

# Bayesian sample size determination for estimating binomial parameters from data subject to misclassification

Elham Rahme\*      Lawrence Joseph†      Theresa W. Gyorkos‡

November 29, 1999

## Summary

We investigate the sample size problem when a binomial parameter is to be estimated, but some degree of misclassification is possible. The problem is especially challenging when the degree to which misclassification occurs is not exactly known. Motivated by a Canadian survey of the prevalence of toxoplasmosis infection in pregnant women, we examine the situation where it is desired that a marginal posterior credible interval for the prevalence of width  $w$  has coverage  $(1 - \alpha)$ , using a Bayesian sample size criterion. The degree to which the misclassification probabilities are known *a priori* can have a very large effect on sample size requirements, and in some cases achieving a coverage of  $(1 - \alpha)$  is impossible, even with an infinite sample size. Therefore, investigators must carefully evaluate the degree to which misclassification can occur when estimating sample size requirements.

**Key words:** Average Coverage Criterion; Bayesian; Binomial distribution, Diagnostic test, Misclassification; Prevalence; Sample size.

## 1 Introduction

Sample size determination is one of the most frequently occurring applied problems, and a vast literature is available on the subject. In many sampling situations encountered in practice, however, specific methodology is lacking. For example, while standard formulae are available for calculating sample size requirements for binomial parameters, it is less clear how to adjust

---

\*Department of Mathematics and Statistics, McGill University, Burnside Hall, 805 Sherbrooke Street West, Montreal, Quebec, Canada, H3A 2K6.

†Department of Epidemiology and Biostatistics, McGill University, Montreal, Canada, and Division of Clinical Epidemiology, Montreal General Hospital, Department of Medicine, 1650 Cedar Avenue, Montreal, Quebec, H3G 1A4, Canada. Email: joseph@binky.ri.mgh.mcgill.ca. Tel: (514)–937-6011 X 4713. Fax: (514)–934-8293.

‡Department of Epidemiology and Biostatistics, McGill University, Montreal, Canada, McGill Centre for Tropical Diseases, McGill University, Montreal, Canada, and Division of Clinical Epidemiology, Montreal General Hospital, Department of Medicine, 1650 Cedar Avenue, Montreal, Quebec, H3G 1A4, Canada.

the sample sizes for the presence of misclassification errors. Consider the following prototypic example, which motivated the development of our methodology.

Accurate estimation of the prevalence of the microscopic parasite *Toxoplasma gondii* is critical for making informed decisions about whether to initiate a prenatal screening program. A study is being planned to estimate the prevalence of this parasite among pregnant women in the province of Quebec, Canada. Suppose one would like to determine the sample size needed such that a 95% interval will be of total width  $w = 0.10$ . A serological kit that detects the presence of antibodies will be used as the diagnostic test. If the diagnostic test is error-free, and therefore the number of infected women in the sample is equal to the number who test positive, then the well known sample size formula based on the normal approximation to the binomial distribution can be used. This gives

$$n = \left( \frac{2Z_{\alpha/2}}{w} \right)^2 \theta(1 - \theta), \quad (1)$$

where  $\theta$  is the prevalence,  $Z_{\alpha/2}$  is the upper  $100(1 - \alpha/2)\%$  percentile of the normal distribution, and  $n$  is the required sample size. For example, using (1) with  $Z_{\alpha/2} = 1.96$ ,  $\theta = 0.5$  and  $w = 0.1$  gives  $n = 384$ . The conservative value of  $\theta = 0.5$  was used, since very little is currently known about the rate of *Toxoplasma gondii* among pregnant women in Quebec, and a rate of 50% is not impossible.

Although the diagnostic test which will be used in the study is among the best available, the assumption of perfect sensitivity (probability of a positive test result in truly positive subjects) and specificity (probability of a negative test result in truly negative subjects) is unrealistic, so that some misclassification is expected. In a previous study using the same diagnostic kit, 65 out of 68 women who had *Toxoplasma gondii* tested positive, while all 22 women who did not have the disease tested negative (Wilson and Ware 1991). Suppose that the sensitivity is exactly  $65/68 = 95.5\%$  and the specificity is  $22/22 = 100\%$ . Let  $p$  be the probability of testing positive for a test with sensitivity  $s$  and specificity  $c$ . Since each positive test is either a true positive or a false positive result,

$$p = \theta s + (1 - \theta)(1 - c). \quad (2)$$

From (2), it is easy to show (Rahme and Joseph 1998) that knowing  $p$  to within a width of  $w(s + c - 1)$  is equivalent to knowing  $\theta$  to within a width of  $w$ . The sample size can then be derived by a small modification to the standard formula (1), giving

$$n_{adj} = \left( \frac{2Z_{\alpha/2}}{w(s + c - 1)} \right)^2 p(1 - p). \quad (3)$$

Using this formula with  $s = 0.955$ ,  $c = 1$  and  $p = 0.5$  raises the sample size to 422. In this case, the values of  $s$  and  $c$  were both equal to or near 1, so the modification causes only a small increase in sample size compared to the standard formula. Smaller  $s + c$  values will lead to larger increases, however, and in the extreme case of a completely uninformative test ( $s + c = 1$ ) even an infinite sample size cannot estimate the prevalence to the desired accuracy.

Equation (3) requires point estimates of  $\theta$  (or  $p$ ),  $s$  and  $c$ . The conservative value of  $\theta$  or  $p$  equal to 0.5 may lead to unnecessarily large sizes for rare conditions, making the choice of  $\theta$  problematic. Another concern is that the normal approximations implied in (1) and (3) may not hold for rare conditions. Moreover, the values for  $s$  and  $c$  were in fact only estimates from relatively small samples, and in general, the sensitivity and specificity of a test are only rarely known exactly. Properly accounting for the uncertainties in  $s$  and  $c$  can have a very substantial effect on the sample size requirements. Consider, for example, a sample size of 384, and suppose one were to observe 100 positive tests. From the usual binomial considerations, assuming perfect tests ( $s = c = 1$ ) gives a prevalence estimate of  $100/384 = 0.26$ , with a 95% equally-tailed credible interval (CI) of (0.22, 0.31). If we instead assume that  $s = 0.955$  and  $c = 1$ , we now estimate the prevalence to be 0.27, with a 95% CI of (0.23, 0.32). Taking into account the uncertainty in the  $s$  and  $c$  values by using Beta prior densities based on the data from Wilson and Ware (1991), however, gives a point estimate of 0.25, with a considerably wider 95% CI of (0.13, 0.31). See Section 3 for the origin of these Beta prior densities.

A final problem with both (1) and (3) is that they fail to consider that the data are unknown at the planning stage of the study, which produces a further random component. Even if  $s$  and  $c$  are fixed constants, different data sets will produce different estimates of  $p$  and  $\theta$ , leading to reports of different interval widths.

In this paper we will discuss a solution to this problem that combines recent research from two different areas. In particular, Bayesian sample size criteria for interval widths (Adcock 1988, Adcock 1997, Gould 1993, Joseph, du Berger and Bélisle 1997) will be integrated with methods for estimating the prevalence of a disease using data from an imperfect diagnostic test (Joseph, Gyorkos and Coupal 1995). Starting from a prior distribution on  $\theta$ ,  $s$  and  $c$  that summarizes what is known about the prevalence and the test properties at the time of study planning, we will calculate the smallest sample size for which a  $(1 - \alpha)$  posterior credible interval for the prevalence will have total width  $w$  with high probability.

Various Bayesian sample size criteria and methods to estimate disease prevalence are reviewed in Section 2, before presenting the methodology to determine sample sizes for diagnostic test studies. An important conclusion is that even small uncertainties about the sensitivity and the specificity of a diagnostic test may lead to a large increase in the sample size needed to reach the desired accuracy, and that in certain cases this accuracy cannot be reached even when  $s + c > 1$ . Section 3 illustrates the methods by returning to the *Toxoplasmosis* study discussed above, and the paper concludes with a discussion.

Although we present our methodology in terms of the diagnostic testing application for which it was developed, the ideas presented here can be directly applied to any situation where misclassified binomial data arise.

## 2 Methods

### 2.1 Bayesian sample size criteria

Let  $\theta$  belonging to the space  $\Theta$  denote the parameter of interest,  $f(\theta)$  the prior density on  $\theta$ , and  $x$  the observed data from data space  $\mathcal{X}$ . Let the likelihood function be given by  $l(x|\theta)$ , and the posterior distribution for  $\theta$  given  $x$  of sample size  $n$  be given by  $f(\theta|x, n)$ . In general, we are looking for the minimum  $n$  such that the posterior credible set of width  $w$  has coverage probability of at least  $(1 - \alpha)$ . For example, letting  $d = w/2$ , we might like to find the smallest  $n$  such that

$$\int_{\hat{\theta}-d}^{\hat{\theta}+d} f(\theta|x, n)d\theta \geq 1 - \alpha,$$

where  $\hat{\theta}$  is the posterior mean. For any given  $n$ , however,  $\hat{\theta}$  and the coverage probability of the credible interval  $[\hat{\theta} - d, \hat{\theta} + d]$  depend on the data  $x$ , which are unknown at the planning stage of the study. To sidestep this problem, we can select  $n$  such that the expected posterior coverage probability is at least  $1 - \alpha$ , where the expectation is over the marginal distribution of  $x$  induced by the prior distribution, given by

$$m(x) = \int_{\Theta} l(x|\theta)f(\theta)d\theta. \tag{4}$$

Therefore, Adcock (1988) suggested seeking the smallest  $n$  satisfying

$$\int_{\mathcal{X}} \left( \int_{\hat{\theta}-d}^{\hat{\theta}+d} f(\theta|x, n)d\theta \right) m(x)dx \geq 1 - \alpha, \tag{5}$$

where the integral becomes a sum if the space  $\mathcal{X}$  is discrete. This criterion involves a weighted average of coverage probabilities of fixed length credible intervals, with weights given by  $m(x)$ , and hence has been termed an average coverage criterion (ACC) by Joseph, Wolfson and du Berger (1995) and Adcock (1997). If  $\theta$  is univariate, then use of (5) is straightforward. Although closed form solutions are often not available, the left hand side of (5) may be calculated for any given  $n$ , and a bisectional search over  $n$  may be used to find the minimum sample size satisfying (5).

The parameter of interest,  $\theta$ , may often be one component from a vector, in which case further integration is necessary to find the marginal posterior density of the component of interest. This occurs in diagnostic testing situations whenever the sensitivity and specificity of the test are not exactly known, which is almost always the case. In the next section, we describe a simple but efficient method to approximate these integrals. Efficiency is important, since the integral will typically be calculated hundreds or thousands of times for each sample size problem.

Although in the sequel we will use the ACC, other criteria could also be examined. For example, an average length criterion (ALC) can be defined (Joseph, du Berger and Bélisle

1997) that, conversely to the ACC, averages variable interval lengths of fixed coverage intervals. The ALC and the ACC often produce similar sample sizes. In some situations, one may prefer a conservative sample size, which instead of averaging over  $\mathcal{X}$ , guarantees the desired coverage and interval length over all possible  $x$  that can arise. While here we have defined the ACC in terms of posterior credible intervals of the form  $\hat{\theta} \pm d$ , highest posterior density (HPD) intervals could also be used. See Joseph, Wolfson, and du Berger (1995) for a comparison of sample sizes from HPD and symmetric intervals in the context of simple binomial sampling. Decision theoretic criteria (Lindley 1997) and sample sizes based on average power of hypothesis tests (Spiegelhalter and Freedman 1986) have also been considered. See Chaloner and Verdinelli (1995) for a recent review of Bayesian optimal design, and Adcock (1997) for an up to date review of both frequentist and Bayesian sample size criteria.

## 2.2 Bayesian estimation of disease prevalence when the sensitivity and the specificity are unknown

If  $x$  positive tests are observed in  $n$  subjects, then from (2), the likelihood function is proportional to

$$l(x|\theta, s, c) \propto [\theta s + (1 - \theta)(1 - c)]^x [\theta(1 - s) + (1 - \theta)c]^{n-x}.$$

From Bayes Theorem, if the joint prior distribution of  $\theta$ ,  $S$ , and  $C$  is given by  $f(\theta, s, c)$ , the joint posterior density becomes

$$f(\theta, s, c|x) = \frac{f(\theta, s, c, x)}{m(x)} = \frac{f(\theta, s, c)l(x|\theta, s, c)}{m(x)},$$

where the marginal (predictive) distribution of  $x$  is

$$m(x) = \int_0^1 \int_0^1 \int_0^1 f(\theta, s, c, x) dsdc d\theta.$$

The marginal posterior density of  $\theta$  is then

$$f(\theta|x) = \int_0^1 \int_0^1 f(\theta, s, c|x) dsdc. \tag{6}$$

It will often be reasonable that  $\theta$ ,  $S$ , and  $C$  are *a priori* independent, given that the test methodology (for example, the cutoff values for continuous tests) remains fixed. This is because the performance of the test within positive and negative subgroups of patients may not be affected by the disease prevalence in the population, and prior knowledge about the sensitivity and specificity given any fixed cutoff usually is gained by independently applying the test to known positive and negative subjects, as in Wilson and Ware (1991). If independent, Beta densities may be conveniently used as prior distributions of  $\theta$ ,  $S$  and  $C$ , which are restricted to a  $[0, 1]$  range, although the usual advantage of conjugacy does not apply. A parameter  $\theta$  follows a Beta density with parameters  $j$  and  $k$  if

$$f(\theta) = \begin{cases} \frac{1}{B(j,k)} \theta^{j-1} (1 - \theta)^{k-1}, & 0 \leq \theta \leq 1, \quad j, k > 0, \quad \text{and} \\ 0, & \text{otherwise,} \end{cases}$$

where  $B(j, k)$  is the Beta function evaluated at  $(j, k)$ . The mean of this density is  $\frac{j}{j+k}$ , while the variance is  $\frac{jk}{(j+k)^2(j+k+1)}$ . If the prior parameters for  $\theta$ ,  $S$  and  $C$  are given by  $(j_\theta, k_\theta)$ ,  $(j_S, k_S)$ , and  $(j_C, k_C)$ , respectively, then the marginal posterior density for  $\theta$  (6) becomes

$$f(\theta|x) \propto [\theta s + (1 - \theta)(1 - c)]^x [\theta(1 - s) + (1 - \theta)c]^{n-x} \times \theta^{j_\theta} (1 - \theta)^{k_\theta} s^{j_S} (1 - s)^{k_S} c^{j_C} (1 - c)^{k_C}. \quad (7)$$

Since in general there is no closed form solution to (6) or even (7), Joseph, Gyorkos and Coupal (1995) used the Gibbs sampler (Gilks, Richardson and Spiegelhalter 1996) to estimate the marginal posterior density of the prevalence given Beta prior distributions for  $\theta$ ,  $S$  and  $C$ . This method is computationally intensive, however, and not suitable for use in algorithms that require repeated use. For example,  $n + 1$  applications of the Gibbs sampler would be required here to estimate the average posterior coverage for each step with sample size  $n$  in the bisectional search. Below we describe a more efficient algorithm.

### 2.3 Bayesian Sample size determination for prevalence studies

We used the non-iterative Monte Carlo algorithm recently described by Ross (1996) to estimate the posterior average coverages. The algorithm proceeds as follows: First, a random sample of size  $k$  is drawn from the joint prior density of  $(\theta, S, C)$ , where  $k$  is typically of size 1,000 or greater. Label these points  $(\theta_i, S_i, C_i)$ , for  $i = 1, 2, \dots, k$ . A weight function,  $\omega_i(x)$ ,  $i = 1, 2, \dots, k$  is attached to each sampled point, where  $\omega_i(x)$  is proportional to the right hand side of (7). The posterior mean given  $x$  is then estimated by

$$\hat{\theta}(x) \approx \frac{\sum_{i=1}^k \theta_i \times \omega_i}{\sum_{i=1}^k \omega_i}.$$

The coverage for each  $x$ ,  $coverage(x; d)$ , is then estimated by the total normalized weights of points  $\sum \theta_i \omega_i$ , where the summation is over points  $i$  with values  $\theta_i$  that fall within the interval  $[\hat{\theta} - d, \hat{\theta} + d]$ . Finally, the average coverage is approximated by the weighted average of the above coverage probabilities

$$coverage(d) \approx \sum_{x=0}^n coverage(x; d) m(x),$$

where the marginal probability function of  $x$  is estimated by

$$m(x) \approx \sum_{i=1}^k \frac{\omega_i}{k}.$$

The ACC sample size is given by the smallest  $n$  such that  $coverage(d)$  is at least  $1 - \alpha$ . Software written in the Splus language that implements this algorithm to calculate average coverages given Beta prior distributions for  $\theta$ ,  $S$ , and  $C$  is available from the authors.

	$n$	$s$	$c$	<i>coverage</i>
1	1473	0.9	0.9	0.950
2	1473	U[0.85,0.95]	0.9	0.947
3	1473	U[0.9,1]	U[0.9,1]	0.618
4	1473	U[0.85,0.95]	U[0.9,1]	0.595
5	1473	U[0.85,0.95]	U[0.85,0.95]	0.589

Table 1: Variation of the average coverage probabilities with increasing uncertainty in the estimation of the sensitivity ( $s$ ) and specificity ( $c$ ), for  $d = 0.02$ . The prior distribution on  $\theta$  is  $U[0, 0.1]$ .

The above algorithm can also be applied when the sensitivity and specificity are known, by simply fixing the sampled points  $S_i$  and  $C_i$  at their true values in each  $(\theta_i, S_i, C_i)$  triplet. This gives the Bayesian analogue to the frequentist sample size given by (3).

While instances with exactly known  $S$  and  $C$  are rare in practice, it is instructive to examine the effect that less than perfect knowledge about the sensitivity and specificity of a test has on average coverage probabilities. To illustrate this effect, consider a Uniform prior density on the range  $[0, 0.1]$  for  $\theta$ , which would be appropriate for a rare condition known to occur in less than 10% of the population. For  $w = 0.04$  and fixed  $s = c = 0.9$ , a sample size of  $n = 1473$  is required such that the average coverage probability is 0.95. Keeping this sample size fixed, we then investigated the effects on the average coverage probability of using Uniform prior densities on  $S$  and  $C$  with varying support.

The results are displayed in Table 1. As expected (see the example in Section 1), uncertainty about the values of  $S$  and  $C$  can substantially reduce the coverage probabilities. It is especially interesting to note that if the prior information implies that the sensitivity and the specificity of the diagnostic test must be larger than a given fixed value (here 0.9), but the exact values are not known, then the coverage probability still substantially decreases even though one is averaging over tests with better properties. This is seen by comparing line 1 to line 3 in Table 1. To gain an intuitive understanding as to why the uncertainty about the test properties has a larger effect than the values of the properties themselves, consider the following two extreme cases:

**Extreme Case I:** Suppose that  $s = c = 0$ , so that the test is guaranteed to never provide the correct diagnosis. While this test is as poor as can be, since we assume that we know its properties exactly, equation (2) reduces to  $p = (1 - \theta)$ , so in fact this test provides as much information as a perfect test with  $s = c = 1$ .

**Extreme Case II:** Now consider a near perfect test, where  $s \in [0.99, 1.0]$  and  $c \in [0.99, 1.0]$  but the exact values are unknown. This test is superior by a wide margin compared to the test in Case I, but clearly provides less information, since some uncertainty is added

by the slight lack of precision in  $s$  and  $c$ . Hence it is easy to see how tests exactly known to be bad can “outperform” excellent tests whose properties are inexactly known.

For low prevalences (here less than 10%), the uncertainty about the specificity has more effect on the coverage probabilities than the uncertainty about the sensitivity. The reason for this is found in equation (2). Since  $S$  is multiplied by  $\theta$ , when  $\theta$  is small the effect of changing values of  $S$  is also small, and therefore the effect of the uncertainty about  $C$  is larger.

### 3 Estimating the prevalence of toxoplasmosis

Consider again the prevalence study of *Toxoplasma gondii* discussed in the introduction. A Uniform prior density on the interval  $[0,1]$  (equivalent to a Beta(1,1) density) was considered appropriate for  $\theta$ , since there had been no recent or representative data from Quebec, and expert opinion varied widely. If we fix  $s = 65/68 = 0.955$  and  $c = 22/22 = 1$ , then a sample size of 309 is adequate to attain 95% average coverage for intervals of width  $w = 0.1$ . The ACC ensures the desired coverage only on average, and so provides a sample size that is less than those given by either (1) or (3). This is because  $\theta = 0.5$  is the most conservative value leading to the smallest coverage probability, while the ACC averages coverage probabilities over prevalence values across the interval  $[0,1]$ . A conservative Bayesian criterion (see Joseph, du Berger and Bélisle 1997) that ensures the desired coverage over all data which may arise in this case will provide a sample size estimate similar to that produced by (3).

Owing to the conjugacy of the Beta family of densities with the binomial likelihood function, the  $j$  and  $k$  prior parameter values from a Beta density (see Section 2.2) can be considered as equivalent to a previously observed number of successes and failures, respectively. See, Gelman et al (1995, page 40). Therefore, from Wilson and Ware (1991), reasonable prior densities for the sensitivity and specificity of this test may be  $S \sim \text{Beta}(65.1, 3.1)$  and  $C \sim \text{Beta}(22.1, 0.1)$ , respectively. These arise from disturbing each count in the observed data by the equivalent of one-tenth of an extra observation, avoiding the  $\text{Beta}(22, 0)$  density, which is degenerate. By the algorithm of Section 2, a sample size of approximately 580 is needed for the average posterior coverage probability of credible intervals of width 0.1 to be at least 0.95. Therefore, taking into account even the relatively small uncertainties about the values of  $S$  and  $C$  leads to a substantially larger sample size, compared to any of the other methods that ignore this source of uncertainty.

Accurately determining what is known about the test properties is crucial, as the final sample sizes (and reported prevalence estimates) can greatly depend on this exercise. For example, suppose that  $S \sim \text{Beta}(66, 4)$  and  $C \sim \text{Beta}(23, 1)$  were used instead of the prior densities given above. These would also be a reasonable choice, since they arise from updating initial uniform prior densities with the data available at the time of the study. For  $w = 0.1$ , and again assuming a uniform prior density for  $\theta$ , we computed the average coverage probability for values of  $n$  ranging from 100 to 2000 with increments of 100, and with Monte Carlo size of  $k = 1000$ .

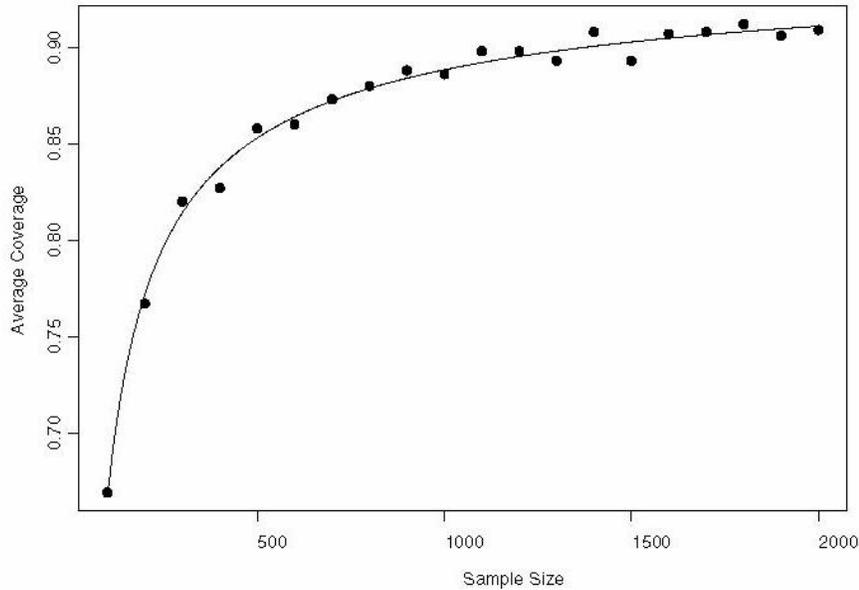


Figure 1: Posterior average coverage probabilities for  $w = 0.10$ , from  $n = 100$  to  $n = 2000$ . The solid line provides the values of the average coverage probabilities predicted by a logistic regression model.

The results, presented in Figure 1, indicate that the posterior average coverage probability does not reach 0.95 even with a sample size of 2000. This is because while increasing the sample size always improves the precision in estimating  $p$ , past a certain point, larger sample sizes do not provide improved estimates of the sensitivity and the specificity of the diagnostic test, essentially because the problem is non-identifiable. While some updating of the prior distributions for  $S$  and  $C$  is possible even in the presence of misclassification, these will not converge to a single point even with an infinite sample size (Joseph 1997). Since estimating  $\theta$  depends on knowledge about  $S$  and  $C$ , increasing the sample size does not always provide increased accuracy in estimating  $\theta$ . While there are three parameters to estimate ( $\theta$ ,  $S$ , and  $C$ ), there is only one degree of freedom available when  $x$  positive tests are observed in a sample size of  $n$  subjects. Walter and Irwig (1988) and others have suggested using several independent tests, and when three or more tests are used, all parameters can be simultaneously estimated without constraints. Increasing the number of diagnostic tests used may not always be feasible, however, and if only one or two tests are used, Bayesian estimation provides the best possible estimates which use all of the available prior information and correctly accounts for all inherent uncertainty (Joseph, Gyorkos and Coupal 1995, Neath and Samaniego 1997). The marginal posterior density for  $\theta$  does not necessarily converge to a point mass as the sample size increases, however, unless  $S$  and  $C$  are *a priori* exactly known (Joseph 1997). Kuha (1997) has discussed

the problem of how many subjects should be verified with a gold standard test in order to validate the test.

The sample size beyond which further sampling provides little additional precision in estimating  $\theta$  is a complex function of the prior information available on  $\theta$ ,  $S$  and  $C$ . Given data such as that presented in Figure 1, however, it may be possible to construct a regression model to estimate the sample size beyond which further sampling does not substantially improve estimation of  $\theta$ , as well as the maximum possible average coverage probability. As the average coverage probabilities must be between 0 and 1, a generalized linear model with logit link may be appropriate. Since the average coverage probability, as a function of  $n$ , seems to have an asymptote with value smaller than 1, we looked for a model of the form

$$\log\left(\frac{\mu}{1-\mu}\right) = u + v/\sqrt{n}, \quad (8)$$

where  $\mu$  is the average posterior coverage probability minus the prior coverage probability of an interval of width  $w$ , and where  $u$  and  $v$  are estimated from the data.

We considered a quasi likelihood model (McCullagh and Nelder 1989) with logit link and constant variance. The latter was considered appropriate, since we hypothesized that a large component of the residual error in this model arises from using the Monte Carlo approximation, whose error varies more with  $k$  than with  $n$ . Of course, the Monte Carlo error can be made as small as desired simply by increasing  $k$ .

Let  $n_0$  be the sample size after which the average coverage probability will not improve by more than a given  $\epsilon > 0$ , even if the sample size were to increase to infinity. From (8) simple algebra leads to

$$n_0 \geq \left( v / \log\left( \frac{\exp(u) - \epsilon(1 + \exp(u))}{\exp(u) + \epsilon(1 + \exp(u)) \exp(u)} \right) \right)^2. \quad (9)$$

Since the prior distribution on  $\theta$  was uniform, the prior coverage probability of the interval of width  $w = 0.1$  is also 0.1. Using the data from Figure 1, the estimated coefficients were  $u = 1.799$  and  $v = -15.303$ , so that the upper limit of the average coverage probability is approximately

$$\exp(1.799)/(1 + \exp(1.799)) + 0.1 = 0.958.$$

From (9), the sample size needed for the posterior average coverage probability to be within  $\epsilon = 0.008$  of this upper limit, that is, to reach an average coverage of 0.950, is 56506. We also calculated the posterior average coverage probability predicted by the model for  $n = 3000, 4000$ , and 5000. These values were 0.920, 0.926, 0.930 respectively, while the corresponding values given by the Monte Carlo approximation were 0.920, 0.920, and 0.921, respectively. Therefore, the model still seems to predict reasonably well, even far outside of the range of the data from which it was estimated.

The main difference between the two sets of prior distributions is the lower limit of the specificity of the test. With data on only 22 subjects, the choice of “noninformative” prior distribution plays a large role. Of course, one cannot be certain of the accuracy of the upper

limit of 0.958 given by the logistic regression model, but it is clear that if the second set of prior densities are used, a very large sample size is needed to achieve the desired accuracy. In the end, a sample size near  $n = 580$  subjects was selected, but a main conclusion was that it will be very worthwhile to gather more data on the properties of the test. For example, if twice as much prior data had been collected on the test properties, then assuming the same proportions were observed,  $S \sim \text{Beta}(130.1, 6.1)$  and  $C \sim (44.1, 0.1)$  prior densities could have been used, leading to a sample size of only 385. If twice the prior information is combined with uniform prior distributions ( $S \sim \text{Beta}(131, 7)$ ,  $C \sim (45, 1)$  and  $\theta \sim \text{Beta}(1, 1)$ ), the sample size is 605, a substantial reduction from 56506 obtained earlier. Prior knowledge about the prevalence, when available, can also be useful. For example, if the prevalence is known *a priori* to be between 10% and 50%, so that a  $\text{Beta}(6, 14)$  prior density on  $\theta$  is appropriate, then the sample size further reduces to 348 from 385.

## 4 Discussion

Estimating a binomial parameter is the aim of many studies. When no misclassification is possible, standard binomial formulae can be used to determine the sample size required to estimate the parameter to any desired accuracy. Unfortunately, misclassification often occurs, and in general, one does not know the exact magnitude of the errors so that simple formulae like (3) cannot be applied. In this case, a Bayesian approach can be used to determine sample size requirements. In particular, we show that it is important when planning a study to estimate the magnitude of misclassification errors as accurately as possible, since the sample size estimates are highly sensitive to uncertainty about their values. In many cases, it may not be possible to estimate a parameter to the desired accuracy with the information currently available, when all uncertainty is fully accounted for. Copas (1988) investigated misclassification errors in binary regression models.

While in this paper we considered data from diagnostic tests, the methods can easily be applied to any similar situation. For example, suppose a political election poll is being planned, but it is suspected that not everyone polled will in fact give their true voting intentions. If one is willing to specify prior distributions on the error rates, the correct sample size for the desired precision can be calculated.

## References

- Adcock, C. A. (1988) Bayesian approach to calculating sample sizes for multinomial sampling. *The Statistician* **36**, 155–159.
- Adcock, C. A. (1997) Sample size determination: a review. *The Statistician* **46**, 261–283.
- Chaloner, K. and Verdinelli, I. (1995) Bayesian experimental design: A review. *Statistical Science* **10**, 273–304.

- Copas, J. B. (1988) Binary regression models for contaminated data (with discussion). *Journal of the Royal Statistical Society, Series B* **50**, 225–265.
- Gelman, A., Carlin, J., Stern, H., and Rubin, D. (1995) *Bayesian Data Analysis*. London: Chapman and Hall.
- Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996) *Markov-Chain Monte Carlo in Practice*. London: Chapman Hall.
- Gould, A. L. (1993) Sample sizes for event rate equivalence trials using prior information. *Statistics in Medicine* **12**, 2009–2023.
- Joseph, L. (1997) Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard – Reply. *American Journal of Epidemiology* **145**, 291.
- 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.
- Joseph, L., Gyorkos, T.W. and Coupal, L. (1995) 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. B. and du Berger, R. (1995) Sample size calculations for binomial proportions via highest posterior density intervals. *The Statistician* **44**, 143–154.
- Kuha, J. (1997) Estimation by data augmentation in regression models with continuous and discrete covariates measured with error. *Statistics in Medicine* **16**, 189–201.
- Lindley, D. V. (1997) The choice of sample size. *The Statistician* **46**, 129–138.
- McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*. New York: Chapman and Hall.
- Neath, A. A. and Samaniego, F. J. (1997) On the efficacy of Bayesian inference for nonidentifiable models. *The American Statistician* **51**, 225–232.
- Rahme, E. and Joseph, L. (1998) Prevalence estimation for a rare disease: Adjusted maximum likelihood. *The Statistician* **47**, 149–158.
- Ross, S. M. (1996) Bayesians should not resample a prior sample to learn about the posterior. *The American Statistician* **50**, 116.
- Spiegelhalter, D. J. and Freedman, L. S. (1986) A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statistics in Medicine* **5**, 1–13.
- Walter, S. D. and Irwig, L. M. (1988) Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *Journal of Clinical Epidemiology* **41**, 923-937.
- Wilson, M. and Ware, D. (1991) Evaluation of the Diagnostic Pasteur Platelia Toxo IgG and Toxo IgM kits for detection of human antibodies to *Toxoplasma gondii*. 91<sup>st</sup> General Meeting, American Society for Microbiology, Dallas, Texas, May 5-9, Abstract number V-23.