Biometrics 60, 388–397
June 2004

Bayesian Sample Size Determination for Prevalence and Diagnostic Test Studies in the Absence of a Gold Standard Test

Nandini Dendukuri,[1,2,*] Elham Rahme,[3,] Patrick Bélisle,[3,***] and Lawrence Joseph[1,3,****]**

[1]Department of Epidemiology and Biostatistics, 1020 Pine Avenue West, McGill University, Montreal, Québec H3A 1A2, Canada
[2]Technology Assessment Unit, Royal Victoria Hospital, R4.14 Ross Pavilion, 687 Pine Avenue West, Montreal, Québec H3A 1A3, Canada
[3]Division of Clinical Epidemiology, Department of Medicine, Montreal General Hospital, 1650 Cedar Avenue, Montreal, Québec H3G 1A4, Canada
[*]*email:* nandini.dendukuri@mcgill.ca
[**]*email:* elham.rahme@mcgill.ca
[***]*email:* belisle@epimgh.mcgill.ca
[****]*email:* lawrence.joseph@mcgill.ca

Summary. Planning studies involving diagnostic tests is complicated by the fact that virtually no test provides perfectly accurate results. The misclassification induced by imperfect sensitivities and specificities of diagnostic tests must be taken into account, whether the primary goal of the study is to estimate the prevalence of a disease in a population or to investigate the properties of a new diagnostic test. Previous work on sample size requirements for estimating the prevalence of disease in the case of a single imperfect test showed very large discrepancies in size when compared to methods that assume a perfect test. In this article we extend these methods to include two conditionally independent imperfect tests, and apply several different criteria for Bayesian sample size determination to the design of such studies. We consider both disease prevalence studies and studies designed to estimate the sensitivity and specificity of diagnostic tests. As the problem is typically nonidentifiable, we investigate the limits on the accuracy of parameter estimation as the sample size approaches infinity. Through two examples from infectious diseases, we illustrate the changes in sample sizes that arise when two tests are applied to individuals in a study rather than a single test. Although smaller sample sizes are often found in the two-test situation, they can still be prohibitively large unless accurate information is available about the sensitivities and specificities of the tests being used.

Key words: Bayesian design; Diagnostic test; Misclassification; Prevalence; Sample size; Sensitivity; Specificity.

1. Introduction

Consider designing a study to estimate the prevalence of *Strongyloides* infection among Cambodian refugees (Joseph, Gyorkos, and Coupal, 1995a). How large a sample is required to accurately estimate the prevalence of infection with a diagnostic test? If a gold standard (or error-free) diagnostic test is used to detect *Strongyloides*, the prevalence can simply be estimated as the proportion of positive responses on the test. Accordingly, sample size calculations can be based on methods for a single proportion (Lemeshow et al., 1990), with variance based on the binomial distribution. For example, if the goal of the study is to estimate a $(1-\alpha)\%$ confidence interval of length l for the prevalence, π, the required sample size, N, may be calculated as

$$N = \frac{4Z^2_{1-\alpha/2}\pi(1-\pi)}{l^2}, \qquad (1)$$

where $Z_{1-\alpha/2}$ is the $1-\alpha/2$ critical value of the standard normal distribution. In order to implement formula (1) a point estimate of π must be provided, which can be obtained from earlier studies or fixed at a conservative estimate of 0.5. An analogous situation arises in diagnostic accuracy studies where the sensitivity (probability of a positive result among truly positive subjects) and specificity (probability that a truly negative subject tests negatively) of a new diagnostic test are to be estimated. If a perfect gold standard test is available, equation (1) can be implemented separately for the number of truly positive and truly negative subjects required.

This idealized situation almost never occurs in practice. Imperfect tests are frequently used in diagnostic studies because an error-free test does not exist or its use is not feasible. Let T be the observed result on a dichotomous diagnostic test and D be the true disease status. Denoting test sensitivity by

$S = P(T = 1 \mid D = 1)$ and test specificity by $C = P(T = 0 \mid D = 0)$, the probability of obtaining a positive test is $p = \pi S + (1 - \pi)(1 - C)$. Thus, estimating a confidence interval of length l for π is equivalent to estimating a confidence interval of length $|l(S + C - 1)|$ for p. Expression (1) can be modified to account for this misclassification, giving

$$N = \frac{4Z^2_{1-\alpha/2}p(1-p)}{(l(S+C-1))^2}. \tag{2}$$

There are at least two serious drawbacks to using equations (1) and (2). First, they require exact point estimates of all parameters involved, while only ranges of possible values for these parameters are usually available from previous studies. Second, in estimating a sample size for π, equation (2) assumes that the binomial distribution will be used in subsequent estimation. This can induce very large biases and confidence intervals that are much too narrow when S and C are not known exactly, as discussed by Joseph et al. (1995a).

In fact, two imperfect tests for detecting *Strongyloides*, a serology test which looks for antibodies in the blood, and microscopy, which directly looks for the nematode in a stool sample, are available. Table 1 summarizes the marginal posterior densities from an earlier study (Joseph et al., 1995) which included $N = 162$ subjects. The width of the 95% credible interval for the prevalence was almost $l = 0.4$, suggesting the need for further study, but how many further subjects should be tested? No sample size methods have appeared in the literature to date for the case of two imperfect tests. For a single gold standard test, equation (1) with the posterior median value of $\pi = 0.76$ and $l = 0.1$ gives $N = 70$, and equation (2) with $S = 0.89$ and $C = 0.67$, the posterior median values for the serology test, gives $N = 168$. As we will show, neither of these sample sizes even approach being adequate, even if two tests are applied, once all uncertainty inherent in the problem is accounted for. Further, π could be better estimated if the test properties were better known, but how should one design a study to estimate test properties in the absence of a gold standard?

Sample size determination for diagnostic studies based on confidence intervals or hypothesis tests has been discussed by Arkin and Wachtel (1990), Lemeshow et al. (1990), Buderer (1996), and Alonzo, Pepe, and Moskowitz (2002). Frequentist methods have also been proposed for efficient study designs to assess diagnostic accuracy (Irwig et al., 1994), for sample size determination of studies aimed at estimating the sensitivity at a given false positive rate (Obuchowski and McClish, 1997), or estimating the area under the receiver operating characteristic curve (Hanley and McNeil, 1982). These and other methods are summarized by Pepe (2003).

Several authors (including Adcock, 1988; Joseph, Wolfson, and du Berger, 1995b; Joseph, du Berger, and Bélisle, 1997) have proposed Bayesian criteria for sample size determination based on posterior variances or credible interval widths. Criteria have also been proposed for studies where the goal is to maximize an expected utility function (Gittins and Pezeshk, 2000). While attractive in theory, these loss function-based approaches are often problematic in practice (Joseph and Wolfson, 1997). Here we focus on Bayesian criteria based on interval widths, recently reviewed by Adcock (1997) and Wang and Gelfand (2002). The criteria we use are summarized in Section 2.

A Bayesian sample size method for prevalence studies using a single non–gold standard diagnostic test was discussed by Rahme, Joseph, and Gyorkos (2000). They showed that prior uncertainty about the sensitivity or specificity of a test can lead to a much larger sample size, compared to sample sizes from formulae (1) or (2). They also showed that the nonidentifiable nature of this problem results in a "plateauing" of the average coverage of posterior credible intervals with increasing sample size. Therefore, in some situations the required coverage may not be attained even with an infinite sample size. Inferences for similar nonidentifiable problems have recently been discussed by Gustafson, Le, and Saskin (2001).

In this article we extend the methods of Rahme et al. (2000) in four directions. First, Section 3 investigates the sample size problem when two conditionally independent binary diagnostic tests are available. Conditional independence implies that the two tests are statistically independent conditional on true disease status, and while it is commonly assumed to hold, it is difficult to verify in practice. As we will demonstrate, the addition of a second non–gold standard test can substantially decrease the sample size required for a given accuracy, although unless test properties are accurately known, very large sample sizes can still result. Second, Section 4 discusses the design of studies whose primary goal is the estimation of test properties rather than prevalence. Third, we employ three different sample size criteria, rather than the single criterion previously discussed, and show that the choice of criterion can have a large impact on the sample size selected. Finally, we examine what happens to the plateau identified by Rahme et al. (2000) when a second test is added.

2. Bayesian Sample Size Criteria

Let $\theta = (\theta_1, \theta_2, \ldots, \theta_m)$ denote the unknown parameters in the study. For example, if we are evaluating two diagnostic

Table 1
Prior distributions for prevalence of Strongyloides and sensitivity and specificity of the serology and microscopy tests

Test	Parameter	Prior median (95% credible interval)	Beta (α, β) priors
	Prevalence (π)	0.76 (0.52, 0.91)	Beta (13.11, 4.59)
Serology test	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)

tests, then we may have $m = 5$ parameters, including the prevalence of the condition in the population and the sensitivities and specificities of each of the two tests. Let Θ denote the parameter space for θ, and let $f(\theta)$ represent the joint prior distribution over Θ. The experiment provides data $x = (x_1, x_2, \ldots, x_N) \in \mathcal{X}$, where N is the sample size, and the possibly vector valued components of x represent the data contributed by each subject. For example, $x_i = (x_{i1}, x_{i2})$ may denote the test results from two diagnostic tests, with $x_{ij} = 1$ or 0, depending on whether the jth diagnostic test ($j = 1, 2$) was positive or negative for the ith subject ($i = 1, \ldots, N$).

The preposterior predictive distribution of x is given by

$$f(x) = \int_{\Theta} f(x \mid \theta) f(\theta)\, d\theta, \tag{3}$$

and the posterior distribution of θ given x is $f(\theta \mid x) = f(x \mid \theta) f(\theta)/f(x)$, where $f(x \mid \theta)$ is the likelihood of the data x. Define $\theta_{-k} = (\theta_1, \theta_2, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_m)$, which takes values in Θ_{-k}. The marginal posterior distribution of the kth component of θ, $k = 1, 2, \ldots, m$, is

$$f(\theta_k \mid x) = \int_{\Theta_{-k}} f(\theta \mid x)\, d\theta_{-k}. \tag{4}$$

Typically, we summarize the marginal posterior density of primary interest with a highest posterior density (HPD) or other posterior credible interval. HPD intervals are the shortest possible intervals for any given coverage probability (Box and Tiao, 1973). At the planning stage, we wish for an interval of length l that covers a particular θ_k, $k = 1, 2, \ldots, m$ with probability $1 - \alpha$. The marginal posterior distribution of θ_k depends on the data vector 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.

Average coverage criterion (ACC): Allowing the coverage probability $1 - \alpha$ to vary with x while holding the credible interval length, l, fixed leads to a sample size defined by the minimum N satisfying $\int_{\mathcal{X}} \{\int_{a(x,N)}^{a(x,N)+l} f(\theta_k \mid x)\, d\theta_k\} f(x)\, dx \geq 1 - \alpha$. Here $f(x)$ is given by (3), $f(\theta_k \mid x)$ is given by (4), and $a(x, N)$ is the lower limit of the HPD interval of length l for the marginal posterior density $f(\theta_k \mid x)$, which in general depends on both x and N.

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 $\mathcal{X}$ we must first find the HPD interval of length $l'(x, N)$ such that $\int_{a(x,N)}^{a(x,N)+l'(x,N)} f(\theta_k \mid x)\, d\theta_k = 1 - \alpha$, and the sample size is the minimum N that satisfies

$$\int_{\mathcal{X}} l'(x, N) f(x)\, dx \leq l, \tag{5}$$

where l is the required average length. The left-hand side of (5) averages the lengths of fixed coverage HPD intervals, weighted by the predictive 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 $\inf_{x \in \mathcal{X}} \{\int_{a(x,N)}^{a(x,N)+l} \times f(\theta_k \mid x)\, d\theta_k\} \geq 1 - \alpha$. 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 (Joseph and Bélisle, 1997), and nonidentifiable models are also often problematic in this sense. Therefore, in this article we use the following modified WOC (MWOC) criterion. Rather than taking the infinum across all possible data sets, we guarantee the desired length and coverage over a subset $\mathcal{S} \in \mathcal{X}$ such that $\mathcal{S}$ has a given probability. For example, we might choose the sample size N such that l and $1 - \alpha$ are guaranteed over 95% of the set $\mathcal{X}$, according to the predictive distribution (3). 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 highly improbable data. As will be discussed below, looking at sample sizes from a variety of criteria exposes the tradeoffs between study cost and accuracy of parameter estimation, and a sample size decision can be based on this information.

3. Sample Size Determination for Prevalence Studies

In this section, we apply the above criteria to prevalence studies using two imperfect diagnostic tests. We discuss the asymptotic limits on the coverage and credible interval lengths, and provide a detailed example. Once the likelihood function and prior distribution are specified, the posterior and predictive distributions are defined, and all criteria of Section 2 are fully specified. Given, however, that closed forms are not available for any of the above quantities, and that for each possible sample size N, averages or maxima over HPD intervals for all possible data sets x must be considered, the numerical challenge is considerable, even by modern computing standards. We briefly outline the algorithms we used to apply these criteria in practice, with full details appearing in the Appendix.

3.1 *ACC, ALC, and MWOC Sample Sizes When Results from Two Tests Are Available*

Let (T_1, T_2) denote the outcomes of two conditionally independent diagnostic tests, and let $N = \sum_{u=0}^{1} \sum_{v=0}^{1} n_{uv}$ be the sample size, where n_{uv} is the number of subjects for whom $(T_1 = u, T_2 = v)$, $u, v = 0, 1$. Define (S_j, C_j), $j = 1, 2$ to be the sensitivity and specificity of the jth test. The likelihood function is then proportional to

$$\begin{aligned} L &= f(n_{11}, n_{10}, n_{01}, n_{00} \mid \pi, S_1, C_1, S_2, C_2) \\ &\propto (\pi S_1 S_2 + (1 - \pi)(1 - C_1)(1 - C_2))^{n_{11}} \\ &\quad \times (\pi S_1 (1 - S_2) + (1 - \pi)(1 - C_1) C_2)^{n_{10}} \\ &\quad \times (\pi (1 - S_1) S_2 + (1 - \pi) C_1 (1 - C_2))^{n_{01}} \\ &\quad \times (\pi (1 - S_1)(1 - S_2) + (1 - \pi) C_1 C_2)^{n_{00}}. \end{aligned}$$

In order to proceed further, the prior distribution $f(\pi, S_1, C_1, S_2, C_2)$ must be specified. Following others (Joseph et al. 1995a; Gustafson et al., 2001; Johnson, Gastwirth, and Pearson, 2001), we use independent beta prior distributions for each of the five parameters. Dichotomous tests are sometimes based on an underlying continuous measure; increasing

values of the cutoff point determining a positive test will simultaneously increase the sensitivity and decrease the specificity of the test. For a given preselected cutoff point, however, it is reasonable to assume that the sensitivity and specificity of the test are independent parameters. This choice of prior distribution is convenient but not unique and may be replaced by other suitable distributions depending on the context of the problem. For an alternative model when the tests are possibly correlated see Dendukuri and Joseph (2001).

With independent beta priors, the marginal posterior density for π becomes

$$\begin{aligned} f(\pi \mid n_{11}, n_{10}, n_{01}, n_{00}) &\propto \int_{S_1,C_1,S_2,C_2} L \times \pi^{\alpha_\pi - 1}(1-\pi)^{\beta_\pi - 1} S_1^{\alpha_{S_1} - 1} \\ &\quad \times (1-S_1)^{\beta_{S_1} - 1} C_1^{\alpha_{C_1} - 1}(1-C_1)^{\beta_{C_1} - 1} \\ &\quad \times S_2^{\alpha_{S_2} - 1}(1-S_2)^{\beta_{S_2} - 1} C_2^{\alpha_{C_2} - 1} \\ &\quad \times (1-C_2)^{\beta_{C_2} - 1}\, dS_1\, dC_1\, dS_2\, dC_2, \end{aligned} \tag{6}$$

where (α_π, β_π) are the parameters of the beta prior distribution for the prevalence, and $(\alpha_{S_j}, \beta_{S_j})$ and $(\alpha_{C_j}, \beta_{C_j})$ are the parameters of the prior distribution over the sensitivity and the specificity of the jth test ($j = 1, 2$), respectively. Similarly, the preposterior predictive distribution for the data is

$$\begin{aligned} f(n_{11}, n_{10}, n_{01}, n_{00}) &= \int_{\pi,S_1,C_1,S_2,C_2} L \times \pi^{\alpha_\pi - 1}(1-\pi)^{\beta_\pi - 1} S_1^{\alpha_{S_1} - 1} \\ &\quad \times (1-S_1)^{\beta_{S_1} - 1} C_1^{\alpha_{C_1} - 1}(1-C_1)^{\beta_{C_1} - 1} \\ &\quad \times S_2^{\alpha_{S_2} - 1}(1-S_2)^{\beta_{S_2} - 1} C_2^{\alpha_{C_2} - 1} \\ &\quad \times (1-C_2)^{\beta_{C_2} - 1}\, d\pi\, dS_1\, dC_1\, dS_2\, dC_2. \end{aligned} \tag{7}$$

Finding the ACC sample size becomes the computational challenge of finding the minimum N such that

$$\sum_{\left\{(n_{11},n_{10},n_{01},n_{00}) \,\middle|\, \sum_{u,v} n_{uv} = N\right\}} \times \left\{ \int_{a(x,N)}^{a(x,N)+l} f(\pi \mid x)\, d\pi \right\} f(x) \geq 1 - \alpha,$$

where $x = (n_{11}, n_{10}, n_{01}, n_{00})$ is the data vector, $f(\pi \mid x)$ is given by (6), and $f(x)$ is given by (7). This implies that for each possible sample size N, and for each data vector $(n_{11}, n_{10}, n_{01}, n_{00})$, we must derive $f(x)$, $f(\pi \mid x)$, and the corresponding HPD interval for π of length l. Finally, the average coverage of these intervals, weighted by $f(x)$, must be calculated, and compared to the desired average coverage. For each data set, x, we used the Gibbs sampler to sample from the posterior density $f(\pi \mid x)$. To estimate the highest posterior density region we tried two methods: (1) An exact method based on the true posterior density, which is a mixture of beta densities, and (2) an approximate method based on a single best fitting beta density. The parameters of this density were found by matching the first two moments of the sample to the mean and variance of a beta distribution. The single beta approximation provided the best compromise between accuracy and efficiency for all criteria, except for the MWOC with very small γ. We found the average coverage corresponding to a range of N values, and fit a curve through these points to estimate the required sample size (Müller and Parmigiani, 1995). Full details of the algorithms used for all criteria are given in the Appendix.

3.2 *Limits on Coverage and Length of Posterior Credible Interval*

The problem of estimating disease prevalence based on the results of two non–gold standard diagnostic tests is nonidentifiable since we have three degrees of freedom but five unknown parameters (Walter and Irwig, 1988). In order to convert the problem to an identifiable one, frequentist methods generally fix two of the five parameters as exactly known. Bayesians are able to obtain the joint posterior density over all five parameters by using informative prior distributions over at least two of the five parameters (Joseph et al., 1995a). The marginal posterior density of each parameter, however, does not converge to a unique point estimate even with an infinite sample size. This issue can be usefully examined by reparameterizing the joint posterior distribution in terms of identifiable and nonidentifiable parameters, as suggested in a different context by Gustafson et al. (2001). Our derivation is similar to that first described by Gustafson (2002). Let $p_{uv} = P(T_1 = u, T_2 = v)$, $u, v = 0, 1$, so that

$$\begin{aligned} p_{11} &= \pi S_1 S_2 + (1-\pi)(1-C_1)(1-C_2), \\ p_{10} &= \pi S_1 (1-S_2) + (1-\pi)(1-C_1)C_2, \\ p_{01} &= \pi (1-S_1) S_2 + (1-\pi) C_1 (1-C_2), \text{ and} \\ p_{00} &= 1 - p_{11} - p_{10} - p_{01}. \end{aligned} \tag{8}$$

The transformation from $(\pi, S_1, C_1, S_2, C_2)$ to $(\pi, S_1, p_{11}, p_{10}, p_{01})$ then results in a joint posterior distribution of the form

$$f(\pi, S_1, p_{11}, p_{10}, p_{01} \mid n_{11}, n_{10}, n_{01}, n_{00}) = \frac{f(\pi, S_1 \mid p_{11}, p_{10}, p_{01}) f(p_{11}, p_{10}, p_{01} \mid n_{11}, n_{10}, n_{01}, n_{00})}{|\pi(1-\pi)(p_{11} + p_{10} - S_1)|},$$

where the denominator arises from the Jacobian of the transformation. As the total sample size $N = n_{11} + n_{10} + n_{01} + n_{00}$ increases to infinity, the true values of (p_{11}, p_{10}, p_{01}) become known. The expression $f(\pi, S_1 \mid p_{11}, p_{10}, p_{01})$ however, remains unaffected by any increase in the sample size, beyond more accurate conditioning on the p_{uv}'s. The joint prior distribution of $(\pi, S_1, C_1, S_2, C_2)$ can be reparameterized as

$$\begin{aligned} f(\pi, S_1, C_1, S_2, C_2) &= f(\pi, S_1, p_{11}, p_{10}, p_{01}) \left| \frac{\partial(p_{11}, p_{10}, p_{01})}{\partial(C_1, S_2, C_2)} \right| \\ &= f_\pi(\pi) f_{S_1}(S_1) f_{C_1}\left(\frac{(1 - p_{10} - p_{11}) - \pi(1 - S_1)}{1 - \pi} \right) \\ &\quad \times f_{S_2}\left(\frac{(p_{01} + p_{11})(p_{10} + p_{11} - S_1\pi) - p_{11}(1-\pi)}{(p_{10} + p_{11} - S_1)\pi} \right) \\ &\quad \times f_{C_2}\left(\frac{p_{10} - (1 - p_{10} - p_{11}) S_1}{(p_{10} + p_{11} - S_1)} \right) \left| \frac{1}{\pi(1-\pi)(p_{11} + p_{10} - S_1)} \right|, \end{aligned}$$

where $f_\pi(\cdot)$ denotes the prior density function of π, and $f_{S_j}(\cdot)$ and $f_{C_j}(\cdot)$ denote the prior densities of the sensitivity and

specificity of the jth test ($j = 1, 2$), respectively. The asymptotic conditional distribution of (π, S_1) given p_{11}, p_{10}, p_{01} is then

$$f(\pi, S_1 \mid p_{11}, p_{10}, p_{01}) = \frac{f(\pi, S_1, p_{11}, p_{10}, p_{01})}{f(p_{11}, p_{10}, p_{01})} \propto f(\pi, S_1, p_{11}, p_{10}, p_{01}) I(\mathcal{S}_{\pi, S_1 \mid p_{11}, p_{10}, p_{01}}), \tag{9}$$

where $I(\cdot)$ is the identity function and $\mathcal{S}_{\pi, S_1 \mid p_{11}, p_{10}, p_{01}}$ is a subset of the unit square determined by the values of (p_{11}, p_{10}, p_{01}). The set $\mathcal{S}_{\pi, S_1 \mid p_{11}, p_{10}, p_{01}}$ is defined in detail in the Appendix. Therefore the main effect of the observed data is the reduction of the range of the prior distribution of (π, S_1) from the unit square to the region defined by $\mathcal{S}_{\pi, S_1 \mid p_{11}, p_{10}, p_{01}}$. Integrating out S_1, we see that even with an infinite sample size, the marginal distribution for π does not converge to a single point. This implies that there is a limit to the accuracy with which π can be estimated, even with an infinite sample size, and this limit largely depends on the prior information about all five unknown parameters. At one extreme, if at least two of the five unknown parameters are given degenerate prior distributions that concentrate on one point, then the problem becomes identifiable, and the prevalence π can be estimated with any desired precision. This is in fact the basis for most frequentist methods for estimation in such problems, as discussed by Walter and Irwig (1988). On the other hand, if uniform prior densities are given for all five parameters, then very little can be said about the prevalence, even with an infinite sample size. The asymptotic value of the average coverage or the average length of HPD intervals for the prevalence can be obtained from the asymptotic posterior distribution of π and the predictive distribution of (p_{11}, p_{10}, p_{01}). We adopt a numerical approach, described in the Appendix. We next provide examples of sample sizes calculated from all three criteria along with a demonstration of the effect of the limits on the possible accuracy of estimation.

3.3 *Sample Size Required to Accurately Estimate the Prevalence of Strongyloides*

Table 1 provides the prior beta parameters we used to find the required sample size. Note that the two tests are complementary, with the high specificity of microscopy pairing well with the high sensitivity of serology. To illustrate the advantage of having results from two non–gold standard diagnostic tests compared to one, we will first calculate the sample sizes required when the serology test alone is available, and then observe the differences when both tests are used.

Table 2 provides the ACC, ALC, and MWOC sample sizes and asymptotic limits of HPD intervals for the prevalence of *Strongyloides* infection for one and two tests. Using $l = 0.2$, the serology test alone provides an asymptotic value of the average coverage and lower 5% quantile of the coverage across all HPD intervals of 88.4% and 87.2%, respectively. The asymptotic average length across all HPD intervals with 95% coverage is 0.259. Thus, when using the serology test alone it is not possible to satisfy the ACC, ALC, or MWOC ($1 - \gamma = 0.95$) criteria with $l = 0.2$ and $1 - \alpha = 0.95$, even with an infinite sample size. The addition of the microscopy test also leads to infinite sample sizes for $l = 0.2$ and $1 - \alpha = 0.95$ for the ACC and MWOC, but a finite sample size is obtained for the ALC. With $l = 0.3$, the asymptotic coverages for the ACC are 97.4% and 99.2% for one and two tests, respectively, leading to sample sizes of 70 for one test and 48 for two tests, a reduction of 31% in sample size requirements. For the ALC with coverage of $1 - \alpha = 0.90$, the

Table 2

Asymptotic values of ACC, ALC, and MWOC criteria for prevalence of Strongyloides and corresponding sample sizes. The single test column indicates that serology testing alone is used, while two tests indicates that both serology and microscopy tests are used. The ∞ indicates that even an infinite sample size is insufficient to attain the desired accuracy.

	Asymptotic coverage			Sample size	
ACC ($1 - \alpha = 0.95$)	Length	Single test	Two tests	Single test	Two tests
	0.1	60.3%	69.7%	∞	∞
	0.2	88.4%	94.2%	∞	∞
	0.3	97.4%	99.2%	70	48
	0.4	99.5%	99.9%	0	0
MWOC ($1 - \alpha = 0.95$)	Length	Single test	Two tests	Single test	Two tests
	0.1	58.3%	49.1%	∞	∞
	0.2	87.2%	81.9%	∞	∞
	0.3	96.9%	96.1%	123	∼90000
	0.4	99.4%	99.6%	0	0
	Asymptotic length			Sample size	
ALC ($l = 0.2$)	Coverage	Single test	Two tests	Single test	Two tests
	99%	0.358	0.257	∞	∞
	95%	0.259	0.198	∞	32809
	90%	0.211	0.167	∞	378
	80%	0.159	0.130	71	46

asymptotic average length for one test is 0.211, meaning even an infinite sample size is insufficient, while the addition of a second test leads to a sample size of 378, and an asymptotic average length of 0.167. Because the addition of a second test leads to many more combinations of possible data sets, depending on the prior distribution used for the properties of the second test, the MWOC asymptotic coverages do not necessarily decrease with the addition of a second test. Because of this, two tests lead to a smaller MWOC sample size with $l = 0.4$, but a larger size with $l = 0.3$. From Table 1, the prior length of the highest density interval with coverage 0.95 for the prevalence is approximately 0.39. Therefore, it is not surprising that no additional sampling is needed to satisfy $1 - \alpha = 0.95$ according to the ACC and MWOC (0.95) criteria with $l = 0.4$, whether one or two tests are used.

Overall, these results illustrate the possible benefits of using a second test in the presence of an imperfect first test. Note, however, that the total number of tests performed doubles when a second test is added, so that if the tests are expensive, using a larger sample size with a single test may be preferred to the smaller sample size that can be achieved with two tests. Researchers are also advised of the severe limitations of some diagnostic tests or combinations of tests, especially if the test properties are not very accurately known a priori. Performing appropriate sample size calculations that fully account for all uncertainty, including imperfect knowledge of test properties, is therefore crucial.

4. Sample Size Calculations for Sensitivity and Specificity

As discussed by Alonzo et al. (2002), studies designed to estimate properties of new diagnostic tests differ from prevalence studies, in that one must first decide whether to use a "case–control" or a "cohort design." Below we discuss methods for cohort designs, where a single sample is assembled and the diagnostic tests are applied to each member of the sample. Care is needed in the choice of prior distributions, since one needs to consider how subjects were recruited to avoid verification bias (Begg, 1987). Technically similar methods to those described below apply to case–control designs, since in the absence of a gold standard test all samples will potentially contain both positive and negative subjects.

To implement the sample size criteria for estimating sensitivity and specificity, we use methods similar to those discussed in Section 3.1, the main difference being that we require the marginal posterior density for S or C, rather than for π. The sample sizes and asymptotic values of the average coverage and average length for these parameters can be estimated using a Monte Carlo algorithm similar to that already discussed for the prevalence, as discussed in the Appendix. The ranges over which the asymptotic marginal distributions of the sensitivity and the specificity are defined are also discussed in the Appendix.

To illustrate our methods we use data from a recently published study on the performance of nucleic acid amplification tests for the detection of *Chlamydia trachomatis* infection in 3639 men (Johnson et al., 2000). In that study, simultaneous results were obtained on three different tests for *Chlamydia trachomatis*: tissue culture, polymerase chain reaction (PCR), and ligase chain reaction. Using latent class analysis (Walter and Irwig, 1988) of the joint results from the three tests we found the estimated mean (and standard deviation) for the sensitivity and specificity of the culture test were 0.70 (0.03) and 0.995 (0.002), respectively. The culture test is known to have poor sensitivity, even though it is widely used as a gold standard reference test for evaluating the sensitivity and specificity of new tests for *Chlamydia* (Hadgu, 1997).

For comparison, we calculate sample sizes required to estimate the sensitivity and specificity of a new test for *Chlamydia* both when culture is treated as a gold standard, and when the imperfect nature of culture is accounted for. Applying a Newton–Raphson method to match 95% intervals from beta densities to the same intervals from the results of the Johnson et al. study described above, we derived Beta(155.63, 66.15) and Beta(906.98, 3.63) prior distributions for the sensitivity and specificity of the culture test, respectively. We assumed a cohort design, and used uniform Beta(1, 1) prior distributions over the prevalence and the sensitivity and specificity of the new test.

Table 3 lists the sample sizes required for estimation of the sensitivity and specificity when culture is treated as a gold standard ($S = C = 1$) and a non–gold standard test (using the above prior distributions). The resultant sample size for a cohort study would be the maximum of the individual sample sizes for the sensitivity and specificity. When culture is assumed to be a gold standard there is no difference in the sample size required for sensitivity or specificity, since a uniform prior for the prevalence was used. The table illustrates that ignoring the imperfect nature of culture will result in unrealistically small sample sizes in all cases. The poor sensitivity of the culture test leads to many false negatives, meaning that a larger sample size is required for estimating the specificity of the new test compared to the sensitivity. The near perfect specificity of the culture test leads to a smaller increase in the sample sizes for the sensitivity compared to assuming culture is a gold standard. It is important to note that despite the reasonably accurate prior information available for the culture test, in many cases even an infinite sample size is insufficient to attain the desired accuracy. This is especially true for the most conservative MWOC criterion.

While we used uniform priors, in practice, researchers may have some idea of the prevalence in their population, and something may be known about the properties of the new diagnostic test from previous work. Other prior distributions may be selected in such cases.

5. Discussion

The vast majority of studies ignore not only the imperfection in "gold standard" reference tests used, but, perhaps more importantly, also ignore the uncertainty in the estimates of the sensitivity and specificity of these tests. This is true both at the design and analysis phases, leading to sample sizes that are typically much too small, and final estimates with confidence intervals that are much too narrow. The situation improves somewhat if three or more conditionally independent tests are available, where latent class models are identifiable (Walter and Irwig, 1988). The identifiability guarantees convergence of each parameter estimate to the true values as the sample size increases, and the extra information provided by

Table 3
Sample size required for estimating sensitivity and specificity of a new test for Chlamydia trachomatis when the reference test, tissue culture, is treated as a gold standard or non–gold standard test. The ∞ indicates that even an infinite sample size is insufficient to attain the desired accuracy.

		Gold standard	Non–gold standard	
ACC ($1 - \alpha = 0.95$)	Length	Sens & spec	Sensitivity	Specificity
	0.1	1298	2261	∞
	0.2	308	492	∞
	0.3	122	194	2014
	0.4	61	94	409
MWOC ($1 - \alpha = 0.95$)	Length	Sens & spec	Sensitivity	Specificity
	0.1	5671	∞	∞
	0.2	1370	2788	∞
	0.3	555	973	∞
	0.4	274	495	∞
ALC ($l = 0.2$)	Coverage	Sens & spec	Sensitivity	Specificity
	99%	403	611	4612
	95%	189	287	935
	90%	129	193	498
	80%	74	109	227

a third test tends to lead to smaller sample sizes compared to one or two tests. Nevertheless, many studies are carried out with only one or two tests. The methods discussed here are important to the planning of such studies, since, as our examples have shown, prohibitively large sample sizes can often result. Narrower priors on the test properties will typically result in smaller sample sizes, so it may be worthwhile to improve knowledge of the test properties before embarking on a large prevalence study. Our methods assume conditional independence between diagnostic tests, i.e., that given a subject's true disease state, the test results from the two tests are independent. This is a reasonable assumption for tests based on different mechanisms, but we caution that even larger sample sizes may be required if this condition does not hold (Dendukuri and Joseph, 2001). Further, we have only investigated binary tests; further work is required on designing studies with imperfect multicategorical or continuous tests.

We have discussed three different sample size criteria, which lead to different sample sizes for any given problem. A natural question, therefore, is which one to use. Clearly, the MWOC criterion is more conservative than either the ACC or ALC, which guarantee the target values for coverage and length only on average. We have found it useful to calculate the sample sizes that result from all criteria, including the MWOC $(1 - \gamma)$ for various values of γ, to develop a fuller understanding of the tradeoffs between sample size and the risk of not meeting target values for l and $1 - \alpha$. Based on this information, a final sample size may be selected. It is especially important for study designers to appreciate that in many cases the desired estimation accuracy cannot be attained even with an infinite sample size. The addition of a second test sometimes alleviates the problem.

Software for calculating the sample sizes for all methods developed in this article is available at `www.medicine.mcgill.ca/epidemiology/Joseph`.

Résumé

La planification des études relatives à un test diagnostic est rendue complexe par le fait que virtuellement aucun test ne fournit de résultats parfaitement précis. Les erreurs de classement induites par l'imperfection des sensibilités et des spécificités des tests diagnostics doivent être prises en considération, que l'objectif principal de l'étude soit d'estimer la prévalence d'une maladie dans une population ou d'investiguer les propriétés d'un nouveau test diagnostic. Un travail antérieur sur la taille d'échantillon requise pour estimer la prévalence d'une maladie dans le cas d'un seul test imparfait a montré des différences très importantes de tailles en comparaison à des méthodes qui supposent un test parfait. Dans cet article nous étendons ces méthodes pour inclure deux tests imparfaits conditionnellement indépendants, et appliquons différents critères pour la détermination bayésienne de la taille d'échantillon pour de telles études. Nous considérons à la fois les études de prévalence de maladies et les études conçues pour estimer la sensibilité et la spécificité des tests diagnostiques. Comme le problème est typiquement non identifiable, nous investiguons les limites de précision dans l'estimation de paramètres quand la taille d'échantillon tend vers l'infini. Au moyen de deux exemples dans les maladies infectieuses, nous illustrons les changements de tailles d'échantillons résultant de l'application de deux tests aux individus de l'étude plutôt que d'un test unique. Bien que l'on trouve souvent des tailles d'échantillons plus faibles dans le cas de deux tests, celles-ci peuvent néanmoins être beaucoup trop grandes, à moins qu'une information précise soit disponible sur la sensibilité et la spécificité des tests utilisés.

References

Adcock, C. J. (1988). A Bayesian approach to calculating sample sizes. *The Statistician* **37,** 433–439.

Adcock, C. J. (1997). Sample size determination: A review. *The Statistician* **46,** 261–283.

Alonzo, T. A., Pepe, M. S., and Moskowitz, C. S. (2002). Sample size calculations for comparative studies of medical tests for detecting presence of disease. *Statistics in Medicine* **21,** 835–852.

Arkin, C. F. and Wachtel, M. S. (1990). How many patients are necessary to assess test performance? *Journal of the American Medical Association* **263,** 275–278.

Begg, C. B. (1987). Biases in the assessment of diagnostic tests. *Statistics in Medicine* **6,** 411–432.

Box, G. and Tiao, G. (1973). *Bayesian Inference in Statistical Analysis*. Boston, Massachusetts: Addison-Wesley.

Buderer, N. M. (1996). Statistical methodology: I. Incorporating the prevalence of disease into the sample size calculation for sensitivity and specificity. *Academic Emergency Medicine* **3,** 895–900.

Dendukuri, N. and Joseph, L. (2001). Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics* **57,** 158–167.

Gittins, J. and Pezeshk, H. (2000). A behavioural Bayes method for determining the size of a clinical trial. *Drug Information Journal* **34,** 355–363.

Gustafson, P. (2002). *On model expansion, model contraction, identifiability, and prior information: Two illustrative scenarios involving mismeasured variables*. Technical Report 203, Department of Statistics, University of British Columbia, Canada.

Gustafson, P., Le, N. D., and Saskin, R. (2001). Case-control analysis with partial knowledge of exposure misclassification probabilities. *Biometrics* **57,** 598–609.

Hadgu, A. (1997). Bias in the evaluation of DNA-amplification tests for detecting *Chlamydia trachomatis*. *Statistics in Medicine* **16,** 1391–1399.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143,** 29–36.

Irwig, L., Glasziou, P., Berry, G., Chock, C., and Mock, P. (1994). Efficient study designs to assess the accuracy of screening tests. *American Journal of Epidemiology* **140,** 759–769.

Johnson, R. E., Green, T. A., Schachter, J., Jones, R. B., Hook, E. W., Black, C. M., Martin, D. H., Louis, M. E. S., and Stamm, W. E. (2000). Evaluation of nucleic acid amplification tests as reference tests for *Chlamydia trachomatis* infections in asymptomatic men. *Journal of Clinical Microbiology* **38,** 4382–4386.

Johnson, W. O., Gastwirth, J. L., and Pearson, L. M. (2001). Screening without a gold-standard: The hui-walter paradigm revisited. *American Journal of Epidemiology* **153,** 921–924.

Joseph, L. and Bélisle, P. (1997). Bayesian sample size determination for normal means and differences between normal means. *The Statistician* **46,** 209–226.

Joseph, L. and Wolfson, D. (1997). Interval-based versus decision theoretic criteria for the choice of sample size. *The Statistician* **46,** 145–149.

Joseph, L., Gyorkos, T., and Coupal, L. (1995a). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology* **141,** 263–272.

Joseph, L., Wolfson, D., and du Berger, R. (1995b). Sample size calculations for binomial proportions via highest posterior density intervals. *The Statistician* **44,** 143–154.

Joseph, L., du Berger, R., and Bélisle, P. (1997). Bayesian and mixed Bayesian/likelihood criteria for sample size determination. *Statistics in Medicine* **16,** 769–781.

Lemeshow, S., Hosmer, D. W., Klar, J., and Lwanga, S. K. (1990). *Adequacy of Sample Size in Health Studies*. Chichester: John Wiley & Sons.

Müller, P. and Parmigiani, G. (1995). Optimal design via curve fitting of Monte Carlo experiments. *Journal of the American Statistical Association* **90,** 1322–1330.

Obuchowski, N. A. and McClish, D. K. (1997). Sample size determination for diagnostic accuracy studies involving binormal ROC curve indices. *Statistics in Medicine* **16,** 1529–1542.

Pepe, M. S. (2003). Study design and hypothesis testing. In *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Chapter 8. Oxford: Oxford University Press.

Rahme, E., Joseph, L., and Gyorkos, T. (2000). Bayesian sample size determination for estimating binomial parameters from data subject to misclassification. *Applied Statistics* **49,** 119–128.

Walter, S. D. and Irwig, L. M. (1988). Estimation of error rates, disease prevalence and relative risks from misclassified data: A review. *Journal of Clinical Epidemiology* **41,** 923–937.

Wang, F. and Gelfand, A. (2002). A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science* **17,** 193–208.

Received April 2003. *Revised October* 2003.
Accepted October 2003.

Appendix

We used the following numerical approach to estimate the required sample sizes. The algorithm presented is for the prevalence π; similar algorithms were used for the sensitivity and specificity.

A.1 *Monte Carlo Algorithm for Estimating Sample Sizes According to ACC, ALC, and MWOC Criteria*

(1) Sample M_1 random values from the joint prior distribution of $(\pi, S_1, C_1, S_2, C_2)$.

(2) For each quintuplet $(\pi_i, S_{1i}, C_{1i}, S_{2i}, C_{2i})$, $i = 1, \ldots, M_1$:

   (a) Calculate the probabilities $p_{uvi} = P(T_{1i} = u, T_{2i} = v)$, $u, v = 0, 1$. These probabilities are given by (8).

   (b) Select a value for N, the required sample size. From the multinomial distribution with parameters $(N, (p_{11i}, p_{10i}, p_{01i}, p_{00i}))$ draw M_2 random values of $(n_{11}, n_{10}, n_{01}, n_{00})$. This is equivalent to sampling

from the preposterior predictive distribution of the data.

(3) For each $(n_{11ij}, n_{10ij}, n_{01ij}, n_{00ij})$, $i = 1, \ldots, M_1$, $j = 1, \ldots, M_2$, divide the range of the prevalence into M_4 intervals of length w (e.g., 0.01). At the center of the kth interval, π_k, estimate the posterior density of π using either the exact or approximate methods described below.

The exact method: The exact posterior density at π_k is a mixture of $(n_{11ij} + 1)(n_{10ij} + 1)(n_{01ij} + 1)(n_{00ij} + 1)$ Beta densities as follows

$$
\begin{aligned}
f_k \propto & \sum_{j_1=0}^{n_{11ij}} \sum_{j_2=0}^{n_{10ij}} \sum_{j_3=0}^{n_{01ij}} \sum_{j_4=0}^{n_{00ij}} \frac{1}{j_1!(n_{11ij}-j_1)!} \\
& \times \frac{1}{j_2!(n_{10ij}-j_2)!} \frac{1}{j_3!(n_{01ij}-j_3)!} \frac{1}{j_4!(n_{00ij}-j_4)!} \\
& \text{Beta}(\alpha_{S_1} + j_1 + j_2, \beta_{S_1} + j_3 + j_4) \\
& \text{Beta}(\alpha_{C_1} + n_{01ij} + n_{00ij} - j_3 - j_4, \\
& \qquad \beta_{C_1} + n_{11ij} + n_{10ij} - j_1 - j_2) \\
& \text{Beta}(\alpha_{S_2} + j_1 + j_3, \beta_{S_2} + j_2 + j_4) \\
& \text{Beta}(\alpha_{C_2} + n_{10ij} + n_{00ij} - j_2 - j_4, \\
& \qquad \beta_{C_2} + n_{01ij} + n_{11ij} - j_1 - j_3) \\
& \pi_k^{\alpha_\pi + j_1 + j_2 + j_3 + j_4} (1 - \pi_k)^{\beta_\pi + N - j_1 - j_2 - j_3 - j_4}, \\
& \qquad k = 1, \ldots, M_4. \qquad \text{(A.1)}
\end{aligned}
$$

The approximate method: This method was developed as an alternative to the exact method which is very time consuming, taking several hours for a single sample size calculation in some cases. The method below is much more efficient, typically taking less than an hour to find a sample size, although this varies depending on the priors and accuracy desired. The posterior mixture of beta distributions is approximated by a single beta distribution. In our experience, this approach is typically sufficiently precise for implementing the ACC and ALC but not the MWOC $(1 - \gamma)$ when γ is small.

(a) For each $(n_{11ij}, n_{10ij}, n_{01ij}, n_{00ij})$, $i = 1, \ldots, M_1$, $j = 1, \ldots, M_2$, obtain a sample of M_3 values from the posterior distribution of π using the Gibbs sampler (Joseph et al., 1995a). Label these values π_{ijk}, $i = 1, \ldots, M_1$, $j = 1, \ldots, M_2$, $k = 1, \ldots, M_3$.

(b) Estimate the mean and variance of the posterior distribution of π given $(n_{11ij}, n_{10ij}, n_{01ij}, n_{00ij})$ as $\mu_{ij} = \sum_{k=1}^{M_3} \pi_{ijk}/M_3$ and $\sigma_{ij}^2 = \sum_{k=1}^{M_3} (\pi_{ijk} - \mu_{ij})^2/M_3 - 1$, respectively.

(c) The posterior distribution of π is approximated by a single beta distribution with parameters $\alpha_{ij} = -\mu_{ij}(\sigma_{ij}^2 + \mu_{ij}^2 - \mu_{ij})/\sigma_{ij}^2$ and $\beta_{ij} = (\mu_{ij} - 1)(\sigma_{ij}^2 + \mu_{ij}^2 - \mu_{ij})/\sigma_{ij}^2$.

(4) From either of the above posterior distributions (exact or approximate), 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 $a + l$, either using standard results from the beta density (approximate method) or Riemann sums (exact method).

(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 to 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. The final sample size selected as the smallest N on the curve satisfying the given criterion (Müller and Parmigiani, 1995a).

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 at $N = 100{,}000$, we reported a sample size of infinity. In practice, studies this large are very rare.

A.2 *Monte Carlo Algorithm to Determine Asymptotic Values of ACC, ALC, and MWOC*

(1) Draw a random sample of size M_1 from $f(\pi, S_1, C_1, S_2, C_2)$.

(2) For each quintuplet of values $(\pi_i, S_{1i}, C_{1i}, S_{2i}, C_{2i})$ calculate $(p_{11i}, p_{10i}, p_{01i}, p_{00i})$, as given by (8).

(3) For each $(p_{11i}, p_{10i}, p_{01i}, p_{00i})$ the joint posterior distribution of (π, S_1) is given by (9). To obtain the marginal posterior distribution of π we need to integrate (9) with respect to S_1. This can be done using standard numerical integration.

(4) For each $(p_{11i}, p_{10i}, p_{01i}, p_{00i})$ the coverage (c_i) of the HPD interval of length l, or the length (l_i) of the HPD interval with coverage $1 - \alpha$ can be obtained as follows: Divide the domain of π into K subintervals of length w. The probability of π being in the kth interval is estimated by $A_k = f(\pi_k \,|\, p_{11i}, p_{10i}, p_{01i}, p_{00i}) \times w$, where π_k is the midpoint of the kth interval. The A_k values are sorted in descending order as $A_{(1)}, A_{(2)}, \ldots, A_{(K)}$. The coverage of the HPD region of length l is then estimated by $c_i = \sum_{k=1}^{K_1^*} A_{(k)}$, where $K_1^* = \max\{k : k \times w \leq l\}$. Similarly, the length of the HPD region of coverage $1 - \alpha$ is estimated by $l_i = K_2^* \times w$ where $K_2^* = \min\{k : \sum_{k'=1}^{k} A_{(k')} \geq 1 - \alpha\}$.

(5) The asymptotic value of the average coverage is estimated by $\frac{1}{M_1} \sum_{i=1}^{M_1} c_i$. The asymptotic value of the average length is estimated by $\frac{1}{M_1} \sum_{i=1}^{M_1} l_i$. The asymptotic value of the MWOC $(1 - \gamma)$ criterion is estimated by the $(1 - \gamma) \times 100$ quantile of the c_i's.

A.3 *Asymptotic Domains of π, S_1, and C_1 When a Single Test Used*

Let p denote the probability of a positive test. The transformation from $(\pi,\ S_1,\ C_1)$ to $(\pi,\ S_1,\ p)$ implies that $C_1 = 1-(p-\pi S_1)/(1-\pi)$. The subset in which the joint posterior distribution, $f(\pi,\ S_1 \mid p)$, is defined is given by

$$\begin{aligned}\mathcal{S}_{\pi,S_1\mid p} &= \{(\pi,S_1): 0<\pi,S_1,C_1<1\}\\ &= \{(\pi,S_1): 0<\pi,S_1<1\}\\ &\quad \cap\{(\pi,S_1):\pi(1-S_1)<1-p\}\cap\{(\pi,S_1):\pi S_1<p\}.\end{aligned}$$

The range of $f(\pi,\ C_1 \mid p)$ can be obtained by replacing S_1 by $(1-C_1)$ and π by $1-\pi$ in $\mathcal{S}_{\pi,S_1\mid p}$ above.

A.4 *Asymptotic Domains of π, S_1, and C_1 When Two Tests Used*

Let $p_{1\cdot} = p_{11}+p_{10}$, $p_{0\cdot} = p_{01}+p_{00}$, $p_{\cdot 1} = p_{11}+p_{01}$, and $p_{\cdot 0} = p_{10}+p_{00}$. The transformation from $(\pi,\ S_1,\ C_1,\ S_2,\ C_2)$ to $(\pi,\ S_1,\ p_{11},\ p_{10},\ p_{01})$ gives

$$C_1 = \frac{(1-p_{1\cdot})-\pi(1-S_1)}{1-\pi},\ S_2 = \frac{p_{\cdot 1}(p_{1\cdot}-S_1\pi)-p_{11}(1-\pi)}{(p_{1\cdot}-S_1)\pi},$$

$$C_2 = \frac{p_{10}-(1-p_{\cdot 1})S_1}{p_{1\cdot}-S_1}.$$

The domain of $(\pi,\ S_1)$ is

$$\begin{aligned}&\mathcal{S}_{\pi,S_1\mid p_{11},p_{10},p_{01}}\\ &\quad=\{(\pi,S_1): 0<\pi,S_1,C_1,S_2,C_2<1\}\\ &\quad=\{(\pi,S_1): 0<\pi,S_1<1\}\\ &\qquad\cap\{(\pi,S_1):\pi(1-S_1)<1-p_{1\cdot}\}\\ &\qquad\cap\{(\pi,S_1):p_{1\cdot}>\pi S_1\}\\ &\qquad\cap\left\{(\pi,S_1):\left\{S_1>\max\left(p_{1\cdot},\frac{p_{1\cdot}(\pi-p_{\cdot 0})+p_{10}(1-\pi)}{\pi p_{\cdot 1}},\right.\right.\right.\\ &\qquad\qquad\left.\left.\left.\frac{p_{1\cdot}}{\pi}-\frac{p_{10}(1-\pi)}{\pi(1-p_{\cdot 1})},\frac{p_{11}}{p_{\cdot 1}},\frac{p_{10}}{p_{\cdot 0}}\right)\right\}\right.\\ &\qquad\quad\cup\left\{S_1<\min\left(p_{1\cdot},\frac{p_{1\cdot}(\pi-p_{\cdot 0})+p_{10}(1-\pi)}{\pi p_{\cdot 1}},\right.\right.\\ &\qquad\qquad\left.\left.\left.\frac{p_{1\cdot}}{\pi}-\frac{p_{10}(1-\pi)}{\pi(1-p_{\cdot 1})},\frac{p_{11}}{p_{\cdot 1}},\frac{p_{10}}{p_{\cdot 0}}\right)\right\}\right\}.\end{aligned}$$

The range of $f(\pi,\ C_1 \mid p_{11},\ p_{10},\ p_{01})$ can be obtained by replacing S_1 by $(1-C_1)$ and π by $1-\pi$ in $S_{\pi,S_1\mid p_{11},p_{10},p_{01}}$ above.

The above algorithms were realized using a combination of Splus 6.0 (Mathsoft, Inc, 2000), Visual C++ (version 6.0, Microsoft), and WinBUGS (version 1.4), linked together through Perl 5.6 scripts (`www.perl.org`).