinction between extremes of health status [204, 205] or as an alternative to using a test dependence structure [206]. Thresholds are chosen to maximize the separation between extreme stages. Both independent [127,207] and correlated [208] tests have been considered in this framework. When three stages are used, these models are similar to the FMM, discussed in Section 2.3, except that residual error is allowed in extreme stages after a threshold is chosen. This use of FMMs was previously discussed in relation to nonconstant accuracy rates in Section 2.6. The choice of dependence structure must be carefully considered as in the case of binary tests, for incorrect specification may result in biased estimates of AUC [209].

In fact, with the use of semiparametric and nonparametric models, thresholding can be avoided entirely. A Dirichlet process model for continuous test outputs has been used [210, 211] to allow data to be drawn from any distribution sufficiently close to a fixed baseline. The process is controlled by the choice of baseline distribution and a precision parameter, where increasing the precision narrows the divergence possible between baseline and sampling distributions. As precision increases to infinity, the model reduces to a parametric model such as the FMMs of Uebersax and Grove [212]. The flexibility of Dirichlet processes can unfortunately result in convergence difficulties when practical sample sizes are used even though their asymptotic convergence properties are well established, for example, Ghosal and van der Vaart [213]. Kernel density estimation methods were used in the past for nonparametric ROC estimation [214, 215] but have received less attention of late.

### 4.3. Partially verified data

It is not uncommon for highly accurate (confirmatory) tests to exist but to be expensive, time-consuming, invasive, and inappropriate for use on the general population. This has lead to interest in finding optimal testing designs [216-221], including pooled samples for rare disease testing and the use of screening tests to determine if confirmatory testing is necessary [12, 222]. These testing designs are efficient but result in only partial verification of disease status within the sample, as individuals testing negative on one or more screening

tests do not have their disease status verified. The missing data from partial verification designs are nonignorable, and so care must be taken when drawing conclusions about test accuracy. Even if the confirmatory test is a gold standard, naive estimates in this design are biased because of the failure to verify the screening test's negative results [10]. The most common design, used in cancer screening and HIV testing, is illustrated in Figure 3

Significant attention has been given to methods for addressing verification bias. Imputation estimators have found use [43, 121, 223, 224] for missing at random (MAR) covariates and are effective if confirmatory testing is carried out completely at random [21, 225]. These methods apply more generally to nonignorable missing data, where verification is given by a parametric model as a function of latent disease and observed covariates independent of disease status using independent tests [11, 193, 196, 226-228]:

$$logit\,\mathbb{P}(V=1|X,D)=\gamma X+\delta D \quad (13)$$

If $\delta = 0$, the verification mechanism is MAR, else there is nonignorable verification bias. In particular, when verification depends upon prior test results, the MAR hypothesis will be invalid unless the prior test is not more accurate than a coin flip. Positive and negative predictive values, however, have been shown to be unbiased for partially verified data under the assumption of independent tests [229]. The probability of verification, given the test score and observed covariates, is called the propensity score and captures the key information about verification in a single value. Propensity scores were used by He and McDermott to stratify the verified sample in order to perform numerical integration of a test's PPV and NPV [230].

As dependence structures and other methods were developed in the literature, applications were made to partial verification designs. Expert knowledge can be incorporated, as in Section 3.2, into a parametric Bayesian model to improve estimation [167]. Dependence structures have been introduced with parametric models for partial verification designs using the Bayesian covariance [14, 18, 95, 162], frequentist random effects and FMM structures [107, 231], or Bayesian random effects [232] models. Partial verification models have been applied to data in which each population receives a subset of a large number of available screening tests [16] and to cost-effective testing for rare diseases through the use of pooled sample testing designs [4, 6, 36, 118, 163]. Tang *et al*. examined the effectiveness of a number of maximum likelihood-derived statistics for prevalence estimation under partial verification using simulated data [233]. When verification is independent of disease status, a verification rate-weighted ratio of verified true positive rates has also been proposed for comparing two tests [234] as well as formulas for necessary sample size to detect a difference from unity in this ratio [235]. This ratio method does not, however, take into account any information from nonverified individuals as it avoids estimation of likelihood functions.

Nonparametric partial verification models have not yet received much attention in the literature. A nonidentifiable semiparametric method was proposed by Pennello [182] to estimate PPV and NPV under the assumption that screen positives have a binomial distribution, and that the triplet of verified negatives, verified positives, and unverified

individuals form a multinomial distribution. Simplifying assumptions can render the model identifiable or even reintroduce the MAR hypothesis. The test ignorance region was proposed by Kosinski and Barnhart [236] to describe all possible accuracy estimates independent of a verification mechanism. This is determined by considering the ranges of sensitivity and specificity possible if the $u_i$ unverified individuals with screening test score $T = i$ had their status verified as positive $u_{iD}$ or negative $u_i - u_{iD}$. The region is then given by the graph of sensitivity against specificity across all combinations of $(u_{1D}, u_{0D})$ in the square $[0, u_1] \times [0, u_0]$. This method allows nonparametric modeling of the verification mechanism and allows one to visually assess the extent to which estimates are robust to the assumptions of a particular verification model.

There is still significant room for development of partial verification screening designs and addressing the resulting estimation issues for test accuracy without gold standard. Devising optimal screening designs and analysis of the resulting partially verified data is an old problem [237] that has not been fully explored. Semiparametric and nonparametric models of verification are underutilized, and the value of the test ignorance region has only begun to be assessed [238]. Models that consider the possibility of error even with a so-called gold standard test have also only begun to be explored. On this last point, while it is almost universally assumed that verification is perfect, it is not strictly necessary to do so, and in practice, it may be preferable to avoid this assumption [78, 239]. For example, biopsy is accepted as a gold standard for cancer diagnosis, but incorrect results are possible if the sample is taken from the wrong site, if not all abnormal sites are biopsied, or if treated as an ordinal test to study rates of inaccurate cancer grading. The low rates of error resulting from imperfect application of gold standards can be modeled by strongly informative priors for sensitivity and specificity of the confirmatory test, leading to a minimal increase in the number of effective parameters (Section 4.1) and therefore not significantly impacting precision of model estimates.

### 4.4. Sample size estimation

When two binary tests are applied to a fully verified population, there exist multiple methods for determining sufficient sample sizes in a power analysis, such as McNemar's test. These estimates are, however, highly biased for incomplete verification designs [240]. In addition, infinite sample sizes may be necessary when the widths of confidence intervals are asymptotically bounded above zero, as in the case of Bayesian nonidentifiable models (Section 3.3). Given these difficulties, this section considers the strengths of the available methods for power analysis from the no gold standard literature.

When powering a study, the minimal sample size $N$ necessary to obtain a fixed interval length $\ell$ for a parameter $\theta_k$ at coverage probability $1-\gamma$ must be determined. The estimate of $\theta_k$, however, depends upon the observed data $x$, which is of course unavailable to the researcher at this stage. In order to proceed, a criterion must be chosen for obtaining sample-independent estimates. Three such criteria are given in the succeeding text.

The average coverage criterion fixes $\ell$ and averages the coverage probability $1 - \gamma$ over data $x$. The value of $N$ chosen is the minimum such that the averaged coverage is at least $1 - \gamma$. The average length criterion, conversely, fixes the coverage probability and allows interval

length to vary with $N$ chosen so that the average interval length over all data $x$ is no larger than $\ell$. The worst outcome criterion (WOC) chooses $N$ such that the infimum of coverage probabilities over all data $x$ is at least $1 - \gamma$. Unfortunately, this infimum tends to be infinite for many applications because of the existence of low probability subsets of possible data with poor coverage. It is recommended that a modified WOC be used instead, which requires only that the WOC holds on some subset $S$ of high probability in the space of all possible data sets [241]. Of these, the average length criterion may be the most commonly used because it corresponds well to standard practice with fixing coverage rates for confidence intervals. The criteria can be estimated by repeated simulation studies across a range of sample sizes in Bayesian models [242, 243].

The existing publications on sample size estimation broadly cover the full range of accuracy models appearing in the literature. For the closely related question of obtaining a minimum number of diseased subjects, ad hoc methods such as prevalence-inflated sample sizes were found to largely agree with formal estimates [244]. This method focused on powering to distinguish unequal sensitivities and specificities and did not consider confidence interval widths. In the ideal case when the HW model applies, by taking advantage of the exact solution in the two-sample two-test case, an Excel spreadsheet for sample size estimation has been developed [245]. If the test population is small, however, prevalence estimates can become biased, and a sampling design must be chosen with care [246]. Sample size estimation for partially verified samples has been considered under the assumption of independent [247] and dependent tests [94]. In the absence of a gold standard and with dependent tests, sample size estimates are available for the inclusion of covariates and a Bayesian power criterion [248] or continuous-valued tests [249]. Nonparametric sample size estimates for ROC curves have been estimated using a Dirichlet process [210].

Nonidentifiable models complicate sample size estimation. Dramatic improvement in credible interval width, and hence necessary sample size, was found to result from the inclusion of a third independent test for identifiability when a single sample was available [250]. It has been suggested for nonidentifiable models that because the credible intervals for some parameters will not shrink arbitrarily small as sample size increases, a comparison should instead be made between prior and posterior variances using a loss function [251]. This is carried out to focus attention on the ratio of credible interval width to its asymptotic lower bound and therefore on the diminishing returns provided when increasing sample size beyond a certain point. The question of how to determine minimum sample sizes for nonidentifiable models has not been fully resolved at this time and will only increase in importance as these models find wider use.

### 4.5. Computing the model

Code is available from the author implementing MLE and the analytic solution of the Hui–Walter model in Stata as well as WinBUGS code for a number of covariance models and data sets from the literature. Hui–Walter estimates are also available from the ‘TAGS’ program [252], which was implemented in R and S-Plus, and an Excel spreadsheet exists for its sample size calculations [245]. SAS code for misclassification in logistic regression is available [5]. Some authors have included their WinBUGS code in an appendix [28, 72, 76,

154, 163, 253-256], and code for multiple models from the papers of Branscum *et al*. [48] and Choi *et al*. [24] is available at http://www.epi.ucdavis.edu/diagnostictests/software.html

## 5. Conclusion

In this review, we have presented all recent work on the subject of measuring diagnostic test accuracy in the absence of a gold standard using latent variable models. Researchers working on this subject tend to approach it within the context of a particular application, such as medical diagnosis, commercial agriculture, or industrial chemistry, and so there is a risk of the limited communication between fields slowing the development and adoption of new methods. To address this issue, we systematically gathered the published literature and identified the major lines of research, their motivation, complications, and how these factors inform the process of modeling for a specific data set.

We recognize several limitations of this review. We restricted ourselves to the English language literature even though a small number of non-English papers were found during our literature search. It is possible that this restriction masked a parallel literature with significant findings. This field is currently active, and so while every effort has been made to include all research meeting our criteria, it is possible that papers published because the last 4 months of 2012 were not included. Finally, we acknowledge that our recommendations for modeling and estimation techniques may be biased by our own experiences. In particular, we have recommended against the use of burn-in or thinning with MCMC because neither have shown to improve estimates in any of the Bayesian models we have considered. Other researchers may reach the opposite conclusion. Ultimately, it is the data that will determine the best approach.

In presenting the current state of the research, it has become apparent that further research is required in multiple areas. Virtually, none of the models in use has analytic solutions or, for Bayesian models, conjugate priors, and so numerical methods must be used. Bayesian methods have flourished despite these conditions, aided by the ease of use of programs such as WinBUGS for numerical estimation, while frequentist methods have suffered from a lack of off-the-shelf code for estimation and the increased interest in studying nonidentifiable models that are generally ill-posed problems from the perspective of frequentist modeling. This state of affairs has also resulted in a lack of comparisons between frequentist and Bayesian methods in the literature [255, 257] with which researchers could assess the merits of each approach.

Some tests have accuracy rates that vary between populations because of observed and unobserved characteristics of the test-subject interaction. It is common in the literature to use a random effects model in these cases, asserting that averging over possible random effects, the variation is accounted for by the model. The robustness of this approach has not been inspected in detail, and alternatives have not been fully considered. We proposed in Section 2.6 that each population be modeled separately using a possibly nonidentifiable model, if modestly informative priors can be elucidated, but this criteria will not always be met. Informative priors may be impossible to obtain or the covariates defining the subpopulations may be unknown, and in these cases, a subtler approach must be found.

There has been a small but significant amount of work carried out with the assumption of an ordinal-valued true disease status. This concept has clear relevance for modeling methods to predict patients' status along an unobserved risk spectrum, as this is a natural goal of clinical medicine that has been largely ignored by studies of diagnostic accuracy [52]. Existing methods for modeling ordinal-valued true status have approached this in a weak sense by including only an intermediate category of ambiguous disease status and then applied a threshold to lump all ambiguous cases into either the nondiseased or diseased subgroups to study the effect this has upon estimates of test accuracy. While this may uncover more information than a strictly binary disease model, in the same sense that ROC curves for continuous-valued tests are an improvement over using a single threshold for binary testing, it does not make full use of the concept of an ordinal disease variable. Furthermore, this approach does not generalize to the case of categorical disease, such as in psychometric testing or rheumatology, where multiple diseases can present with similar symptoms for which the available tests are not always able to provide a unique diagnosis. If one is to assess the accuracy of a test for schizophrenia, say, that can give 'false positives' by also returning positive when the patient has bipolar disorder, then our latent disease variable has three states of healthy, bipolar, and schizophrenia, but without an ordering between the two disease states. Assessing the accuracy of such a test must take into account the differential diagnosis because the two types of false positive, for healthy and bipolar subjects, should not receive the same statistical treatment [258]. If the tests are continuous, then an ROC (hyper)surface could be constructed by comparing accuracy of the test for all diseases as the threshold varies. This will raise problems of interpretation when more than two diseases are considered as the resulting hypersurface cannot be easily displayed.

Finally, extensive work is needed on the topics of screening designs and combinations of tests. The existing publications on screening focus on choice of assumptions necessary to the model the partially verified data provided by these designs. Little work has been carried out by models that do not assume full verification to determine optimal frequency of screening or by improving prediction through weighted combinations of tests. The latter issue has been considered at length in the genetic biomarker literature under the assumption of full verification [259-261] but has not fully explored in the absence of full verification.

## Acknowledgments

This research was supported by the Intramural Research Program of the NIH, Clinical Research Center and through an Inter-Agency Agreement with the Social Security Administration.

## References

1. Hui S, Walter S. Estimating the error rates of diagnostic tests. Biometrics. 1980; 36:167–171. [PubMed: 7370371]
2. Albert P. Estimating diagnostic accuracy of multiple binary tests with an imperfect reference standard. Statistics in Medicine. 2009; 28:780–797. [PubMed: 19101935]
3. Gastwirth J, Johnson W, Reneau D. Bayesian analysis of screening data: application to AIDS in blood donors. Canadian Journal of Statistics. 1991; 19:135–150.
4. Johnson W, Pearson L. Dual screening. Biometrics. 1999; 55:867–873. [PubMed: 11315019]

5. Lyles R, Tang L, Superak H, King C, Celentano D, Lo Y, Sobel J. Validation data-based adjustments for outcome misclassification in logistic regression: an illustration. Epidemiology. 2011; 22:589–598. [PubMed: 21487295]

6. Mendoza-Blanco J, Tu X, Iyengar S. Bayesian inference on prevalence using a missing-data approach with simulation-based techniques: applications to HIV screening. Statistics in Medicine. 1996; 15:2161–2176. [PubMed: 8910961]

7. Qu Y, Tan M, Kutner M. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. Biometrics. 1996; 52:797–810. [PubMed: 8805757]

8. Tu X, Kowalski J, Jia G. Bayesian analysis of prevalence with covariates using simulation-based techniques: applications to HIV screening. Statistics in Medicine. 1999; 18:3059–3073. [PubMed: 10544306]

9. Albert P, McShane L, Shih J. NCI Tumor Marker Network. Latent class modeling approaches for assessing diagnostic error without a gold standard: with applications to p53 immunohistochemical assays in bladder tumors. Biometrics. 2001; 57:610–619. [PubMed: 11414591]

10. Alonzo T, Brinton J, Ringham B, Glueck D. Bias in estimating accuracy of a binary screening test with differential disease verification. Statistics in Medicine. 2011; 30:1852–1864. [PubMed: 21495059]

11. Baker S. Evaluating multiple diagnostic tests with partial verification. Biometrics. 1995; 51:330–337. [PubMed: 7539300]

12. Baker S, Connor R, Kessler L. The partial testing design: a less costly way to test equivalence for sensitivity and specificity. Statistics in Medicine. 1998; 17:2219–2232. [PubMed: 9802180]

13. Bernatsky S, Joseph L, Bélisle P, Boivin J, Rajan R, Moore A, Clarke A. Bayesian modelling of imperfect ascertainment methods in cancer studies. Statistics in Medicine. 2005; 24:2365–2379. [PubMed: 15977290]

14. Berry G, Smith C, Macaskill P, Irwig L. Analytic methods for comparing two dichotomous screening or diagnostic tests applied to two populations of differing disease prevalence when individuals negative on both tests are unverified. Statistics in Medicine. 2002; 21:853–862. [PubMed: 11870821]

15. Chen J, Kodell R. A decision-tree strategy for combining diagnostic tests for prediction. Biometrical Journal. 1999; 41:235–250.

16. Chen S, Watson P, Parmigiani G. Accuracy of MSI testing in predicting germline mutations of MSH2 and MLH1: a case study in Bayesian meta-analysis of diagnostic tests without a gold standard. Biostatistics. 2005; 6:450–464. [PubMed: 15831578]

17. Iversen E, Giovanni P, Chen S. Multiple model evaluation absent the gold standard through model combination. Journal of the American Statistical Association. 2008; 103:897–909.

18. Martinez E, Achcar J, Louzada-Neta F. Bayesian estimation of diagnostic tests accuracy for semi-latent data with covariates. Journal of Biopharmaceutical Statistics. 2005; 15:809–821. [PubMed: 16078387]

19. Shen Y, Wu D, Zelen M. Testing the independence of two diagnostic tests. Biometrics. 2001; 57:1009–1017. [PubMed: 11764239]

20. Van der Merwe L, Maritz J. Estimating the conditional false-positive rate for semi-latent data. Epidemiology. 2002; 13:424–430. [PubMed: 12094097]

21. Walter S. Estimation of test sensitivity and specificity when disease confirmation is limited to positive results. Epidemiology. 1999; 10:67–72. [PubMed: 9888282]

22. Walter S, Macaskill P, Lord S, Irwig L. Effect of dependent errors in the assessment of diagnostic or screening test accuracy when the reference standard is imperfect. Statistics in Medicine. 2012; 31:1129–1138. [PubMed: 22351623]

23. Zhou X, Castelluccio P, Zhou C. Nonparametric estimation of ROC curves in the absence of a gold standard. Biometrics. 2005; 61:600–609. [PubMed: 16011710]

24. Choi Y, Johnson W, Collins M, Gardner I. Bayesian inferences for receiver operating characteristic curves in the absence of a gold standard. Journal of Agricultural, Biological, and Environmental Statistics. 2006; 11:210–229.

25. Clegg T, Duignan A, Whelan C, Gormley E, Good M, Clarke J, Toft N, More S. Using latent class analysis to estimate the test characteristics of the $\gamma$-interferon test, the single intradermal

comparative tuberculin test and a multiplex immunoassay under Irish conditions. Veterinary Microbiology. 2011; 151:68–76. [PubMed: 21470800]

26. Kostoulas P, Leontides L, Enøe C, Billinis C, Flourou M, Sofia M. Bayesian estimation of sensitivity and specificity of serum ELISA and faecal culture for diagnosis of paratuberculosis in Greek dairy sheep and goats. Preventive Veterinary Medicine. 2006; 76:56–73. [PubMed: 16806541]
27. Norris M, Johnson W, Gardner I. Modeling bivariate longitudinal diagnostic outcome data in the absence of a gold standard. Statistics and its Interface. 2009; 2:171–185.
28. Norton S, Johnson W, Jones G, Heuer C. Evaluation of diagnostic tests for Johne's disease (mycobacterium avium subspecies paratuberculosis) in New Zealand dairy cows. Journal of Veterinary Diagnostic Investigation. 2010; 22:341–351. [PubMed: 20453206]
29. Wang C, Turnbull B, Gröhn Y, Nielsen S. Estimating receiver operating characteristic curves with covariates when there is no perfect reference test for diagnosis of Johne's disease. Journal of Dairy Science. 2006; 89:3038–3046. [PubMed: 16840620]
30. Wang C, Turnbull B, Gröhn Y, Nielsen S. Nonparametric estimation of ROC curves based on Bayesian models when the true disease state is unknown. Journal of Agricultural, Biological, and Environmental Statistics. 2007; 12:128–146.
31. Wang C, Turnbull B, Nielsen S, Gröhn Y. Bayesian analysis of longitudinal Johne's disease diagnostic data without a gold standard test. Journal of Dairy Science. 2011; 94:2320–2328. [PubMed: 21524521]
32. Adel A, Saegerman C, Speybroeck N, Praet N, Victor B, De Deken R, Soukehal A, Berkvens D. Canine leishmaniasis in Algeria: true prevalence and diagnostic test characteristics in groups of dogs of different functional type. Veterinary Parasitology. 2010; 172:204–213. [PubMed: 20627416]
33. Bazarusanga T, Geysen D, Vercruysse J, Marcotty T. The sensitivity of PCR and serology in different Theileria parva epidemiological situations in Rwanda. Veterinary Parasitology. 2008; 154:21–31. [PubMed: 18384961]
34. Gonçalves L, Subtil A, Rosário de Oliveira M, do Rosário V, Lee P, Shaio M. Bayesian latent class models in malaria diagnosis. PLoS One. 2012; 7(e40633):1–13.
35. Pereira G, Lousada F, Barbosa V, Ferreira-Silva M, Moraes-Souza H. A general latent class model for performance evaluation of diagnostic tests in the absence of a gold standard: an application to Chagas disease. Computational and Mathematical Methods in Medicine. 2012; 2012:1–12.
36. Speybroeck N, Williams C, Lafia K, Devleesschauwer B, Berkvens D. Estimating the prevalence of infections in vector populations using pools of samples. Medical and Veterinary Entomology. 2012; 26:361–371. [PubMed: 22486773]
37. Gastwirth J. The statistical precision of medical screening procedures: application to polygraph and AIDS antibodies test data. Statistical Science. 1987; 2:213–222.
38. Bruner L, Carr G, Harbell J, Curren R. An investigation of new toxicity test method performance in validation studies: 3 sensitivity and specificity are not independent of prevalence or distribution of toxicity. Human & Experimental Toxicity. 2002; 21:325–334.
39. Benítez-Silva H, Buchinsky M, Chan H, Cheidvasser S, Rust J. How large is the bias in self-reported disability? Journal of Applied Econometrics. 2004; 19:649–670.
40. Benítez-Silva, H.; Buchinsky, M.; Rust, J. Technical Report NBER Working Paper Series, 10219. National Bureau of Economic Research; Cambridge, MA: 2008. How large are the classification errors in the social security disability award process?.
41. Spencer B. Estimating the accuracy of jury verdicts. Journal of Empirical Legal Studies. 2007; 4:305–329.
42. Blick, J.; Hagen, P. Technical Report NPAFC Doc 370. Alaska Dept. of Fish and Game; Juneau, AK: 1998. The use of agreement measures and latent class models to assess the reliability of thermally-marked otolith classifications.
43. Conn P, Diefenbach D. Adjusting age and stage distributions for misclassification errors. Ecology. 2007; 88:1977–1983. [PubMed: 17824429]
44. Weichenthal S, Joseph L, Bélisle P, Dufresne A. Bayesian estimation of the probability of asbestos exposure from lung fiber counts. Biometrics. 2010; 66:603–612. [PubMed: 19508240]

45. Bernatsky S, Joseph L, Pineau C, Bélisle P, Boivin J, Banerjee D, Clarke A. Estimating the prevalence of polymyositis and dermatomyositis from administrative data: age, sex and regional differences. Annals of Rheumatic Disease. 2009; 68:1192–1196.
46. Bernatsky S, Joseph L, Pineau C, Tamblyn R, Feldman D, Clarke A. A population-based assessment of systemic lupus erythematosus incidence and prevalence - results and implications of using administrative data for epidemiological studies. Rheumatology. 2007; 46:1814–1818. [PubMed: 18032538]
47. Yan T, Kreuter F, Tourangeau R. Latent class analysis of response inconsistencies across modes of data collection. Social Science Research. 2012; 41:1017–1027. [PubMed: 23017914]
48. Branscum A, Gardner I, Johnson W. Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling. Preventive Veterinary Medicine. 2005; 68:145–163. [PubMed: 15820113]
49. Enøe C, Georgiadis M, Johnson W. Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. Preventive Veterinary Medicine. 2000; 45:61–81. [PubMed: 10802334]
50. Toft N, Jørgensen E, Højsgaard S. Diagnosing diagnostic tests: evaluating the assumptions underlying the estimation of sensitivity and specificity in the absence of a gold standard. Preventive Veterinary Medicine. 2005; 68:19–33. [PubMed: 15795013]
51. Grant M, Booth A. A typology of reviews: an analysis of 14 review types and associated methodologies. Health Information and Libraries Journal. 2009; 26:91–108. [PubMed: 19490148]
52. Rutjes A, Reitsma J, Coomarasamy A, Khan K, Bossuyt P. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technology Assessment. 2007; 11(50)
53. Baughman A, Bisgard K, Cortese M, Thompson W, Sanden G, Strebel P. Utility of composite reference standards and latent class analysis in evaluating the clinical accuracy of diagnostic tests for pertussis. Clinical and Vaccine Immunology. 2008; 15:106–114. [PubMed: 17989336]
54. Hadgu A. Discrepant analysis: a biased and unscientific method for estimating test sensitivity and specificity. Journal of Clinical Epidemiology. 1999; 52:1231–1237. [PubMed: 10580787]
55. Hadgu A, Dendukuri N, Wang L. Evaluation of screening tests for detecting Chlamydia trachomatis. Epidemiology. 2012; 23:72–82. [PubMed: 22157304]
56. Lipman H, Astles J. Quantifying the bias associated with use of discrepant analysis. Clinical Chemistry. 1998; 44:108–115. [PubMed: 9550567]
57. McAdam A. Discrepant analysis: how can we test a test? Journal of Clinical Microbiology. 2000; 38:2027–2029. [PubMed: 10834948]
58. Weng T. Evaluation of diagnostic tests: measuring degree of agreement and beyond. Drug Information Journal. 2001; 35:577–588.
59. Walter S, Irwig L. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. Journal of Clinical Epidemiology. 1988; 41:923–937. [PubMed: 3054000]
60. Hui S, Zhou X. Evaluation of diagnostic tests without gold standards. Statistical Methods in Medical Research. 1998; 7:354–370. [PubMed: 9871952]
61. Zhou X. Correcting for verification bias in studies of a diagnostic test's accuracy. Statistical Methods in Medical Research. 1998; 7:337–353. [PubMed: 9871951]
62. Jones G, Johnson W, Hanson T, Christensen R. Identifiability of models for multiple diagnostic testing in the absence of a gold standard. Biometrics. 2010; 66:855–863. [PubMed: 19764953]
63. Pepe M, Janes H. Insights into latent class analysis of diagnostic test performance. Biostatistics. 2007; 8:474–484. [PubMed: 17085745]
64. 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. [PubMed: 22017371]
65. Uebersax J. Validity inferences from interobserver agreement. Psychological Bulletin. 1988; 104:405–416.
66. Brenner H. Use and limitations of dual measurements in correcting for nondifferential exposure misclassification. Epidemiology. 1992; 3:216–222. [PubMed: 1591320]

67. Vacek P. The effect of conditional dependence on the evaluation of diagnostic tests. Biometrics. 1985; 41:959–968. [PubMed: 3830260]
68. Sinclair M, Gastwirth J. On procedures for evaluating the effectiveness of reinterview survey methods: application to labor force data. Journal of the American Statistical Association. 1996; 91:961–969.
69. Torrance-Rynard V, Walter S. Effects of dependent errors in the assessment of diagnostic test performance. Statistics in Medicine. 1997; 16:2157–2175. [PubMed: 9330426]
70. Johnson W, Hanson T. Comment on “on model expansion, model contraction, identifiability and prior information: two illustrative scenarios involving mismeasured variables”. Statistical Science. 2005; 20:111–140.
71. Boelaert M, El-Safi S, Hailu A, Mukhtar M, Rijal S, Sundar S, Wasunna M, Aseffa A, Mbui J, Menten J, Desjeux P, Peeling R. 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; 102:32–40. [PubMed: 17942129]
72. Nérette P, Stryhn H, Dohoo I, Hammell L. Using pseudogold standards and latent-class analysis in combination to evaluate the accuracy of three diagnostic tests. Preventive Veterinary Medicine. 2008; 85:207–225. [PubMed: 18355935]
73. Toft N, Åkerstedt J, Tharaldsen J, Hopp P. Evaluation of three serological tests for diagnosis of Maedi-Visna virus infection using latent class analysis. Veterinary Microbiology. 2007; 120:77–86. [PubMed: 17118583]
74. Berkvens D, Speybroeck N, Praet N, Adel A, Lesaffre E. Estimating disease prevalence in a Bayesian framework using probabilistic constraints. Epidemiology. 2006; 17:145–153. [PubMed: 16477254]
75. Fablet C, Marois C, Kobisch M, Madec F, Rose N. Estimation of the sensitivity of four sampling methods for Mycoplasma hyopneumoniae in live pigs using a Bayesian approach. Veterinary Microbiology. 2010; 143:238–245. [PubMed: 20036079]
76. Geurden T, Claerebout E, Vercruysse J, Berkvens D. Estimation of diagnostic test characteristics and prevalence of Giardia duodenalis in dairy calves in Belgium using a Bayesian approach. International Journal for Parasitology. 2004; 34:1121–1127. [PubMed: 15380683]
77. Habib I, Sampers I, Uyttendaele M, De Zutter L, Berkvens D. A Bayesian modelling framework to estimate Campy-lobacter prevalence and culture methods sensitivity: application to a chicken meat survey in Belgium. Journal of Applied Microbiology. 2008; 105:2002–2008. [PubMed: 19120647]
78. Lu Y, Dendukuri N, Schiller I, Joseph L. A Bayesian approach to simultaneously adjusting for verification and reference standard bias in diagnostic test studies. Statistics in Medicine. 2010; 29:2532–2543. [PubMed: 20799249]
79. Praud A, Gimenez O, Zanella G, Dufour B, Pozzi N, Antras V, Meyer L, Garin-Bastuji B. Estimation of sensitivity and specificity of five serological tests for the diagnosis of porcine brucellosis. Preventive Veterinary Medicine. 2012; 104:94–100. [PubMed: 22153032]
80. Anisur Rahman A, Saegerman C, Berkvens D, Fretin D, Osman Gani M, Ershaduzzaman M, Uddin Ahmed M, Emmanuel A. Bayesian estimation of true prevalence, sensitivity and specificity of indirect ELISA, Rose Bengal test and slow agglutination test for the diagnosis of brucellosis in sheep and goats in Bangladesh. Preventive Veterinary Medicine. 2013; 110:242–252. [PubMed: 23276401]
81. Hadgu A, Qu Y. A biomedical application of latent class models with random effects. Journal of the Royal Statistical Society. 1998; 47:603–616.
82. Albert P, Dodd L. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. Biometrics. 2004; 60:427–435. [PubMed: 15180668]
83. Dendukuri N, Hadgu A, Wang L. Modeling conditional dependence between diagnostic tests: a multiple latent variable model. Statistics in Medicine. 2009; 28:441–461. [PubMed: 19067379]
84. Goetghebeur E, Liinev J, Boelaert M, Van der Stuyft P. Diagnostic test analyses in search of their gold standard: latent class analyses with random effects. Statistical Methods in Medical Research. 2000; 9:231–248. [PubMed: 11084706]
85. Sadatsafavi M, Shahidi N, Marra F, FitzGerald M, Elwood K, Guo N, Marra C. A statistical method was used for the meta-analysis of tests for latent TB in the absence of a gold standard,

combining random-effect and latent-class methods to estimate accuracy. Journal of Clinical Epidemiology. 2010; 63:257–269. [PubMed: 19692208]

86. Xu H, Craig B. A probit latent class model with general correlation structures for evaluating accuracy of diagnostic tests. Biometrics. 2009; 65:1145–1155. [PubMed: 19210729]
87. Shih J, Albert P. Latent model for correlated binary data with diagnostic error. Biometrics. 1999; 55:1232–1235. [PubMed: 11315074]
88. Xie Y, Chen Z, Albert P. A crossed random effects modeling approach for estimating diagnostic accuracy from ordinal ratings without a gold standard. Statistics in Medicine. 2013; 32:3472–3485. [PubMed: 23529923]
89. Ünlü A. Estimation of careless error and lucky guess probabilities for dichotomous test items: a psychometric application of a biometric latent class model with random effects. Journal of Mathematical Psychology. 2006; 50:309–328.
90. Tan M, Qu Y, Rao J. Robustness of the latent variable model for correlated binary data. Biometrics. 1999; 55:258–263. [PubMed: 11318164]
91. Fujisawa H, Izumi S. Inference about misclassification probabilities from repeated binary responses. Biometrics. 2000; 56:706–711. [PubMed: 10985206]
92. Baadsgaard N, Jørgensen E. A Bayesian approach to the accuracy of clinical observations. Preventive Veterinary Medicine. 2003; 59:189–206. [PubMed: 12835004]
93. Drews C, Flanders W, Kosinski A. Use of two data sources to estimate odds ratios in case-control studies. Epidemiology. 1993; 4:327–335. [PubMed: 8347743]
94. Roldán Nofuentes J, Luna del Castillo J, Femia Marzo P. Computational methods for comparing two binary diagnostic tests in the presence of partial verification of the disease. Computational Statistics and Data Analysis. 2009; 24:695–718.
95. Böhning D, Patilea V. A capture-recapture approach for screening using two diagnostic tests with availability of disease status for the test positives only. Journal of the American Statistical Association. 2008; 103:212–221.
96. Flanders W, Drews C, Kosinski A. Methodology to correct for differential misclassification. Epidemiology. 1995; 6:152–156. [PubMed: 7742401]
97. Yang I, Becker M. Latent variable modeling of diagnostic accuracy. Biometrics. 1997; 53:948–958. [PubMed: 9290225]
98. Hanson T, Johnson W, Gardner I. Log-linear and logistic modeling of dependence among diagnostic tests. Preventive Veterinary Medicine. 2000; 45:123–137. [PubMed: 10802337]
99. Baker S, Freedman L, Parmar M. Using replicate observations in observer agreement studies with binary assessments. Biometrics. 1991; 47:1327–1338. [PubMed: 1786322]
100. Walter S, Franco E. Use of latent class models to accommodate inter-laboratory variation in assessing genetic polymorphisms associated with disease risk. BMC Genetics. 2008; 9(51):1–10. [PubMed: 18173855]
101. Verma-Kumar S, Abraham D, Dendukuri N, Cheeran J, Sukumar R, Balaji K. Serodiagnosis of tuberculosis in Asian elephants (Elephas maximus) in southern India: a latent class analysis. PLoS One. 2012; 7(e49548):1–8.
102. Sepúlveda R, Vicente-Villardón J, Galindo M. The biplot as a diagnostic tool of local dependence in latent class models. A medical application. Statistics in Medicine. 2008; 27:1855–1869. [PubMed: 18265437]
103. Garrett E, Zeger S. Latent class model diagnosis. Biometrics. 2000; 56:1055–1067. [PubMed: 11129461]
104. Subtil A, Oliveira M, Gonçalves L. Conditional dependence diagnostic in the latent class model: a simulation study. Statistics and Probability Letters. 2012; 82:1407–1412.
105. Gardner I, Stryhn H, Lind P, Collins M. Conditional dependence between tests affects the diagnosis and surveillance of animal diseases. Preventive Veterinary Medicine. 2000; 45:107–122. [PubMed: 10802336]
106. Chu H, Chen S, Louis T. 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; 104:512–523. [PubMed: 19562044]

107. Albert P, Dodd L. On estimating diagnostic accuracy from studies with multiple raters and partial gold standard evaluation. Journal of the American Statistical Association. 2008; 103:61–73. [PubMed: 19802353]
108. Chu H, Zhou Y, Cole S, Ibrahim J. On the estimation of disease prevalence by latent class models for screening studies using two screening tests with categorical disease status verified in test positives only. Statistics in Medicine. 2010; 29:1206–1218. [PubMed: 20191614]
109. Zhang B, Chen Z, Albert P. Estimating diagnostic accuracy of raters without a gold standard by exploiting a group of experts. Biometrics. 2012; 68:1294–1302. [PubMed: 23006010]
110. Bonde M, Toft N, Thomsen P, Sørensen J. Evaluation of sensitivity and specificity of routine meat inspection of Danish slaughter pigs using latent class analysis. Preventive Veterinary Medicine. 2010; 94:165–169. [PubMed: 20132995]
111. De Waele V, Berzano M, Berkvens D, Speybroeck N, Lowery C, Mulcahy G, Murphy T. Age-stratified Bayesian analysis to estimate sensitivity and specificity of four diagnostic tests for detection of cryptosporidium oocysts in neonatal calves. Journal of Clinical Microbiology. 2011; 49:76–84. [PubMed: 21048012]
112. Jones G, Johnson W, Vink W. Evaluating a continuous biomarker for infection by using observed disease status with covariate effects on disease. Applied Statistics. 2009; 58:705–717.
113. Kosinski A, Flanders W. Evaluating the exposure and disease relationship with adjustment for different types of exposure misclassification: a regression approach. Statistics in Medicine. 1999; 18:2795–2808. [PubMed: 10521867]
114. Krogh M, Toft N, Enevoldsen C. Latent class evaluation of a milk test, a urine test, and the fat-to-protein percentage ratio in milk to diagnose ketosis in dairy cows. Journal of Dairy Science. 2011; 94:2360–2367. [PubMed: 21524525]
115. Epstein L, Muñoz A, He D. Bayesian imputation of predictive values when covariate information is available and gold standard diagnosis is unavailable. Statistics in Medicine. 1996; 15:463–476. [PubMed: 8668872]
116. Gao S, Hui S. Estimating the incidence of dementia from two-phase sampling with non-ignorable missing data. Statistics in Medicine. 2000; 19:1545–1554. [PubMed: 10844717]
117. Gao S, Hui S, Hall K, Hendrie H. Estimating disease prevalence from two-phase surveys with non-response at the second phase. Statistics in Medicine. 2000; 19:2101–2114. [PubMed: 10931514]
118. Lewis F, Gunn G, McKendrick I, Murray F. Bayesian inference for within-herd prevalence of Leptospira interrogans serovar Hardjo using bulk milk antibody testing. Biostatistics. 2009; 10:719–728. [PubMed: 19628639]
119. Nérette P, Dohoo I, Hammell L. Estimation of specificity and sensitivity of three diagnostic tests for infectious salmon anaemia virus in the absence of a gold standard. Journal of Fish Diseases. 2005; 28:89–99. [PubMed: 15705154]
120. Frössling J, Bonnett B, Lindberg A, Björkman C. Validation of a Neospora caninum iscom ELISA without a gold standard. Preventive Veterinary Medicine. 2003; 57:141–153. [PubMed: 12581597]
121. Alonzo T, Pepe M, Lumley T. Estimating disease prevalence in two-phase studies. Biostatistics. 2003; 4:313–326. [PubMed: 12925524]
122. Zhou X, Castelluccio P, Hui S, Rodenberg C. Comparing two prevalence rates in a two-phase design study. Statistics in Medicine. 1999; 18:1171–1182. [PubMed: 10363338]
123. Johnson W, Gardner I, Metoyer C, Branscum A. On the interpretation of test sensitivity in the two-test two-population problem: assumptions matter. Preventive Veterinary Medicine. 2009; 91:116–121. [PubMed: 19651450]
124. Sinclair M, Gastwirth J. Properties of the Hui and Walter and related methods for estimating prevalence rates and error rates of diagnostic testing procedures. Drug Information Journal. 2000; 34:605–615.
125. Joseph L, Gyorkos T, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. American Journal of Epidemiology. 1995; 141:263–272. [PubMed: 7840100]

126. Brenner H, Gefeller O. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. Statistics in Medicine. 1997; 16:981–991. [PubMed: 9160493]

127. Coste J, Jourdain P, Pouchot J. A gray zone assigned to inconclusive results of quantitative diagnostic tests: application to the use of brain natriuretic peptide for diagnosis of heart failure in acute dyspneic patients. Clinical Chemistry. 2006; 52:2229–2235. [PubMed: 17053156]

128. Garrett E, Eaton W, Zeger S. 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. [PubMed: 12111879]

129. Wang Z, Zhou X. Random effects models for assessing diagnostic accuracy of traditional Chinese doctors in absence of a gold standard. Statistics in Medicine. 2012; 31:661–671. [PubMed: 21626532]

130. Wang Z, Zhou X, Wang M. Evaluation of diagnostic accuracy in detecting ordered symptom statuses without a gold standard. Biostatistics. 2011; 12:567–581. [PubMed: 21209155]

131. Nielsen L, Toft N, Ersbøll A. Evaluation of an indirect serum ELISA and a bacteriological faecal culture test for diagnosis of Salmonella serotype Dublin in cattle using latent class models. Journal of Applied Microbiology. 2004; 96:311–319. [PubMed: 14723692]

132. Boelaert M, Aoun K, Liinev J, Goetghebeur E, Van der Stuyft P. The potential of latent class analysis in diagnostic test validation for canine Leishmania infantum infection. Epidemiology and Infection. 1999; 123:499–506. [PubMed: 10694163]

133. Qu Y, Hadgu A. A model for evaluating sensitivity and specificity for correlated diagnostic tests in efficacy studies with an imperfect reference test. Journal of the American Statistical Association. 1998; 93:920–928.

134. Lau T. The latent class model for multiple binary screening tests. Statistics in Medicine. 1997; 16:2283–2295. [PubMed: 9351165]

135. Carabin H, Balolong E, Joseph L, McGarvey S, Johansen M, Fernandez T, Willingham A, Olveda R. Estimating sensitivity and specificity of a faecal examination methods for Schistosoma japonicum infection in cats, dogs, water buffaloes, pigs, and rats in Western Samar and Sorgoson provinces, the Philippines. International Journal for Parasitology. 2005; 35:1517–1524. [PubMed: 16188261]

136. Erkanli A, Soyer R, Costello E. Bayesian inference for prevalence in longitudinal two-phase studies. Biometrics. 1999; 55:1145–1150. [PubMed: 11315060]

137. Baadsgaard N, Højsgaard S, Gröhn Y, Schukken Y. Forecasting clinical disease in pigs: comparing a naive and a Bayesian approach. Preventive Veterinary Medicine. 2004; 64:85–100. [PubMed: 15325764]

138. Billiouw M, Brandt J, Vercruysse J, Speybroeck N, Marcotty T, Mulumba M, Berkvens D. Evaluation of the indirect fiuorescent antibody test as a diagnostic tool for East Coast fever in eastern Zambia. Veterinary Parasitology. 2005; 127:189–198. [PubMed: 15710519]

139. Cook R, Ng E, Meade M. Estimation of operating characteristics for dependent diagnostic tests based on latent Markov models. Biometrics. 2000; 56:1109–1117. [PubMed: 11129468]

140. Wolfe R, Carlin J, Patton G. Transitions in an imperfectly observed binary variable: depressive symptomatology in adolescents. Statistics in Medicine. 2003; 22:427–440. [PubMed: 12529873]

141. Enøe C, Christensen G, Andersen S, Willeberg P. The need for built-in validation of surveillance data so that changes in diagnostic performance of post-mortem meat inspection can be detected. Preventive Veterinary Medicine. 2003; 57:117–125. [PubMed: 12581595]

142. Engel B, Backer J, Buist W. Evaluation of the accuracy of diagnostic tests from repeated measurements without a gold standard. Journal of Agricultural, Biological, and Environmental Statistics. 2010; 15:83–100.

143. Engel B, Buist W, Orsel K, Dekker A, de Clercq K, Grazioli S, van Roermund H. A Bayesian evaluation of six diagnostic tests for foot-and-mouth disease for vaccinated and non-vaccinated cattle. Preventive Veterinary Medicine. 2008; 86:124–138. [PubMed: 18455817]

144. Brown E. Bayesian estimation of the time-varying sensitivity of a diagnostic test with application to mother-to-child transmission of HIV. Biometrics. 2010; 66:1266–1274. [PubMed: 20222936]

145. Pauler D, Laird N. A mixture model for longitudinal data with application to assessment of noncompliance. Biometrics. 2000; 56:464–472. [PubMed: 10877305]

146. Jones G, Johnson W, Vink W, French N. A framework for the joint modeling of longitudinal diagnostic outcome data and latent infection status: application to investigating the temporal relationship between infection and disease. Biometrics. 2012; 68:371–379. [PubMed: 22004274]

147. Geisser S, Johnson W. Optimal administration of dual screening tests for detecting a characteristic with special reference to low prevalence diseases. Biometrics. 1992; 48:839–852. [PubMed: 1330025]

148. Viana M, Ramakrishnan V. Bayesian estimates of predictive value and related parameters of a diagnostic test. Canadian Journal of Statistics. 1992; 20:311–321.

149. Smith A, Roberts G. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. Journal of the Royal Statistical Society Series B. 1993; 55:3–23.

150. Roberts G, Smith A. Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. Stochastic Processes and their Applications. 1994; 49:207–216.

151. Toft N, Innocent G, Gettinby G, Reid S. Assessing the convergence of Markov chain Monte Carlo methods: an example from evaluation of diagnostic tests in absence of a gold standard. Preventive Veterinary Medicine. 2007; 79:244–256. [PubMed: 17292499]

152. Link W, Eaton M. On thinning of chains in MCMC. Methods in Ecology and Evolution. 2012; 3:112–115.

153. Geweke, J. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In: Bernado, J.; Berger, J.; Dawid, A.; Smith, A., editors. Bayesian Statistics 4. Clarendon Press; Oxford, UK: 1992. p. 169-193.

154. Choi Y, Johnson W, Thurmond M. Diagnosis using predictive probabilities without cut-offs. Statistics in Medicine. 2006; 25:699–717. [PubMed: 16220514]

155. Georgiadis M, Johnson W, Gardner I, Singh R. Correlation-adjusted estimation of sensitivity and specificity of two diagnostic tests. Journal of the Royal Statistical Society. 2003; 52:63–76.

156. Gustafson P, Le N, Saskin R. Case-control analysis with partial knowledge of exposure misclassification probabilities. Biometrics. 2001; 57:598–609. [PubMed: 11414590]

157. Ladouceur M, Rahme E, Pineau C, Joseph L. Robustness of prevalence estimates derived from misclassified data from administrative databases. Biometrics. 2007; 63:272–279. [PubMed: 17447953]

158. Speybroeck N, Praet N, Claes F, Hong N, Torres K, Mao P, Thinh T, Gamboa D, Sochantha T, Thang N, Coosemans M, Büscher P, D'Alessandro U, Berkvens D, Erhart A. True versus apparent malaria infection prevalence: the contribution of a Bayesian approach. PLoS ONE. 2011; 6(e16705):1–7.

159. Brenner H. How independent are multiple 'independent' diagnostic classifications? Statistics in Medicine. 1996; 15:1377–1386. [PubMed: 8841648]

160. Andritsos N, Matargas M, Paramithiotis S, Nychas G, Drosinos E. Estimating the diagnostic accuracy of three culture-dependent methods for the Listeria monocytogenes detection from a Bayesian perspective. International Journal of Food Microbiology. 2012; 156:181–185. [PubMed: 22507629]

161. Hanson T, Johnson W, Gardner I. Hierarchical models for estimating herd prevalence and test accuracy in the absence of a gold standard. Journal of Agricultural, Biological, and Environmental Statistics. 2003; 8:223–239.

162. Engel B, Swildens B, Stegeman A, Buist W, de Jong M. Estimation of sensitivity and specificity of three conditionally dependent diagnostic tests in the absence of a gold standard. Journal of Agricultural, Biological, and Environmental Statistics. 2006; 11:360–380.

163. Su CL, Gardner IA, Johnson WO. Bayesian estimation of cluster-level test accuracy based on different sampling schemes. Journal of Agricultural, Biological, and Environmental Statistics. 2007; 12:250–271.

164. Johnson WO, Gastwirth JL. Bayesian inference for medical screening tests: approximations useful for the analysis of acquired immune deficiency syndrome. Journal of the Royal Statistical Society. 1991; 53:427–439.

165. deC Bronsvoort BM, Koterwas B, Land F, Handel IG, Tucker J, Morgan KL, Tanya VN, Abdoel TH, Smits HL. Comparison of a flow assay for Brucellosis antibodies with the reference cELISA test in West African Bos indicus. PLoS ONE. 2009; 4(e5221):1–7.

166. Orr KA, O'Reilly KL, Scholl DT. Estimation of sensitivity and specificity of two diagnostics tests for bovine immunodeficiency virus using Bayesian techniques. Preventive Veterinary Medicine. 2003; 61:79–89. [PubMed: 14519338]

167. Buzoianu M, Kadane JB. Adjusting for verification bias in diagnostic test evaluation: a Bayesian approach. Statistics in Medicine. 2008; 27:2453–2473. [PubMed: 17979150]

168. Andersen S. Re: "Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard". American Journal of Epidemiology. 1997; 144:290. [PubMed: 9012602]

169. Johnson WO, Gastwirth JL, Pearson LM. Screening without a "gold standard": the Hui-Walter paradigm revisited. American Journal of Epidemiology. 2001; 153:921–924. [PubMed: 11323324]

170. Neath AA, Samaniego FJ. On the efficacy of Bayesian inference for nonidentifiable models. The American Statistician. 1997; 51:225–232.

171. Black MA, Craig BA. Estimating disease prevalence in the absence of a gold standard. Statistics in Medicine. 2002; 21:2653–2669. [PubMed: 12228883]

172. Gustafson P. On model expansion, model contraction, identifiability and prior information: two illustrative scenarios involving mismeasured variables. Statistical Science. 2005; 20:111–140.

173. Gustafson P. The utility of prior information and stratification for parameter estimation with two screening tests but no gold standard. Statistics in Medicine. 2005; 24:1203–1217. [PubMed: 15558709]

174. Gustafson P. What are the limits of posterior distributions arising from nonidentified models, and why should we care? Journal of the American Statistical Association. 2009; 104:1682–1695.

175. Xie Y, Carlin BP. Measures of Bayesian learning and identifiability in hierarchical models. Journal of Statistical Planning and Inference. 2006; 136:3458–3477.

176. Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society Series B. 2002; 64:583–639.

177. Spiegelhalter, D.; Best, TA.; Gilks, W. Bugs: Bayesian Inference Using Gibbs Sampling Version 0.50. MRC Biostatistics Unit; Cambridge: 1996.

178. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. Biometrics. 2001; 57:158–167. [PubMed: 11252592]

179. Claeskens, G.; Hjort, NL. Model Selection and Model Averaging Cambridge. University Press; Cambridge: 2008.

180. Chu H, Cole SR, Wei Y, Ibrahim JG. Estimation and inference for case-control studies with multiple non-gold standard exposure assessments: with an occupational health application. Biostatistics. 2009; 10:591–602. [PubMed: 19515637]

181. Mossman D, Bowen MD, Vanness DJ, Bienenfeld D, Correll T, Kay J, Klykylo WM, Lehrer DS. Quantifying the accuracy of forensic examiners in the absence of a "gold standard". Law and Human Behavior. 2010; 34:402–417. [PubMed: 19771499]

182. Pennello GA. Bayesian analysis of diagnostic test accuracy when disease state is unverified for some subjects. Journal of Biopharmaceutical Statistics. 2011; 21:954–970. [PubMed: 21830925]

183. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial. Statistical Science. 1999; 14:382–401.

184. Uhler C. Mastitis in dairy production: estimation of sensitivity, specificity and disease prevalence in the absence of a gold standard. Journal of Agricultural, Biological, and Environmental Statistics. 2009; 14:79–98.

185. Geisser S. Comparing two tests for diagnostic or screening purposes. Statistics & Probability Letters. 1998; 40:113–119.

186. Skaltsa K, Jover L, Carrasco JL. Estimation of the diagnostic threshold accounting for decision costs and sampling uncertainty. Biometrical Journal. 2010; 52:676–697. [PubMed: 20976697]

187. Irion A, Beck HP, Smith T. Assessment of positivity in immuno-assays with variability in background measurements: a new approach applied to the antibody response to Plasmodium falciparum MSP2. Journal of Immunological Methods. 2002; 259:111–118. [PubMed: 11730846]

188. Abbey CK, Eckstein MP, Boone JM. Estimating the relative utility of screening mammography. Medical Decision Making. 2013; 33:510–520. [PubMed: 23295543]
189. Walter SD, Irwig L, Glasziou PP. Meta-analysis of diagnostic tests with imperfect reference standards. Journal of Clinical Epidemiology. 1999; 10:943–951. [PubMed: 10513757]
190. Liu D, Zhou XH. Semiparametric estimation of the covariate-specific ROC curve in presence of ignorable verification bias. Biometrics. 2011; 67:906–916. [PubMed: 21361890]
191. Branscum AJ, Johnson WO, Hanson TE, Gardner IA. Bayesian semiparametric ROC curve estimation and disease diagnosis. Statistics in Medicine. 2008; 27:2474–2496. [PubMed: 18300333]
192. Fluss R, Reiser B, Faraggi D, Rotnitsky A. Estimation of the ROC curve under verification bias. Biometrical Journal. 2009; 51:475–490. [PubMed: 19588455]
193. Liu D, Zhou XH. A model for adjusting for nonignorable verification bias in estimation of the ROC curve and its area with likelihood-based approach. Biometrics. 2010; 66:1119–1128. [PubMed: 20222937]
194. Rodenberg C, Zhou XH. ROC curve estimation when covariates affect the verification process. Biometrics. 2000; 56:1256–1262. [PubMed: 11129488]
195. Zhou XH, Castelluccio P. Nonparametric analysis for the ROC areas of two diagnostic tests in the presence of nonignorable verification bias. Journal of Statistical Planning and Inference. 2003; 115:193–213.
196. Zhou XH, Castelluccio P. Adjusting for non-ignorable verification bias in clinical studies for Alzheimer's disease. Statistics in Medicine. 2004; 23:221–230. [PubMed: 14716724]
197. He H, Lyness JM, McDermott MP. Direct estimation of the area under the receiver operating characteristic curve in the presence of verification bias. Statistics in Medicine. 2009; 28:361–376. [PubMed: 18680124]
198. 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. [PubMed: 22568612]
199. Jafarzadeh SR, Johnson WO, Utts JM, Gardner IA. Bayesian estimation of the receiver operating characteristic curve for a diagnostic test with a limit of detection in the absence of a gold standard. Statistics in Medicine. 2010; 29:2090–2106. [PubMed: 20603894]
200. Su JQ, Liu JS. Linear combinations of multiple diagnostic markers. Journal of the American Statistical Association. 1993; 88:1350–1355.
201. Liu A, Schisterman EF, Zhu Y. On linear combinations of biomarkers to improve diagnostic accuracy. Statistics in Medicine. 2005; 24:37–47. [PubMed: 15515132]
202. Yu B, Zhou C, Bandinelli S. Combining multiple continuous tests for the diagnosis of kidney impairment in the absence of a gold standard. Statistics in Medicine. 2011; 30:1712–1721. [PubMed: 21432889]
203. Jin H, Lu Y. A procedure for determining whether a simple combination of diagnostic tests may be noninferior to the theoretical optimum combination. Medical Decision Making. 2008; 28:909–916. [PubMed: 18556633]
204. Alvord WG, Drummond JE, Arthur LO, Biggar RJ, Goedert JJ, Levine PH, Murphy EL, Weiss SH, Blattner WA. A method for predicting individual HIV infection status in the absence of clinical information. Aids Research and Human Retroviruses. 1988; 4:295–304. [PubMed: 3207513]
205. Rindskopf D, Rindskopf W. The value of latent class analysis in medical diagnosis. Statistics in Medicine. 1986; 5:21–28. [PubMed: 3961312]
206. Formann AK. Measurement errors in caries diagnosis: some further latent class models. Biometrics. 1994; 50:865–871. [PubMed: 7981408]
207. Caraguel C, Stryhn H, Gagné N, Dohoo I, Hammell L. Use of a third class in latent class modelling for the diagnostic evaluation of five infectious salmon anaemia virus detection tests. Preventive Veterinary Medicine. 2012; 104:165–173. [PubMed: 22051529]
208. Xu H, Black MA, Craig BA. Evaluating accuracy of diagnostic tests with intermediate results in the absence of a gold standard. Statistics in Medicine. 2013; 32:2571–2584. [PubMed: 23212851]

209. Albert PS. Random effects modeling approaches for estimating ROC curves from repeated ordinal tests without a gold standard. Biometrics. 2007; 63:593–602. [PubMed: 17688512]
210. Cheng D, Branscum AJ, Johnson WO. Sample size calculations for ROC studies: parametric robustness and Bayesian nonparametrics. Statistics in Medicine. 2012; 31:131–142. [PubMed: 22139729]
211. Ladouceur M, Rahme E, Bélisle P, Scott AN, Schwartzman K, Joseph L. Modeling continuous diagnostic test data using approximate Dirichlet process distributions. Statistics in Medicine. 2011; 30:2648–2662. [PubMed: 21786286]
212. Uebersax JS, Grove WM. A latent trait mixture model for the analysis of rating agreement. Biometrics. 1993; 49:823–835. [PubMed: 10798855]
213. Ghosal S, van der Vaart A. Posterior convergence rates of Dirichlet mixtures at smooth densities. The Annals of Statistics. 2007; 35:697–723.
214. Lloyd CJ. Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. Journal of the American Statistical Association. 1998; 93:1356–1364.
215. Zou KH, Hall WJ, Shapiro DE. Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. Statistics in Medicine. 1997; 16:2143–2156. [PubMed: 9330425]
216. Chock C, Irwig L, Berry G, Glasziou P. Comparing dichotomous screening tests when individuals negative on both tests are not verified. Journal of Clinical Epidemiology. 1997; 50:1211–1217. [PubMed: 9393377]
217. Gastwirth JL, Johnson WO. Screening with cost-effective quality control: potential applications to HIV and drug testing. Journal of the American Statistical Association. 1994; 89:972–981.
218. Irwig L, Glasziou PP, Berry G, Chock C, Mock P, Simpson JM. Efficient study designs to assess the accuracy of screening tests. American Journal of Epidemiology. 1994; 140:759–769. [PubMed: 7942777]
219. McNamee R. Optimal designs of two-stage studies for estimation of sensitivity, specificity and positive predictive value. Statistics in Medicine. 2002; 21:3609–3625. [PubMed: 12436459]
220. Warnick LD, Nielsen LR, Nielsen J, Greiner M. Simulation model estimates of test accuracy and predictive values for Danish Salmonella surveillance program in dairy herds. Preventive Veterinary Medicine. 2006; 77:284–303. [PubMed: 16979767]
221. Zelen M, Haitovsky Y. Testing hypotheses with binary data subject to misclassification errors: analysis and experimental design. Biometrika. 1991; 78:857–865.
222. Alonzo TA, Kittelson JM. A novel design for estimating relative accuracy of screening tests when complete disease verification is not feasible. Biometrics. 2006; 62:605–612. [PubMed: 16918926]
223. Alonzo TA, Pepe MS. Assessing accuracy of a continuous screening test in the presence of verification bias. Journal of the Royal Statistical Society Series C. 2005; 54:173–190.
224. Zheng Y, Barlow WE, Cutter G. Assessing accuracy of mammography in the presence of verification bias and intrareader correlation. Biometrics. 2005; 61:259–268. [PubMed: 15737102]
225. Alonzo TA. Verification bias-corrected estimators of the relative true and false positive rates of two binary screening tests. Statistics in Medicine. 2005; 24:403–417. [PubMed: 15543634]
226. Kosinski AS, Barnhart HX. Accounting for nonignorable verification bias in assessment of diagnostic tests. Biometrics. 2003; 59:163–171. [PubMed: 12762453]
227. Paliwal P, Gelfand AE. Estimating measures of diagnostic accuracy when some covariate information is missing. Statistics in Medicine. 2006; 25:2981–2993. [PubMed: 16345056]
228. Zhou XH. Maximum likelihood estimators of sensitivity and specificity corrected for verification bias. Communications in Statistics; theory and methods. 1993; 22:3177–3198.
229. Zhou XH. Effect of verification bias on positive and negative predictive values. Statistics in Medicine. 1994; 13:1737–1745. [PubMed: 7997707]
230. He H, McDermott MP. A robust method using propensity score stratification for correcting verification bias for binary tests. Biostatistics. 2012; 13:32–47. [PubMed: 21856650]
231. Albert PS. Imputation approaches for estimating diagnostic accuracy for multiple tests from partially verified designs. Biometrics. 2007; 63:947–957. [PubMed: 17825024]

232. Stock EM, Stamey JD, Sankaranarayanan R, Young DM, Muwonge R, Arbyn M. Estimation of disease prevalence, true positive rate, and false positive rate of two screening tests when disease verification is applied on only screen-positives: a hierarchical model using multi-center data. Cancer Epidemiology. 2012; 36:153–160. [PubMed: 21856264]

233. Tang ML, Qui SF, Poon WY, Tang NS. Test procedures for disease prevalence with partially validated data. Journal of Biopharmaceutical Statistics. 2012; 22:368–386. [PubMed: 22251180]

234. Alonzo TA. Comparing accuracy in an unpaired post-market device study with incomplete disease assessment. Biometrical Journal. 2009; 51:491–503. [PubMed: 19572317]

235. Alonzo TA, Pepe MS, Moskowitz CS. Sample size calculations for comparative studies of medical tests for detecting presence of disease. Statistics in Medicine. 2002; 21:835–852. [PubMed: 11870820]

236. Kosinski AS, Barnhart HX. A global sensitivity analysis of performance of a medical diagnostic test when verification bias is present. Statistics in Medicine. 2003; 22:2711–2721. [PubMed: 12939781]

237. Baker SG, Pinsky PF. A proposed design and analysis for comparing digital and analog mammography. Journal of the American Statistical Association. 2001; 96:421–428.

238. van Geloven N, Broeze KA, Opmeer BC, Mol BM, Zwinderman AH. How to deal with double partial verification when evaluating two index tests in relation to a reference test? Statistics in Medicine. 2012; 31:1265–1276. [PubMed: 22161741]

239. Wacholder S, Armstrong B, Hartge P. Validation studies using an alloyed gold standard. American Journal of Epidemiology. 1993; 137:1251–1258. [PubMed: 8322765]

240. Rahme E, Joseph L, Gyorkos TW. Bayesian sample size determination for estimating binomial parameters from data subject to misclassification. Applied Statistics. 2000; 49:119–128.

241. Dendukuri N, Rahme E, Bélisle P, Joseph L. Bayesian sample size determination for prevalence and diagnostic test studies in the absence of a gold standard test. Biometrics. 2004; 60:388–397. [PubMed: 15180664]

242. Branscum AJ, Johnson WO, Gardner IA. Sample size calculations for studies designed to evaluate diagnostic test accuracy. Journal of Agricultural, Biological, and Environmental Statistics. 2007; 12:112–127.

243. Cheng D, Stamey JD, Branscum AJ. A general approach to sample size determination for prevalence surveys that use dual test protocols. Biometrical Journal. 2007; 49:694–706. [PubMed: 17722203]

244. Li J, Fine J. On sample size for sensitivity and specificity in prospective diagnostic accuracy studies. Statistics in Medicine. 2004; 23:2537–2550. [PubMed: 15287083]

245. Georgiadis MP, Johnson WO, Gardner IA, Singh R. Sample size determination for estimation of the accuracy of two conditionally independent tests in the absence of a gold standard. Preventive Veterinary Medicine. 2005; 71:1–10. [PubMed: 16076507]

246. Su CL, Gardner IA, Johnson WO. Diagnostic test accuracy and prevalence inferences based on joint and sequential testing with finite population sampling. Statistics in Medicine. 2004; 23:2237–2255. [PubMed: 15236428]

247. Kosinski AS, Chen Y, Lyles RH. Sample size calculations for evaluating a diagnostic test when the gold standard is missing at random. Statistics in Medicine. 2010; 29:1572–1579. [PubMed: 20552570]

248. Beavers DP, Stamey JD. Bayesian sample size determination for binary regression with a misclassified covariate and no gold standard. Computational Statistics and Data Analysis. 2012; 56:2574–2582.

249. Cheng D, Branscum AJ, Stamey JD. A Bayesian approach to sample size determination for studies designed to evaluate continuous medical tests. Computational Statistics and Data Analysis. 2010; 54:298–307.

250. Dendukuri N, Bélisle P, Joseph L. Bayesian sample size for diagnostic test studies in the absence of a gold standard: comparing identifiable with non-identifiable models. Statistics in Medicine. 2010; 29:2688–2697. [PubMed: 20803558]

251. Gustafson P. Sample size estimation when biases are modelled rather than ignored. Journal of the Royal Statistical Society. 2006; 169:865–881.

252. Pouillot R, Gerbier G, Gardner IA. “TAGS,” a program for the evaluation of test accuracy in the absence of a gold standard. Preventive Veterinary Medicine. 2002; 53:67–81. [PubMed: 11821138]

253. Brochier B, De Blander H, Hanosset R, Berkvens D, Losson B, Saegerman C. Echinococcus multilocularis and Toxocara canis in urban red foxes (Vulpes vulpes) in Brussels, Belgium. Preventive Veterinary Medicine. 2007; 80:65–73. [PubMed: 17324480]

254. Geurden T, Berkvens D, Casaert S, Vercruysse J, Claerebout E. A Bayesian evaluation of three diagnostic assays for the detection of Giardia duodenalis in symptomatic and asymptomatic dogs. Veterinary Parasitology. 2008; 157:14–20. [PubMed: 18723290]

255. Liu P, Yang HT, Qiang LY, Jin H, Xiao S, Shi ZX. Evaluation of 30 commercial assays for the detection of antibodies to HIV in China using classical and Bayesian statistics. Journal of Virological Methods. 2010; 170:73–79. [PubMed: 20833204]

256. Marcotty T, Simukoko H, Berkvens D, Vercruysse J, Praet N, Van den Bossche P. Evaluating the use of packed cell volume as an indicator of trypanosomal infections in cattle in eastern Zambia. Preventive Veterinary Medicine. 2008; 87:288–300. [PubMed: 18586340]

257. Liu P, Yang HT, Qiang LY, Xiao S, Shi ZX. Estimation of the sensitivity and specificity of assays for screening antibodies to HIV: a comparison between the frequentist and Bayesian approaches. Journal of Virological Methods. 2012; 186:89–93. [PubMed: 22981458]

258. Obuchowski NA, Goske MJ, Applegate KE. Assessing physicians’ accuracy in diagnosing paediatric patients with acute abdominal pain: measuring accuracy for multiple diseases. Statistics in Medicine. 2001; 20:3261–3278. [PubMed: 11746317]

259. Lin H, Zhou L, Peng H, Zhou XH. Selection and combination of biomarkers using ROC method for disease classification and prediction. The Canadian Journal of Statistics. 2011; 39:324–343.

260. McIntosh MW, Pepe MS. Combining several screening tests: optimality of the risk score. Biometrics. 2002; 58:657–664. [PubMed: 12230001]

261. Pfeiffer RM, Bura E. A model free approach to combining biomarkers. Biometrical Journal. 2008; 50:558–570. [PubMed: 18663762]

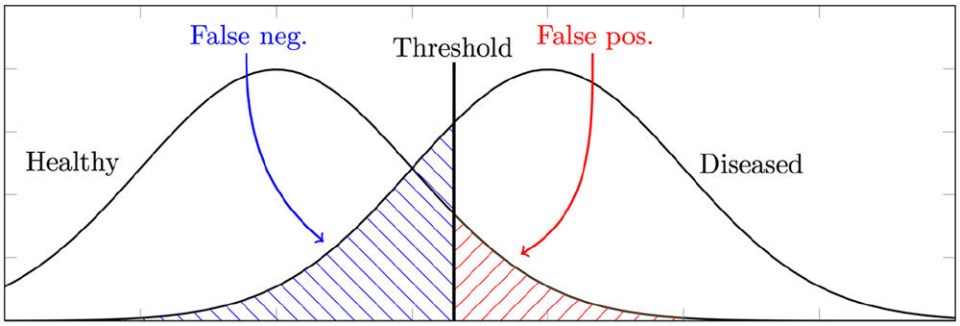


**Figure 1.**
Threshold determines false positive and negative rates.

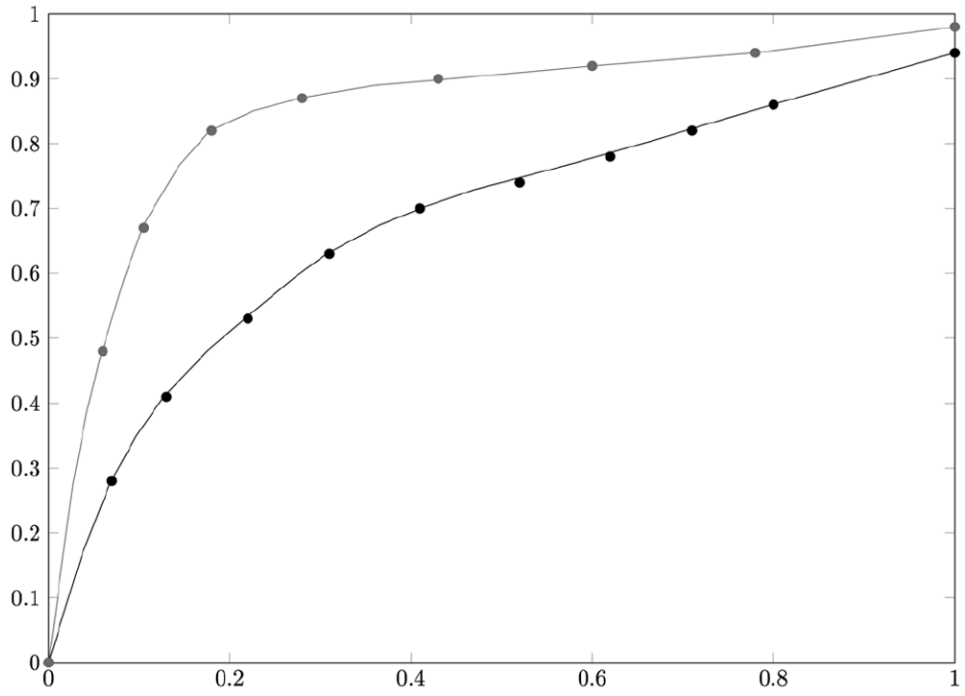


**Figure 2.**
Two ROC curves.

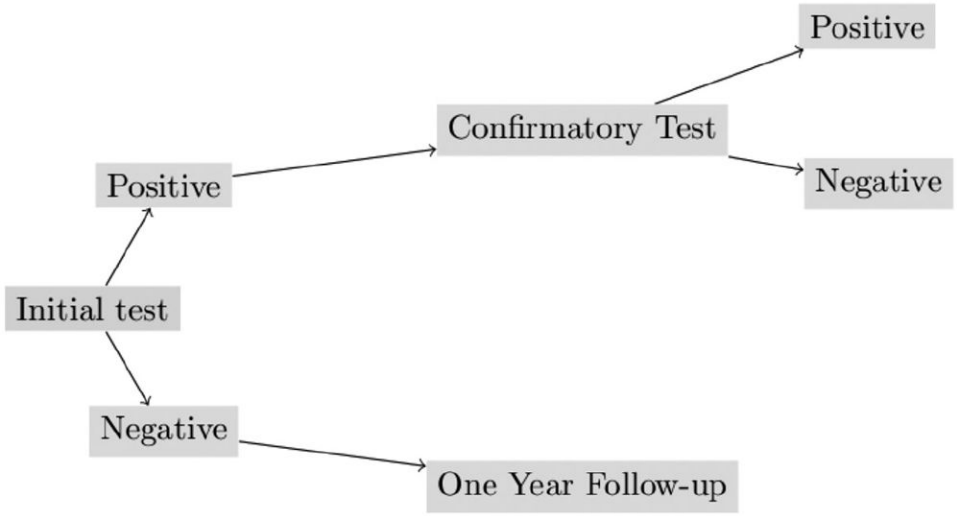


**Figure 3.**
A typical screening design.

**Table I**

Results of two tests for tuberculosis in two populations.

| | | $T_2$ | | | | | $T_2$ | | |
|---|---|---|---|---|---|---|---|---|---|
| **Pop. 1** | $T_1$ | **Positive** | **Negative** | **Total** | **Pop. 2** | $T_1$ | **Positive** | **Negative** | **Total** |
| | Positive | 14 | 4 | 18 | | Positive | 887 | 31 | 918 |
| | Negative | 9 | 528 | 537 | | Negative | 37 | 367 | 404 |
| | Total | 23 | 532 | 555 | | Total | 924 | 398 | 1322 |

*Stat Med*. Author manuscript; available in PMC 2015 October 30.

**Table II**

Testing for true disability status using SIPP variables.

| | | Beneficiary status | | | | Beneficiary status | |
|---|---|---|---|---|---|---|---|
| **High risk** | | + | − | **Low risk** | | + | − |
| Self-report | + | 974 | 583 | Self-report | + | 263 | 117 |
| | − | 384 | 6789 | | − | 785 | 34,703 |

SIPP, Survey of Income and Program Participation.

**Table III**

Two data sets with population-dependent sensitivity [49].

| | First data set | | | | | Second data set | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Sensitivity | | | | | Sensitivity | |
| | $\pi_1$ | $\pi_2$ | Specificity | Pop. 1 | Pop. 2 | $\pi_1$ | $\pi_2$ | Specificity | Pop. 1 | Pop. 2 |
| True values | 0.05 | 0.02 | 0.85 | 0.95 | 0.85 | 0.05 | 0.02 | 0.85 | 0.95 | 0.65 |
| MLE | 0.044 | 0.014 | 0.84 | 1 | 1 | 0.044 | 0.007 | 0.847 | 1 | 1 |
| MCMC | 0.087 | 0.035 | 0.859 | 0.739 | 0.739 | 0.073 | 0.017 | 0.854 | 0.797 | 0.797 |

**Table IV**

Posterior medians and 95% CI using data from [113].

| | Uninformative | | Moderately informative | | Highly informative | |
|---|---|---|---|---|---|---|
| **Variable** | **Median** | **95% CI** | **Median** | **95% CI** | **Median** | **95% CI** |
| $\pi$ | 0.135 | 0.032–0.362 | 0.136 | 0.056–0.255 | 0.142 | 0.080–0.218 |
| $\alpha_1$ | 0.447 | 0.203–0.954 | 0.434 | 0.267–0.654 | 0.435 | 0.317–0.566 |
| $\beta_1$ | 0.918 | 0.868–0.971 | 0.919 | 0.875–0.957 | 0.923 | 0.885–0.955 |
| $\alpha_2$ | 0.873 | 0.569–0.989 | 0.904 | 0.743–0.979 | 0.885 | 0.783–0.951 |
| $\beta_2$ | 0.717 | 0.629–0.916 | 0.727 | 0.651–0.823 | 0.728 | 0.665–0.794 |
| $\alpha_3$ | 0.390 | 0.168–0.839 | 0.409 | 0.185–0.739 | 0.420 | 0.198–0.698 |
| $\beta_3$ | 0.929 | 0.876–0.994 | 0.932 | 0.882–0.987 | 0.935 | 0.885–0.985 |
| $p_D$ | 5.476 | | 5.499 | | 5.003 | |
| DIC | 42.110 | | 41.315 | | 40.302 | |

DIC, deviance information criterion.