The International Journal of Biostatistics

Volume 8, *Issue* 1 2012 *Article* 31

Partial Identification arising from Nondifferential Exposure Misclassification: How Informative are Data on the Unlikely, Maybe, and Likely Exposed?

Dongxu Wang, *University of British Columbia*
Tian Shen, *QLT Inc.*
Paul Gustafson, *University of British Columbia*

Recommended Citation:
Wang, Dongxu; Shen, Tian; and Gustafson, Paul (2012) "Partial Identification arising from Nondifferential Exposure Misclassification: How Informative are Data on the Unlikely, Maybe, and Likely Exposed?," *The International Journal of Biostatistics*: Vol. 8: Iss. 1, Article 31.
DOI: 10.1515/1557-4679.1397

Partial Identification arising from Nondifferential Exposure Misclassification: How Informative are Data on the Unlikely, Maybe, and Likely Exposed?

Dongxu Wang, Tian Shen, and Paul Gustafson

Abstract

There is quite an extensive literature on the deleterious impact of exposure misclassification when inferring exposure-disease associations, and on statistical methods to mitigate this impact. Virtually all of this work, however, presumes a common number of states for the true exposure status and the classified exposure status. In the simplest situation, for instance, both the true status and the classified status are binary. The present work diverges from the norm, in considering classification into three states when the actual exposure status is simply binary. Intuitively, the classification states might be labeled as `unlikely exposed,' `maybe exposed,' and `likely exposed.' While this situation has been discussed informally in the epidemiological literature, we provide some theory concerning what can be learned about the exposure-disease relationship, under various assumptions about the classification scheme. We focus on the challenging situation whereby no validation data is available from which to infer classification probabilities, but some prior assertions about these probabilities might be justified.

KEYWORDS: Bayesian methods, case-control analysis, exposure misclassification, partial identification.

Author Notes: Work supported by a grant from the Natural Sciences and Engineering Research Council of Canada

1 Introduction

It is well known that unacknowledged exposure misclassification can bias estimates of exposure-disease association. Consider, for example, a binary exposure subject to misclassification, with the extent of misclassification described by sensitivity and specificity, i.e., the probability of correct classification for exposed and unexposed subjects respectively. For a specific study design, and under assumptions about how the misclassification mechanism depends on outcome and confounding variables, the impact of misclassification can be assessed. For instance, consider unmatched case-control sampling, where the goal is to infer exposure prevalences in case and control populations, and thereby infer the exposure-disease odds-ratio. A readily understood situation arises if the misclassification is *nondifferential*, i.e., the classification probabilities are unaffected by disease status. The bias in estimating the target parameter by wrongly assuming there is no misclassification can then be expressed as a function of sensitivity and specificity (see, for instance, Gustafson 2004, Ch. 3). In sufficiently simple settings, nondifferential misclassification which is not accounted for will bias estimation toward the null of no exposure-disease association.

Some understanding of how the bias induced by unacknowledged misclassification depends on various aspects of the problem at hand can spawn *informal* strategies for mitigating this bias. This is taken up in Dosemeci and Stewart (1996), who interpret their results as follows:

> These findings suggest that if, in the exposure assessment process of a case-control study, where the exposure prevalence is low, an occupational hygienist is not sure about the exposure status of a subject, it is judicious to classify that subject as unexposed.

This recommendation arises since, in the presence of low exposure prevalences, the magnitude of the bias increases much faster as the specificity drops from one than it does as the sensitivity drops from one. Thus keeping the exposure group 'pure,' by limiting the misclassification of truly unexposed subjects into it, becomes paramount.

The form of such a recommendation suggests thinking of the exposure classification, at least initially, as being made into one of *three* categories. For sake of definiteness, we label these categories as *unlikely exposed*, *maybe exposed*, and *likely exposed*. Then, depending on the context, some mitigation of bias could be achieved by collapsing the observed exposure data from three categories down to two, e.g., merging the first two categories if expo-

sure prevalence is low. (In the face of high exposure prevalences, analogous considerations would suggest instead merging the last two categories.) After such a merge, data analysis can follow along the routine lines of inferring the odds-ratio from a 2×2 exposure-disease table of counts.

It is natural to ask whether a more *formal* statistical scheme might better mitigate bias and/or better reflect *a posteriori* uncertainty about the target parameter. Particularly, we investigate directly modelling the exposure classification into the unlikely, maybe, and likely categories. Thus the sensitivity and specificity of the classification scheme are supplanted by probability distributions across the three categories, for the truly exposed and the truly unexposed respectively. While non-differential misclassification with more than two categories has been considered in the literature (see, for instance, Dosemeci, Wacholder, and Lubin 1990, Birkett 1992, Weinberg, Umbach, and Greenland 1994, Correa-Villaseor, Stewart, Franco-Marina, and Seacat 1995), this is typically considered when the same set of labels for more than two ordered states is used for both the true and observed exposure status (e.g. none, low, high). 'Non-square' situations, such as two states for the true status and three states for the observed status, do not seem to have garnered attention.

In our framework, we quantify the information about exposure prevalences, and hence the odds-ratio, in a large-sample sense. In situations where classification probabilities are known, or can be consistently estimated from validation data, then inferential options for consistent estimation of the odds-ratio are available; see, for instance, Gustafson (2004, Ch. 5) or Buonaccorsi (2010, Ch. 3). This paper focusses on the more challenging setting where classification probabilities cannot be estimated consistently. Given this, we cannot expect to consistently estimate the exposure-disease odds-ratio as the case and control sample sizes increase. We may, however, be able to rule out some values for the odds-ratio.

First, in Section 2, we focus on determining *identification regions* from *prior regions*. Particularly, given assumptions about the possible values of classification probabilities, we show what values of exposure prevalances are compatible with the distribution of the observable data. This falls within the rubric of partially identified models (e.g., Manski 2003), whereby even the observation of an infinite amount of data does not reveal the true values of the target parameters, but does rule out some values. We consider two prior regions based on different assumptions about *a priori* plausible values of the misclassification probabilities. The first is a weak assumption that the exposure classification scheme is 'better than random,' in a particular sense.

The second is a stronger assumption of monotonicity, in the sense that for any two categories, and either level of true exposure, the worse classification is less likely. We also compare the resulting identification regions to those which arise from collapsing the three categories down to two categories and then acknowledging misclassification. Note that identification regions for partially identified models are relevant whether non-Bayesian or Bayesian inference is to be pursued.

Having established the form of the identification regions, we turn to Bayesian inference. First, in Section 3 , we determine the behaviour of the posterior distributions over the control and case exposure prevalences, as the control and case sample sizes go to infinity. This is pursued via the general approach to determining the limiting posterior distribution in partially identified models outlined in Gustafson (2005) (see also Gustafson 2010). Necessarily, the support of the limiting posterior distribution is the corresponding identification region. Then, in Section 4, we give some examples of posterior distributions arising from finite samples approaching their limits as the sample size grows. We also compare these with posterior distributions arising from the informal approach of collapsing to two categories and then ignoring misclassification.

2 Identification Regions

Let Y, X, and X^* represent an individual's disease status, actual exposure status, and apparent exposure status respectively. The disease status Y and the actual exposure status X each have two categories, coded as zero for 'absence' and one for 'presence.' The apparent exposure status X^* has three categories, coded as zero for unlikely exposed, one for maybe exposed, and two for likely exposed. Thus observed data take the form of a 2×3 (Y, X^*) data table. For clarity of exposition, we assume case-control sampling throughout, whereby the data table is obtained from sampling the population distribution of $(X^*|Y)$, and the task is to infer the (X, Y) odds-ratio. However, our findings are also directly applicable to cross-sectional or prospective sampling of (X^*, Y) or $(Y|X^*)$ respectively.

Define r_0 and r_1 to be the prevalences of exposure among controls and cases:

$$r_0 = Pr\{X = 1 \mid Y = 0\},$$
$$r_1 = Pr\{X = 1 \mid Y = 1\}.$$

The non-differential misclassification assumption is invoked, under which X^* and Y are conditionally independent given X. Thus the misclassification is described via p_{ij} denoting the probability of classifying a subject truly having the i-th exposure level into the j-th category:

$$\boldsymbol{p_0} = \begin{pmatrix} p_{00} \\ p_{01} \\ p_{02} \end{pmatrix} = \begin{pmatrix} Pr\{X^* = 0 \mid X = 0\} \\ Pr\{X^* = 1 \mid X = 0\} \\ Pr\{X^* = 2 \mid X = 0\} \end{pmatrix},$$

$$\boldsymbol{p_1} = \begin{pmatrix} p_{10} \\ p_{11} \\ p_{12} \end{pmatrix} = \begin{pmatrix} Pr\{X^* = 0 \mid X = 1\} \\ Pr\{X^* = 1 \mid X = 1\} \\ Pr\{X^* = 2 \mid X = 1\} \end{pmatrix}.$$

Then, the prevalences of apparent exposure among controls and cases, say $\boldsymbol{\theta_0}$ and $\boldsymbol{\theta_1}$, can be expressed as combinations of r_0, r_1, $\boldsymbol{p_0}$, and $\boldsymbol{p_1}$:

$$\boldsymbol{\theta_0} = \begin{pmatrix} \theta_{00} \\ \theta_{01} \\ \theta_{02} \end{pmatrix} = \begin{pmatrix} r_0 p_{10} + (1 - r_0) p_{00} \\ r_0 p_{11} + (1 - r_0) p_{01} \\ r_0 p_{12} + (1 - r_0) p_{02} \end{pmatrix} = \begin{pmatrix} Pr\{X^* = 0 \mid Y = 0\} \\ Pr\{X^* = 1 \mid Y = 0\} \\ Pr\{X^* = 2 \mid Y = 0\} \end{pmatrix},$$

$$\boldsymbol{\theta_1} = \begin{pmatrix} \theta_{10} \\ \theta_{11} \\ \theta_{12} \end{pmatrix} = \begin{pmatrix} r_1 p_{10} + (1 - r_1) p_{00} \\ r_1 p_{11} + (1 - r_1) p_{01} \\ r_1 p_{12} + (1 - r_1) p_{02} \end{pmatrix} = \begin{pmatrix} Pr\{X^* = 0 \mid Y = 1\} \\ Pr\{X^* = 1 \mid Y = 1\} \\ Pr\{X^* = 2 \mid Y = 1\} \end{pmatrix}.$$

Clearly the distribution of the observed data depends on the unknown parameters only via $\boldsymbol{\theta} = (\boldsymbol{\theta_0}, \boldsymbol{\theta_1})$, and only functions of $\boldsymbol{\theta}$ are consistently estimable.

Going forward, it is useful to note that $\boldsymbol{p_i}$ and $\boldsymbol{\theta_i}$ belong to the probability simplex on three categories, which we denote as $\mathbb{S}_3$. When useful, we visualize points in $\mathbb{S}_3$ by plotting the probability assigned to the first (third) category on the vertical (horizontal) axis, so that $\mathbb{S}_3$ is represented as the lower-left triangle in the unit square $(0,1)^2$.

We study the situation where the only direct information about the classification probabilities $\boldsymbol{p} = (\boldsymbol{p_0}, \boldsymbol{p_1})$ is *a priori* knowledge that they lie in a particular subset of $\mathbb{S}_3^2$. We write this as $\boldsymbol{p} \in \mathbb{P}$, and refer to $\mathbb{P}$ as the *prior region*. In concept we could also consider a prior region for the exposure prevalences $\boldsymbol{r} = (r_0, r_1)$ which would be a subset of $(0,1)^2$, but in fact in this paper we only consider the situation where this prior region is all of $(0,1)^2$.

Following the general approach to partial identification, as described for instance by Manski (2003), we start with the prior region and the values of

$\boldsymbol{\theta}$ (thinking of the latter as equivalent to observation of an infinite number of controls and cases). Then we would like to know all the values of the unknown parameters (particularly the target parameters $\boldsymbol{r}$) which could have produced this value of $\boldsymbol{\theta}$. Formally, let the *identification region* $\mathbb{Q}(\boldsymbol{\theta})$ be all values of the target parameters (r_0, r_1) which yield this value of $\boldsymbol{\theta}$ for some choice of $\boldsymbol{p} \in \mathbb{P}$. Since we learn the values of $\boldsymbol{\theta}$ as the sample size increases, the identification region can be regarded as all values of (r_0, r_1) which are still plausible after having observed an infinite amount of data, presuming that the classification probabilities are indeed inside the prior region. Note that the identification region $\boldsymbol{r} \in \mathbb{Q}(\boldsymbol{\theta})$ will in turn generate an identification region (typically an interval) for the exposure-disease odds-ratio $OR = \{r_1/(1-r_1)\}/\{r_0/(1-r_0)\}$.

2.1 Constraint A

To motivate a realistic prior region, note that merging the maybe and unlikely categories together would result in a binary classification scheme having sensitivity p_{12} and specificity $1-p_{02}$, so that an assumption of 'better than random' classification is expressed as $p_{12} > p_{02}$. Similarly, if the maybe and likely categories were merged, the assumption $p_{00} > p_{10}$ would hold sway. Therefore, if categories are not actually collapsed, it is natural to assume that both inequalities hold. We refer to this as constraint A, and express the prior region as $\boldsymbol{p} \in \mathbb{P}_A$. With respect to our visualization scheme, $\boldsymbol{p_0}$ and $\boldsymbol{p_1}$ can be anywhere in the lower-left triangle representing $\mathbb{S}_3$, so long as $\boldsymbol{p_1}$ is south-east of $\boldsymbol{p_0}$.

Consider a value of $\boldsymbol{\theta}$ which is *compatible* with constraint A, i.e., this $\boldsymbol{\theta}$ arises for some value of $\boldsymbol{p} \in \mathbb{P}_A$ along with some value of $\boldsymbol{r} \in (0,1)^2$. Geometrically, it is immediate that $\boldsymbol{\theta}$ is compatible with constraint A if and only if the line connecting $\boldsymbol{\theta_0}$ and $\boldsymbol{\theta_1}$ has negative slope. Below we refer to this line simply as 'the connecting line.' Without loss of generality, but for ease of exposition, we assume $\boldsymbol{\theta}_1$ lies south-east of $\boldsymbol{\theta}_0$ (as must arise if in fact $r_0 < r_1$, so that $OR > 1$). Then the identification region in terms of $\boldsymbol{p}$ is immediately visualized as follows: $\boldsymbol{p}_0$ can lie anywhere on the connecting line above $\boldsymbol{\theta_0}$ and $\boldsymbol{p}_1$ can lie anywhere on the connecting line below $\boldsymbol{\theta_1}$ (though of course each $\boldsymbol{p_i}$ must remain within $\mathbb{S}_3$).

This geometry lends itself to simple algebraic description of the identification region in terms of $\boldsymbol{z} = (z_0, z_1)$, where

$$\boldsymbol{p_0} = \boldsymbol{\theta_0} + z_0(\boldsymbol{\theta_0} - \boldsymbol{\theta_1}), \quad (1)$$

$$\boldsymbol{p_1} = \boldsymbol{\theta_1} + z_1(\boldsymbol{\theta_1} - \boldsymbol{\theta_0}), \quad (2)$$

i.e., z_i indicates how far $\boldsymbol{p_i}$ lies beyond $\boldsymbol{\theta_i}$. Thus each z_i is non-negative, but cannot exceed the value which maps $\boldsymbol{p_i}$ onto the boundary of $\mathbb{S}_3$. Consequently, the identification region for (z_0, z_1) is rectangular, given by $0 \leq z_i \leq \bar{z}_i(\boldsymbol{\theta})$, for $i = 0, 1$, where

$$\bar{z}_0(\boldsymbol{\theta}) = \begin{cases} min(\frac{\theta_{02}}{\theta_{12}-\theta_{02}}, \frac{\theta_{01}}{\theta_{11}-\theta_{01}}) & \text{if } \theta_{11} - \theta_{01} > 0, \\ \frac{\theta_{02}}{\theta_{12}-\theta_{02}} & \text{if } \theta_{11} - \theta_{01} \leq 0. \end{cases} \quad (3)$$

$$\bar{z}_1(\boldsymbol{\theta}) = \begin{cases} min(\frac{\theta_{10}}{\theta_{00}-\theta_{10}}, \frac{\theta_{11}}{\theta_{01}-\theta_{11}}) & \text{if } \theta_{01} - \theta_{11} > 0, \\ \frac{\theta_{10}}{\theta_{00}-\theta_{10}} & \text{if } \theta_{01} - \theta_{11} \leq 0. \end{cases} \quad (4)$$

From here, it is easy to verify that the rectangular identification region for (z_0, z_1) maps to a polygonal identification region for (r_0, r_1), via the map $(r_0, r_1) = (1 + z_0 + z_1)^{-1}(z_0, 1 + z_0)$. In particular,

$$\mathbb{Q}_A(\boldsymbol{\theta}) = \left\{(r_0, r_1) \in (0,1)^2 : r_1 > \frac{\bar{z}_0(\boldsymbol{\theta}) + 1}{\bar{z}_0(\boldsymbol{\theta})} r_0, r_1 > \frac{\bar{z}_1(\boldsymbol{\theta})}{\bar{z}_1(\boldsymbol{\theta}) + 1} r_0 + \frac{1}{\bar{z}_1(\boldsymbol{\theta}) + 1}\right\}.$$

The situation described thus far is illustrated in the left panels of Figure 1, for the arbitrarily chosen example values of $\boldsymbol{\theta_0} = (0.645, 0.200, 0.155)$ and $\boldsymbol{\theta_1} = (0.567, 0.200, 0.233)$. The top panel illustrates these $\boldsymbol{\theta_i}$ values and the identification region for $\boldsymbol{p}$, within $\mathbb{S}_3$. The middle panel shows this identification region expressed in terms of $\boldsymbol{z}$, and finally the bottom panel visualizes this region as $\boldsymbol{r} \in \mathbb{Q}_A(\boldsymbol{\theta})$.

2.2 Constraint B

Sometimes a stronger assumption than constraint A may be justified, making explicit reference to the chance of 'maybe' classification. The monotonicity of $\boldsymbol{p_0}$ and $\boldsymbol{p_1}$ might be assumed, whereby the worse a classification is, the less likely it is. This constraint, henceforth referred to as constraint B, can be expressed as $p_{00} > p_{01} > p_{02}$ and $p_{10} < p_{11} < p_{12}$, which we denote as $\boldsymbol{p} \in \mathbb{P}_B$. The visual representation of $\mathbb{P}_B$ is given in the upper-right panel of Figure 1, in which $\boldsymbol{p_0}$ must lie in the upper shaded triangle and $\boldsymbol{p_1}$ must lie in the lower shaded triangle.

Say that $\boldsymbol{\theta}$ is compatible with constraint A, and again assume, without loss of generality, that $\boldsymbol{\theta_1}$ is south-east of $\boldsymbol{\theta_0}$. Taking the geometric view, the identification region under constraint B will be non-empty if and only if the portion of the connecting line above $\boldsymbol{\theta_0}$ intersect the prior region for $\boldsymbol{p_0}$ *and* the portion below $\boldsymbol{\theta_1}$ intersects the prior region for $\boldsymbol{p_1}$. We will say that $\boldsymbol{\theta}$

is compatible with constraint B if the identification region is non-empty. If $\boldsymbol{\theta}$ arises from a true value of $\boldsymbol{p} \in \mathbb{P}_B$ then by definition $\boldsymbol{\theta}$ is compatible with constraint B. However, if $\boldsymbol{\theta}$ arises from a true value of $\boldsymbol{p} \in \mathbb{P}_A - \mathbb{P}_B$, then $\boldsymbol{\theta}$ may or may not be compatible with constraint B. Upon inspection of the upper-right panel of Figure 1, we see that compatibility with constraint B arises if and only if the connecting line intersects the vertical axis between 0.5 and 1, and also intersects the horizontal axis between 0.5 and 1.

For a $\boldsymbol{\theta}$ compatible with constraint B, we can again express the identification region in terms of $\boldsymbol{z}$. The upper bounds on z_0 and z_1 must correspond to the intersection of the connecting line with the vertical and horizontal axes respectively, and therefore be the same upper bounds (3) and (4) that apply under constraint A. Note as well though that if $\boldsymbol{\theta}$ is compatible with constraint B, then (3) and (4) specialize to $\bar{z}_0(\boldsymbol{\theta}) = \theta_{02}/(\theta_{12} - \theta_{02})$ and $\bar{z}_1(\boldsymbol{\theta}) = \theta_{10}/(\theta_{00} - \theta_{10})$. It is also clear from the geometric view (upper-right panel of Figure 1 again) that z_i will have a positive lower bound if and only if $\boldsymbol{\theta_i}$ lies outside the prior region for $\boldsymbol{p_i}$. Thus our identification region is now expressed as $\underline{z}_i(\boldsymbol{\theta}) \leq z_i \leq \bar{z}_i(\boldsymbol{\theta})$, for $i = 0, 1$, where

$$\begin{aligned} \underline{z}_0(\boldsymbol{\theta}) &= \max\left(0, \frac{\theta_{01} - \theta_{02}}{(\theta_{11} - \theta_{12}) - (\theta_{01} - \theta_{02})}\right), \\ \underline{z}_1(\boldsymbol{\theta}) &= \max\left(0, \frac{\theta_{10} - \theta_{11}}{(\theta_{00} - \theta_{01}) - (\theta_{10} - \theta_{11})}\right). \end{aligned}$$

Again this rectangular identification region for (z_0, z_1) induces a polygonal identification region $(r_0, r_1) \in \mathbb{Q}_B(\boldsymbol{\theta})$, as illustrated in the middle-right and lower-right panels of Figure 1. Formally,

$$\begin{aligned} \mathbb{Q}_B(\boldsymbol{\theta}) = \Bigg\{ & (r_0, r_1) \in (0,1)^2 : r_1 > \frac{\bar{z}_0(\boldsymbol{\theta}) + 1}{\bar{z}_0(\boldsymbol{\theta})} r_0, r_1 > \frac{\bar{z}_1(\boldsymbol{\theta})}{\bar{z}_1(\boldsymbol{\theta}) + 1} r_0 + \frac{1}{\bar{z}_1(\boldsymbol{\theta}) + 1}, \\ & r_1 < \frac{\underline{z}_0(\boldsymbol{\theta}) + 1}{\underline{z}_0(\boldsymbol{\theta})} r_0, r_1 < \frac{\underline{z}_1(\boldsymbol{\theta})}{\underline{z}_1(\boldsymbol{\theta}) + 1} r_0 + \frac{1}{\underline{z}_1(\boldsymbol{\theta}) + 1} \Bigg\}. \end{aligned}$$

Note that if $\boldsymbol{\theta}$ is compatible with constraint B, the identification regions for $\boldsymbol{z}$ under constraints A and B are both rectangular with the same north-east vertex. Consequently, in terms of $\boldsymbol{r}$, the lower boundary of $\mathbb{Q}_B(\boldsymbol{\theta})$ is guaranteed to be a subset of the lower boundary of $\mathbb{Q}_A(\boldsymbol{\theta})$.

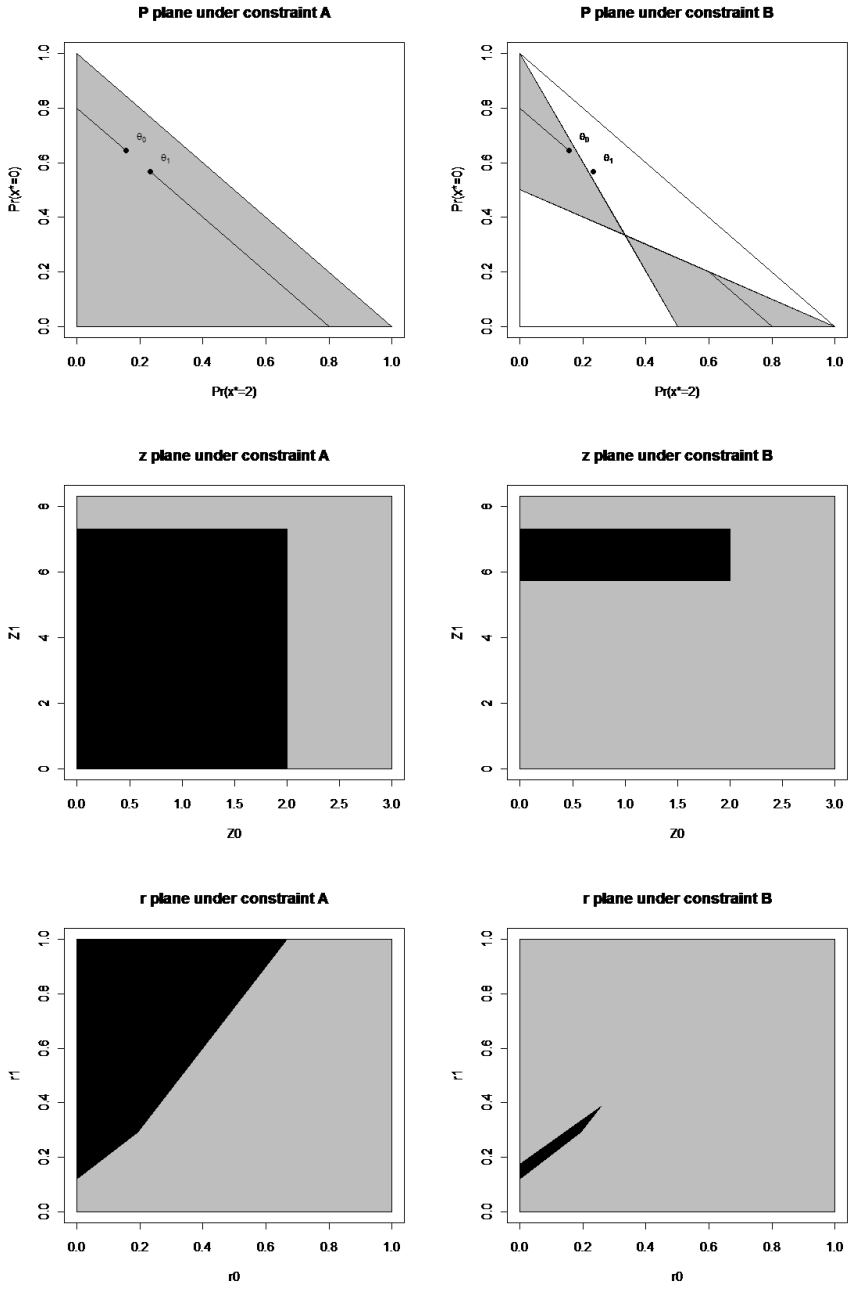

Figure 1: Prior and identification regions for example values $\boldsymbol{\theta_0} = (0.645, 0.200, 0.155)$ and $\boldsymbol{\theta_1} = (0.567, 0.200, 0.233)$. The left (right) plots correspond to constraint A (B). The top, middle and bottom plots correspond to the probability simplex, the $\boldsymbol{z}$ plane and the $\boldsymbol{r}$ plane respectively. In all cases prior regions are in grey and identification regions are in black.

2.3 Comparison Between Constraint A and B

We can summarize some salient features of the identification regions under constraints A and B in the following theorem.

Theorem 2.1 *Say that $\boldsymbol{\theta}$ is compatible with constraint A. Also, assume without loss of generality that $\theta_{00} > \theta_{10}$ and $\theta_{02} < \theta_{12}$ (as must arise if $r_0 < r_1$, so that $OR > 1$, and consequently a lower bound on OR will be a bound away from the null). Then:*

(i) If $\boldsymbol{\theta} \in \mathbb{P}_B$, then $\mathbb{Q}_A(\boldsymbol{\theta}) = \mathbb{Q}_B(\boldsymbol{\theta})$. That is, both constraints give rise to the same identification region if $\boldsymbol{\theta_0}$ and $\boldsymbol{\theta_1}$ fall in the prior regions for $\boldsymbol{p_0}$ and $\boldsymbol{p_1}$ under constraint B. Otherwise, $\mathbb{Q}_B(\boldsymbol{\theta}) \subset \mathbb{Q}_A(\boldsymbol{\theta})$.

(ii) Constraint A yields an infinite upper bound on the odds-ratio.

(iii) If $\boldsymbol{\theta}$ is compatible with constraint B, then constraint B yields a finite upper bound on the odds-ratio if and only if $\boldsymbol{\theta_0}$ lies outside the prior region for $\boldsymbol{p_0}$ and *$\boldsymbol{\theta_1}$ lies outside the prior region for $\boldsymbol{p_1}$.*

(iv) Constraint A yields a lower bound on the odds-ratio achieved by $r_0 = \bar{z}_0(\boldsymbol{\theta})/(\bar{z}_0(\boldsymbol{\theta}) + \bar{z}_1(\boldsymbol{\theta}) + 1)$ and $r_1 = (\bar{z}_0(\boldsymbol{\theta}) + 1)/(\bar{z}_0(\boldsymbol{\theta}) + \bar{z}_1(\boldsymbol{\theta}) + 1)$. If $\boldsymbol{\theta}$ is compatible with constraint B, then the same lower bound applies under constraint B.

Proof:
(i) : If $\boldsymbol{\theta} \in \mathbb{P}_B$ then $\underline{z}_0(\boldsymbol{\theta}) = \underline{z}_1(\boldsymbol{\theta}) = 0$, and the result follows immediately.

$(ii)\&(iii)$: The odds-ratio tends to infinity as r_0 goes to zero or r_1 goes to one. This corresponds to $\boldsymbol{p_0}$ going to $\boldsymbol{\theta_0}$ (from above/left) or $\boldsymbol{p_1}$ going to $\boldsymbol{\theta_1}$ (from below/right), along the connecting line. Geometrically, the prior region under constraint A can never preclude either possibility, simply because $(\boldsymbol{\theta_0}, \boldsymbol{\theta_1}) \in \mathbb{P}_A$. Both possibilities are precluded under constraint B, however, if and only if the line segment between $\boldsymbol{\theta_0}$ and $\boldsymbol{\theta_1}$ does not intersect the union of the two components of $\mathbb{P}_B$ (i.e., the line segment does not intersect the shaded region in the upper-right panel of Figure 1).

(iv) : Clearly the maximum value of r_0 and the minimum value of r_1 correspond to the two intersections of the connecting line with the $\mathbb{S}_3$ boundary. This can also be visualized as the north-east vertex of the identification rectangle for $\boldsymbol{z}$, and as the middle of the three vertices which give the lower boundary for $\mathbb{Q}_A(\boldsymbol{\theta})$ or $\mathbb{Q}_B(\boldsymbol{\theta})$. ∎

2.4 Collapsing Exposure to Two Categories

As alluded to in the Introduction, it is informative to compare the identification regions described above to the identification region arising when exposure is collapsed from three to two categories. Particularly, if low exposure prevalences are anticipated, the 'unlikely' and 'maybe' categories could be merged. Then the binary apparent exposure status would be

$$X^{**} = \begin{cases} 0 & \text{if } X^* \in \{0,1\}, \\ 1 & \text{if } X^* = 2, \end{cases}$$

with the quality of classification described by specificity $1 - p_0^*$ and sensitivity p_1^*, where $p_0^* = p_{02}$ and $p_1^* = p_{12}$. A weak and commonly invoked assumption is that $p_0^* < p_1^*$, stating that the classification scheme is better than simply choosing an exposure status completely at random. Thus we take the prior region $\mathbb{P}^*$ to be the triangular region on $(0,1)^2$ for which this inequality holds. The information gleaned from an infinite data sample would be the value of $\boldsymbol{\theta^*}$, where

$$\theta_0^* = r_0 p_1^* + (1 - r_0) p_0^* = Pr\left\{X^{**} = 1 \mid Y = 0\right\},$$
$$\theta_1^* = r_1 p_1^* + (1 - r_1) p_0^* = Pr\left\{X^{**} = 1 \mid Y = 1\right\}.$$

The identification region for this problem is determined by Gustafson, Le, and Saskin (2001). However, we express their results in a form more amenable for comparison with the results in Sections 2.1 and 2.2. Assume without loss of generality that $\theta_0^* < \theta_1^*$ (as must arise if $r_0 < r_1$). As per (1) and (2), we can define $\boldsymbol{z^*} = (z_0^*, z_1^*)$, where $p_0^* = \theta_0^* + z_0(\theta_0^* - \theta_1^*)$ and $p_1^* = \theta_1^* + z_1(\theta_1^* - \theta_0^*)$. Then by the same geometric argument as earlier, we have a rectangular identification region of the form $0 < z_i^* < \bar{z}_i^*(\boldsymbol{\theta^*})$ for $i = 0, 1$, with

$$\begin{aligned} \bar{z}_0^*(\boldsymbol{\theta^*}) &= \frac{\theta_0^*}{\theta_1^* - \theta_0^*}, \\ \bar{z}_1^*(\boldsymbol{\theta^*}) &= \frac{1 - \theta_1^*}{\theta_1^* - \theta_0^*}, \end{aligned}$$

where $\theta_0^* = \theta_{02}$ and $\theta_1^* = \theta_{12}$. Moreover, the identification region maps to $\boldsymbol{r}$ just as before, according to $(r_0, r_1) = (1 + z_0^* + z_1^*)^{-1}(z_0^*, 1 + z_0^*)$. Thus we again have a polygonal boundary for the identification region $\boldsymbol{r} \in \mathbb{Q}^*(\boldsymbol{\theta^*})$.

It is very easy to compare the effect of collapsing to the use of three categories and constraint A.

Theorem 2.2 *Say that $\boldsymbol{\theta}$ is compatible with constraint A. Also, assume without loss of generality that $\theta_{00} > \theta_{10}$ and $\theta_{02} < \theta_{12}$ (as must arise if $r_0 < r_1$). Then:*

$\mathbb{Q}_A(\boldsymbol{\theta}) \subset \mathbb{Q}^*(\boldsymbol{\theta^*})$. *Consequently, $\mathbb{Q}^*(\boldsymbol{\theta^*})$ cannot produce a finite upper bound on the odds-ratio, while the lower bound cannot exceed that corresponding to $\mathbb{Q}_A(\boldsymbol{\theta})$.*

Proof: When $\boldsymbol{\theta}$ is compatible with constraint A, $\bar{z}_0^*(\boldsymbol{\theta^*}) \geq \bar{z}_0(\boldsymbol{\theta})$ and , $\bar{z}_1^*(\boldsymbol{\theta^*}) > \bar{z}_1(\boldsymbol{\theta})$. Also, the lower bound of $z_i(\boldsymbol{\theta})$ for both collapsed case and constraint A are zero ($\underline{z}_i(\boldsymbol{\theta^*}) = \underline{z}_i(\boldsymbol{\theta^*}) = 0$, for $i = 0, 1$). By mapping the identification region for (z_0, z_1) to (r_0, r_1), we can directly get the conclusion that $\mathbb{Q}_A(\boldsymbol{\theta}) \subset \mathbb{Q}^*(\boldsymbol{\theta^*})$. Since the upper bound on the odds-ratio under constraint A is infinite, the collapsed case will also yields an infinite upper bound on the odds-ratio. Also, the lower bound on the odds-ration for the collapsed case will always smaller or equal to the lower bound under constraint A. ∎

2.5 Examples

To examine identification regions under some realistic scenarios, we use two settings of exposure prevalence among controls (r_0), two settings of the odds-ratio (OR), and two settings of the classification probabilities ($\boldsymbol{p_0}, \boldsymbol{p_1}$). The value of r_1 is determined from r_0 and OR. We use '-' and '+' to label the first and second settings for each of the three factors. For instance, then, $(-++)$ would label the scenario with the first setting for exposure prevalence but the second setting for both the odds-ratio and the classification probabilities.

For the exposure prevalence among controls, the settings are $r_0^- = 0.05$ and $r_0^+ = 0.15$. For the exposure-disease association, the settings are $OR^- = 1.2$ and $OR^+ = 2.0$. Note that we do not consider the null situation ($OR = 1$), for the behaviour is quite different in this situation. Clearly $OR = 1$ if and only if $\boldsymbol{\theta_0} = \boldsymbol{\theta_1}$, so that if the null is true, the identification interval for the odds-ratio reduces to a single point. For the classification probabilities, the first setting is 'symmetric,' in the sense that exposed and unexposed subjects are classified equally well. Particularly, $\boldsymbol{p_0^-} = (0.75, 0.20, 0.05)$ and $\boldsymbol{p_1^-} = (0.05, 0.20, 0.75)$. The second setting corresponds to highly specific but not sensitive classification. That is, exposure is hard to detect, with $\boldsymbol{p_0^+} = (0.900, 0.075, 0.025)$ and $\boldsymbol{p_1^+} = (0.2, 0.3, 0.5)$. There are $2^3 = 8$ values of $\boldsymbol{\theta}$ arising from all combinations of these factors. The identification regions for the four $\boldsymbol{p} = \boldsymbol{p^-}$ scenarios are displayed in Figures 2 through 5. The identifi-

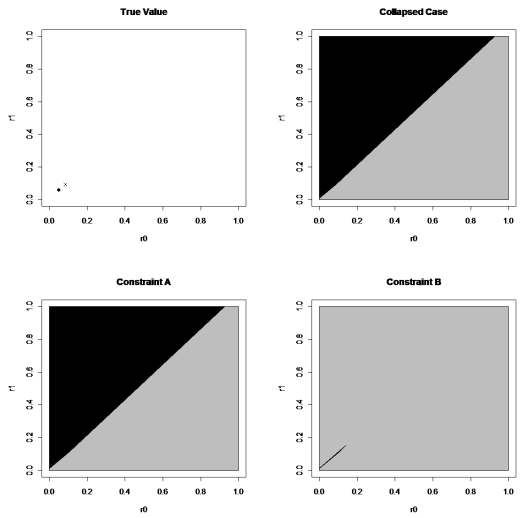

Figure 2: Identification regions for the combination $(---)$. In the upper-left panel, the dot indicates the true exposure prevalences (r_0, r_1), while the cross indicates apparent exposure prevalences upon collapsing to two categories and ignoring misclassification, (θ_0^*, θ_1^*). In the other three panels, prior and identification regions are indicated in grey and black respectively. The upper-right panel is the $\boldsymbol{r}$ plane in collapsed case, the lower-left panel is under constraint A, and the lower-right panel is under constraint B. In the collapsed case, θ_0^*=0.0850 and θ_1^*=0.0916. Under constraint A and B, $\boldsymbol{\theta_0}$=(0.7150, 0.2000, 0.0850) and $\boldsymbol{\theta_1}$=(0.7084, 0.2000, 0.0916).

cation regions for the $\boldsymbol{p} = \boldsymbol{p}^+$ scenarios are available as supplementary figures (`www.stat.ubc.ca/~gustaf`).

From the figures, we see that, in all cases, collapsing and using constraint A yield very similar identification regions for $\boldsymbol{r}$. The identification region using constraint B is typically very much smaller, though of course constraints A and B are guaranteed to yield the same lower bound for the odds-ratio. Moreover, while some values of $\boldsymbol{\theta}$ produce a finite upper bound on the odds-ratio under constraint B, this does not happen for any of the eight scenarios considered here. To the extent that our scenarios are typical, this suggests that a finite upper bound is uncommon. In fact, we can see that low exposure prevalences will tend to produce values of $\boldsymbol{\theta_0}$ close to $\boldsymbol{p_0}$, and therefore inside the prior region under constraint B, unless $\boldsymbol{p_0}$ happens to be very close to the boundary of the prior region. Thus we can intuit that a finite upper bound for the odds-ratio will not commonly arise. More specifically, for given classification

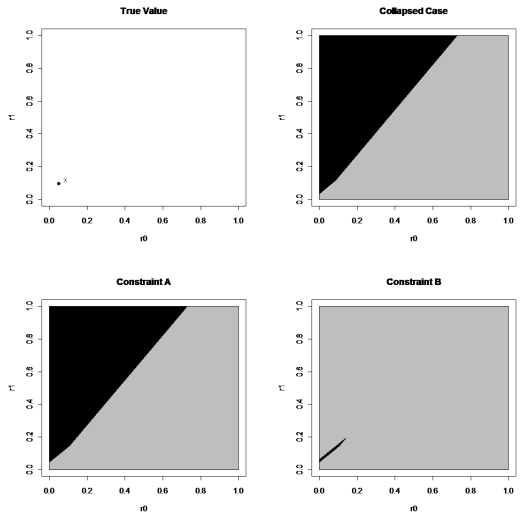

Figure 3: Identification regions for the combination $(- + -)$. The layout is the same as Figure 2. In the collapsed case, θ_0^*=0.0850 and θ_1^*=0.1167. Under constraint A and B, $\boldsymbol{\theta_0}$=(0.7150, 0.2000, 0.0850) and $\boldsymbol{\theta_1}$=(0.6833, 0.2000, 0.1167).

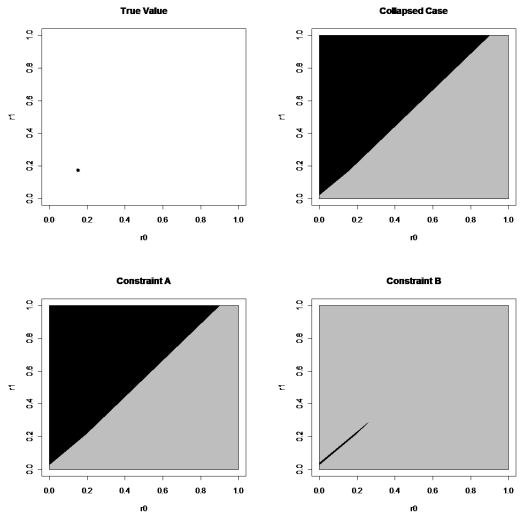

Figure 4: Identification regions for the combination $(+ - -)$. The layout is the same as Figure 2. In the collapsed case, θ_0^*=0.1550 and θ_1^*=0.1723. Under constraint A and B, $\boldsymbol{\theta_0}$=(0.6450, 0.2000, 0.1550) and $\boldsymbol{\theta_1}$=(0.6277, 0.2000, 0.1723).

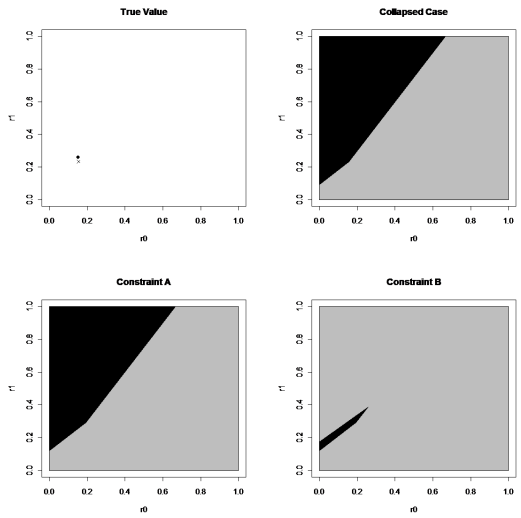

Figure 5: Identification regions for the combination $(+ + -)$. The layout is the same as Figure 2. In the collapsed case, θ_0^*=0.1550 and θ_1^*=0.2326. Under constraint A and B, $\boldsymbol{\theta_0}$=(0.6450, 0.2000, 0.1550) and $\boldsymbol{\theta_1}$=(0.5674, 0.2000, 0.2326).

probabilities $\boldsymbol{p}$ it is a simple matter to characterize how large r_0 must be (and how small r_1 must be) to produce $\boldsymbol{\theta}$ for which there is a finite upper bound on OR. This will happen if

$$\frac{p_{01} - p_{02}}{(p_{10} + 2p_{12}) - (p_{00} + 2p_{02})} < \quad r_i \quad < \frac{2p_{00} + p_{02} - 1}{(2p_{00} + p_{02}) - (2p_{10} + p_{12})},$$

for $i = 0, 1$. For instance, with $\boldsymbol{p} = \boldsymbol{p}^-$, the upper bound is finite if $r_i \in (0.214, 0.786)$, and with $\boldsymbol{p} = \boldsymbol{p}^+$, this bound is finite if $r_i \in (0.200, 0.892)$.

The lower bounds of the odds-ratio in our eight scenarios are given in Table 1. As guaranteed by theory, the lower bound is the same for constraints A and B, but lower for the collapsed case. In most scenarios, the collapsed case bound is only very slightly lower. In a practical sense the bounds are useful. For instance, in the $(+ - -)$ and $(+ - +)$ scenarios one can rule out an odds-ratio below 1.14 when the true value is 1.2, and in the $(+ + -)$ and $(+ + +)$ scenarios one can rule out an odds-ratio below 1.7 when the true value is 2. It is also worth noting that the lower bound in the collapsed case corresponds to the large-sample limit of the raw odds-ratio in the collapsed data table. Thus the extent to which constraints A and B produce a higher lower bound than this reflects the utility of a formal adjustment approach over

collapsing the unlikely and maybe categories together and treating this as the unexposed category, without adjustment for misclassification.

Table 1: The lower bound of odds-ratio for collapsed case and for constraint A and B in each scenario.

Scenarios	Lower bound of odds-ratio for collapsed case	Lower bound of odds-ratio for constraint A and B
$(---)$	1.085	1.087
$(--+)$	1.097	1.100
$(-+-)$	1.422	1.436
$(-++)$	1.474	1.496
$(+--)$	1.135	1.143
$(+-+)$	1.137	1.147
$(++-)$	1.652	1.706
$(+++)$	1.643	1.715

3 Limiting Posterior Distribution

In a partially identified context such as that faced here, determining the identification region is only part of the inferential story. From a Bayesian perspective, as the sample size goes to infinity, the investigator learns more than just the identification region. The posterior distribution of the target parameter will tend to a limiting distribution over the identification region, so an obvious issue to address is the extent to which the limiting posterior distribution is flat or peaked across the identification region.

3.1 Principle

Suppose r_0, r_1, $\boldsymbol{p_0}$, and $\boldsymbol{p_1}$ are independent of each other *a priori*. We assume that $r_0 \sim U(0,1)$, $r_1 \sim U(0,1)$, $\boldsymbol{p_0} \sim \text{Dirichlet}(c_{00}, c_{01}, c_{02})$, and $\boldsymbol{p_1} \sim \text{Dirichlet}(c_{10}, c_{11}, c_{12})$, with the additional truncation of $(\boldsymbol{p_0}, \boldsymbol{p_1})$ to the

assumed prior region $\mathbb{P}$. Under these assumptions, the joint prior density can be written as:

$$f(r_0, r_1, \boldsymbol{p_0}, \boldsymbol{p_1}) \;\; \propto \;\; \left(\prod_{i=0}^{1} \prod_{j=0}^{2} p_{ij}^{c_{ij}-1} \right) I_{(0,1)}(r_0) I_{(0,1)}(r_1) I_{\mathbb{P}}(\boldsymbol{p}_0, \boldsymbol{p}_1).$$

Since the value of $\boldsymbol{\theta}$ is estimable from data, and r_0 and r_1 are target parameters, a reparameterization from $(r_0, r_1, \boldsymbol{p_0}, \boldsymbol{p_1})$ to $(r_0, r_1, \boldsymbol{\theta_0}, \boldsymbol{\theta_1})$ is helpful. By change of variables, the transformation gives the joint prior density as:

$$\begin{aligned} f(r_0, r_1, \boldsymbol{\theta_0}, \boldsymbol{\theta_1}) \propto & \left(\frac{r_0\theta_{10} - r_1\theta_{00}}{r_0 - r_1} \right)^{c_{00}-1} \times \\ & \left(\frac{r_0\theta_{11} - r_1\theta_{01}}{r_0 - r_1} \right)^{c_{01}-1} \times \\ & \left(\frac{r_0\theta_{12} - r_1\theta_{02}}{r_0 - r_1} \right)^{c_{02}-1} \times \\ & \left(\frac{(1-r_1)\theta_{00} - (1-r_0)\theta_{10}}{r_0 - r_1} \right)^{c_{10}-1} \times \\ & \left(\frac{(1-r_1)\theta_{01} - (1-r_0)\theta_{11}}{r_0 - r_1} \right)^{c_{11}-1} \times \\ & \left(\frac{(1-r_1)\theta_{02} - (1-r_0)\theta_{12}}{r_0 - r_1} \right)^{c_{12}-1} \times \\ & \frac{1}{(r_0 - r_1)^2} \; I_{\mathbb{Q}(\boldsymbol{\theta})}(r_0, r_1), \end{aligned}$$

where a non-zero density is obtained only when $\boldsymbol{r} \in \mathbb{Q}(\boldsymbol{\theta})$.

The joint posterior density of all the parameters given the data can be expressed as:

$$f(r_0, r_1, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1 \mid X^*, Y) = f(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1 \mid X^*, Y) f(r_0, r_1 \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1).$$

The distribution of the data (X^*, Y) gives direct information on parameters $\boldsymbol{\theta_0}$ and $\boldsymbol{\theta_1}$ only. As the sample sizes of the control and case groups increases, the conditional density $f(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1 \mid X^*, Y)$ will become narrower, converging to a point mass at the true values of $\boldsymbol{\theta_0}$ and $\boldsymbol{\theta_1}$ in the limit. Also, it is easy to point out that for fixed $(\boldsymbol{\theta_0}, \boldsymbol{\theta_1})$, the conditional prior density $f(r_0, r_1 \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$ is simply proportional to the joint prior density $f(r_0, r_1, \boldsymbol{\theta_0}, \boldsymbol{\theta_1})$. Thus the limiting posterior distribution of (r_0, r_1) can simply be 'read off' from the expression given above.

As a final step, the limiting posterior distribution of (r_0, r_1) induces a limiting posterior distribution on the exposure-disease odds-ratio. By change of variables and marginalization, we have the limiting posterior density of the log odds-ratio, $s = \text{logit } r_1 - \text{logit } r_0$, as

$$f(s|\boldsymbol{\theta_0}, \boldsymbol{\theta_1}) \;=\; \int g(r_0; s) f(r_0, \text{expit}(s + \text{logit } r_0)|\boldsymbol{\theta_0}, \boldsymbol{\theta_1})\, dr_0, \qquad (5)$$

where

$$g(r_0; s) \;=\; \text{expit}(s + \text{logit } r_0)\{1 - \text{expit}(s + \text{logit } r_0)\}.$$

Note that the support of the integrand in (5) is those r_0 for which $\{r_0, \text{expit}(s+\text{logit } r_0)\} \in \mathbb{Q}(\boldsymbol{\theta})$. By inspection (e.g., see the bottom rows of Figure 1), for given $(s, \boldsymbol{\theta})$ this could be either an interval of r_0 values or a pair of disjoint intervals. Particularly, we can think of the support as arising from intersecting the identification region in the $\boldsymbol{r}$ plane with the level curve $\text{logit } r_1 - \text{logit } r_0 = s$. It is also easy to note that provided the prior density of $(\boldsymbol{r}|\boldsymbol{\theta})$ is bounded on $\mathbb{Q}(\boldsymbol{\theta})$, the limiting density $f(s|\boldsymbol{\theta})$ will tend to zero as s approaches the lower bound on the log-OR, since the support of the integrand in (5) is readily seen to shrink to a single point in this limit, i.e., a unique r_0 value gives rise to the lower bound value of s. For given values of $\boldsymbol{\theta}$, we can readily evaluate (5) using one-dimensional numerical integration.

3.2 Examples

The eight scenarios from Section 2 are revisited, in combination with two different settings of the prior distribution according to hyperparameters $(\boldsymbol{c_0}, \boldsymbol{c_1})$. The first setting is $\boldsymbol{c_0^-} = \boldsymbol{c_1^-} = (1, 1, 1)$, corresponding to uniform distributions for $\boldsymbol{p_0}$ and $\boldsymbol{p_1}$ across the prior region. As a second setting we take $\boldsymbol{c_0^+} = (6, 4, 2)$, $\boldsymbol{c_1^+} = (2, 4, 6)$ which assigns more prior weight to better classifications (henceforth we refer to this setting as the 'weighted' prior). We can mimic these hyperparameter settings for the collapsed case as well, via a Beta($\boldsymbol{c_{0*}}$) prior on specificity and a Beta($\boldsymbol{c_{1*}}$) prior on specificity. Then we take $\boldsymbol{c_{0*}^-} = \boldsymbol{c_{1*}^-} = (1, 1)$ as an instance of uniform priors. In light of the collapsibility property of Dirichlet distributions, the analogous 'weighted prior' setting when the maybe and unlikely categories are combined is $\boldsymbol{c_{0*}^+} = (10, 2)$, $\boldsymbol{c_{1*}^+} = (6, 6)$.

For the four scenarios involving 'symmetric' classification probabilities, the limiting posterior distributions of $\log OR$ appear in Figures 6 through 9.

The corresponding figures for the other four scenarios are available as supplementary figures (`www.stat.ubc.ca/~gustaf`). In the case of uniform priors, we consistently see constraint B lead to a more peaked limiting posterior distribution than constraint A, even though the identification region is unchanged. Thus, if it can be invoked, there is a benefit associated with the stronger assumption about misclassification probabilities. In turn, posteriors under constraint A are more peaked than their collapsed-case counterparts, even though the identification regions are only very marginally bigger for the collapsed case analysis. Thus we see a benefit associated with directly adjusting for misclassification into the three exposure categories, rather than collapsing to two categories and then adjusting.

The behaviour of the posteriors arising from the weighted priors is more nuanced. Under constraint A, moving from the uniform prior to the weighted prior tends to result in a more concentrated posterior, as one might expect. However, and surprisingly, under constraint B, moving to the weighted prior tends to flatten the posterior. Consequently, with the weighted prior, the constraint A and constraint B posterior distributions tend to be very similar. We have further investigated this surprising 'interaction' between using the more concentrated prior and the stronger constraint, and it seems to persist quite generally if exposure prevalences are low and the odds-ratio is modest. If we start with uniform priors and constraint A, the resulting posterior induces a negative dependence between $\log OR$ and $\pi_W(\boldsymbol{p})$, where $\pi_W()$ is the weighted prior density on the classification probabilities. Thus moving from the uniform prior to π_W 'downweights' the long right tail, and thereby sharpens the posterior distribution of $\log OR$. However, upon 'removing points' that do not satisfy constraint B, the dependence is seen to become positive. Thus the constraint B analysis has this curious feature of a more concentrated prior leading to a less concentrated posterior. We also note that with the weighted prior constraint A or B again leads to a more concentrated posterior than the 'collapse then adjust' strategy.

4 Finite-Sample Posteriors

Until now, we have only considered limiting behaviour in the infinite sample-size limit. Under this situation, the posterior on $\boldsymbol{\theta_0}$ and $\boldsymbol{\theta_1}$ reduces to a point mass at the true values. It is instructive to see how the finite-sample posterior distribution of the log odds-ratio moves toward the limiting posterior distribution when the sample size increases, by simulating data under several of the

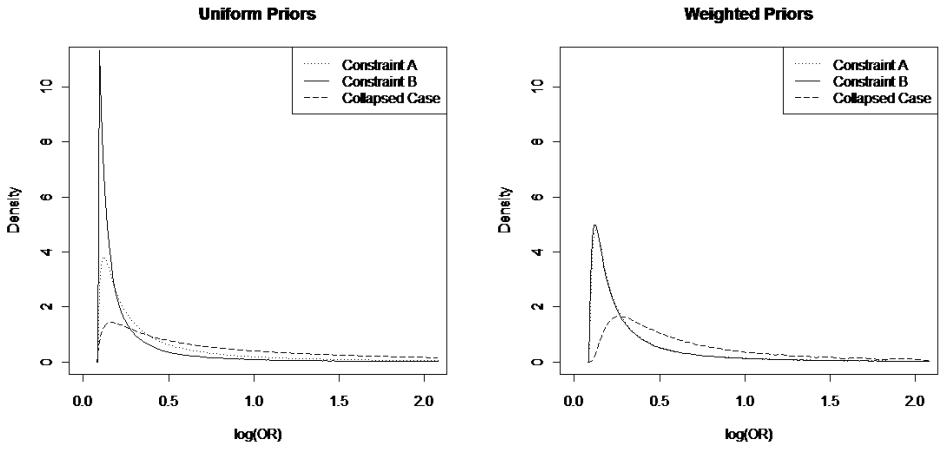

Figure 6: Limiting posterior distributions under the combination $(- - -)$. The left panel gives the limiting posterior distributions of the log odds-ratio under constraint A, constraint B, and the collapsed case, when $\boldsymbol{p_0}$ and $\boldsymbol{p_1}$ have uniform priors. The right panel gives the limiting posterior distributions when the prior distribution gives more weight to better exposure classifications. In this scenario, the true log odds-ratio is 0.1823. The lower bound of the log odds-ratio is 0.0839 under both constraint A and B, and 0.0818 under collapsed case.

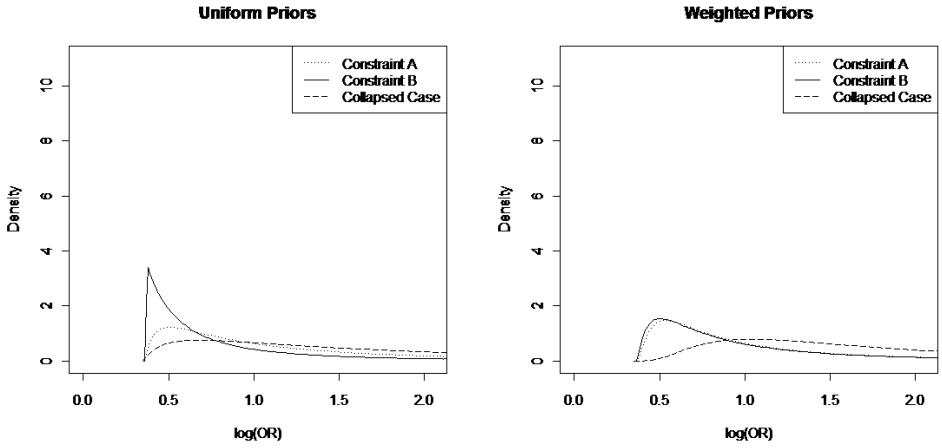

Figure 7: Limiting posterior distributions under the combination $(- + -)$. The layout is the same as Figure 6. In this scenario, the true log odds-ratio is 0.6932. The lower bound of the log odds-ratio is 0.3620 under both constraint A and B, and 0.3519 under the collapsed case.

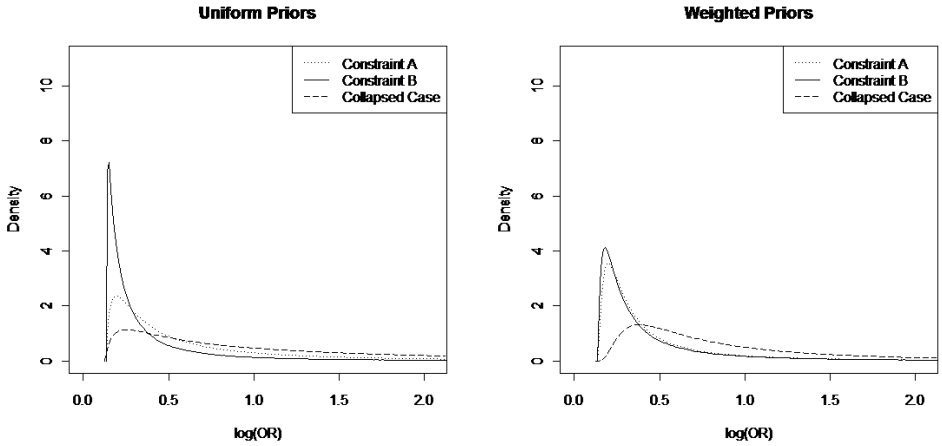

Figure 8: Limiting posterior distributions under the combination $(+\ -\ -)$. The layout is the same as Figure 6. In this scenario, the true log odds-ratio is 0.1823. The lower bound of the log odds-ratio is 0.1332 under both constraint A and B, and 0.1267 under the collapsed case.

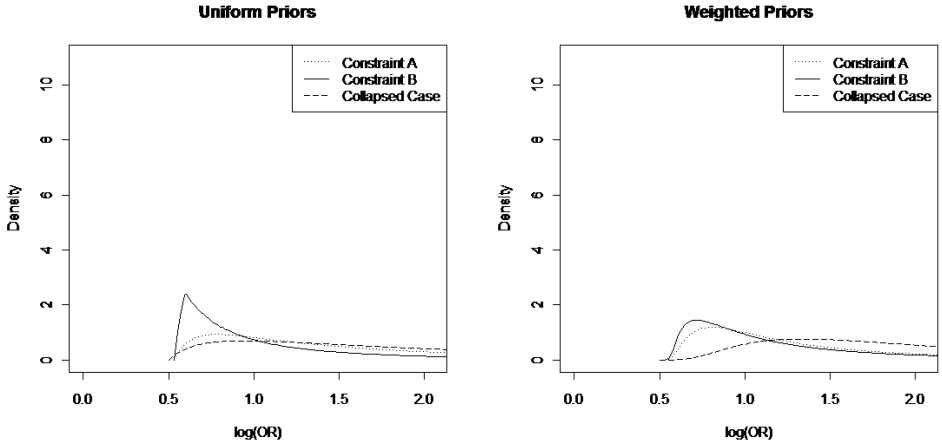

Figure 9: Limiting posterior distributions under the combination $(+\ +\ -)$. The layout is the same as Figure 6. In this scenario, the true log odds-ratio is 0.6932. The lower bound of the log odds-ratio is 0.5341 under both constraint A and B, and 0.5023 under the collapsed case.

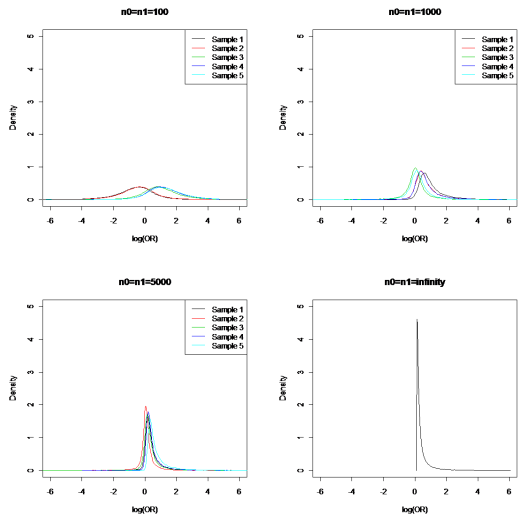

Figure 10: Posterior distributions under the combination $(---)$. From the upper-left panel to the lower-right panel, the posterior distributions of the log odds-ratio for the sample sizes 100, 1000, 5000, and the limiting posterior distribution are listed.

previous scenarios. The prior distributions are taken as $\boldsymbol{p}_0 \sim Dirichlet(6,4,2)$ and $\boldsymbol{p}_1 \sim Dirichlet(2,4,6)$, truncated according to constraint A. We simulate five independent data sequences with equal numbers of controls and cases ($n_i = n$, for $i = 0, 1$), and then determine the posterior distribution of log OR after $n = 100$, $n = 1000$, and $n = 5000$ observations, using WinBUGS (Lunn, Thomas, Best, and Spiegelhalter, 2000). We generically write $\boldsymbol{D_n}$ for the observed data. Posterior densities arising under the $(---)$ and $(+++)$ scenarios appear in Figures 10 and 11, with the limiting posterior densities also given.

In both scenarios we see the sampling variation in the posterior distribution diminish with sample size. We also see, however, that the posterior approaches its limit much more quickly in the $(+++)$ scenario then the $(---)$ scenario. In fact, this is readily understood, particularly if we contemplate how the posterior variance approaches its limit. We write the posterior variance as $\mathrm{Var}\{s(\boldsymbol{r})|\boldsymbol{D_n}\}$, where $s(\boldsymbol{r}) = \mathrm{logit} r_1 - \mathrm{logit} r_0$ is the log odds-ratio, and note that

$$\mathrm{Var}\{s(\boldsymbol{r})|\boldsymbol{D_n}\} \quad = \quad E\left[\mathrm{Var}\{s(\boldsymbol{r})|\boldsymbol{\theta}\}|\boldsymbol{D_n}\right] + \mathrm{Var}\left[E\{s(\boldsymbol{r})|\boldsymbol{\theta}\}|\boldsymbol{D_n}\right], \qquad (6)$$

where the first term tends to a positive constant as n increases, but the second term is of the order n^{-1}. In our general experience with partially identified

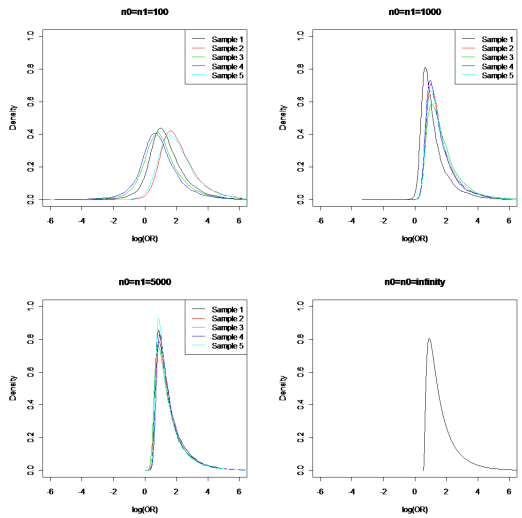

Figure 11: Posterior distributions under the combination $(+++)$. The layout is the same as Figure 10.

models, the first term can vary widely with the true parameter values. For instance, here it is far larger under the $(+++)$ parameters settings than the $(---)$ settings. On the other hand, the second (order n^{-1}) term, which is governed by the Fisher information in the model for $(\boldsymbol{D_n}|\boldsymbol{\theta})$, can vary much less with the parameter values. Thus getting 'close to convergence,' which corresponds to the second term becoming small compared to the first, can arise at a much smaller n when the first term is large, i.e., when the limiting posterior distribution is wide. Variance decompositions such as (6) in partially identified models are studied at length by Gustafson (2006).

The simulated datasets are also analyzed via the informal method alluded to in Section 1. That is, 'unlikely' and 'maybe' subjects are merged and taken as 'unexposed,' while the 'likely' subjects are taken to be 'exposed'. Then a standard analysis, without any adjustment for misclassification, is applied to the resulting 2×2 data table. A Bayesian instantiation of the standard analysis is applied, whereby the exposure prevalances for controls and cases are assigned independent uniform priors, leading to independent Beta posterior distributions. The corresponding posterior distributions for the log odds-ratio appear in Figures 12 and 13. In fact, these work quite well. By ignoring misclassification, markedly more peaked posterior distributions are obtained. Yet even when $n = 5000$, the resulting bias does not yet dominate. That is, the posterior does not yet rule out the true value of $OR = 1.2$ in the $(---)$

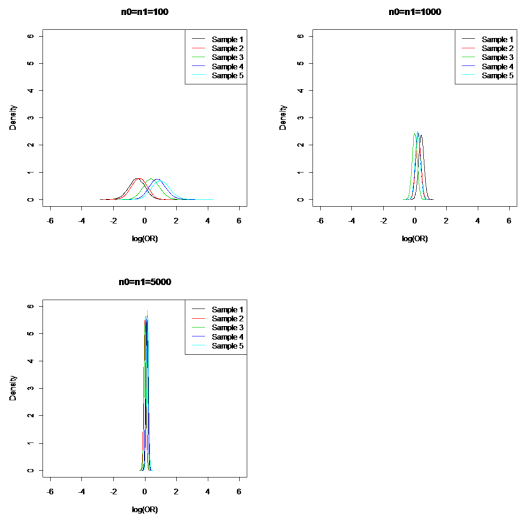

Figure 12: Posterior distributions via informal analysis under the combination $(- - -)$. From the upper-left panel to the lower-right panel, the posterior distributions of the log odds-ratio for the sample sizes 100, 1000, and 5000.

setting or $OR = 2.0$ in the $(+ + +)$ setting. Thus the informal strategy of choosing to treat 'maybe' subjects as being unexposed in light of low exposure prevalence proves to be useful. Of course we know that with enough data we would eventually be led astray. That is, from Table 1 we know that the posterior will tend to a point mass at $OR = 1.09$ in the $(- - -)$ case and a point mass at $OR = 1.72$ in the $(+ + +)$ case. Thus in concept, if not in practice, the informal scheme is unappealing.

5 Discussion

We have considered non-differential classification of a truly binary exposure into three categories. In this setting, inference about the exposure-disease association could be based on collapsing of categories as implicitly advocated by Dosemeci and Stewart (1996). Then the data could be analyzed without acknowledging misclassification, or perhaps binary misclassification with unknown sensitivity and specificity could be acknowledged. More formally, and as investigated here, the classification into three states can be modelled explicitly. This yields a partially identified inference problem, for which the first-order issue in the efficacy of inference is the size of identification region.

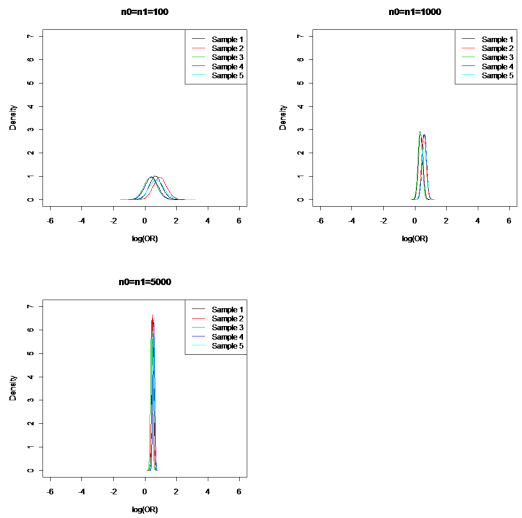

Figure 13: Posterior distributions via informal analysis under the combination $(+++)$. The layout is the same as Figure 13.

Regardless of whether Bayesian or non-Bayesian inference is pursued, the size of the identification region summarizes how much uncertainty about target parameters would remain if an infinite amount of data could be collected. In our case, particularly, Section 2 illustrates how an infinite amount of data can rule out near-null values of the exposure-disease association. We saw that the choice of prior region for the classification probabilities can have a marked effect on the bivariate identification region for the control and case exposure prevalences, but little or no effect on the resulting identification interval for the odds-ratio.

The second-order issue, investigated in Section 3, is the extent to which the posterior distribution, in the large-sample limit, is flat or concentrated across the identification region. We saw that in many circumstances the limiting posterior distribution of the log odds-ratio is indeed quite peaked. In Section 4 we also illustrated briefly how this limiting posterior distribution is approached with finite data sets, and drew comparisons with the informal approach of collapsing to two exposure categories and not adjusting for misclassification.

References

Birkett, N. (1992): "Effect of nondifferential misclassification on estimates of odds ratios with multiple levels of exposure," *American Journal of Epidemiology*, 136, 356–362.

Buonaccorsi, J. P. (2010): *Measurement Error: Models, Methods, and Applications*, Chapman and Hall, CRC Press.

Correa-Villaseor, A., W. F. Stewart, F. Franco-Marina, and H. Seacat (1995): "Bias from nondifferential misclassification in case-control studies with three exposure levels," *Epidemiology*, 6, pp. 276–281.

Dosemeci, M. and P. A. Stewart (1996): "Recommendations for reducing the effects of misclassification on relative risk estimates." *Occupational Hygiene*, 3, 169–176.

Dosemeci, M., S. Wacholder, and J. H. Lubin (1990): "Does nondifferential misclassification of exposure always bias a true effect toward the null value?" *American Journal of Epidemiology*, 19, 746–748.

Gustafson, P. (2004): *Measurement Error and Misclassification in Statistics and Epidemiology: Impact and Bayesian Adjustments*, Chapman and Hall, CRC Press.

Gustafson, P. (2005): "On model expansion, model contraction, identifiability, and prior information: two illustrative scenarios involving mismeasured variables (with discussion)," *Statistical Science*, 20, 111–140.

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

Gustafson, P. (2010): "Bayesian inference for partially identified models," *International Journal of Biostatistics*, 6, issue 2 article 17.

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

Lunn, D. J., A. Thomas, N. Best, and D. Spiegelhalter (2000): "WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility." *Statistics and Computing*, 10, 325–337.

Manski, C. F. (2003): *Partial Identification of Probability Distributions*, Springer.

Weinberg, C. R., D. M. Umbach, and S. Greenland (1994): "When will nondifferential misclassification of an exposure preserve the direction of a trend? (with discussion)," *American Journal of Epidemiology*, 140, 565–571.