Research Article

Statistics in Medicine

Received 13 August 2009

Published online 27 August 2010 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4037

Bayesian sample size for diagnostic test studies in the absence of a gold standard: Comparing identifiable with non-identifiable models

Nandini Dendukuri,[a*†] Patrick Bélisle[b] and Lawrence Joseph[c]

Diagnostic tests rarely provide perfect results. The misclassification induced by imperfect sensitivities and specificities of diagnostic tests must be accounted for when planning prevalence studies or investigations into properties of new tests. The previous work has shown that applying a single imperfect test to estimate prevalence can often result in very large sample size requirements, and that sometimes even an infinite sample size is insufficient for precise estimation because the problem is non-identifiable. Adding a second test can sometimes reduce the sample size substantially, but infinite sample sizes can still occur as the problem remains non-identifiable. We investigate the further improvement possible when three diagnostic tests are to be applied. We first develop methods required for studies when three conditionally independent tests are available, using different Bayesian criteria. We then apply these criteria to prototypic scenarios, showing that large sample size reductions can occur compared to when only one or two tests are used. As the problem is now identifiable, infinite sample sizes cannot occur except in pathological situations. Finally, we relax the conditional independence assumption, demonstrating in this once again non-identifiable situation that sample sizes may substantially grow and possibly be infinite. We apply our methods to the planning of two infectious disease studies, the first designed to estimate the prevalence of Strongyloides infection, and the second relating to estimating the sensitivity of a new test for tuberculosis transmission. The much smaller sample sizes that are typically required when three as compared to one or two tests are used should encourage researchers to plan their studies using more than two diagnostic tests whenever possible. User-friendly software is available for both design and analysis stages greatly facilitating the use of these methods.

Keywords: Bayesian design; diagnostic test; Latent class model; misclassification; sample size

1. Introduction

Virtually no diagnostic test is error free, which complicates both the analysis and design of research studies involving diagnostic testing data. A large statistical literature has addressed these problems from an analytic viewpoint. Models include latent class analysis assuming both independent [1–5] and possibly correlated tests [6–10]. This is a particularly important issue, because it can be difficult to determine whether two tests are conditionally independent, although the final inferences can depend on the choice of the model. If tests are based on biologically very different mechanisms, then at least approximate conditional independence may hold. For example, three conditionally independent tests might be based on serology, microscopy, and genetics. Conversely, if the mechanisms of action of two tests are similar, one can expect some correlation from substantive reasons alone. Biologically, one can look for common elements of tests and consider these to be additional latent class factors [11]. Analytically, one way to address this problem is to run both independent and correlated models and check the robustness of important parameter inferences to model choice [12, 13] or to use model selection criteria such as Bayes factors [14] to statistically evaluate the model that may provide the best

[a]*Department of Epidemiology and Biostatistics, 1020 Pine Avenue West, McGill University, Montreal, Que., Canada H3A 1A2, and Technology Assessment Unit, Royal Victoria Hospital, R4.09, 687 Pine Avenue West, Que., Canada H3A 1A1*

[b]*Division of Clinical Epidemiology, McGill University Health Centre, Royal Victoria Hospital, 687 Pine Avenue West, V Building, Room V2.08, Montreal, Que., Canada H3A 1A1*

[c]*Department of Epidemiology and Biostatistics, 1020 Pine Avenue West, McGill University, Montreal, Que., Canada H3A 1A2, and Division of Clinical Epidemiology, McGill University Health Centre, Royal Victoria Hospital, 687 Pine Avenue West, V Building, Room V2.10, Montreal, Que., Canada H3A 1A1*

[*]*Correspondence to: Nandini Dendukuri, Department of Epidemiology and Biostatistics, 1020 Pine Avenue West, McGill University, Montreal, Que., Canada H3A 1A2, and Technology Assessment Unit, Royal Victoria Hospital, R4.09, 687 Pine Avenue West, Que., Canada H3A 1A1.*

[†]*E-mail: nandini.dendukuri@mcgill.ca*

Table I. Prior distributions for prevalence of *Strongyloides* and sensitivity and specificity of the serology and microscopy tests.

Test	Parameter	Prior median (95 per cent credible interval)	Beta(α, β) priors
	Prevalence (π)	0.76 (0.52, 0.91)	Beta(13.11,4.59)
Serology	Sensitivity (S_1)	0.89 (0.80, 0.95)	Beta(58.97,7.59)
	Specificity (C_1)	0.67 (0.36, 0.95)	Beta(5.23,2.17)
Microscopy	Sensitivity (S_2)	0.31 (0.22, 0.44)	Beta(22.15,45.97)
	Specificity (C_2)	0.96 (0.91, 0.99)	Beta(84.09,3.53)

fit [8, 10]. Some authors [7, 8, 13] have suggested that if dependence is not strong, then inferences obtained from models that assume conditional independence may not be very different from inferences from models that account for small correlations between tests.

Latent class models have been extended to include hierarchical [15] and regression components [16], and have been applied to data from clinical epidemiology [15–18], public health [3–5, 19–21], and veterinary medicine [22–27]. Other methods for analysis of diagnostic test results have been proposed that do not rely on latent class models, such as using a composite gold standard derived from combinations of test results [28]. There are now several books reviewing all these methods [29–31].

The design of diagnostic studies in the absence of a gold standard test is a more complex problem that has received little attention [32, 33]. Nevertheless, the work to date has shown that there can be orders of magnitude of difference between sample size requirements suggested by methods that account for imperfect tests compared with those that naively assume perfect properties [34, 35]. Therefore, it is important to consider the properties of the tests being used to avoid studies that are much too small after all inherent uncertainties are accounted for. Furthermore, correlation among tests may further increase sample size requirements, because a second test may add less information to a first test if the tests are not conditionally independent.

Consider Dendukuri *et al.* [33], who described the sample size requirements for estimating the prevalence of *Strongyloides* infection in a refugee population. There is no gold-standard test available for *Strongyloides* infection, but two routine tests are typically applied, serology and microscopy. Information about the sensitivity, specificity, and disease prevalence was available from an earlier study [3], reproduced here in Table I for convenience. Note the considerable uncertainty in the prevalence estimate, with a 95 per cent credible interval width of approximately 0.4. Dendukuri *et al.* [33] sought to design a study that would reduce this uncertainty to a more acceptable width of 0.1, but found that this level of precision could not be attained even with an infinite sample size. Rather, only a lower precision corresponding to a credible interval width of 0.3 could be achieved with a finite sample, the size being lower when microscopy and serology were used together compared with either test used alone.

While several methods have been proposed when no gold standard is available [28–31], latent class models are the most common choice at the analysis stage, so that it makes sense to use these same models when planning a study and calculating sample size requirements. These models allow for simultaneous estimation of disease prevalence and test properties such as the sensitivity and specificity without naively assuming that one of the tests is a gold standard, resulting in more realistic estimates of all parameters. Under conditional independence, data from three or more binary tests result in an identifiable model, but at least four tests are required if tests are possibly dependent.

When only one or two tests are available, parameters from the resulting non-identifiable problem can be estimated by a Bayesian approach [3]. Here, the data are augmented by the external information in the form of a prior distribution. Unless sufficient prior information is available on a subset of parameters (at least two parameters must have substantive prior distributions for both the one and two test situations), marginal posterior densities can remain wide. Gustafson [36] reviews issues surrounding the identifiability of models for diagnostic tests, concluding that non-identifiable models with small amounts of substantive prior information often outperform simpler identifiable models. This is because it is still possible to learn from data in non-identifiable models.

When the problem is non-identifiable, the joint posterior distribution of the parameters does not converge asymptotically to a single point as the sample size increases. This implies that there is a limiting value to the maximum coverage of fixed width credible intervals and minimum length of fixed coverage credible intervals. Therefore, whether a finite sample size is possible depends on whether the desired coverage and interval widths are equal to or less than the maximum coverage and/or equal to or greater than the minimum widths. From a design perspective, this lack of identifiability implies that very large sample sizes are often required to achieve even moderate accuracy in the parameter estimation [32]. The sample sizes are largely driven by the degree to which test properties are *a priori* known, and in some cases, even an infinite sample size will not lead to the desired accuracy in estimation. The use of a second test helps to reduce the

required sample size in some cases, and sometimes it is possible to achieve the desired precision with a finite sample size using two tests, when the sample size is infinite for any single test (see examples in [33]). The addition of a third conditionally independent test renders the problem identifiable, so that infinite sample sizes should not generally occur, and the number of subjects required for any given accuracy should decrease, at the expense of an additional test per subject. This raises the question: Would the addition of a third test, such as eosinophil counts, to the *Strongyloides* infection study described above have allowed us to achieve a 95 per cent credible interval of length 0.1 with a finite and feasible sample size?

The issues described in the above prevalence study also apply when estimating the properties of a new diagnostic test. Consider sample size determination for a study of the molecular epidemiology of tuberculosis [18]. Select mycobacterial DNA sequences provide clues about which cases of active tuberculosis are likely clustered, implying recent transmission between these cases, versus reactivation of previously acquired infection. The proportion of the recently transmitted cases is important for public health, as different control methods are implemented as transmission rates increase. The standard typing method is IS*6110* Restriction Fragment Length Polymorphism (IS*6110* RFLP), but the recently developed polymerase chain reaction (PCR)-based genotyping modalities, including MIRU-VNTR (mycobacterial interspersed repetitive units-variable number of tandem repeats) and spoligotyping provide quicker results. Investigating the properties of these new tests, however, is rendered difficult by the lack of a gold standard method for classifying cases as clustered or not. As many of these tests are relatively new, their properties have not been extensively investigated. What sample size would be necessary, for example, to learn about the properties of MIRU, if all three tests were to be used in a study?

In this paper, we investigate the impact of the addition of a third test on the sample size of studies designed to estimate either disease prevalence or properties of a diagnostic test. In Section 2, we describe the application of three Bayesian sample size criteria to the problem of designing a study using three diagnostic tests. Section 3 presents a series of prototypic examples, designed to illustrate the degree to which sample size requirements may be decreased when a third test is added compared with a study using only two tests, and also investigates the effect of varying the amount of prior information available. We return to the two applications described above in Section 4, and end with a summary and discussion in Section 5.

2. Bayesian sample size criteria applied to diagnostic studies involving three tests

Bayesian interval-based criteria are ideally suited for design of diagnostic studies in the absence of a gold-standard test, since these methods allow for the specification of prior distributions that account for all inherent uncertainties in the parameters at the planning stage. See [37, 38] for general reviews of Bayesian sample size methods. Below we specify our model in terms of a likelihood function and joint prior distribution over all unknown parameters. These in turn lead to the marginal distribution of the data that will eventually be collected, which will be used for calculating the required sample size.

2.1. Likelihood function and prior and marginal distributions

When three conditionally independent, binary diagnostic tests are available, the likelihood function L of the observed data $x=(x_1, x_2, \ldots, x_N)$ of sample size N can be written in terms of the prevalence, sensitivity, and specificity parameters. Let π denote the prevalence of the condition under study, and S_j and C_j, $j=1,2,3$, denote the sensitivities and specificities of three tests, then we have (as in [3]):

$$L=L(x|\pi,S_1,S_2,S_3,C_1,C_2,C_3)\propto\prod_{i=1}^{N}\left(\pi\prod_{j=1}^{3}S_j^{x_{ij}}(1-S_j)^{1-x_{ij}}+(1-\pi)\prod_{j=1}^{3}C_j^{1-x_{ij}}(1-C_j)^{x_{ij}}\right), \tag{1}$$

where $x_i=(x_{i1},x_{i2},x_{i3})$ is the vector of results on the three tests for the ith subject, such that $x_{ij}=1$ or 0 depending on whether the ith subject had a positive or negative result on the jth test.

Let θ be the vector of unknown parameters $(\pi,S_1,S_2,S_3,C_1,C_2,C_3)$, and let θ belong to the parameter space Θ. Although in theory any joint prior density over Θ can be used, it is convenient to use independent marginal beta densities for each parameter. Indeed, this has been the almost universal choice for models using dichotomous tests in the past [3–5]. Assuming all parameters follow independent beta(α,β) prior distributions, the prior marginal distribution of the data is given by

$$f(x)\propto\int_{\Theta}L\times\pi^{\alpha_\pi-1}(1-\pi)^{\beta_\pi-1}\prod_{j=1}^{3}S_j^{\alpha_{S_j}-1}(1-S_j)^{\beta_{S_j}-1}C_j^{\alpha_{C_j}-1}(1-C_j)^{\beta_{C_j}-1}\,\mathrm{d}\theta, \tag{2}$$

where L is as defined in (1).

Applying Bayes theorem, the marginal posterior distribution of the prevalence is given by

$$f(\pi|x) \propto \int_{\Theta_{-\pi}} L \times \pi^{\alpha_\pi - 1}(1-\pi)^{\beta_\pi - 1} \prod_{j=1}^{3} S_j^{\alpha_{S_j}-1}(1-S_j)^{\beta_{S_j}-1} C_j^{\alpha_{C_j}-1}(1-C_j)^{\beta_{C_j}-1} \, \mathrm{d}\theta_{-\pi}, \tag{3}$$

where $\theta_{-\pi}$ denotes the vector of unknown parameters excluding π, which belongs to the parameter space $\Theta_{-\pi}$. The remaining marginal posterior distributions can be expressed in a similar fashion.

When tests are not conditionally independent, a wide variety of models have been proposed [7–11], using both fixed and random effects to accommodate correlations between tests. For simplicity, we will investigate the effects of conditional dependence on sample size using the fixed-effects model [7]. All the methods described above carry over to this case, except for a change to the likelihood function (1). For example, if tests 1 and 2 are correlated, then two covariance parameters need to be added to the model, say covs_{12} for the sensitivity and covc_{12} for the specificity, and the probabilities of test results that comprise the likelihood function change accordingly. For example, the probability of tests 1 and 2 both being positive given that true disease status is positive changes from S_1S_2 in the independent case to $S_1S_2+\mathrm{covs}_{12}$ in the conditionally dependent case, and the probability of tests 1 and 2 both being negative given a true negative status changes from C_1C_2 to $C_1C_2+\mathrm{covc}_{12}$. Similar changes are needed for each term in (1), see [7] for details.

2.2. *Bayesian sample size criteria*

Typically, we summarize the marginal posterior density of primary interest with a highest posterior density (HPD) or other posterior credible intervals. At the planning stage, we may wish for an interval of length l that covers a particular parameter, say π, with probability $1-\alpha$. The marginal posterior distribution of π depends on the data vector $x \in \mathscr{X}$, which is of course unknown at the planning stages of the experiment. We can eliminate this uncertainty in different ways, leading to the following three criteria [39].

Average Coverage Criterion (*ACC*): Allowing the coverage probability to vary with x while holding the credible interval length l fixed, leads to a sample size defined by the minimum N satisfying

$$\int_{\mathscr{X}} \left\{ \int_{a(x,N)}^{a(x,N)+l} f(\pi|x)\,\mathrm{d}\pi \right\} f(x)\,\mathrm{d}x \geqslant 1-\alpha, \tag{4}$$

where $1-\alpha$ is the required average coverage, $f(x)$ is given by (2), $f(\pi|x)$ is given by (3), and $a(x,N)$ is the lower limit of the HPD interval of length l for the marginal posterior density $f(\pi|x)$.

Average Length Criterion (*ALC*): Conversely, we can allow the HPD interval length to vary while fixing the coverage probability. In this case, for each x in $\mathscr{X}$ we must first find the HPD interval of length $l'(x,N)$ such that $\int_{a(x,N)}^{a(x,N)+l'(x,N)} f(\pi|x)\,\mathrm{d}\pi = 1-\alpha$, and the sample size is the minimum N that satisfies

$$\int_{\mathscr{X}} l'(x,N) f(x)\,\mathrm{d}x \leqslant l, \tag{5}$$

where l is the required average length. The left-hand side of (5) averages the lengths of fixed coverage HPD intervals, weighted by the marginal distribution $f(x)$.

Worst Outcome Criterion (*WOC*): A conservative approach is to ensure a maximum length of l and a minimum coverage probability of $1-\alpha$, regardless of the data x that occur. Thus, we choose the minimum N such that

$$\inf_{x \in \mathscr{X}} \left\{ \int_{a(x,N)}^{a(x,N)+l} f(\pi|x)\,\mathrm{d}\pi \right\} \geqslant 1-\alpha. \tag{6}$$

In practice, there is often at least one data set that leads to very poor accuracy, so that the WOC sample size is infinite. For example, this is always the case when sampling from a Normal distribution [39], and non-identifiable models are also often problematic in this sense. Therefore, in this paper we use the following modified WOC (MWOC) criterion. Rather than taking the infimum across all possible data sets, we guarantee the desired length and coverage over a subset $\mathscr{S} \in \mathscr{X}$ such that $\mathscr{S}$ has a given probability. For example, we might choose the sample size N such that l and $1-\alpha$ are guaranteed over 95 per cent of the set $\mathscr{X}$, according to the marginal distribution (2). We denote this by MWOC(0.95) or more generally, MWOC($1-\gamma$). Thus, we can avoid the situation of having to select an unnecessarily large sample size to guard against improbable data. Other criteria have also been defined, see [38, 40] for recent summaries.

Some authors [38, 41] have distinguished between sampling priors, used for creating the marginal distribution for the data (see equation (2)) and analysis priors, used to derive the posterior distributions once data are available. The usual motivation for this is to use the best available prior information for the marginal distribution of the data in planning the

study, but assume that low information priors will be used at the analysis stage, to 'let the data speak for themselves'. When models may be non-identifiable, however, one must use substantive prior information at the analysis stage to derive reasonable posterior inferences, so that there is less reason to use different prior distributions at the design and analysis phases. Therefore, throughout this paper, we assume that the sampling prior is equal to the analysis prior, although it is straightforward to extend these methods to accommodate different sampling and analysis prior distributions if desired.

The integration required to find the final sample size is non-trivial, since one needs to integrate not only over the parameter space, as in standard problems in Bayesian inference, but also over the sample space $\mathscr{X}$ of the three test results. The numerical techniques and algorithms we used are described in the Appendix.

A user-friendly program that implements the sample size methods discussed in this paper is available from http://www.medicine.mcgill.ca/epidemiology/Joseph/. Similar software for analysis of data from diagnostic tests in the absence of a gold standard test is also available from this source.

3. Variations of sample size requirements across prototypic scenarios

We now illustrate our methods via two sets of sample size calculations, discussed in Sections 3.1 and 3.2, respectively. In Section 3.1, we investigate the sample size reductions that can occur when estimating the prevalence of a condition if three rather than two conditionally independent tests are used. We investigate two opposing situations, where the third test has either better or poorer sensitivity and specificity compared with the first two tests. As a robustness check to the assumption of conditional independence, we investigate how the sample sizes may change when the second and third tests are weakly (correlation $\rho=0.1$), moderately ($\rho=0.25$) or strongly ($\rho=0.5$) correlated. Throughout, we assumed that the correlations apply equally to positive and negative cases, or, in the notation of Section 2, $\text{covs}_{23}=\text{covc}_{23}$. Note that the degree of possible correlation depends on the test properties [7], and 0.5 is close to the highest possible correlation in our scenarios.

Section 3.2 investigates sample size requirements for diagnostic test studies, where the properties of a new test rather than the prevalence are of primary interest. Using a uniform prior over the sensitivity and specificity of the new test, we again compare sample sizes from three tests with those from just two tests across differing amounts of prior information, and we again check the robustness to the assumption of conditional independence using the same choice of correlation parameters as described above.

Throughout, regardless of the parameter of primary interest, we set the desired precision and coverage for the different sample size criteria to commonly used values $1-\alpha=0.95$ and $l=0.1$, respectively. There are too many parameters and choices of prior distributions to cover all possibilities that may arise in practice, but our results are representative of situations that commonly occur.

3.1. Prevalence studies: comparing sample size estimates when three rather than two diagnostic tests are used

We denote the three diagnostic tests by X1, X2, and X3, and our main interest here is to compare sample sizes when X1 and X2 are used alone with the case when X3 is added.

Table II contains the results from 18 different scenarios, distinguished by the prior distributions used for the prevalence and sensitivities and specificities of X1, X2, and X3, and the sample size criterion used. Note that the Low (L), Moderate (M), and High (H) labels refer to the location of the mean for each prior distribution for the prevalence, sensitivity or specificity, and not to the degree of uncertainty in these prior distributions. As discussed in [33], both the location and the degree of uncertainty in the prior distributions can impact the sample size requirements. In low prevalence situations across all criteria (first nine lines of Table II), the reductions in sample size requirements when a third test is added are variable. When the third test has better properties compared with the first two tests, sample size reductions can be very large, over 95 per cent in some cases, and at least 66 per cent across all scenarios investigated. Adding a third test that is not as well performing compared with the two already in use results in only very modest or no reductions in the sample size.

Under higher prevalence conditions, when the test properties remain better for the third test, infinite sizes for two tests are reduced to non-infinite sizes in all the cases we simulated, with sample sizes for three tests falling well below 1000 in many cases. Adding a third test that is not as well performing decreases the sample sizes by smaller but still meaningful amounts. Theoretically, a main factor driving the sample size is how close the prevalence is to 50 per cent, where binomial variance is maximized and each subject contributes the minimum amount of information.

Assuming conditional independence among all tests, it is clear that the addition of a third well performing test is highly desirable in diagnostic testing studies where the prevalence is to be estimated. This addition results in very large decrease in sample size requirements, in many cases avoiding the infinite sample sizes that arise when using just two tests. When the third test has poorer properties compared with the two tests already in use, much smaller but still often substantial

Table II. Comparing sample size requirements when two or three conditionally independent diagnostic tests are available for estimating the prevalence of a disease or condition.

	Prior distributions							Sample size	
Sample size criterion	π	S_1	C_1	S_2	C_2	S_3	C_3	Two tests	Three tests
ALC	L	M	M	M	M	H	H	3368	222
ALC	L	M	M	M	M	H	M	3368	502
ALC	L	H	H	H	H	M	M	132	133
ACC	L	M	M	M	M	H	H	5302	264
ACC	L	M	M	M	M	H	M	5302	583
ACC	L	H	H	H	H	M	M	160	156
MWOC(0.95)	L	M	M	M	M	H	H	∞	627
MWOC(0.95)	L	M	M	M	M	H	M	∞	1732
MWOC(0.95)	L	H	H	H	H	M	M	338	316
ALC	M	M	M	M	M	H	H	∞	654
ALC	M	M	M	M	M	H	M	∞	2305
ALC	M	H	H	H	H	M	M	463	397
ACC	M	M	M	M	M	H	H	∞	666
ACC	M	M	M	M	M	H	M	∞	2455
ACC	M	H	H	H	H	M	M	466	402
MWOC(0.95)	M	M	M	M	M	H	H	∞	1177
MWOC(0.95)	M	M	M	M	M	H	M	∞	7700
MWOC(0.95)	M	H	H	H	H	M	M	682	505

Prior distribution for the prevalence (π) is either Low (L = beta(2.5, 22.5), with 95 per cent prior credible interval (CrI) = (0.02, 0.24)) or Moderate (M = beta(36.05, 54.53), with 95 per cent CrI = (0.3, 0.5)). Prior distribution for the sensitivities (S_i) and specificities (C_i) of each test $i = 1, 2, 3$ are either Moderate (M = beta(55.21, 22.11), with 95 per cent CrI = (0.6, 0.8)) or High (H = beta(116.06, 12.05), with 95 per cent CrI = (0.85, 0.95)).

reductions in sample size requirements can generally be expected, particularly as the prevalence approaches 0.5. Using three tests will result in finite sample sizes, except if one or more of the tests perform no better than chance. For example, if a test has sensitivity = specificity = 50 per cent, then it will provide no information about the prevalence regardless of the sample size, since a positive test can be a true or a false positive result with equal probability.

When two of the three tests are correlated the required ALC and ACC sample sizes rise, but remain reasonable throughout the cases we investigated. For example, the ALC sample size from the first line of Table II is 222 with three independent tests, but rises to 242, 280, and 249, respectively, for small, moderate, and large correlations between tests 2 and 3. Similarly, the ACC sample size for the same choice of prior distributions and independent tests is 264 (line 4 from Table II), and this rises to 304, 338, and 318, respectively, for small, moderate, and large correlations between tests 2 and 3. The MWOC(0.95) sample sizes however, were much larger when correlations were added, approaching infinity. This occurs because a small minority of cases with poor results has a much larger effect on the MWOC sample sizes compared with sizes from criteria that do not search for the worst possible outcomes.

3.2. New diagnostic test studies: comparing sample size estimates when three rather than two diagnostic tests are used

Table III contains the results from nine different scenarios for comparing the sample sizes that result from three compared to two tests when estimating the sensitivity of a test to detect a given condition or disease. Throughout, we used uniform prior distributions for X1, which we assumed as the new test under study, whose sensitivity is to be estimated by the study being designed. We assumed a uniform prevalence, in other words, we assumed the researchers would not know the prevalence of the condition in their test subjects, which may occur, for example, if they were volunteers with an unclear history. Tests X2 and X3 were assumed to be routinely used but imperfect diagnostic tests whose properties are known to within a certain accuracy. For X2 and X3, we used various different prior distributions, with properties of X2 sometimes poorer than those of X3, and vice versa.

Under conditional independence, we found that when using two tests it was not possible to satisfy the criteria with a finite sample size in all but one case. With the addition of a third test all criteria were satisfied with a finite size, but large sample sizes occurred particularly for the MWOC criteria. When the third test had high sensitivity and specificity the sample size required was at least 50 per cent less compared with when the specificity alone or both sensitivity and specificity were moderate. Of course, from the symmetry of the problem similar results would be expected for estimating the specificity.

When tests two and three are correlated, sample sizes rise very substantially. For example, when the second test has moderate and the third test has high properties, sample sizes were 2120, 4777, and 30 000 for the ALC, ACC, and MWOC(0.95), respectively, as seen in Table III. However, regardless of the degree of correlation between tests, the sample

Table III. Comparing sample size requirements when two or three conditionally independent diagnostic tests are available for estimating the sensitivity of a diagnostic test for a given disease or condition.

	Prior distributions							Sample size	
Sample size criterion	π	S_1	C_1	S_2	C_2	S_3	C_3	Two tests	Three tests
ALC	U	U	U	M	M	H	H	∞	2120
ALC	U	U	U	M	M	H	M	∞	4233
ALC	U	U	U	H	H	M	M	31 002	2120
ACC	U	U	U	M	M	H	H	∞	4777
ACC	U	U	U	M	M	H	M	∞	11 236
ACC	U	U	U	H	H	M	M	∞	4777
MWOC(0.95)	U	U	U	M	M	H	H	∞	30 000
MWOC(0.95)	U	U	U	M	M	H	M	∞	67 500
MWOC(0.95)	U	U	U	H	H	M	M	∞	30 000

Prior distribution for the sensitivities (S_i) and specificities (C_i) of each test $i = 2, 3$ can be Moderate (M = beta(55.21, 22.11), with 95 per cent CrI = (0.6, 0.8)) or High (H = beta(116.06, 12.05), with 95 per cent CrI = (0.85, 0.95)). U indicates a uniform prior density over the interval [0, 1], which was used for the test of interest $i = 1$ and the prevalence of the condition, assumed unknown within the testing population.

sizes rose close to 10 000 for the ALC and approached infinity for both the ACC and MWOC. Therefore, when tests are correlated, one needs extremely large sizes to obtain 95 per cent posterior intervals of width 0.1 or less for the sensitivity.

Overall, we found higher sample sizes for estimating test properties compared with the lower sizes required for similar accuracy in estimating the prevalence. This is in part explained by the fact that every subject contributes toward estimating prevalence, but only positive subjects contribute toward estimating the sensitivity (and only negative subjects contribute toward estimating the specificity). Therefore, when prevalence is high, one should need smaller sample sizes for estimating the sensitivity. To check this, we used the same prior choices as in line 1 of Table III, but changed the prior distribution of the prevalence to center at 70 per cent, using a beta(70,30) distribution. Assuming three independent tests, we find an ALC sample size of 416, which rises only slightly to 427, 433, and 435, respectively, for small, medium and large correlations between tests 2 and 3, similar to the results found for the prevalence. In fact, even two independent tests are sufficient in this case, giving a sample size of 440, only slightly larger than the three test sample sizes. This is because the high prevalence ensures that most subjects are positive, so that the properties of the new test can be efficiently evaluated. The key message here is that if test properties are the target of the investigation, it is best to choose a group of subjects whose prevalence is well known.

In the following section, we illustrate the use of our methods in practice through the two motivating examples introduced in Section 1.

4. Sample size for studies of *Strongyloides* prevalence and tuberculosis transmission tests

Section 4.1 discusses designing a study to estimate the prevalence of *Strongyloides* infection, and Section 4.2 concerns planning a study to accurately estimate the sensitivities and specificities of new tests for detecting tuberculosis transmission. In each case, we will compare the sample sizes from a design using two diagnostic tests to a study that adds a third test. For example, for *Strongyloides* infection the third test may be based on eosinophil counts, and hence the three tests may be assumed conditionally independent, at least approximately.

4.1. Planning a study to estimate the prevalence of Strongyloides infection

As discussed in [33], when using the prior distributions in Table I, the sample sizes for two tests are infinite regardless of the criterion used if we desire a posterior credible interval coverage of 95 per cent to be of length $l = 0.1$ or smaller. We now assume that a third test is available, with sensitivity and specificity with prior 95 per cent credible interval range of (0.6, 0.8), in other words, the 'moderate' test as defined in Tables II and III. Assuming that a study will use all three tests to estimate the prevalence, to attain the desired accuracy would require sample sizes of 5660, 5818, and 24 038 for the ALC, ACC, and MWOC(0.95), respectively.

These results show that a study that was impossible to carry out with only two tests becomes feasible if one can find a third test, even if it has only moderate to good properties. Nevertheless, a relatively high sample size of over 5000 subjects is needed to guarantee estimation accuracy of ± 0.05 on average when using a 95 per cent credible interval, and almost five times that number is required to guarantee the desired interval width and coverage using the MWOC criterion. These sizes are much larger than would be suggested by naive use of a binomial sample size criterion that

ignores the inevitable errors in the test results. For example, attaining an accuracy of ± 0.05 with a 95 per cent confidence interval requires a sample size of only 384 subjects or fewer, depending on the assumed prevalence [42]. This comparison underlines the importance of carefully assessing the test properties in any study involving diagnostic tests.

4.2. Planning a study to estimate the sensitivity of PCR-based genotyping modalities for the detection of tuberculosis transmission

Based on the results of Scott *et al.* [18], we estimate that the sensitivity of RFLP is in the range (0.19, 0.39), and the specificity ranges from (0.81, 0.95). Similarly, for spoligotyping the sensitivity is in the range (0.78, 0.99), whereas the range for the specificity is (0.44, 0.98). We converted these ranges into beta prior inputs, by assuming the mid-point of the range to be the mean, and taking the standard deviation to be one fourth of the range. Using uniform prior inputs for the prevalence of clustering and the sensitivity and specificity of MIRU, we find sample sizes of infinity across all criteria for a two-test design, which uses the standard RFLP and MIRU. However, if all the three tests are used, the sample sizes become 17 300, 51 000, and infinity, for the ALC, ACC, and MWOC(0.95), respectively.

In this case, it is not clear that adding a third test renders the problem feasible in practice, even if the test is conditionally independent from the first two tests. While the sample sizes are no longer infinite, they remain very large, and it is unlikely that any study will be able to find sufficient numbers for an accurate estimation, unless a worldwide collaborative effort is made. This is especially true since two simplifying assumptions are made here, neither of which may hold in practice. First, it is possible that MIRU and spoligotyping may be conditionally dependent, as both rely on PCR-based methods, albeit on different regions of the genome, rendering any correlation small or zero. Second, the model used here is really a simplification to reality, since we assumed clustering of each case occurs independently from other cases, which may not be true since one case may have infected another, implying that a complex infectious disease model with many parameters may be required for accurate modeling. Overall then, one must either admit that accurate estimation of the properties of MIRU is not possible, or, more constructively, one must first plan studies to more accurately estimate the properties of RFLP and spoligotyping. If the prior information about these two tests could be sufficiently sharpened, much lower sample sizes would be achieved.

5. Discussion

In this paper, we have described several methods for sample size calculations when three conditionally independent diagnostic tests are available, which leads to an identifiable model. These methods generally provide finite sample sizes, as compared with the often infinite sample sizes given by similar criteria for the non-identifiable models that result when only one or two diagnostic tests are available. When tests may be conditionally dependent, however, sample sizes may again increase substantially. As discussed by Johnson *et al.* [5], another route toward an identifiable model when only two tests are available is adding a second population with a different prevalence. It would be interesting to compare sample size requirements for this scenario with the three-test situation, although that is not done here.

In practice, most diagnostic studies currently use naive sample size calculations that ignore the imperfect sensitivity and specificity of virtually all diagnostic tests. We have shown this to be a serious problem, as either the estimates are likely to be biased, if the imperfections of the tests used are ignored at the analysis stage as well as the planning stage, or the final posterior credible intervals will be very wide, if one accounts for the imperfect tests at the analysis stage only.

Using three diagnostic tests rather than one or two adds to the cost per subject of a study, while allowing for fewer subjects overall. Our methods and user-friendly software allow study planners to accurately assess the accuracy gained by adding more tests, following which a final decision about the number of tests and sample size can be made. We expect, however, that many researchers may be surprised by the generally high sample sizes required by diagnostic test studies once all uncertainties are included in the model. We hope that our work will encourage researchers to carefully consider all design issues, and evaluate the tests that will be used in their studies, as the best route toward smaller sample sizes is the use of tests whose properties are more accurately known.

Appendix A

A software package which implements the methods described in Section 2 of this paper called PropMisclassSampleSize is available from the author's web page www.medicine.mcgill.ca/epidemiology/Joseph/. This program calculates sample sizes for one, two or three conditionally independent diagnostic tests using the Bayesian criteria. This appendix describes the numerical algorithms the software package uses to calculate the sample sizes. While we present algorithms for

estimating the prevalence, very similar algorithms can be used for estimating sample sizes for the sensitivity and specificity of any test. As detailed in Section 2 above, only small changes to the likelihood function are required to account for possibly conditionally dependent tests.

(1) Sample M_1 random values from the joint prior distribution of $(\pi, S_1, C_1, S_2, C_2, S_3, C_3)$.
(2) For each vector $(\pi_i, S_{1i}, C_{1i}, S_{2i}, C_{2i}, S_{3i}, C_{3i}), i=1, \ldots, M_1$:

   (a) The three tests define a multinomial probability function with eight categories, as each test can independently be positive or negative, given either of two true disease states.
   (b) Select a value for N, the required sample size. From the multinomial distribution from the previous step, draw M_2 random values such that the total number in the eight cells equals N. This is equivalent to sampling from the preposterior predictive distribution of the data.

(3) For each of these multinomial vectors, estimate the posterior density of π. This is done as follows:

   (a) For each of the sampled vectors, obtain a sample of M_3 values from the posterior distribution of π using the Gibbs sampler [3]. Label these values π_{ijk}, $i=1, \ldots, M_1$, $j=1, \ldots, M_2$, $k=1, \ldots, M_3$.
   (b) Estimate the mean (μ_{ij}) and variance (σ^2_{ij}) of the posterior distribution of π_{ij} using the output from this Gibbs sampler.
   (c) The posterior distribution of π_{ij} is approximated by a single Beta distribution with parameters $\alpha_{ij} = -\mu_{ij}(\sigma^2_{ij}+\mu^2_{ij}-\mu_{ij})/\sigma^2_{ij}$ and $\beta_{ij}=(\mu_{ij}-1)(\sigma^2_{ij}+\mu^2_{ij}-\mu_{ij})/\sigma^2_{ij}$.

(4) For each posterior distribution, we used a Newton–Raphson-type algorithm to find the location of the HPD interval. This involved choosing a lower limit for the interval, say a, calculating the height of the density curve for π at a and $a+l$, and iterating until $f(a)=f(a+l)$. Coverages were then given by the area under the curve between a and $a+l$, using standard results from the beta density.
(5) To implement the ACC criterion, compare the average coverage of HPD intervals of length l to the predetermined value of $1-\alpha$. If the average coverage is greater (smaller) than $1-\alpha$ we return to Step 1 and repeat the algorithm with a smaller (greater) value for N until the criterion is met. Similarly, to implement the ALC criterion the average length of the HPD intervals with coverage $1-\alpha$ is compared with l. To implement the MWOC$(1-\gamma)$ criterion we compare the $(1-\gamma)\times 100$ percentile of the coverages to $1-\alpha$.

For sample sizes $N_1, N_2, \ldots, N_T$ covering a range near the correct sample size, we generated coverages (c_i) and lengths (l_i) using the above algorithms. We then fit the quadratic model $\log(l_i \text{ or } c_i)=\alpha+\beta_1\log(N_i)+\beta_2\{\log(N_i)\}^2$ to the points (N_i, l_i) or (N_i, c_i), for the ALC and ACC, respectively. This curve allows us to quickly zero in on the required sample size, which is defined as the smallest N for which the given criterion is satisfied.

Increasing the values of M_1, M_3, and M_4 increases the precision of the sample size estimate, but increasing M_2 while keeping M_1, M_3, and M_4 fixed has little effect on the precision. If the required coverage or length criterion was not met by a size of N=100 000, we stopped our search and reported a sample size of infinity. As studies of this dimension are very rare, for all practical purposes, the desired accuracy cannot be reached.

References

1. Walter S, Irwig L. Estimation of error rates, disease prevalence and relative risks from misclassified data: a review. *Journal of Clinical Epidemiology* 1988; **41**:923–937.
2. Espeland M, Handelman S. Using latent class models to characterize and assess relative error in discrete measurements. *Biometrics* 1989; **45**:587–599.
3. 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.
4. Demissie K, White N, Joseph L, Ernst P. Bayesian estimation of asthma prevalence, and comparison of exercise and questionnaire diagnostics in the absence of a gold standard. *Annals of Epidemiology* 1998; **8**:201–208.
5. Johnson W, Gastwirth J, Pearson L. Screening without a 'gold standard': the Hui–Walter paradigm revisited. *American Journal of Epidemiology* 2001; **153**:921–924.
6. Qu Y, Tan M, Kutner M. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics* 1996; **52**:797–810.
7. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics* 2000; **57**:208–217.
8. Black M, Craig B. Estimating disease prevalence in the absence of a gold standard. *Statistics in Medicine* 2002; **21**:2653–2669.
9. Branscum S, Gardner I, Johnson W. Estimation of diagnostic-test sensitivity and specificity through bayesian modeling. *Preventive Veterinary Medicine* 2005; **68**:145–163.
10. Menten J, Boelaert M, Lesaffre E. Bayesian latent class models with conditionally dependent diagnostic tests: a case study. *Statistics in Medicine* 2008; **27**:4469–4488.

11. Dendukuri N, Hadgu A, Wang L. Modeling conditional dependence between diagnostic tests: a multiple latent variable model. *Statistics in Medicine* 2009; **28**:441–461.
12. Weber MF, Verhoeff J, van Schaik G, van Maanen C. Evaluation of Ziehl–Neelsen stained faecal smear and ELISA as tools for surveillance of clinical paratuberculosis in cattle in the netherlands. *Preventive Veterinary Medicine* 2009; **92**:256–266.
13. Monti G, Frankena K, Engel B, Buist W, Tarabla H, de Jong M. Evaluation of a new antibody-based enzyme-linked immunosorbent assay for the detection of bovine leukemia virus infection in dairy cattle. *Journal of Veterinary Diagnostic Investigation* 2005; **17**:451–457.
14. Kass R, Raftery A. Bayes factors. *Journal of the American Statistical Association* 2005; **90**:773–795.
15. Bernatsky S, Joseph L, Bélisle P, Boivin J, Rajan R, Moore A, Clarke A. A Bayesian hierarchical model for estimating the properties of cancer ascertainment methods in cohort studies. *Statistics in Medicine* 2005; **24**:2365–2379.
16. Carabin H, Marshall C, Joseph L, Riley S, Olveda R, McGarvey S. Estimating and modelling the dynamics of the intensity of infection with Schistosoma japonicum in villagers of Leyte, Philippines. Part I: a Bayesian cumulative logit model. *American Journal of Tropical Medicine and Hygiene* 2005; **72**:745–753.
17. Moayyedi P, Duffy J, Delaney B. New approaches to enhance the accuracy of the diagnosis of reflux disease. *Gut* 2004; **53**:iv55–iv57.
18. Scott A, Joseph L, Bélisle P, Behr M, Schwartzman K. Bayesian estimation of tuberculosis clustering rates from DNA sequence data. *Statistics in Medicine* 2008; **27**:140–156.
19. Weichenthal S, Joseph L, Bélisle P, Dufresne A. Bayesian estimation of the probability of asbestos exposure from lung fibre counts. *Biometrics* 2010; **66**(2):603–612. DOI: 10.1111/j.1541-0420.2009.01279.x.
20. Tarafder M, Carabin H, Joseph L, Balolong E, Olveda R, McGarvey S. Estimating the sensitivity and specificity of Kato–Katz stool examination technique for detection of hookworms, *Ascaris lumbricoides*, and *Trichuris trichiura* infections in humans in the absence of a gold standard. *International Journal for Parasitology* 2010; **40**:399–404.
21. Ladouceur M, Rahme E, Pineau C, Joseph L. Robustness of prevalence estimates derived from misclassified data from administrative databases. *Biometrics* 2007; **63**:272–279.
22. Rose N, Boutrouille A, Fablet C, Madec F, Eloit M, Pavio N. The use of Bayesian methods for evaluating the performance of a virus-like particles-based ELISA for serology of Hepatitis E virus infection in swine. *Journal of Virological Methods* 2010; **163**:329–335.
23. Benitoa A, Carmena D, Joseph L, Martineza J, Guisantesa J. Dog echinococcosis in northern Spain: comparison of coproantigen and serum antibody assays with coprological exam. *Veterinary Parasitology* 2006; **142**:102–111.
24. 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.
25. 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.
26. Boelaert M, El-Safi S, Hailu A, Mukhtar M, Rijal S, Sundar S, Wasunna M, Aseffa A, Mbui J, Menten J, Desjeux P, Peeling RW. 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.
27. Kostoulas P, Leontides L, Ene C, Billinis C, Florou 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.
28. Alonzo T, Pepe M. Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Statistics in Medicine* 1999; **18**:2987–3003.
29. Pepe M. *The Statistical Evaluation of Medical Tests For Classification and Prediction*. Oxford University Press: Oxford, 2003.
30. Gustafson P. *Measurement Error and Misclassification in Statistics and Epidemiology*: *Impacts and Bayesian Adjustments*. Chapman & Hall: New York, 2003.
31. Broemelling L. *Bayesian Biostatistics and Diagnostic Medicine*. Wiley: New York, 2007.
32. Rahme E, Joseph L, Gyorkos T. Bayesian sample size determination for estimating binomial parameters from data subject to misclassification. *Applied Statistics* 2000; **49**:119–128.
33. Dendukuri N, Rahme E, Bélisle P, Joseph L. Bayesian sample size determination for prevalence and diagnostic studies in the absence of a gold standard test. *Biometrics* 2004; **60**:388–397.
34. Arkin CF, Wachtel MS. How many patients are necessary to assess test performance? *JAMA* 1990; **263**:275–278.
35. 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.
36. Gustafson P. On model expansion, model contraction, identifiability, and prior information: two illustrative scenarios involving mismeasured variables (with Discussion). *Statistical Science* 2005; **20**:111–140.
37. Adcock CJ. Sample size determination: a review. *The Statistician* 1997; **46**:261–283.
38. Wang F, Gelfand AE. A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science* 2002; **17**:193–208.
39. Joseph L, Bélisle P. Bayesian sample size determination for normal means and differences between normal means. *The Statistician* 1997; **46**:209–226.
40. M'Lan C, Joseph L, Wolfson D. Bayesian sample size determination for case-control studies. *Journal of the American Statistical Association* 2006; **101**:760–772.
41. Joseph L, du Berger R, Bélisle P. Bayesian and mixed Bayesian/likelihood criteria for sample size determination. *Statistics in Medicine* 1997; **16**:769–781.
42. Desu M, Raghavarao D. *Sample Size Methodology*. Academic Press: Boston, 1990.