To Appear in American Journal of Epidemiology

Bayesian sample size determination for case-control studies when exposure may be misclassified

Running Head: Planning case-control studies with exposure errors

Lawrence Joseph* Patrick Bélisle†

July 5, 2013

*Corresponding author: Division of Clinical Epidemiology, McGill University Health Centre, Royal Victoria Hospital, 687 Pine Avenue West, V Building, Room V2.10, Montreal, Quebec, H3A 1A1, Canada, and Department of Epidemiology and Biostatistics, 1020 Pine Avenue West, McGill University, Montreal, Quebec, H3A 1A2, Canada. Email: lawrence.joseph@mcgill.ca, Tel: 514-934-1934 ext 44713, Fax: 514-934-8293.

†Division of Clinical Epidemiology, McGill University Health Centre, Royal Victoria Hospital, 687 Pine Avenue West, V Building, Room V2.08, Montreal, Quebec, H3A 1A1, Canada. Email: patrick.belisle@clinepi.mcgill.ca

Abstract

The odds ratio (OR) is frequently used for estimating the effect of an exposure on the probability of disease in case-control studies. In planning such studies, methods for sample size determination are required to ensure sufficient accuracy in estimating the OR once the data are collected. Very often, the exposure used in epidemiological studies is not perfectly ascertained. This can arise from recall bias, use of a proxy exposure measurement, uncertain work exposure history, and laboratory or other errors. The resulting misclassification can have large impacts on the accuracy and precision of estimators, and specialized estimation techniques have been developed to adjust for these biases. However, much less work has been done to account for the anticipated decrease in precision of estimators at the design stage. We develop methods for sample size determination for ORs in the presence of exposure misclassification, using several interval-based Bayesian criteria. Using a series of prototypic examples, we compare sample size requirements after adjusting for misclassification to those required when this problem is ignored. We illustrate the methods by planning a case-control study of the effect of late introduction of peanut to the diet of children to the subsequent development of peanut allergy.

Key words: Bayesian methods; case-control study; misclassification error; sample size determination; study design.

List of Abbreviations

ACC	Average Coverage Criterion
ALC	Average Length Criterion
HPD	Highest Posterior Density
MWOC	Modified Worst Outcome Criterion
OR	Odds Ratio
WOC	Worst Outcome Criterion

Introduction

Statistical techniques that adjust for possible biases in observational studies are increasingly common in epidemiology (1-4). Methods for designing studies that will eventually need to be adjusted for such biases, however, have lagged behind. This paper addresses this gap in presenting a method for adjusting sample size requirements for case-control studies with possible misclassification bias.

As discussed in the classic text by Schlesselman (5), case-control designs can be used to estimate the effect of an exposure on the probability of a disease or condition. For example, it is hypothesized that late introduction of peanut to the diet may increase the probability of peanut allergy in children (6). Suppose that a case-control study is to be conducted to estimate the effect of introduction of the food prior to one year old compared to later introduction. Groups of peanut allergic and non-peanut allergic children will be surveyed, with their parents providing information about when their children were first introduced to peanut or products containing peanut. Exposure misclassification may arise from inaccurate recall of exposure information, which may also differ in magnitude between cases and controls. Under these circumstances, what sample size is required for accurate estimation of the odds ratio, once statistical methods which adjust for the misclassification due to recall bias are applied?

Ignoring misclassification, Wickramaratne (7) and Lemeshow et al (8) review classical sample size methods for hypothesis testing and interval estimation for odds ratios. Frequentist sample size methods depend on accurate point estimates of the required inputs, which here include not only the exposure rates, but also the rates of misclassification within each disease class. As we will show, the estimated sample sizes can be very sensitive to minor changes to these inputs. It is therefore advantageous to consider Bayesian methods, where prior densities not only allow for uncertainty in the inputs, but incorporate this uncertainty into the sample size requirements. This is especially important in the presence of misclassification, which induces a non-identified model. As discussed by various authors (9-12), calculating sample sizes within non-identified models is inherently different from regular

problems, since the posterior density does not converge to a single point as the sample size increases. Therefore, even infinite sample sizes may not guarantee sufficient accuracy.

From a Bayesian viewpoint, sample size for case-control studies, including examination of the optimal control-to-case ratio, was addressed by De Santis et al (13) and M'Lan et al (14). Neither, however, considered the change in sample size resulting from possible exposure misclassification. Devine and Smith (15) addressed the change in sample size requirements induced by misclassification using frequentist power based criteria. Gustafson (16) reviewed general Bayesian methods to adjust for misclassified exposure data, and Stamey and Gerlach (17) considered misclassification for case-control studies, but their method requires a validation sample that is not always available.

The outline of this paper is as follows. Section 2 summarizes various Bayesian sample size criteria based on highest posterior density (HPD) credible interval lengths, and applies them to estimating odds ratios from case-control studies in the presence of misclassification. We consider not only the uncertainty in misclassification rates, but also allow a search for the optimal control-to-case ratio, an important component of study design. Sample sizes from a series of prototypic examples are given in Section 3, comparing the change in sample size with and without consideration of misclassification. Section 4 returns to the peanut allergy study, providing the required sample sizes first without misclassification and then under a plausible range of misclassification rates. Since peanut allergy is relatively rare, we also consider control-to-case ratios greater than one, and check the effect on overall sample sizes. We end with a discussion in Section 5.

Sample size determination for odds ratios in the presence of misclassification

We begin by describing a model for adjusting odds ratios for exposure misclassification, similar to that used by Gustafson et al (18). We next define several Bayesian sample size criteria, and indicate how they can be applied to help design studies with bias adjusted odds ratios. Technical details and numerical algorithms are given in the appendix.

Let $i = 1, 2$ index the case and control populations, respectively. For any given sample size N, we observe the two-by-two layout given in Table 1. Let the true probability of exposure among cases be given by p_1, and the true probability of exposure among controls be p_2. Let $p_1^{'}$ and $p_2^{'}$ be the observed probabilities in the presence of misclassification. The two sets of probabilities are related by the equations

$$\begin{aligned} p_1^{'} &= p_1 * S_1 + (1 - p_1) * (1 - C_1) \quad \text{and} \\ p_2^{'} &= p_2 * S_2 + (1 - p_2) * (1 - C_2) \ , \end{aligned} \tag{1}$$

where S_i, and C_i are correct classification rates within case ($i = 1$) or control ($i = 2$) populations, defined by

$$\begin{aligned} S_i &= Pr\{\text{classified as exposed} \mid \text{truly exposed}\} \quad \text{and} \\ C_i &= Pr\{\text{classified as not exposed} \mid \text{truly not exposed}\} \ . \end{aligned}$$

If $S_1 = S_2$ and $C_1 = C_2$ then there are 2 correct classification rate parameters to estimate, and otherwise there are four, giving a total of $m = 4$ or $m = 6$ unknown parameters, when added to p_1 and p_2. As all parameters have range [0,1], beta densities may be used as prior distributions. The available data in Table 1 provide only three degrees of freedom, meaning in practice that the problem is non-identifiable, and that informative prior distributions need to be placed over a subset of parameters in order to obtain reasonable inferences. Typically one supposes some knowledge or limits about the classification rates, after which one can use less informative priors over p_1 and p_2.

The likelihood function for the observed data is a product of two binomial functions, and using the notation from Table 1 is proportional to

$$(p_1^{'})^a (1 - p_1^{'})^b (p_2^{'})^c (1 - p'_2)^d. \tag{2}$$

The posterior density function is proportional to the product of the likelihood function and the prior density over the unknown parameters $\{p_1, p_2, S_1, S_2, C_1, C_2\}$. The odds ratio is then estimated by integrating the full posterior density to eliminate the nuisance parameters $\{S_1, S_2, C_1, C_2\}$, and introducing the change of variable $OR = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}$. As there is no

closed form solution, inference proceeds by Markov Chain Monte Carlo methods. We used WinBUGS (19) to implement a Gibbs sampler algorithm.

The marginal posterior density of the OR is typically summarized by an HPD credible interval. In planning a study, suppose we desire an interval of length l that includes the OR with probability $1-\alpha$. For example, we may wish to estimate the OR to an accuracy of $l = 0.5$ with a $1-\alpha = 95\%$ interval. The marginal posterior density of the OR depends on the data, which are of course unknown at the planning stage. We can account for this uncertainty in different ways, leading different sample size criteria. We consider three such criteria, the Average Coverage Criterion (ACC), the Average Length Criterion (ALC), and the Modified Worst Outcome Criteria (MWOC), defined in detail in M'Lan et al (14).

Allowing the posterior probability $1-\alpha$ to vary with each potential data set while holding the credible interval length l fixed, leads to a sample size that guarantees the desired posterior probability on average, that is, the ACC sample size. The average is taken over the set of all possible data sets, weighted by the probability that each data may arise, as determined by the prior densities over all parameters.

Conversely, we can allow the HPD interval length l to vary while fixing the posterior probability at $1-\alpha$. This ALC sample size averages the lengths of fixed probability HPD intervals over all possible data sets, again weighted by the prior distributions.

Rather than averaging over potential data sets, a conservative approach would be to ensure a maximum length of l and a minimum probability of $1-\alpha$, regardless of the data set that occurs, termed the WOC sample size. 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 the posterior density of an odds ratio where the rate in the denominator might be close to zero, causing instability in the odds ratio estimate. Therefore, in this paper we use the MWOC the desired length and posterior probability are guaranteed over a subset of all data sets with a given probability. For example, we might choose the sample size such that the desired l and $1-\alpha$ are guaranteed over 95% of all data sets. We denote this by MWOC(95), or more generally, MWOC($100 \times (1-\gamma)$), where

γ represents the probability that a randomly selected data set will not satisfy the length and/or posterior probability requirements. Use of the MWOC avoids the situation of having to select an unnecessarily large sample size to guard against highly improbable data.

Since there are no closed form solutions, we used the numerical algorithm detailed in the appendix to estimate the optimal sample size. A user-friendly program called "SSCOR" (Sample Size Calculations for Odds Ratios) that implements all of the above methods is available from the first author's web site, at `www.medicine.mcgill.ca/epidemiology/Joseph/`. We next use this software to determine sample sizes for various scenarios that may occur in the planning of case-control studies, comparing situations with and without misclassification.

Sample sizes for prototypic scenarios

Misclassification of exposure and the subsequent need to adjust the odds ratio to account for these errors has important implications for the design of case-control studies. In general, the larger the misclassification error and the more uncertain one is of the magnitude of this error, the more uncertainty there will be in the final odds ratio estimate, and consequently, the larger the sample size requirements will be. In this section we present some prototypic scenarios that will illustrate the degree to which sample size needs increase with increasing amounts of misclassification.

Throughout we will assume that a 95% HPD interval is desired. We will consider three values for the true odds ratio, approximately centered on OR values of 0.7, 1 and 1.5, with desired total HPD interval lengths of 0.4, 0.2 and 0.8, respectively. The latter were chosen to be sufficiently small such that definitive inferences can be made. For example, if the true OR = 0.7 and the length of the HPD interval is 0.4, then a 95% interval close to (0.5, 0.9) can be expected, sufficiently far from the null value to convincingly demonstrate a protective effect. On the other hand, if the OR = 1, then a total length 0.2 will result in an interval similar to (0.9, 1.1), which will often be close enough to the null value to conclude no clinically important effect. Similarly, when the OR is 1.5, the interval will be approximately (1.1, 1.9), which we assume is far enough from the null value to conclude a positive effect.

We assumed a beta(10, 90) prior for the exposure rate within the case group. This density provides a mean rate of 10%, typical of many exposures, and the parameters sum to 100, providing knowledge equivalent to 100 prior observations. To obtain ORs of 0.7, 1, and 1.5, we modeled the exposure rate amongst controls by beta(13.7, 86.3), beta(10, 90) and beta(6.9, 93.1) densities, respectively. These provide median ORs (95% prior credible intervals) of 0.694 (0.281, 1.65), 1 (0.387, 2.58), and 1.52 (0.550, 4.47), centered close to the target values.

For each of these scenarios we calculated sample sizes under no misclassification, and assuming low, moderate and high degrees of misclassification. For moderate and high degrees of misclassification, we also considered narrower and wider prior densities around the central value, since knowledge about the misclassification rate can have as much of an effect on the sample size as the misclassification rate itself. Low misclassification was defined by a beta(681.5, 13) prior probability of correct classification, a density with 95% range from 0.97 to 0.99, implying a misclassification rate between 1% and 3%. We used beta(116, 12) and beta(214.35, 70.8) densities to represent the moderate and high misclassification rates with wider prior ranges, respectively, implying error rates between 5% and 15% and between 20% to 30%. For narrower prior ranges, we used misclassification rates between 9% and 11% for the moderate error rate, and between 24% and 26% for high rates, corresponding to beta(3103.9, 344) and beta(5400.3, 1799.5) densities for correct classification, respectively.

For each of the above $3 \times 6 = 18$ scenarios, we considered four criteria, the ACC, ALC, MWOC(50) and MWOC(90). We calculated two sample sizes for each scenario (except no misclassification), depending on whether one assumes the same or allows for different misclassification rates within the diseased and non-diseased populations. This determines whether there are four or six unknown parameters to estimate. Thus, we considered a total of 132 different scenarios. An upper limit of 100,000 subjects per group was set, as larger studies would usually not be practical.

Although any joint density over $\{p_1, p_2, S_1, S_2, C_1, C_2\}$ can be used, aside from the four parameter model that sets $S_1 = S_2$ and $C_1 = C_2$, we have chosen independent priors for

each parameter. While equation (1) exposes the relationship between $p_1^{'}$ and $p_2^{'}$ and the parameters from which they are derived, there is no particular reason to suspect that the correct classification probabilities are dependent on p_1 and p_2. Nevertheless, should this be the case, our methods are easily modified by using a different joint prior over the parameter space.

Table 2 presents the resulting sample sizes assuming $g = 1$. When there is no misclassification, the sample sizes range from a low of 813 for the MWOC(50) criterion when the OR is 0.7, to a high of 32,072 for the most strict criterion, the MWOC(90) with OR=1. The ACC and ALC sample sizes are intermediate to these extremes. In general, the sample sizes were largest for the narrowest interval around an OR=1, and smallest for OR=0.7 with an interval length of 0.4.

Under the $m = 4$ parameter model, even low rates of misclassification greatly increase the desired sample sizes. For example, there was an approximately 20% increase in sample sizes when OR=0.7, but often more than a doubling of the sample size when OR=1, including two cases where the 100,000 ceiling was reached, under the MWOC(90). As expected, moderate and high levels of misclassification require even larger sample sizes, and having a better knowledge of the misclassification rate decreases the sample size compared to when this rate is less well known, *a priori*.

The situation is considerably worse when the misclassification rates are not assumed identical across groups, that is, when $m = 6$. Indeed, even low degrees of misclassification often create sample sizes above 100,000, except for the scenarios with OR=0.7. High degrees of misclassification almost always leads to very large sizes.

Gustafson et al (18) derives the posterior density of the OR under misclassification assuming an infinite sample size. Typically, the posterior density of the OR substantially narrows up to a certain sample size, past which there are diminishing returns. When combined with the Monte Carlo methods given in the Appendix of Dendukuri et al (10), it is possible to determine whether an infinite sample size will satisfy a given criterion. For example, Table 2 indicates that the ACC sample size for an OR of 1 for moderate misclassification with

a narrow prior is $> 100,000$, but a sample size of just under 200,000 per group (400,000 in total) is sufficient to satisfy the ACC, reaching an average posterior probability of 0.952 at 200,000. On the other hand, even an infinite sample size is not sufficient for this same situation under high misclassification with a wide prior.

Sample size required to accurately estimate the effect of late introduction to peanut on peanut allergy

The overall prevalence of peanut allergy among Montreal children is approximately 1.5% (20). Suppose we anticipate 60% of cases have late exposure (95% prior interval from 55% to 65%, represented by a beta(229.8, 153.2) density), compared to 30% of controls with late exposure (95% prior interval from 25% to 35%, represented by a beta(100.5, 234.5) density), giving an OR close to 3.5 (95% CrI (2.58, 4.80)). If a case-control study is being planned, what should the sample size be so that the OR is estimated to within a total HPD interval length of 1?

The sample size will depend on many factors, including which sample size criterion will be used, how many controls will be selected for each case, whether one allows for possible misclassification errors in parental information about the timing of first introduction to peanut in the diet, and the degree to which this misclassification is assumed to be known.

Table 3 provides the required sample sizes under a wide variety of possible design choices. Somewhat unrealistically, if no misclassification is assumed, for a control-to-case ratio of $g = 1$, the sample sizes range from under 3000 to over 6000, depending on the criterion selected. If one is content to ensure a posterior HPD length of 1 only on average (or median), sample sizes close to 3000 are needed. If one wants to be 95% certain of obtaining an HPD interval of 1 or less, then sample sizes in the 5000 to 6000 range are required. The final choice can be based on the trade-off between the certainty of obtaining an HPD interval of length 1 versus the costs associated with the larger sample sizes. The effect of the control-to-case ratio can also be gleaned from Table 3, where higher values of g raise the total sample sizes requirements by a few percentage points. While a ratio of $g = 1$ is optimal, difficulties in

finding cases, which are relatively rare, may lead to other choices.

If one more realistically assumes a low misclassification rate of 1% to 3%, input as a beta(681.5, 13) density for the correct classification probability, and assumes the rate to be equal in the two groups ($m = 4$), then sample sizes rise by roughly 10% across all criteria. Under a moderate rate of misclassification of 9% to 11%, as represented by a beta(3103.852, 344.0302) density for the correct classification probability, the sample sizes rise by about 60% compared to the no misclassification case, although they may remain feasible.

It is interesting to consider what may happen if one plans a sample size for a study ignoring measurement error, but later analyses the data considering measurement error. For example, if there are equal numbers of cases and controls ($g = 1$), the MWOC(50) from Table 3 suggests a total sample size of 2654 assuming no measurement error. If this size is used for the study which is subsequently analyzed using a moderate rate of misclassification of between 9% and 11%, then the length of an HPD interval of probability 0.95 will be 1.2, about 20% wider than the original planned length.

We can also evaluate the effect of the $m = 6$ parameter model. If we use a moderate rate of misclassification of 9% to 11% in both groups, but allow distinct parameters for these rates, the ALC sample size under $g = 1$ and an HPD length of 1 is 5248. However, it is also reasonable to assume that misclassification rates are lower in the peanut allergic group (cases) compared to the non-allergic group (controls), since cases may make more effort to recall, or may remember the history more accurately, given the likely reaction that would have occurred in the child upon early ingestion. For example, we might assume a moderate misclassification rate of 9% to 11% for cases, and a larger misclassification rate of 15% to 25% for controls. While the latter interval is wide, it remains plausible that many controls might not accurately remember this history, and the exact recall rate would typically not be accurately known. With $g = 1$ and using the ALC criterion with a total HPD interval length of 1, the total sample size is greater than 200,000 (over 100,000 per group). Under these conditions, obtaining HPD lengths of 1 may not be feasible, even though the inputs are entirely plausible. Since an OR of 3.5 is far from the null value of OR = 1, a study with

lower accuracy is still informative. Doubling the width from 1 to 2 reduces the sample size considerably, to a very manageable 362.

All of the above sample sizes are fully Bayesian, in the sense that relatively strong prior information is assumed for each parameter, and used not only for the purposes of planning the study, but also within the eventual analysis. One cannot use non-informative (say, beta(1,1) or uniform prior densities) across all parameters, since the problem is non-identifiable. However, it is possible to use relatively weak prior information for the exposure prevalences within the case and control groups, provided good prior information is available for the misclassification parameters. This strategy can be used by researchers who prefer to "let the data speak for themselves" at the analysis stage, while still planning their studies to accommodate possible misclassification errors.

For example, we can consider the prior densities used above, but divide each parameter by 10, reducing the prior effective sample size. Thus, the beta density for the probability of exposure in cases changes from beta(229.8, 153.2) to a beta(22.98, 15.32), and that for the controls changes from beta(100.5, 234.5) to beta(10.05, 23.45). Summing across beta parameters given a prior effective sample size of 71.8, compared to the previous size of 718. With $g = 1$, an ALC sample size of 5202 is required to attain a total HPD length of 1, even in the absence of misclassification. Under moderate misclassification of 9% to 11%, the sample size roughly doubles to 9626. Reducing the effective prior sample size by another factor of 10 to 7.18 returns sample sizes that are larger than 200,000, even with no misclassification.

Another strategy is to separate "design priors," used to generate the data, from "analysis priors," used at the analysis stage (21). While the non-identified nature of our problem requires at least some informative priors at both design and analysis stages, one can place uniform priors on p_1 and p_2 at the analysis stage. This will provide sample sizes for researchers wanting to use minimal prior information for the OR at the analysis stage. Under this strategy, the ALC sample sizes under moderate misclassification are 5908, 6483, and 7564 for $g = 1$, 2, and 3, respectively.

Discussion

The vast majority of case-control studies ignore misclassification, and even those that might consider this possibility ignore the uncertainty in *a priori* knowledge of these misclassification rates. This is true both at the design and analysis phases, leading to sample sizes that are typically much too small, and final estimates with credible intervals that are much too narrow. The methods presented here are important to the planning of such studies, and serve as a warning that ignoring misclassification in study planning and analysis may lead to wildly optimistic interval estimates.

We have discussed 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 for low values of γ is more conservative than either the ACC or ALC, which guarantee the target values for posterior probability and length only on average. As we have done for Tables 2 and 3 here, 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 inherent trade-offs between sample size and the risk of not meeting target values for interval length and posterior probability. Based on this information, a final sample size may be selected that balances statistical rigor with practical concerns. 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. Clearly, designing studies to have the lowest possible misclassification rates is important, if possible. In some cases, it must be admitted that no study design will result in misclassification rates low enough to derive sufficiently accurate estimates.

While we have applied our methods to the design of case-control studies with exposure misclassification, similar methods can be developed for other study designs and different sources of biases. As bias adjustments methods gain in popularity at the analysis stage (1-4), methods for designing such studies will also increase in importance.

Acknowledgements

Author Affiliations: Division of Clinical Epidemiology, McGill University Health Centre, Montreal, Canada (Lawrence Joseph & Patrick Bélisle); Department of Epidemiology and Biostatistics, McGill University, Montreal, Canada (Lawrence Joseph).

References

1. Greenland S. Multiple bias modeling for analysis of observational data (with discussion). *Journal of the Royal Statistical Society, Series A*. 2005;168(2):267-308.

2. Greenland S, Lash TL. Bias analysis. In: *Modern Epidemiology,* 3rd ed. Philadelphia: Lippincott-Wolters-Kluwer, 2008;345-380.

3. Hoggatt KJ, Greenland S, Ritz B. Adjustment for response bias via two phase analysis: An application. *Epidemiology.* 2009;20(6):872-879.

4. Greene N, Greenland S, Olsen J, Nohr EA. Estimating bias from loss to follow-up in the Danish National Birth Cohort. *Epidemiology.* 2011;22(6):815-822.

5. Schlesselman JJ. *Case-control studies: Design, conduct, analysis.* New York: Oxford University Press; 1982.

6. Du Toit G, Katz Y, Sasieni P, Mesher D, Maleki S, Fisher H, Fox A, Turcanu V, Amir T, Zadik-Mnuhin G, Cohen A, Livne I, Lack G. Early consumption of peanuts in infancy is associated with a low prevalence of peanut allergy. *Journal of Allergy and Clinical Immunology.* 2008;122(5):984-991.

7. Wickramaratne PJ. Sample Size Determination in Epidemiologic Studies. *Statistical Methods in Medical Research.* 1995;4(4):311-337.

8. Lemeshow S, Hosmer D, Klar J, Lwanga S. *Adequacy of sample size in health studies.* Chichester, England: John Wiley & Sons Ltd; 1990.

9. Rahme E, Joseph L, Gyorkos T. Bayesian sample size determination for estimating binomial parameters from data subject to misclassification. *Journal of the Royal Statistical Society, Series C, Applied Statistics*. 2000;49(1):119-228.

10. Dendukuri N, Rahme E, Blisle P, Joseph L. Bayesian sample size determination for prevalence and diagnostic studies in the absence of a gold standard test *Biometrics*. 2004;60(2):388-397.

11. Gustafson, P. Sample size implications when biases are modelled rather than ignored. *Journal of the Royal Statistical Society, Series A*. 2006:169(4);883-902.

12. Stamey J, Seaman J, Young D. Bayesian sample size determination for inference on two binomial populations with no gold standard classifier. *Statistics in Medicine*. 2005:24(19);2963-2976.

13. De Santis F, Perone Pacifico M, Sambucini V. Optimal predictive sample size for case-control studies. *Journal of the Royal Statistical Society, Series C, Applied Statistics*. 2004;53(3):427-441.

14. M'Lan E, Joseph L, Wolfson D. Bayesian sample size determination for case-control studies. *Journal of the American Statistical Association.* 2006:101(474):760-772.

15. Devine, OJ, Smith JM. Estimating sample size for epidemiologic studies: the impact of ignoring exposure measurement uncertainty. *Statistics in Medicine.* 1998:17(12);1375-1389.

16. Gustafson P. *Measurement error and misclassification in statistics and epidemiology.* New York: Chapman and Hall; 2004.

17. Stamey J, Gerlach R. Bayesian sample size determination for case-control stidies with misclassification. *Computational Statistics and Data Analysis*. 2007:51(6);2982-2992.

18. Gustafson P, Le ND, Saskin R. Case-control analysis with partial knowledge of exposure misclassification probabilities. *Biometrics*. 2001:57(2);598-609.

19. Spiegelhalter D, Thomas A. Best N. *WinBUGS Version 1.2 User Manual.* MRC Biostatistics Unit: Cambridge U. K.; 1999.

20. Ben-Shoshan M, Kagan R, Alizadehfar R. Joseph L. Turnbull E, St-Pierre Y, Clarke A. Is the prevalence of peanut allergy increasing? A five-year follow-up study on the prevalence of peanut allergy in primary school children in Montreal. *The Journal of Allergy and Clinical Immunology.* 2009;123(4):783-788.

21. Gelfand A, Wang F. A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science.* 2002;17(2):193-208.

22. Chen M, Shao Q. Monte Carlo Estimation of Bayesian Credible and HPD Intervals. *Journal of Computational and Graphical Statistics.* 1999;8(1):69-92.

Appendix

We now present the details of our numerical algorithm that determines the required sample sizes for estimating odds ratios in the presence of exposure misclassification. The algorithm is illustrated for the most general case where $S_1 \neq S_2$ and $C_1 \neq C_2$, but similar steps can be followed when the exposure classification probabilities are equal within case and control populations.

1. Sample M_1 random values from the joint prior density of $\{p_1, p_2, S_1, S_2, C_1, C_2\}$. This involves selecting values from a beta density for each parameter.

2. For each of the M_1 sets of parameters sampled in step 1, use equations (1) and (2) from Section 2 to calculate the probabilities of falling into each of the four cells defined by Table 1.

3. Select a tentative value for the sample size N, keeping in mind that a control-to-case ratio other than $g = 1$ may be selected. For each of the M_1 random situations, draw

M_2 random two-by-two tables a, b, c and d, using the probabilities calculated in step 2. This is equivalent to sampling data from the marginal distribution of the data. In practice, $M_2 = 1$ is sufficient.

4. For each of the $M_1 \times M_2$ data sets, run the Gibbs sampler algorithm via WinBUGS (15), to derive samples from the posterior densities for p_1 and p_2. Using these values, calculate $OR = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}$, a sample from the posterior density of the OR adjusted for misclassification error.

5. Use the method of Chen and Shao (22) to calculate an HPD interval from each posterior sample from step 4, and hence calculate the length or posterior probability of each sample, as required by the chosen criterion. This method assumes unimodality of the posterior density, which in our experience is satisfied. If not, symmetric intervals can be substituted.

6. To implement the ACC criterion, compare the average coverage of HPD intervals of length l to the predetermined value of the coverage $1 - \alpha$. If the average coverage is greater (smaller) than the desired $1 - \alpha$, return to step 1 using a smaller (greater) sample size N. Continue until the criterion is met, using, for example, a bisectional search or model based strategy to select the next N. A model based strategy can use the pairs of N and average coverage values to create a fitted curve to predict the most likely value of N required, refining the model after each step. In practice, we have found a model of the form

$$\text{avg cov} = \alpha + \beta\Phi\left(\frac{\log(N) - \mu}{\sigma}\right)$$

fits the data well, where α and β are regression parameters to be estimated, μ and σ are measures of central tendency and spread of the logarithms of the sample sizes selected, respectively, and $\Phi(\cdot)$ is the cumulative normal density. To implement the ALC, compare the average length of the HPD intervals with fixed coverage $1-\alpha$, using a similar search strategy as for the ACC until the desired average length is attained.

Finally, to implement the MWOC, for each N we must compare the proportion of samples which satisfy both the desired length and coverage, stopping when the proportion matches the desired $1 - \gamma$.

The ratio of controls to cases is another important design choice. Sample sizes can be calculated across a range of values for g, selecting the value that leads to the smallest sample size that is feasible in a given study, considering the availability of cases and controls.

	Disease +	Disease −	Total
Exposure +	a	c	$a + c$
Exposure −	b	d	$b + d$
Total	$a + b$	$c + d$	N

Table 1: Two-by-two table of observed data.

Criterion	OR	Interval Length	No Misclassification	Low Misclassification	Moderate Misclassification Narrow Prior	Moderate Misclassification Wide Prior	High Misclassification Narrow Prior	High Misclassification Wide Prior
Identical Misclassification Parameters - 4 Parameter Model								
ACC	0.7	0.4	1,341	1,665	3,422	3,698	12,326	14,223
ALC	0.7	0.4	1,016	1,254	2,588	3,053	9,567	12,519
MWOC(50)	0.7	0.4	813	1,004	2,055	2,344	7,495	9,210
MWOC(90)	0.7	0.4	2,696	3,288	6,688	7,288	25,018	28,721
ACC	1	0.2	15,992	40,760	$> 100,000$	$> 100,000$	> 100000	$> 100,000$
ALC	1	0.2	11,427	20,690	53,802	$> 100,000$	$> 100,000$	$> 100,000$
MWOC(50)	1	0.2	8,595	10,745	23,454	57,437	92,838	$> 100,000$
MWOC(90)	1	0.2	32,072	$> 100,000$	$> 100,000$	$> 100,000$	> 100000	$> 100,000$
ACC	1.5	0.8	3,514	8,936	41,237	$> 100,000$	$> 100,000$	$> 100,000$
ALC	1.5	0.8	2,176	3,920	10,862	$> 100,000$	72,435	$> 100,000$
MWOC(50)	1.5	0.8	1,434	1,962	4,751	9,273	20,604	47,454
MWOC(90)	1.5	0.8	7,353	$> 100,000$	$> 100,000$	$> 100,000$	$> 100,000$	$> 100,000$
Distinct Misclassification Parameters - 6 Parameter Model								
ACC	0.7	0.4	1,341	2,391	6,226	$> 100,000$	$> 100,000$	$> 100,000$
ALC	0.7	0.4	1,016	1,653	3,957	$> 100,000$	$> 100,000$	$> 100,000$
MWOC(50)	0.7	0.4	813	1,234	2,797	$> 100,000$	24,810	$> 100,000$
MWOC(90)	0.7	0.4	2,696	9,793	$> 100,000$	$> 100,000$	$> 100,000$	$> 100,000$
ACC	1	0.2	15,992	$> 100,000$	$> 100,000$	$> 100,000$	$> 100,000$	$> 100,000$
ALC	1	0.2	11,427	$> 100,000$	$> 100,000$	$> 100,000$	$> 100,000$	$> 100,000$
WMOC(50)	1	0.2	8,595	$> 100,000$	$> 100,000$	$> 100,000$	$> 100,000$	$> 100,000$
MWOC(90)	1	0.2	32,072	$> 100,000$	$> 100,000$	$> 100,000$	$> 100,000$	$> 100,000$
ACC	1.5	0.8	3,514	$> 100,000$	$> 100,000$	$> 100,000$	$> 100,000$	$> 100,000$
ALC	1.5	0.8	2,176	$> 100,000$	$> 100,000$	$> 100,000$	$> 100,000$	$> 100,000$
MWOC(50)	1.5	0.8	1,434	4,095	17,569	$> 100,000$	$> 100,000$	$> 100,000$
MWOC(90)	1.5	0.8	7,353	$> 100,000$	$> 100,000$	$> 100,000$	$> 100,000$	$> 100,000$

Table 2: Sample sizes per group for the prototypic scenarios described in Section 3, assuming control to case ratio $g = 1$.

ACC = Average Coverage Criterion, ALC = Average Length Criterion, g = case to control ratio, MWOC = Modified Worst Outcome Criterion, OR = Odds Ratio.

Criterion and case to control ratio	No Misclassification	Low Misclassification	Moderate Misclassification
ACC			
1	2,908	3,242	4,918
2	3,078	3,441	5,160
3	3,468	3,876	5,848
ALC			
1	2,758	3,056	4,600
2	2,925	3,249	4,914
3	3,280	3,664	5,516
MWOC(50)			
1	2,654	2,958	4,440
2	2,843	3,141	4,734
3	3,196	3,564	5,336
MWOC(95)			
1	5,014	5,784	8,912
2	5,427	6,159	9,468
3	6,200	7,048	10,868

Table 3: Sample sizes for both groups combined for the early introduction of peanut example of Section 4.

ACC = Average Coverage Criterion, ALC = Average Length Criterion, MWOC = Modified Worst Outcome Criterion.