Topics on the Effect of Non-differential Exposure Misclassification

by

Dongxu Wang

B.Sc., The University of British Columbia, 2010

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate Studies

(Statistics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

July 2012

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. When the exposure is a continuous variable or a binary variable, a general mismeasurement phenomenon is attenuation in the strength of the relationship between exposure and outcome. However, few have investigated the effect of misclassification on a polychotomous variable. Using Bayesian methods, we investigate how misclassification affects the exposure-disease associations under different settings of classification matrix. Also, we apply a trend test and understand the effect of misclassification according to the power of the test. In addition, since virtually all of work on the impact of exposure misclassification presumes the simplest situation where both the true status and the classified status are binary, my work diverges from the norm, in considering classification into three categories 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 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.

Table of Contents

List of Tables

List of Figures

Acknowledgements

First and foremost, I would like to express my sincerest gratitude to my supervisor, Paul Gustafson. During my graduate study, his expertise and suggestions help me all the time of research and writing of this thesis, and at the same time, his patience and understanding allows me the room to work in my own way. This thesis would not have been completed without the guidance and the help of him. He is the best supervisor anyone could wish for. I would like to thank Professor Lang Wu for taking time out from his busy schedule to be the second reader of this thesis.

I must also thank Professor Rollin Brant, John Petkau, Matias Salibian-Barrera, and Ruben Zamar for their teaching. They are the best teachers I have had in my life.

I would like to thank all the student colleagues, Ardavan Saeedi, Chao Xiong, Hao Luo, Hongyang Zhang, James Proudfoot, Jessica Chen, Jing Dong, Lang Qin, Meijiao Guan, Qian Ye, Yang Liu, Yongliang Zhai, Yuqing Wu, for their various forms of support during my graduate study. They provide a stimulating and fun environment in which to learn and grow.

I would like to thank all the staffs in the department office, Peggy Ng, Elaine Salameh, and Andrea Sollberger, for their consistent help during these two years. Because of them, I can focus on my research.

Last but not the least, I wish to thank my parents, Jian Wang and Guang Yu, for supporting me throughout all my life. You are always the most important persons to me. Also I need to thank my girlfriend, Zhaozhao Qin, for her love and care during these years.

Chapter 1

Introduction

In many studies, especially in the area of epidemiology, researchers intend to measure both the outcome Y and the exposure X for given experimental units and use statistical models to investigate the relationship between Y and X according to the recorded data. Depending on the study design, the number of exposure and outcome variables vary a lot. In this thesis, we will focus on the simplest case which contains only one outcome Y and one exposure X. In general, both Y and X can be either binary, polychotomous, or continuous. For example, the outcome can be death status which is a binary variable, health condition measured in five-point Likert scale which is a polychotomous variable, or blood pressure which is continuous.

In the ideal case, when the outcome and the exposure are exactly the target variables and are precisely measured, the measurement is considered as 'gold standard'. Under this situation, as both Y and X are measured precisely, there is no doubt that the relationship inferred from the recorded data is the intended relationship. However, in almost all studies, it is impossible to have the gold standard measurement because of various types of limitations. These limitations include selection bias and information bias, which are discussed by Gordis (1996) [7]. For example, Krisbnaiab, et al (2005) [15], carried out a cross-sectional study to investigate the association between smoking and cataracts. To access cumulative smoking dose, subjects were classified based on cigarette pack-years, which was calculated by multiplying the number of packs of cigarettes smoked per day by the number of years the person had smoked. However, as data were gathered through questionnaires, both the number of packs of cigarettes smoked per day and the number of years of smoking were recorded imprecisely because of the reporting bias. It is easy to understand that many heavy smokers were not expected to report as many cigarettes as they smoked, while some smokers might even over report. Therefore, the recorded variables might not be exactly the actual target variables, but some corresponding surrogate variables, a surrogate outcome Y^* in place of Y or a surrogate exposure X^* in place of X.

In some cases, however, several measurements exist but none of them is

viewed as gold standard. Mismeasurement will not be concerned in these cases, and the issues discussed in this thesis will not apply. Instead, these measurements might be combined into one or more dimensions which is useful to describe the main features of them, as mentioned in Kraemer, et al (2003) [14].

1.1 Exposure Mismeasurement

The goal of a study is to investigate the relationship between Y and X, while the relationship gained from the recorded data is actually between the surrogates, but not the target variables. The analysis which pretends the surrogate variables are exactly the same as the variables of interest is referred to as 'naive'. Because of the mismeasurement, the recorded data might lead to misleading results, which affect the conclusion much. Therefore, it is worth discussing the effect of mismeasurement. For example, Kenkel, et al (2004) [13], investigated the effect of mismeasurement in retrospective smoking data. They showed that failing to account for mismeasurement would have led to incorrect inferences about the association between cigarette price and smoking participation. As the conclusion, it is important to take into account the impact of mismeasurement.

In real cases, both the outcome and the exposure are possibly subject to mismeasurement. In Gustafson (2004) [8], it shown that the main focus is on the situations where the exposure X is subject to mismeasurement. The equally realistic scenarios where the outcome Y is subject to mismeasurement are reasonable to be ignored. Under the situation where only exposure mismeasurement happens, doing the analysis by pretending the surrogate of the exposure X^* is actually the variable of interest can lead to biased inferences, not just less precise inferences. Therefore, in this thesis we will only consider the situation where the exposure X is subject to mismeasurement but the outcome Y is precisely measured. That is, the discussion in the following part is about the studies where both a surrogate exposure X^* and an actual outcome Y are recorded. Moreover, many studies have been done when the outcome is subject to misclassification, such as in Garcia-Zattera, et al (2010) [6], and in Savoca (2011) [19].

If the exposure X is categorical, imprecise measurements of X are called misclassifications, while if it is continuous, they are called measurement errors, as described in Shen (2009) [20]. When X is continuous, set the measurement error to be Z. The surrogate X^* is defined to be a linear form of X such that $X^* = X + Z$. When X is categorical, misclassification

happens when classifying an item into a wrong exposure level.

Also, it is worth mentioning that mismeasurement may occur in two forms: differential and non-differential. In differential mismeasurement, the distribution of the surrogate X^* depends on both the outcome Y and the true exposure X. For example, mismeasurement of exposure may occur such that subjects in case group are mismeasured more often than those in control group, as explained in Gordis (1996) [7]. In contrast, non-differential mismeasurement refers to the situation where the surrogate X^* depends only on the true exposure X but not on the outcome Y. More formally, the conditional distribution of $(X^* \mid X, Y)$ is identically the conditional distribution of $(X^* \mid X)$. Non-differential mismeasurement results from the degree of inaccuracy that characterizes how information is obtained from any study group, as explained in Gordis (1996) [7]. In this thesis, all the discussions are under the assumption of non-differential exposure mismeasurement.

1.2 The Impact of Mismeasurement

Many have discussed how well naive estimation performs in the face of mismeasurement, for both continuous and binary exposures, i.e., in Gustafson (2004) [8]. A general mismeasurement phenomenon is attenuation in the strength of the relationship between Y and X induced by mismeasuring X as X^*. The attenuation factor is used to represent the magnitude of the impact of mismeasurement. However, even polychotomous exposure is much more common than binary exposure in real cases, not many works have done on this kind of situations. To adjust misclassification in those studies, it is worth discussing the effect of misclassification for polychotomous exposure first. We will briefly state the effect of mismeasurement for continuous and binary exposures, and mainly discuss the effect for polychotomous exposure in Chapter 2.

1.2.1 Mismeasured Continuous Variables

Consider the case where both the outcome Y and the exposure X are continuous. The true relationship between Y and X is assumed to be $E(Y \mid X) = \beta_0 + \beta_1 X$, where the coefficients β_0 and β_1 are unknown. Under the situation when only exposure mismeasurement happens, Y is measured correctly but X^* is recorded as a surrogate of the target exposure X. If the mismeasurement is not adjusted by the researchers, the naive analysis will investigate the relationship $E(Y \mid X^*) = \beta_0^* + \beta_1^* X^*$ instead of

$E(Y \mid X) = \beta_0 + \beta_1 X$. In this way, β_0^* and β_1^* will be interpreted, but not β_0 and β_1.

Assume that both X and X^* follow normal distributions. Also, assume that X^* is a non-differential and unbiased surrogate of X. Denote that $X \sim N(\mu, \sigma^2)$ and $X^* \sim N(X, \tau^2\sigma^2)$, where the parameter τ is chosen as $SD(X^* \mid X)/SD(X)$. Under the standard assumption, standard distributional theory implies that $(X^* \mid X)$ has a bivariate normal distribution. Using the non-differential property and the properties of this bivariate normal distribution, the relationship between β^* and β can be written as:

$$\beta_0^* = \beta_0 + \frac{\mu\beta_1}{1+\tau^2},$$

$$\beta_1^* = \frac{\beta_1}{1+\tau^2}.$$

The attenuation factor is referred as $\beta_1^*/\beta_1 = 1/(1+\tau^2)$, which is positive but smaller than one. It shows that the relationship between Y and X^* has the same direction of trend as between Y and X. However, since $|\beta_1^*| < |\beta_1|$, the slope of the relationship between Y and X^* will be always flatter. Espino-Hernandez, et al (2010) [5], carried out a study about how to adjust measurement error in case-control studies with continuous exposure using the Bayesian method.

1.2.2 Misclassified Binary Variables

A categorical exposure can be either binary or polychotomous. When X is binary, the impact of misclassification is easy to intuit. The magnitude of the misclassification can be described by the sensitivity and the specificity of X^*, i.e., the probability of correct classification for exposed and unexposed subjects respectively. Note that the sensitivity is defined as $SN = Pr(X^* = 1 \mid X = 1)$, and the specificity is defined as $SP = Pr(X^* = 0 \mid X = 0)$.

Consider the case where the outcome Y is continuous and the exposure X is binary. Without loss of generality, the relationship between Y and X can be written as $E(Y \mid X) = a + bX$, such that $E(Y \mid X = 0) = a$ and $E(Y \mid X = 1) = a + b$. When misclassifications arise, $E(Y \mid X^*) = a^* + b^*X^*$ is estimated instead of $E(Y \mid X) = a + bX$, where

$$a^* = a + bPr(X = 1 \mid X^* = 0)$$

and

$$\frac{b^*}{b} = 1 - Pr(X = 0 \mid X^* = 1) - Pr(X = 1 \mid X^* = 0).$$

The attenuation factor is referred as b^*/b. Since the attenuation factor is always smaller than 1, larger probability of misclassification leads to more attenuated result, which is the same for continuous exposure. Chu (2010) [2] carried out a study about how to adjust misclassification in case-control studies with binary exposure using the Bayesian method.

The attenuation factor b^*/b can also be represented in terms of sensitivity and specificity. Note that prevalence is defined as the number of exposed persons present in the population at the specific time divided by the number of persons in the population at that time. Setting the actual and apparent prevalences of exposure to be $\alpha = Pr(X = 1)$ and $\alpha^* = Pr(X^* = 1)$ respectively, the attenuation factor can be written as:

$$\frac{b^*}{b} = (SN + SP - 1)\frac{\alpha(1-\alpha)}{\alpha^*(1-\alpha^*)}.$$

Since it is obvious that α^* is a function of (α, SN, SP), the attenuation factor is also a function of (α, SN, SP).

1.2.3 Misclassified Polychotomous Variables

In a lot of research, the outcome Y is still a continuous variable as described in Section 1.2.1 and 1.2.2, but the exposure X becomes an ordinary polychotomous variable, a categorical variable with more than two categories. Under this situation, if misclassification happens, the impact is more complex and hard to predict. For example, Lindblad, et al (2005) [16], carried out a study about the association between the intensity of smoking and cataracts. In this study, the exposure is the intensity of smoking, which is an ordinary variable with more than 2 categories. Subjects were classified into one of many exposure levels according to the number of cigarettes smoked per day. As the data is based on the Swedish Mammography Cohort, exposure misclassification is unavoidable. Without the adjustment of misclassification, biased results may influence the conclusion. It will be informative to investigate how misclassification affects the results in such studies.

As stated in Section 1.2.1 and 1.2.2, when the exposure X is either continuous or binary, the mismeasurement always attenuates the strength of the exposure-disease association. Is there a same conclusion for a polychotomous exposure? When the exposure X is an ordinary polychotomous variable, both the distribution of the exposure and the outcome will affect the final results. We will treat subjects with the same exposure level as a group, and use the group mean outcomes as responses. In Chapter 2, we

will first discuss the effect of misclassification with different distributions of actual group mean outcomes. In this part, we will mainly investigate two situations, in which subjects cannot be misclassified more than one category off the true exposure level and in which subjects cannot be misclassified more than two categories. In each of these two situations, we will start with the simplest case where the group mean outcomes are monotone, and then expand to the general case. Consequently, we will describe the effect of misclassification with different distributions of the actual exposure. After all, to compare with the continuous and binary exposure, we will use a trend test to analyze the effect of misclassification for polychotomous exposure.

1.3 Case-Control Studies with a "Maybe" Exposed Group

In epidemiology, there are two main types of designs for clinical studies, randomized studies and observational studies. Three types of observational studies are mostly applied: cohort study, case-control study, and cross-sectional study. Because of the limitations for a cross-sectional study in establishing a relationship between exposure and outcome, cohort and case-control studies are relied on in epidemiologic investigations and clinical research to estimate etiologic relationships. In Chapter 3, we will focus on unmatched case-control study, where the goal is to infer exposure prevalences in case and control populations, and thereby infer the exposure-disease odds ratio.

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, et al (1996) [3]. For instance, in interpreting their findings they say:

> 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 exposure prevlance 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. Particulary, we investigate directly modelling the exposure classification into the 'unlikely exposed', 'maybe exposed', and 'likely exposed' 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, et al (1990) [4], and Weinberg, et al (1994) [21]), 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 my 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) [8] or Buconaccorsi (2010) [1]. This thesis 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 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 [18]), 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.

Having established the form of the identification regions, we turn to determining 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) [9] (also see Gustafson 2010 [11]). Necessarily, the support of the limiting posterior distribution is the corresponding identification region. Finally, we also show, via simulation, how the posterior distribution approaches its limit as the sample size grows.

Chapter 2

The Impact of Misclassified Polychotomous Variable

As introduced in Section 1.2.1 and 1.2.2, mismeasurement will attenuate the exposure-disease association for both continuous exposure and binary exposure. When the actual exposure X is a polychotomous ordered variable, the impact of misclassification is more complex and hard to intuit. Denote Y as the outcome of experimental units, we only consider the situation when Y is continuous. Also, let X and X^* represent actual exposure status and apparent exposure status, which are both polychotomous ordered variables. It means that non-differential misclassification gives rise to X^* as a surrogate for X. Without loss of generality, let X and X^* take values in 1, 2, $\cdots$, k.

Let P denote the classification matrix, where p_{ij} is the probability of classifying a subject into the j-th exposure level given this subject is actually in the i-th exposure level ($p_{ij} = P(X^* = j \mid X = i)$). Some weak assumptions can be made in order to ensure that the classification is better than randomly assign subjects into exposure levels. We name it the monotonicity assumption of p_{ij} such that:

- p_{ij} is maximized when $j = i$, for any fixed i from 1 to k;
- when $1 \leq j \leq i$, p_{ij} decreases as j decreases;
- when $i \leq j \leq k$, p_{ij} decreases as j increases.

Treating subjects with the same exposure level as a group, we will use the group mean outcomes as responses. Let $\boldsymbol{\gamma}$ be the mean outcome for actual exposure status ($\gamma_i = E(Y \mid X = i)$), $\boldsymbol{\gamma}^*$ be the mean outcome for apparent exposure status ($\gamma_i^* = E(Y \mid X^* = i)$), and $\boldsymbol{\alpha}$ be the distribution of actual exposure ($\alpha_i = Pr(X = i)$). According to Bayes' theorem, the relationship between $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^*$ can be expressed based on $\boldsymbol{\alpha}$ and P, such that:

$$\gamma_j^* = \frac{\sum_{i=1}^{k} \gamma_i p_{ij} \alpha_i}{\sum_{i=1}^{k} p_{ij} \alpha_i}. \tag{2.1}$$

Given the values of P and $\boldsymbol{\alpha}$, it is informative to investigate the mapping from $\boldsymbol{\gamma}$ to $\boldsymbol{\gamma}^*$ to understand the impact of misclassification. To have a deeper understanding about the behavior of X and X^*, we can treat the classification matrix P as a transition matrix of a Markov chain and $\boldsymbol{\alpha}$ as the probability distribution of the starting position. Then, the distribution of the surrogate X^* can be viewed as the distribution of X with one step toward the stationary distribution, which can be expressed as:

$$Pr(X^* = j) = \sum_{i=1}^{k} p_{ij}\alpha_i.$$

When calculating γ_j^*, the denominator of (2.1) is exactly the form of $Pr(X^* = j)$, but the numerator of (2.1) is distributed as one step toward the stationary distribution with the starting distribution proportional to $\boldsymbol{\gamma} \odot \boldsymbol{\alpha}$, i.e., $\boldsymbol{\gamma} \odot \boldsymbol{\alpha}$ is a vector with k elements such that $(\gamma \odot \alpha)_i = \boldsymbol{\gamma}_i\boldsymbol{\alpha}_i$ for any i from 1 to k. To investigate how $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^*$ behave, we can compare the change between the mean outcomes of each pair of adjacent levels for both X and X^*.

2.1 Subjects Cannot Be Misclassified More Than Two Categories off the True Exposure Level

When no misclassification happens at any exposure level, the classification matrix P is an identity matrix, resulting in exactly the same $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^*$. Under the monotonicity assumption of p_{ij}, let's first consider the least severe situation when misclassification happens, where subjects cannot be misclassified more than one category off the true exposure level. Under this situation, the classification matrix is restricted such that $p_{ij} = 0$ if $|i - j| \geq 2$. We will analyze the effect of misclassification for polychotomous exposure based on two different distributions of $\boldsymbol{\gamma}$.

2.1.1 When $E(Y \mid X)$ is Monotone

Let's first have a look at the effect of misclassification for polychotomous exposure under the simplest situation where $\boldsymbol{\gamma}$ is monotone. For any type of exposure distribution, we can generate a theorem.

Theorem 2.1.1 *Under the situation when $p_{ij} = 0$ if $|i - j| \geq 2$, if $\boldsymbol{\gamma}$ is monotone, $\boldsymbol{\gamma}^*$ will also be monotone with the same direction of trend as $\boldsymbol{\gamma}$.*

Proof:

(i): First, assume that X is uniformly distributed. That is, $\alpha_i = 1/k$ for any i from 1 to k. Then, the changes of the mean outcomes between adjacent levels can be expressed as:

- When $j = 1$,

$$\begin{aligned}\gamma_2^* - \gamma_1^* &= \frac{\gamma_1 p_{12} + \gamma_2 p_{22} + \gamma_3 p_{32}}{p_{12} + p_{22} + p_{32}} - \frac{\gamma_1 p_{11} + \gamma_2 p_{21}}{p_{11} + p_{21}} \\ &\propto (\gamma_2 - \gamma_1) p_{11} p_{22} + (\gamma_3 - \gamma_1) p_{11} p_{32} \\ &\quad + (\gamma_1 - \gamma_2) p_{21} p_{12} + (\gamma_3 - \gamma_2) p_{21} p_{32}.\end{aligned}$$

- When $1 < j < k - 2$,

$$\begin{aligned}\gamma_{j+1}^* - \gamma_j^* &= \frac{\gamma_j p_{j,j+1} + \gamma_{j+1} p_{j+1,j+1} + \gamma_{j+2} p_{j+2,j+1}}{p_{j,j+1} + p_{j+1,j+1} + p_{j+2,j+1}} \\ &\quad - \frac{\gamma_{j-1} p_{j-1,j} + \gamma_j p_{jj} + \gamma_y p_{j+1,j}}{p_{j-1,j} + p_{jj} + p_{j+1,j}} \\ &\propto (\gamma_j - \gamma_{j-1}) p_{j-1,j} p_{j,j+1} \\ &\quad + (\gamma_{j+1} - \gamma_{j-1}) p_{j-1,j} p_{j+1,j+1} \\ &\quad + (\gamma_{j+2} - \gamma_{j-1}) p_{j-1,j} p_{j+2,j+1} \\ &\quad + (\gamma_{j+1} - \gamma_j) p_{jj} p_{j+1,j+1} \\ &\quad + (\gamma_{j+2} - \gamma_j) p_{jj} p_{j+2,j+1} \\ &\quad + (\gamma_j - \gamma_{j+1}) p_{j,j+1} p_{j+1,j} \\ &\quad + (\gamma_{j+2} - \gamma_{j+1}) p_{j+1,j} p_{j+2,j+1}.\end{aligned}$$

- When $j = k - 1$,

$$\begin{aligned}\gamma_k^* - \gamma_{k-1}^* &= \frac{\gamma_{k-1} p_{k-1,k} + \gamma_k p_{kk}}{p_{k-1,k} + p_{kk}} \\ &\quad - \frac{\gamma_{k-2} p_{k-2,k-1} + \gamma_{k-1} p_{k-1,k-1} + \gamma_k p_{k,k-1}}{p_{k-2,k-1} + p_{k-1,k-1} + p_{k,k-1}} \\ &\propto (\gamma_{k-1} - \gamma_{k-2}) p_{k-2,k-1} p_{k-1,k} + (\gamma_k - \gamma_{k-2}) p_{k-2,k-1} p_{kk} \\ &\quad + (\gamma_k - \gamma_{k-1}) p_{k-1,k-1} p_{kk} + (\gamma_{k-1} - \gamma_k) p_{k,k-1} p_{k-1,k}.\end{aligned}$$

Under the monotonicity assumption of p_{ij}, it is easy to show $p_{jj} p_{j+1,j+1} > p_{j,j+1} p_{j+1,j}$ for any j. Therefore, when $\boldsymbol{\gamma}$ monotonically increases ($\gamma_{j-1} \leq \gamma_j \leq \gamma_{j+1} \leq \gamma_{j+2}$), we can prove that $\gamma_j^* \leq \gamma_{j+1}^*$ for any j from 1 to $k - 1$.

It means that $\boldsymbol{\gamma}^*$ also monotonically increases. Similarly, when $\boldsymbol{\gamma}$ monotonically decreases, $\boldsymbol{\gamma}^*$ also monotonically decreases ($\gamma_j^* \geq \gamma_{j+1}^*$ for any j from 1 to $k-1$). Notice that $\gamma_j^* = \gamma_{j+1}^*$ holds if and only if $\gamma_{j-1} = \gamma_j = \gamma_{j+1} = \gamma_{j+2}$.

(*ii*): In general, for any $\boldsymbol{\alpha}$ which satisfies $0 < \alpha_i < 1$ and $\sum_{i=1}^{k} \alpha_i = 1$, this theorem still holds. Take the situation when $1 < j < k-1$ as an example. The relationship between γ_j^* and γ_{j+1}^* can be represented as:

$$\begin{aligned}
\gamma_{j+1}^* - \gamma_j^* = & \frac{\gamma_j\alpha_j p_{j,j+1} + \gamma_{j+1}\alpha_{j+1}p_{j+1,j+1} + \gamma_{j+2}\alpha_{j+2}p_{j+2,j+1}}{\alpha_j p_{j,j+1} + \alpha_{j+1}p_{j+1,j+1} + \alpha_{j+2}p_{j+2,j+1}} \\
& - \frac{\gamma_{j-1}\alpha_{j-1}p_{j-1,j} + \gamma_j\alpha_j p_{jj} + \gamma_{j+1}\alpha_{j+1}p_{j+1,j}}{\alpha_{j-1}p_{j-1,j} + \alpha_j p_{jj} + \alpha_{j+1}p_{j+1,j}} \\
\propto & (\gamma_j - \gamma_{j-1})\alpha_{j-1}\alpha_j p_{j-1,j}p_{j,j+1} \\
& + (\gamma_{j+1} - \gamma_{j-1})\alpha_{j-1}\alpha_{j+1}p_{j-1,j}p_{j+1,j+1} \\
& + (\gamma_{j+2} - \gamma_{j-1})\alpha_{j-1}\alpha_{j+2}p_{j-1,j}p_{j+2,j+1} \\
& + (\gamma_{j+1} - \gamma_j)\alpha_j\alpha_{j+1}p_{jj}p_{j+1,j+1} \\
& + (\gamma_{j+2} - \gamma_j)\alpha_j\alpha_{j+2}p_{jj}p_{j+2,j+1} \\
& + (\gamma_j - \gamma_{j+1})\alpha_j\alpha_{j+1}p_{j,j+1}p_{j+1,j} \\
& + (\gamma_{j+2} - \gamma_{j+1})\alpha_{j+1}\alpha_{j+2}p_{j+1,j}p_{j+2,j+1}.
\end{aligned}$$

Based on (*i*), when $\boldsymbol{\gamma}$ is monotone, $\boldsymbol{\gamma}^*$ will still be monotone. Similarly, this conclusion also holds when $j = 1$ and $j = k-1$. ∎

To visualize the effect of misclassification, we will show an example where monotone increasing $\boldsymbol{\gamma}$ leads to monotone increasing $\boldsymbol{\gamma}^*$ in Example 2.1.1.

Example 2.1.1 *Take $k = 6$, $\gamma_i = i^2$, and*

$$P = \begin{pmatrix} 0.6 & 0.4 & 0 & 0 & 0 & 0 \\ 0.2 & 0.6 & 0.2 & 0 & 0 & 0 \\ 0 & 0.2 & 0.6 & 0.2 & 0 & 0 \\ 0 & 0 & 0.2 & 0.6 & 0.2 & 0 \\ 0 & 0 & 0 & 0.2 & 0.6 & 0.2 \\ 0 & 0 & 0 & 0 & 0.4 & 0.6 \end{pmatrix}$$

as an example. Both $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^*$ are displayed in Figure 2.1 when X is uniformly distributed ($\alpha_i = 1/6$), and in Figure 2.2 when X is unimodal ($\boldsymbol{\alpha} = (1,2,3,3,2,1)/12$). It shows that an increasing $\boldsymbol{\gamma}$ leads to an increasing $\boldsymbol{\gamma}^*$, as expected according to Theorem 2.1.1. ∎

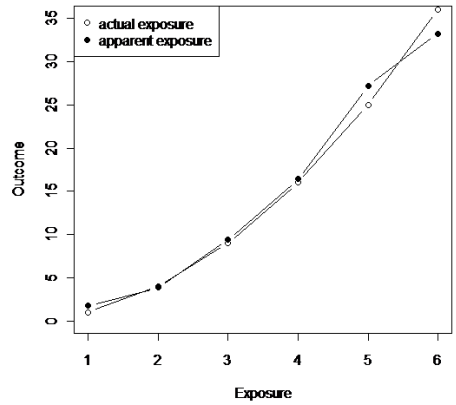

Figure 2.1: $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^*$ when X is uniformly distributed and $\boldsymbol{\gamma}$ is increasing.

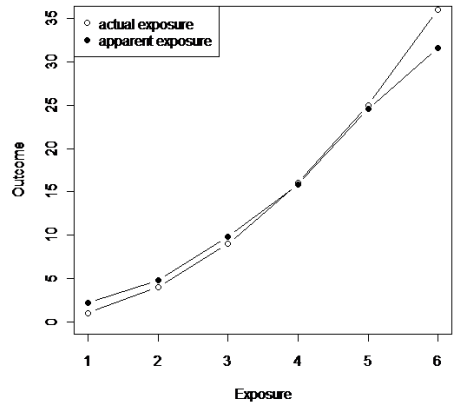

Figure 2.2: $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^*$ when X is unimodal and $\boldsymbol{\gamma}$ is increasing.

2.1.2 When $E(Y \mid X)$ is not Monotone

When $\boldsymbol{\gamma}$ is not monotone, a similar conclusion as in Theorem 2.1.1 does not hold. Let $sign(\gamma_{j+1} - \gamma_j)$ and $sign(\gamma^*_{j+1} - \gamma^*_j)$ represent the change of mean outcomes of actual exposure and apparent exposure at the j-th adjacent levels. For example, when $sign(\gamma_{j+1} - \gamma_j)$ is positive, the mean outcome increases at the j-th level; and when $sign(\gamma_{j+1} - \gamma_j)$ is negative, the mean outcome decreases. When $\boldsymbol{\gamma}$ is monotone, $sign(\gamma_{j+1} - \gamma_j)$ and $sign(\gamma^*_{j+1} - \gamma^*_j)$ will always be the same, according to Theorem 2.1.1. However, $sign(\gamma_{j+1} - \gamma_j)$ might differ from $sign(\gamma^*_{j+1} - \gamma^*_j)$ when $\boldsymbol{\gamma}$ is not monotone.

Consider the simplest case where X is uniformly distributed and $\boldsymbol{\gamma}$ is unimodal with the lowest value taken at $i = l$, where l is between 2 and $k - 1$. Under this situation, γ_l is the minimum value among $\boldsymbol{\gamma}$. Using the same procedure as in Theorem 2.1.1, it is easy to prove that when $j \neq l - 1$ or l, $sign(\gamma_{j+1} - \gamma_j)$ and $sign(\gamma^*_{j+1} - \gamma^*_j)$ are still always the same; however, either $\gamma^*_{l-1} < \gamma^*_l$ or $\gamma^*_l > \gamma^*_{l+1}$ might be the case. It means that only when $j = l - 1$ or l, it is possible that $sign(\gamma_{j+1} - \gamma_j)$ and $sign(\gamma^*_{j+1} - \gamma^*_j)$ are different. The relationships of the mean outcome among the $(l-1)$-th, l-th, and $(l+1)$-th levels highly depend on the magnitudes of γ_{l-1} and γ_{l+1}.

To visualize the effect of misclassification, we will show two examples where $\boldsymbol{\gamma}$ is unimodal in Example 2.1.2.

Example 2.1.2 *Take $k = 6$, X is uniformly distributed, $\boldsymbol{\gamma}$= $(50, 49, 48, 1, 5, 6)$, and*

$$P = \begin{pmatrix} 0.9 & 0.1 & 0 & 0 & 0 & 0 \\ 0.1 & 0.8 & 0.2 & 0 & 0 & 0 \\ 0 & 0.2 & 0.5 & 0.3 & 0 & 0 \\ 0 & 0 & 0.2 & 0.5 & 0.3 & 0 \\ 0 & 0 & 0 & 0.2 & 0.7 & 0.1 \\ 0 & 0 & 0 & 0 & 0.1 & 0.9 \end{pmatrix}$$

as an example. Both $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^*$ are displayed in Figure 2.3. Under this setting, $\gamma_4 < \gamma_5$ but $\gamma^*_4 > \gamma^*_5$. The minimum mean outcome of actual exposure is taken at the 4-th level, however the minimum mean outcome of apparent exposure is taken at the 5-th level. Because of misclassification, the minimum value moves one category to the right.

In contrast, take $k = 6$, X is uniformly distributed, $\boldsymbol{\gamma} = (6, 5, 3, 2, 1, 30)$,

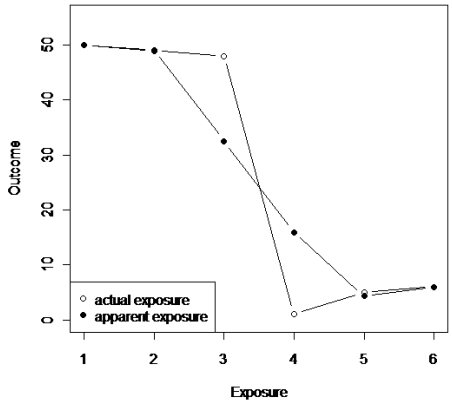

Figure 2.3: $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^*$ when X is uniformly distributed and $\boldsymbol{\gamma}$ is unimodal.

and

$$P = \begin{pmatrix} 0.9 & 0.1 & 0 & 0 & 0 & 0 \\ 0.1 & 0.8 & 0.2 & 0 & 0 & 0 \\ 0 & 0.2 & 0.5 & 0.3 & 0 & 0 \\ 0 & 0 & 0.3 & 0.5 & 0.2 & 0 \\ 0 & 0 & 0 & 0.3 & 0.5 & 0.2 \\ 0 & 0 & 0 & 0 & 0.4 & 0.6 \end{pmatrix}$$

as an example. Both $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^*$ are displayed in Figure 2.4. Under this setting, $\gamma_4 > \gamma_5$ but $\gamma_4^* < \gamma_5^*$. The minimum mean outcome of actual exposure is taken at the 5-th level, however the minimum mean outcome of apparent exposure is taken at the 4-th level. Misclassification can also move the minimum value one category to the left. ∎

Based on Example 2.1.2, we can have a conclusion for the situation where $\boldsymbol{\gamma}$ is unimodal under the assumptions we made. When $j \neq l-1$ or l, the directions of changes of $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^*$ will be the same as described in Theorem 2.1.1. When $j = l-1$ or l, although it is possible that $sign(\gamma_{j+1} - \gamma_j)$ and $sign(\gamma_{j+1}^* - \gamma_j^*)$ are different, the overall shape of $\boldsymbol{\gamma}^*$ is still unimodal. The overall shape of $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^*$ will stay unchanged with the mode moving only one category either to the left or to the right.

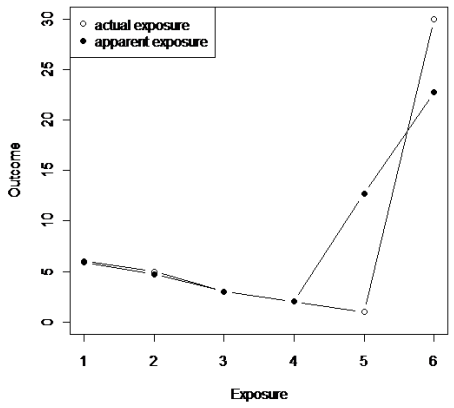

Figure 2.4: $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^*$ when X is uniformly distributed and $\boldsymbol{\gamma}$ is unimodal.

2.2 Subjects Cannot Be Misclassified More Than Three Categories off the True Exposure Level

In Section 2.1, we have discussed the situation with the least severe misclassification. As one step further, let's consider a worse situation where the surrogate X^* can be misclassified two categories off the true exposure level, which is still under the monotonicity assumption of p_{ij}. That is, the classification matrix is more complicated such that $p_{ij} = 0$ if $|i - j| \geq 3$. We will also analyze the effect of misclassification for polychotomous exposure based on two different distributions of $\boldsymbol{\gamma}$ as in Section 2.1.

2.2.1 When $E(Y \mid X)$ is Monotone

Still, let's first consider the simplest case when $\boldsymbol{\gamma}$ is monotone. Under the situation where X is uniformly distributed, the relationship between γ_j^* and

γ^*_{j+1} can be expressed as:

$$\begin{aligned}
\gamma^*_{j+1} - \gamma^*_j = & \frac{\gamma_{j-1}p_{j-1,j+1} + \gamma_j p_{j,j+1} + \gamma_{j+1}p_{j+1,j+1} + \gamma_{j+2}p_{j+2,j+1} + \gamma_{j+3}p_{j+3,j+1}}{p_{j-1,j+1} + p_{j,j+1} + p_{j+1,j+1} + p_{j+2,j+1} + p_{j+3,j+1}} \\
& - \frac{\gamma_{j-2}p_{j-2,j} + \gamma_{j-1}p_{j-1,j} + \gamma_j p_{jj} + \gamma_{j+1}p_{j+1,j} + \gamma_{j+2}p_{j+1,j}}{p_{j-2,j} + p_{j-1,j} + p_{jj} + p_{j+1,j} + p_{j+2,j}} \\
\propto & (\gamma_{j-1} - \gamma_{j-2})p_{j-2,j}p_{j-1,j+1} + (\gamma_j - \gamma_{j-2})p_{j-2,j}p_{j,j+1} \\
& + (\gamma_{j+1} - \gamma_{j-2})p_{j-2,j}p_{j+1,j+1} + (\gamma_{j+2} - \gamma_{j-2})p_{j-2,j}p_{j+2,j+1} \\
& + (\gamma_{j+3} - \gamma_{j-2})p_{j-2,j}p_{j+3,j+1} \\
& + (\gamma_j - \gamma_{j-1})p_{j-1,j}p_{j,j+1} + (\gamma_{j+1} - \gamma_{j-1})p_{j-1,j}p_{j+1,j+1} \\
& + (\gamma_{j+2} - \gamma_{j-1})p_{j-1,j}p_{j+2,j+1} + (\gamma_{j+3} - \gamma_{j-1})p_{j-1,j}p_{j+3,j+1} \\
& + (\gamma_{j-1} - \gamma_j)p_{jj}p_{j-1,j+1} + (\gamma_{j+1} - \gamma_j)p_{jj}p_{j+1,j+1} \\
& + (\gamma_{j+2} - \gamma_j)p_{jj}p_{j+2,j+1} + (\gamma_{j+3} - \gamma_j)p_{jj}p_{j+3,j+1} \\
& + (\gamma_{j-1} - \gamma_{j+1})p_{j+1,j}p_{j-1,j+1} + (\gamma_j - \gamma_{j+1})p_{j+1,j}p_{j,j+1} \\
& + (\gamma_{j+2} - \gamma_{j+1})p_{j+1,j}p_{j+2,j+1} + (\gamma_{j+3} - \gamma_{j+1})p_{j+1,j}p_{j+3,j+1} \\
& + (\gamma_{j-1} - \gamma_{j+2})p_{j+2,j}p_{j-1,j+1} + (\gamma_j - \gamma_{j+2})p_{j+2,j}p_{j,j+1} \\
& + (\gamma_{j+1} - \gamma_{j+2})p_{j+2,j}p_{j+1,j+1} + (\gamma_{j+3} - \gamma_{j+2})p_{j+2,j}p_{j+3,j+1}.
\end{aligned}$$

From the expression of apparent mean outcome change above, it is hard to determine $sign(\gamma^*_{j+1} - \gamma^*_j)$ given $sign(\gamma_{j+1} - \gamma_j)$. Therefore, Theorem 2.1.1 dose not hold under this situation. Even when $\boldsymbol{\gamma}$ is monotone, it is impossible to predict the changes of the mean outcomes of apparent exposure. For example, given $\boldsymbol{\gamma}$ is monotone increasing, the mean outcome of $\boldsymbol{\gamma}^*$ at some levels might decrease. Here is an example where $\boldsymbol{\gamma}$ is monotone but $\boldsymbol{\gamma}^*$ is not in Example 2.2.1.

Example 2.2.1 *Take $k = 6$, X is uniformly distributed,*
$\boldsymbol{\gamma} = (1, 2, 50, 51, 52, 53)$*, and*

$$P = \begin{pmatrix}
0.9 & 0.09 & 0.01 & 0 & 0 & 0 \\
0.01 & 0.34 & 0.33 & 0.32 & 0 & 0 \\
0.085 & 0.4 & 0.5 & 0.01 & 0.005 & 0 \\
0 & 0.01 & 0.48 & 0.5 & 0.006 & 0.004 \\
0 & 0 & 0.32 & 0.33 & 0.34 & 0.01 \\
0 & 0 & 0 & 0.01 & 0.09 & 0.9
\end{pmatrix}$$

as an example. Both $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^*$ are displayed in Figure 2.5. Given $\gamma_3 < \gamma_4$, misclassification leads to $\gamma^*_3 > \gamma^*_4$. Therefore, although $\boldsymbol{\gamma}$ is monotone increasing, it is possible that $\boldsymbol{\gamma}^*$ is not. ∎

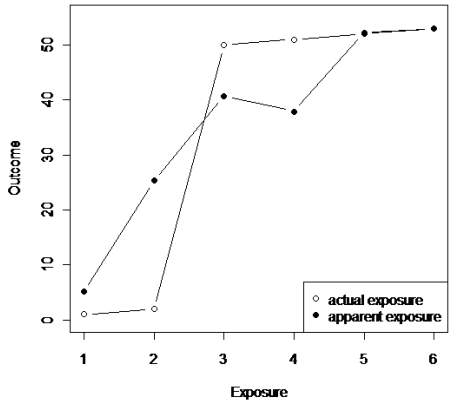

Figure 2.5: $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^*$ when X is uniformly distributed and $\boldsymbol{\gamma}$ is increasing.

2.2.2 When $E(Y \mid X)$ is not Monotone

In Section 2.1.2, we concluded that, under the situation where $p_{ij} = 0$ if $|i - j| \geq 2$, a unimodal $\boldsymbol{\gamma}$ will still lead to a unimodal $\boldsymbol{\gamma}^*$. However, for a worse classification matrix such that $p_{ij} = 0$ if $|i - j| \geq 3$, the same conclusion does not hold. It is possible that $\boldsymbol{\gamma}$ is unimodal but $\boldsymbol{\gamma}^*$ is not, as shown in Example 2.2.2.

Example 2.2.2 *Take $k = 6$, X is uniformly distributed,*
$\boldsymbol{\gamma}= (180, 30, 10, 1, 80, 81)$*, and*

$$P = \begin{pmatrix} 0.8 & 0.101 & 0.099 & 0 & 0 & 0 \\ 0.107 & 0.8 & 0.047 & 0.046 & 0 & 0 \\ 0.13 & 0.22 & 0.3 & 0.22 & 0.13 & 0 \\ 0 & 0.015 & 0.23 & 0.26 & 0.25 & 0.245 \\ 0 & 0 & 0.235 & 0.245 & 0.265 & 0.255 \\ 0 & 0 & 0 & 0.24 & 0.35 & 0.41 \end{pmatrix}$$

as an example. Both $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^*$ are displayed in Figure 2.6. It shows that even though $\boldsymbol{\gamma}$ is unimodal with γ_i minimized at $i = 4$, $\boldsymbol{\gamma}^*$ is not unimodal but bimodal where both $i = 2$ ($\gamma_1^* > \gamma_2^*$ and $\gamma_3^* > \gamma_2^*$) and $i = 4$ ($\gamma_3^* > \gamma_4^*$ and $\gamma_5^* > \gamma_4^*$) are local minimum points. ∎

In general, with other assumptions unchanged as in Section 2.1.2, even $\boldsymbol{\gamma}$ is unimodal, a worse misclassification in which subjects can be misclassified

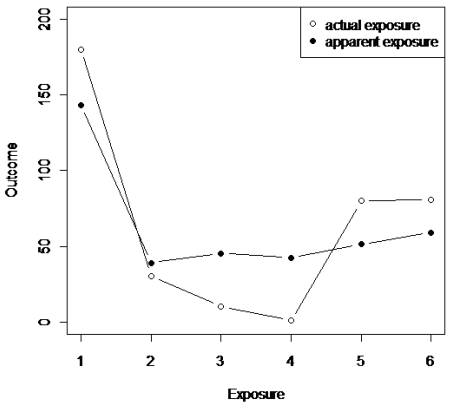

Figure 2.6: $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^*$ when X is uniformly distributed and $\boldsymbol{\gamma}$ is unimodal.

more than one category away from the true exposure level will influence the overall shape of $\boldsymbol{\gamma}^*$ such that it might not be unimodal.

2.3 The Effect of the Exposure Distribution

After discussing the effect of misclassification with different classification matrices and distributions of mean outcomes, we will focus on how $\boldsymbol{\gamma}^*$ will be influenced by the distribution of exposure X. Remind that we denote $\boldsymbol{\alpha}$ as the distribution of X such that $\alpha_i = Pr(X = i)$. Here, we only consider the simplest situation where subject can only be misclassified one category away from the true exposure level ($p_{ij} = 0$ if $|i - j| \geq 2$) under the monotonicity assumption of p_{ij} and $\boldsymbol{\gamma}$ is monotone.

In Section 2.1.1, we have proved that if $\boldsymbol{\gamma}$ is monotone, then $\boldsymbol{\gamma}^*$ will also be monotone with the same direction of trend, which does not depend on the value of $\boldsymbol{\alpha}$. The distribution of X does not affect the overall trend under this situation. Although the trend of adjacent levels will stay the same, the magnitude of apparent mean outcome $\boldsymbol{\gamma}^*$ will depend on $\boldsymbol{\alpha}$. We will discuss how $\boldsymbol{\gamma}^*$ is influenced by the value of $\boldsymbol{\alpha}$.

Given the definition of P and $\boldsymbol{\gamma}$ unchanged, let $\boldsymbol{\gamma}^{*1}$ be the vector of mean outcomes when X is uniformly distributed, and $\boldsymbol{\gamma}^{*2}$ be the vector of mean outcomes when X follows a non-uniform distribution. When $1 < j < k - 1$, the relationship between γ_j^{*1} and γ_j^{*2} given the same values of $\boldsymbol{\gamma}$ is shown

as:

$$\begin{aligned}\gamma_j^{*2} - \gamma_j^{*1} &= \frac{\gamma_{j-1}\alpha_{j-1}p_{j-1,j} + \gamma_j\alpha_j p_{jj} + \gamma_{j+1}\alpha_{j+1}p_{j+1,j}}{\alpha_{j-1}p_{j-1,j} + \alpha_j p_{jj} + \alpha_{j+1}p_{j+1,j}} \\ &\quad - \frac{\gamma_{j-1}p_{j-1,j} + \gamma_j p_{j,j} + \gamma_{j+1}p_{j+1,j}}{p_{j-1,j} + p_{jj} + p_{j+1,j}} \\ &\propto (\gamma_j - \gamma_{j-1})(\alpha_j - \alpha_{j-1})p_{j-1,j}p_{jj} \\ &\quad + (\gamma_{j+1} - \gamma_{j-1})(\alpha_{j+1} - \alpha_{j-1})p_{j-1,j}p_{j+1,j} \\ &\quad + (\gamma_{j+1} - \gamma_j)(\alpha_{j+1} - \alpha_j)p_{jj}p_{j+1,j}.\end{aligned}$$

Since subject can only be misclassified one category away from the true exposure level, it is easy to understand that the relationship between γ_j^{*2} and γ_j^{*1} is only influenced by three terms of $\boldsymbol{\alpha}$, (α_{j-1}, α_j, and α_{j+1}). Similarly, we can write out the expression when $j = 1$ and $j = k$, such that:

$$\gamma_1^{*2} - \gamma_1^{*1} \propto (\gamma_2 - \gamma_1)(\alpha_2 - \alpha_1)p_{11}p_{21},$$

$$\gamma_k^{*2} - \gamma_k^{*1} \propto (\gamma_k - \gamma_{k-1})(\alpha_k - \alpha_{k-1})p_{k-1,k}p_{kk}.$$

As other values of $\boldsymbol{\alpha}$ except α_{j-1}, α_j, and α_{j+1} do not have any effect on the magnitude of γ_j^*, a theorem can be generated based on only these three terms of $\boldsymbol{\alpha}$.

Theorem 2.3.1 *When $p_{ij} = 0$ if $|i - j| \geq 2$ under the monotonicity assumption of p_{ij}:*

- *If $\boldsymbol{\gamma}$ and $(\alpha_{j-1}, \alpha_j, \alpha_{j+1})$ are both monotone with the same direction of trend, $\gamma_j^{*2} \geq \gamma_j^{*1}$.*
- *If $\boldsymbol{\gamma}$ and $(\alpha_{j-1}, \alpha_j, \alpha_{j+1})$ are both monotone but with different directions of trend, $\gamma_j^{*2} \leq \gamma_j^{*1}$.*

Only when $(\alpha_{j-1}, \alpha_j, \alpha_{j+1})$ is monotone, we can determine the relationship between γ_j^{*1} and γ_j^{*2}. It allows us to compare the effect of misclassification between the cases when exposure is uniformly distributed and non-uniformly distributed. However, when $(\alpha_{j-1}, \alpha_j, \alpha_{j+1})$ is not monotone, the relationship between γ_j^{*1} and γ_j^{*2} will be hard to predict.

2.4 Test for Trend Across Categories

When exposure X is a monotone continuous variable and the mean actual outcome $E(Y \mid X)$ has a linear relationship with X, the slope of $E(Y \mid X)$

will always be steeper than the slope of the mean apparent outcome $E(Y \mid X^*)$. Moreover, the relationship between the intercepts of $E(Y \mid X)$ and the intercepts of $E(Y \mid X^*)$ is influenced by the direction of the trend. When $E(Y \mid X)$ is monotone increasing, the starting value of $E(Y \mid X^*)$ will always higher than that of $E(Y \mid X)$, and vice versa. What if X is a polychotomous variable instead of a continuous variable? Let's have a look at the simplest case where misclassification is least severe under the monotonicity assumption of p_{ij} and $\boldsymbol{\gamma}$ is monotone.

In Theorem 2.1.1, we have proved that monotone $\boldsymbol{\gamma}$ will always lead to monotone $\boldsymbol{\gamma}^*$ with the same direction of trend. Therefore, misclassification does not influence the overall shape of the apparent mean outcome for polychotomous exposure, which is the same as that for continuous exposure. Besides that, we can generate a theorem on the values of the mean outcomes at the starting and ending levels.

Theorem 2.4.1 *Under the situation where $p_{ij} = 0$ if $|i - j| \geq 2$ and for any value of $\boldsymbol{\alpha}$:*
When $\boldsymbol{\gamma}$ is monotone increasing,

- *γ_1^* is larger than γ_1,*
- *γ_k^* is smaller than γ_k.*

When $\boldsymbol{\gamma}$ is monotone decreasing, the contrary is the case.

Proof:

(i): When $\boldsymbol{\gamma}$ is monotone increasing, γ_1 is the smallest element among γ_i, for any i from 1 to k. The relationship between γ_1^* and γ_1 can be expressed as:

$$\begin{aligned}
\gamma_1^* - \gamma_1 &= \frac{\sum_{i=1}^{k} \alpha_i p_{i1} \gamma_i}{\sum_{i=1}^{k} \alpha_i p_{i1}} - \gamma_1 \\
&\propto \sum_{i=1}^{k} \alpha_i p_{i1} \gamma_i - \sum_{i=1}^{k} \alpha_i p_{i1} \gamma_1 \\
&> 0.
\end{aligned}$$

As it can be proved that $\gamma_1^* - \gamma_1 > 0$, γ_1^* is larger than γ_1 for sure.

(ii): When there are k exposure levels and $\boldsymbol{\gamma}$ is monotone increasing, γ_k is the largest element among γ_i, for any i from 1 to k. The relationship

between γ_k^* and γ_k can be expressed as:

$$\begin{aligned}\gamma_k^* - \gamma_k &= \frac{\sum_{i=1}^{k} \alpha_i p_{ik} \gamma_i}{\sum_{i=1}^{k} \alpha_i p_{ik}} - \gamma_k \\ &\propto \sum_{i=1}^{k} \alpha_i p_{ik} \gamma_i - \sum_{i=1}^{k} \alpha_i p_{ik} \gamma_k \\ &< 0.\end{aligned}$$

As it can be proved that $\gamma_k^* - \gamma_k < 0$, γ_k^* is smaller than γ_k for sure.

Similarly, when $\boldsymbol{\gamma}$ is monotone decreasing, using the same procedure, it is easy to prove that $\gamma_1^* - \gamma_1 < 0$ and $\gamma_k^* - \gamma_k > 0$. ∎

In summary, from Theorem 2.4.1, we can state that the smallest mean outcome for apparent exposure will always be larger than that for actual exposure; and the largest mean outcome will always be smaller. Therefore, misclassification lead to the same conclusion on the smallest and the largest values of mean outcomes for both continuous exposure and polychotomous exposure.

From both Theorem 2.1.1 and Theorem 2.4.1, we can summarize that the largest changes between the mean outcomes of any two levels for apparent exposure will always be smaller than that for actual exposure. However, it is possible that the change of $\boldsymbol{\gamma}^*$ at some adjacent levels is larger than that of $\boldsymbol{\gamma}$ for polychotomous exposure, which will not happen for continuous exposure. Therefore, we still cannot conclude that misclassification attenuate the exposure-disease association for polychotomous exposure.

Example 2.4.1 *Let's consider the situation when $k = 6$, X is uniformly distributed, $\boldsymbol{\gamma} = (1, 2, 50, 51, 100, 110)$ which is monotone increasing, and*

$$P = \begin{pmatrix} 0.9 & 0.1 & 0 & 0 & 0 & 0 \\ 0.1 & 0.8 & 0.2 & 0 & 0 & 0 \\ 0 & 0.2 & 0.5 & 0.3 & 0 & 0 \\ 0 & 0 & 0.2 & 0.5 & 0.3 & 0 \\ 0 & 0 & 0 & 0.2 & 0.7 & 0.1 \\ 0 & 0 & 0 & 0 & 0.1 & 0.9 \end{pmatrix}.$$

Both $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^*$ are displayed in Figure 2.7. It is obvious that the change between γ_3^* and γ_4^* is larger than the change between γ_3 and γ_4, although the change of apparent mean outcomes between any other adjacent levels is smaller than that of actual apparent mean outcomes. ∎

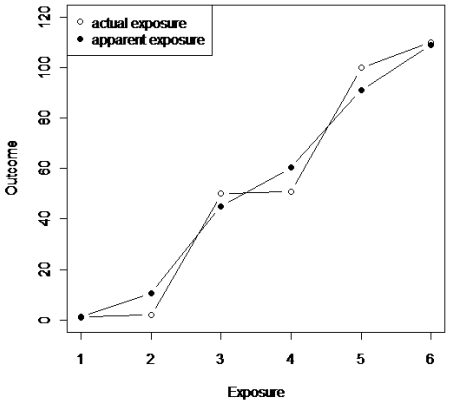

Figure 2.7: $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^*$ when X is uniformly distributed and $\boldsymbol{\gamma}$ is increasing.

To investigate the effect of misclassification for polychotomous exposure, we will check how the overall trend of mean outcomes changes from $\boldsymbol{\gamma}$ to $\boldsymbol{\gamma}^*$ by a trend test. We will first introduce the trend test for actual exposure, and then for apparent exposure. Under both parts, we only consider the simplest situation where X is uniformly distributed.

2.4.1 Trend Test for Actual Exposure

Treating all subjects with the same exposure level as a group, we can consider the group mean outcomes as the target responses. Assume that each group has the same expected sample size, denoted as n. Also, we make an equal variance assumption such that the variance of the response for each subject is the same, denoted as σ^2. Moreover, assume that these subjects are all independent from each other.

Denote variable $\bar{Y}$ to be the vector of sample mean outcomes for the k groups. Assume that $\bar{Y}$ follows a normal distribution, which can be written as:

$$\bar{Y} \sim N(\boldsymbol{\beta}, \Sigma),$$

where $\boldsymbol{\beta}$ is the true mean outcome and Σ is the covariance matrix. As for group mean outcomes, the exposure level X is the group index, the design matrix will be an identity matrix I_k. Under the assumptions stated above, the covariance matrix can be expressed as $\Sigma = n^{-1}\sigma^2 I_k$.

For actual exposure X, the sample mean outcomes can be represented as $\hat{\boldsymbol{\beta}} = \boldsymbol{\gamma}$. Tests for linear trend across categories are best performed using a linear contrast in the coefficients corresponding to the various levels of X, say $\boldsymbol{c}'\boldsymbol{\beta}$. With the null hypothesis $H_0 : \boldsymbol{c}'\boldsymbol{\beta} = 0$, we can test if X has a non-zero trend, and then have a concise view about the slope of overall mean outcomes. Given there are k exposure levels, the coefficient $\boldsymbol{c}$ can be calculated using the least squares approach. In this way, $\boldsymbol{c}'\boldsymbol{\beta}$ is the slope of the least squares line of the group mean outcomes on the group index, i.e., $\boldsymbol{\beta}$ on $(1, \ldots, k)$. Fixed number of categories leads a same value of $\boldsymbol{c}$.

According to Ordinary Least Squares method,

$$\boldsymbol{c}'\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{c}'\boldsymbol{\beta}, \boldsymbol{c}'\Sigma\boldsymbol{c}).$$

The test statistic of a one-sample t-test can be written as:

$$T = \frac{\boldsymbol{c}'\hat{\boldsymbol{\beta}}}{\sqrt{\boldsymbol{c}'\Sigma\boldsymbol{c}}}.$$

Note that the power of a test for significance level 0.05 is the probability that the test will reject the null hypothesis when the null hypothesis is false. For this one-sample t-test, the power calculation can be expressed as:

$$Pr(|T| > 1.96) = \Phi\left(-1.96 + \frac{|\Delta|\sqrt{n_t}}{\sigma_t}\right),$$

where

- n_t is the total sample size (here is k since we use the group index as the observation);
- $|\Delta|$ is the expected mean difference ($\boldsymbol{c}'\boldsymbol{\beta}$);
- σ_t is the standard deviation ($\sqrt{\boldsymbol{c}'\Sigma\boldsymbol{c}}$).

Given the number of exposure levels to be k, the power calculation will only depend on two terms: the expected mean difference and the standard deviation. The larger $|\Delta|/\sigma_t$ is, the higher the power will be.

2.4.2 Trend Test for Apparent Exposure

Although the apparent exposure level of a certain subject might differ from the actual exposure level, the number of exposure levels will always be the same for both actual exposure and apparent exposure. Therefore, if treating

X^* as the group index, the design matrix for X^* is still an identity matrix I_k, which is the same as for X. However, the variance of each group is influenced by misclassification. Denote the variance of the i-th apparent exposure level as σ_i^2. Because of misclassification, σ_i^2 will not be equal for each group, which differ from that for actual exposure. In general, σ_i^2 can be written as:

$$\begin{aligned}\sigma_i^2 &= Var(Y \mid X^* = i) \\ &= E[Var(Y \mid X) \mid X^* = i] + Var[E(Y \mid X) \mid X^* = i] \\ &= \sigma^2 + Var\{\beta_1 I(X = 1) + \cdots + \beta_k I(X = k) \mid X^* = i\}.\end{aligned}$$

Given a subject is apparently classified into the i-th exposure level, the actual exposure level of this subject can be one of a fixed number of levels with fixed probabilities. Therefore, it is reasonable to treat that $(I(X = 1), \cdots, I(X = k) \mid X^* = i)$ follows a multinomial distribution. Denote P^* to be the transition matrix, where $p_{ij}^* = Pr(X = j \mid X^* = i)$. Each element of this transition matrix can be calculated based on the classification matrix P such that $p_{ij}^* = p_{ji}/\sum_{j=1}^k p_{ji}$. From the properties of the multinomial distribution, we can state that:

$$Var(I(X = i) \mid X^* = h) = p_{hi}^*(1 - p_{hi}^*), \tag{2.2}$$

$$Cov(I(X = i), I(X = j) \mid X^* = h) = -p_{hi}^* p_{hj}^*. \tag{2.3}$$

Based on the expression in (2.2) and (2.3), we can express σ_i^2 in terms of σ^2 and p_{ij}^*:

- when $i = 1$,

$$\begin{aligned}\sigma_1^2 &= \sigma^2 + Var(\beta_1 I(X = 1) + \beta_2 I(X = 2) \mid X^* = 1) \\ &= \sigma^2 + \beta_1^2 p_{11}^*(1 - p_{11}^*) + \beta_2^2 p_{12}^*(1 - p_{12}^*) - 2\beta_1\beta_2 p_{11}^* p_{12}^*;\end{aligned}$$

- when $1 < i < k$,

$$\begin{aligned}\sigma_i^2 &= \sigma^2 + Var(\beta_{i-1} I(X = i - 1) + \beta_i I(X = i) \\ &\quad + \beta_{i+1} I(X = i + 1) \mid X^* = i) \\ &= \sigma^2 + \beta_{i-1}^2 p_{i,i-1}^*(1 - p_{i,i-1}^*) + \beta_i^2 p_{ii}^*(1 - p_{ii}^*) + \beta_{i+1}^2 p_{i,i+1}^*(1 - p_{i,i+1}^*) \\ &\quad - 2\beta_{i-1}\beta_i p_{i,i-1}^* p_{ii}^* - 2\beta_{i-1}\beta_{i+1} p_{i,i-1}^* p_{i,i+1}^* - 2\beta_i\beta_{i+1} p_{ii}^* p_{i,i+1}^*;\end{aligned}$$

- when $i = k$,

$$\begin{aligned}\sigma_k^2 &= \sigma^2 + Var(\beta_{k-1} I(X = k - 1) + \beta_k I(X = k) \mid X^* = k) \\ &= \sigma^2 + \beta_{k-1}^2 p_{k,k-1}^*(1 - p_{k,k-1}^*) + \beta_k^2 p_{kk}^*(1 - p_{kk}^*) \\ &\quad - 2\beta_{k-1}\beta_k p_{k,k-1}^* p_{kk}^*.\end{aligned}$$

From the expression above, it is obvious that $\sigma_i^2 \geq \sigma^2$ for any i. In conclusion, the group variance for apparent exposure will not be equal and always be larger than that for actual exposure.

Additionally, the sample size of each group also changes because of misclassification. Each group does not have the same number of subjects for apparent exposure. Let n_i denote the expected sample size of the i-th apparent group. Given the value of P, n_i can be calculated as:

- when $i = 1$, $n_1 = (p_{11} + p_{21})n$;
- when $1 < i < k$, $n_i = (p_{i-1,i} + p_{ii} + p_{i+1,i})n$;
- when $i = k$, $n_k = (p_{k-1,k} + p_{kk})n$.

Let Σ^* be the covariance matrix for apparent exposure. Then, Σ^* can be written based on both σ_i^2 and n_i:

$$\Sigma^* = \begin{pmatrix} \frac{\sigma_1^2}{n_1} & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \frac{\sigma_2^2}{n_2} & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & \frac{\sigma_3^2}{n_3} & \cdots & 0 & 0 & 0 \\ & & & \cdots & & & \\ 0 & 0 & 0 & \cdots & \frac{\sigma_{k-2}^2}{n_{k-2}} & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & \frac{\sigma_{k-1}^2}{n_{k-1}} & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & \frac{\sigma_k^2}{n_k} \end{pmatrix}.$$

For apparent exposure X^*, the sample mean outcomes can be represented as $\hat{\boldsymbol{\beta}} = \boldsymbol{\gamma}^*$. The form of the test statistic and the power calculation are exactly the same as for actual exposure, but with Σ replaced by Σ^*.

2.4.3 Power of the Trend Test

To investigate the effect of misclassification, one possible way is to compare the power of the trend test between X and X^*. Larger power of the test represents stronger exposure-disease association. As misclassification does not affect the number of exposure levels, given the total sample size n_t unchanged, power calculation only depend on two terms,

$$\frac{|\Delta|}{\sigma_t} = \frac{\boldsymbol{c}'\boldsymbol{\beta}}{\sqrt{\boldsymbol{c}'\Sigma\boldsymbol{c}}}. \tag{2.4}$$

As shown in Example 2.4.1, even when $\boldsymbol{\gamma}$ is monotonically increasing, the change between γ_i^* and γ_{i+1}^* is not always smaller than that between γ_i and

γ_{i+1}. As the result, we can not guarantee that for the numerators of (2.4), $\boldsymbol{c}'\boldsymbol{\gamma} > \boldsymbol{c}'\boldsymbol{\gamma}^*$ is always the case. As a result, it is possible that misclassification increases the power of the trend test. To compare the powers between actual exposure and apparent exposure, we can simply calculate the ratio of $|\Delta|/\sigma_t$ between actual exposure and apparent exposure, and compare it with 1. The ratio of $|\Delta|/\sigma_t$ can be expressed as:

$$\frac{\boldsymbol{c}'\boldsymbol{\gamma}/\sqrt{\boldsymbol{c}'\Sigma\boldsymbol{c}}}{\boldsymbol{c}'\boldsymbol{\gamma}^*/\sqrt{\boldsymbol{c}'\Sigma^*\boldsymbol{c}}}. \tag{2.5}$$

The numerator of (2.5) is part of the power calculation for actual exposure and the denominator of (2.5) is for apparent exposure. If $(\boldsymbol{c}'\boldsymbol{\gamma}/\sqrt{\boldsymbol{c}'\Sigma\boldsymbol{c}})/(\boldsymbol{c}'\boldsymbol{\gamma}^*/\sqrt{\boldsymbol{c}'\Sigma^*\boldsymbol{c}}) > 1$, misclassification decreases the power of the trend test, and vice versa.

We will first give an example in which $\boldsymbol{c}'\boldsymbol{\gamma} < \boldsymbol{c}'\boldsymbol{\gamma}^*$ is the case but misclassification decreases the power of the trend test in Example 2.4.2.

Example 2.4.2 *Let's consider a case when $k = 6$. By least squares approach, the coefficient of linear contrast $\boldsymbol{c}$ can be calculated as:*

$$\boldsymbol{c}' = (-5, -3, -1, 1, 3, 5).$$

Then, the null hypothesis can be written as:

$$H_0 : -5\beta_1 - 3\beta_2 - \beta_3 + \beta_4 + 3\beta_5 + 5\beta_6 = 0.$$

Take X is uniformly distributed, $\sigma = 2$, $\boldsymbol{\gamma} = (4.5, 4.6, 5.0, 5.1, 5.5, 5.6)$, and

$$P = \begin{pmatrix} 0.6 & 0.4 & 0 & 0 & 0 & 0 \\ 0.1 & 0.5 & 0.4 & 0 & 0 & 0 \\ 0 & 0.1 & 0.5 & 0.4 & 0 & 0 \\ 0 & 0 & 0.4 & 0.5 & 0.1 & 0 \\ 0 & 0 & 0 & 0.4 & 0.5 & 0.1 \\ 0 & 0 & 0 & 0 & 0.4 & 0.6 \end{pmatrix}$$

as an example. Both γ and γ^* are displayed in Figure 2.8. Under this setting, $\boldsymbol{c}'\boldsymbol{\gamma}^* = 8.34$ which is larger than $\boldsymbol{c}'\boldsymbol{\gamma} = 8.30$. Misclassification can increase the value of the linear contrast. However, because $\boldsymbol{c}'\Sigma\boldsymbol{c} < \boldsymbol{c}'\Sigma^*\boldsymbol{c}$, it is still possible that misclassification decreases the power of the trend test. Under this setting,

$$\frac{\boldsymbol{c}'\boldsymbol{\gamma}/\sqrt{\boldsymbol{c}'\Sigma\boldsymbol{c}}}{\boldsymbol{c}'\boldsymbol{\gamma}^*/\sqrt{\boldsymbol{c}'\Sigma^*\boldsymbol{c}}} = 1.1359.$$

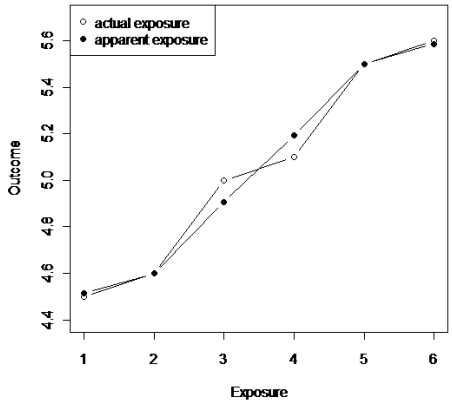

Figure 2.8: $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^*$ when X is uniformly distributed and $\boldsymbol{\gamma}$ is increasing.

As $(\boldsymbol{c}'\boldsymbol{\gamma}/\sqrt{\boldsymbol{c}'\Sigma\boldsymbol{c}})/(\boldsymbol{c}'\boldsymbol{\gamma}^*/\sqrt{\boldsymbol{c}'\Sigma^*\boldsymbol{c}}) > 1$*, the power of the trend test for actual exposure is larger than that for apparent exposure. It means that misclassification attenuates the exposure-disease association. In this example, misclassification has the same effect for polychotomous exposure as for continuous or binary exposure (i.e., when there are 6 subjects and the actual exposure levels of these subjects are all different, the power of the trend test for* γ *is 0.2877, and the power for* γ^* *is 0.2347 under significance level 0.05).* ∎

Beside the numerator of (2.4), the denominator of (2.4) for γ can also not be guaranteed smaller than that for γ^*. In conclusion, misclassification will not always attenuate the exposure-disease association. Here is an example where misclassification increases the power of the trend test in Example 2.4.3,

Example 2.4.3 *Let's consider the situation where all the other settings are the same as in Example 2.4.2, but with a different classification matrix:*

$$P = \begin{pmatrix} 0.9 & 0.1 & 0 & 0 & 0 & 0 \\ 0.25 & 0.7 & 0.05 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0.05 & 0.7 & 0.25 \\ 0 & 0 & 0 & 0 & 0.1 & 0.9 \end{pmatrix}.$$

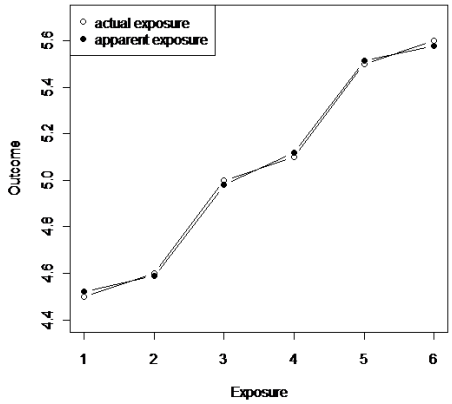

Figure 2.9: $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^*$ when X is uniformly distributed and $\boldsymbol{\gamma}$ is increasing.

Both γ and γ^* are displayed in Figure 2.9. Under this setting, although $\boldsymbol{c}'\boldsymbol{\gamma} > \boldsymbol{c}'\boldsymbol{\gamma}^*$, $\boldsymbol{c}'\Sigma\boldsymbol{c} > \boldsymbol{c}'\Sigma^*\boldsymbol{c}$ at the same time. It can be calculated that

$$\frac{\boldsymbol{c}'\boldsymbol{\gamma}/\sqrt{\boldsymbol{c}'\Sigma\boldsymbol{c}}}{\boldsymbol{c}'\boldsymbol{\gamma}^*/\sqrt{\boldsymbol{c}'\Sigma^*\boldsymbol{c}}} = 0.9977.$$

As $(\boldsymbol{c}'\boldsymbol{\gamma}/\sqrt{\boldsymbol{c}'\Sigma\boldsymbol{c}})/(\boldsymbol{c}'\boldsymbol{\gamma}^*/\sqrt{\boldsymbol{c}'\Sigma^*\boldsymbol{c}}) < 1$, misclassification increases the power of the trend test. The power of the trend test for γ^* will be larger than that for γ (i.e., when there are 6 subjects and the actual exposure levels of these subjects are all different, the power of the trend test for γ is 0.2877, and the power for γ^* is 0.2888 under significance level 0.05). Therefore, misclassification strengthen the exposure-disease association. Under this setting, misclassification does not have the same effect on polychotomous exposure as on continuous or binary exposure. ∎

2.5 Example: The Effect of Vitamin C on Tooth Growth in Guinea Pigs

To examine the effect of Vitamin C on tooth growth in guinea pigs, 60 guinea pigs were observed in a study. Each guinea pig was assigned one of three dose levels of Vitamin C (0.5, 1.0 and 2.0mg) and the length of odontoblasts

was recorded as the response to represent tooth growth condition. Define that guinea pigs applied 0.5mg Vitamin C were in group 1, 1.0mg in group 2, and 2.0mg in group 3, and there were 20 guinea pigs in each group. This data set (ToothGrowth) can be found in the "datasets" package in R.

Suppose that the original data set in R package represented the real situation, named as the actual data set. Then, the vector of group mean outcomes for true exposure can be calculated based on this actual data set and denoted as $\boldsymbol{\gamma}$. Assume that researchers measured the outcomes (the length of odontoblasts) accurately but did not always classify the guinea pigs into their true exposure levels (dose level of Vitamin C) during the study. As exposure misclassification happened, the recorded exposure levels of guinea pigs is the surrogate X^* but not the actual exposure X.

Assume that the actual exposure is uniformly distributed, and the exposure level cannot be misclassified more than one category away from the true exposure level. Let the classification matrix be:

$$P = \begin{pmatrix} 0.6 & 0.4 & 0 \\ 0.2 & 0.6 & 0.2 \\ 0 & 0.4 & 0.6 \end{pmatrix},$$

i.e., $p_{12} = 0.4$ means forty percents of guinea pigs in the first group which were applied dose level 0.5mg were wrongly recorded as being applied dose level 1.0mg. Using this classification matrix, we can create a new data set, named as the recorded data set. Assuming that researchers only have this recorded data set but not the actual data set, all the results are analyzed based on this recorded data set. If researchers did not adjust for misclassification, they treated the apparent exposure as the true exposure, which led to biased exposure-disease association. In the recorded data set, guinea pigs in the same group would not have the same levels of Vitamin C in reality. Even there were actually 20 guinea pigs received one of three dose levels of Vitamin C, the sample size of each group would not be equal in the recorded data set.

As the actual data set is unknown, the recorded data set is the only available information to analyze. Denote the vector of group mean outcomes of this recorded data set as $\boldsymbol{\gamma}^*$, which corresponding to the surrogate exposure. Based on the recorded data, the group sample sizes n_i and the covariance matrix Σ^* can also be calculated. Then, it is possible to investigate the effect of misclassification in this study.

Based on both the actual data set and the recorded data set, the vector of the actual group mean outcome can be calculated as
$\boldsymbol{\gamma} = (10.065, 19.735, 26.100)$, and the vector of the apparent group mean

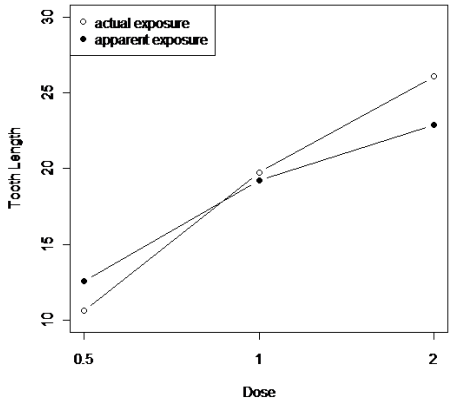

Figure 2.10: Relationship between tooth length and dose levels.

outcome can be calculated as $\boldsymbol{\gamma}^* = (12.585, 19.206, 22.900)$. Both $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^*$ are displayed in Figure 2.10.

This example obeys the conclusions in both Theorem 2.1.1 and 2.4.1. Since the exposure levels cannot be misclassified more than one category away from the true exposure levels and $\boldsymbol{\gamma}$ is monotonically increasing, $\boldsymbol{\gamma}^*$ is also monotonically increasing. The mean outcome of the guinea pigs which were applied 0.5mg Vitamin C for apparent exposure is larger that that for actual exposure, and the mean outcome of those which were applied 2.0mg Vitamin C for apparent exposure is smaller than that for actual exposure. To investigate the effect of misclassification for the overall mean outcomes, we will use a trend test and compare the power of the test between actual exposure and apparent exposure as in Section 2.4.

The number of exposure levels for both actual exposure and apparent exposure is three. Using the least squares approach, the corresponding coefficient of the linear contrast c can be calculated as:

$$\boldsymbol{c}' = (-1, 0, 1).$$

As discussed in Section 2.4.3, power calculation depends on only two terms: $|\Delta|$ and σ_t. Given both the actual data set and the recorded data set, all values of n, n_i, Σ, and Σ^* can be worked out. Then, we can calculate

the value of the ratio of $|\Delta|/\sigma_t$:

$$\frac{\boldsymbol{c}'\boldsymbol{\gamma}/\sqrt{\boldsymbol{c}'\Sigma\boldsymbol{c}}}{\boldsymbol{c}'\boldsymbol{\gamma}^*/\sqrt{\boldsymbol{c}'\Sigma^*\boldsymbol{c}}} = 3.56.$$

In this example, the ratio of $|\Delta|/\sigma_t$ between actual exposure and apparent exposure is 3.56, which is much larger than 1. The power of the trend test for γ will be larger than the power of the trend test for γ^*. It means that misclassification attenuate the exposure-disease association, which has the same effect as for continuous or binary exposure.

Chapter 3

Case-Control Study with "Maybe" Exposed Group

In normal case-control studies, there are always two categories for disease, presence and absence, and two categories for exposure, exposed and unexposed. Then, based on the collected data set, researchers can create a 2×2 table and use this 2×2 table to analyze the exposure-disease association. In this thesis, we only consider the non-differential exposure misclassification. It means that misclassification only happens during classifying exposure status of subjects. The outcomes status are always accurately classified. We will use a surrogate variable X^* to represent the apparent exposure status. When the exposure prevalence is low and the exposure status of a subject is not sure, researchers always classify this subject as unexposed. Let's have a look at what if keeping all this kind of subjects as a new group, but not merging them into the unexposed group. In this way, there are three exposure levels instead of two, 'likely exposed', 'maybe exposed', and 'unlikely exposed'.

3.1 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 has two categories, coded as zero for 'absence' and one for 'presence'. The actual exposure status X also has two categories, coded as zero for 'exposed' and one for 'unexposed'. However, 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.

The target parameters in the study are the prevalences of true exposure. Define r_0 and r_1 to be the prevalences of true exposure among controls and

cases such that:

$$\begin{aligned} r_0 &= Pr\{X = 1 \mid Y = 0\}, \\ r_1 &= Pr\{X = 1 \mid Y = 1\}. \end{aligned}$$

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

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

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

Since the apparent exposure has three categories, both $\boldsymbol{p_0}$ and $\boldsymbol{p_1}$ are three dimensional.

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

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

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

Both $\boldsymbol{\theta_0}$ and $\boldsymbol{\theta_1}$ are also three dimensional.

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 is denoted as $\mathbb{S}_3$. When useful, it is possible to 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 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) [18], 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 caes). 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 the values of $\boldsymbol{\theta}$ will be learnt 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 odds ratio $OR = \{r_1/(1-r_1)\}/\{r_0/(1-r_0)\}$.

3.1.1 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 the 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 this line is simply referred as 'the connecting line'. Without loss of generality, but for ease of exposition, I assume $\boldsymbol{\theta}_1$ lies south-east of $\boldsymbol{\theta}_0$ (as must arise if $r_0 < r_1$).

Then, the identification region for $\boldsymbol{p}$ can be expressed as:

$$\begin{aligned}\{\boldsymbol{p_0}, \boldsymbol{p_1} :& \theta_{00} < p_{00} < 1, 0 < p_{02} < \theta_{02}, \\ & 0 < p_{10} < \theta_{10}, \theta_{12} < p_{12} < 1, \\ & p_{00} = ap_{02} + b, p_{10} = ap_{12} + b\},\end{aligned}$$

where $a = (\theta_{00} - \theta_{10})/(\theta_{02} - \theta_{12})$ and $b = (\theta_{02}\theta_{10} - \theta_{12}\theta_{00})/(\theta_{02} - \theta_{12})$.

It means that $\boldsymbol{p}_0$ 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 $\mathbb{S}_3$).

To transfer the identification region to (r_0, r_1), we introduce two more parameters z_0 and z_1. Setting $z_0 = r_0/(r_1 - r_0)$ and $z_1 = (1 - r_1)/(r_1 - r_0)$, this geometry lends itself to simple algebraic description of the identification region in terms of $\boldsymbol{z} = (z_0, z_1)$, where

$$\boldsymbol{p_0} = \boldsymbol{\theta_0} + z_0(\boldsymbol{\theta_0} - \boldsymbol{\theta_1}), \tag{3.1}$$

$$\boldsymbol{p_1} = \boldsymbol{\theta_1} + z_1(\boldsymbol{\theta_1} - \boldsymbol{\theta_0}), \tag{3.2}$$

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

$$\bar{z}_0(\boldsymbol{\theta}) = \begin{cases} min(\frac{\theta_{02}}{\theta_{12}-\theta_{02}}, \frac{\theta_{01}}{\theta_{11}-\theta_{01}}) & \text{if } \theta_{11} - \theta_{01} > 0, \\ \frac{\theta_{02}}{\theta_{12}-\theta_{02}} & \text{if } \theta_{11} - \theta_{01} \leq 0. \end{cases} \tag{3.3}$$

$$\bar{z}_1(\boldsymbol{\theta}) = \begin{cases} min(\frac{\theta_{10}}{\theta_{00}-\theta_{10}}, \frac{\theta_{11}}{\theta_{01}-\theta_{11}}) & \text{if } \theta_{01} - \theta_{11} > 0, \\ \frac{\theta_{10}}{\theta_{00}-\theta_{10}} & \text{if } \theta_{01} - \theta_{11} \leq 0. \end{cases} \tag{3.4}$$

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, the identification region for (r_0, r_1) can be expressed as:

$$\begin{aligned}\mathbb{Q}_A(\boldsymbol{\theta}) = \Big\{(r_0, r_1) :& r_1 > \frac{\bar{z}_0(\boldsymbol{\theta}) + 1}{\bar{z}_0(\boldsymbol{\theta})} r_0, r_1 > \frac{\bar{z}_1(\boldsymbol{\theta})}{\bar{z}_1(\boldsymbol{\theta}) + 1} r_0 + \frac{1}{\bar{z}_1(\boldsymbol{\theta}) + 1}, \\ & 0 < r_0, r_1 < 1\Big\}.\end{aligned}$$

The situation described thus far is illustrated in the left panels of Figure 3.1, for the 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 $\boldsymbol{r} \in \mathbb{Q}_A(\boldsymbol{\theta})$.

3.1.2 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 $p_{10} < p_{11} < p_{12}$. The prior region is defined to be $\boldsymbol{p} \in \mathbb{P}_B$. The visual representation of $\mathbb{P}_B$ is given in the upper-right panel of Figure 3.1, 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 $\boldsymbol{p_1}$. 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 3.1, Thus 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. Compared with constraint A, there are four additional restrictions defining the identification region. The identification region for $\boldsymbol{p}$ can be expressed as:

$$\begin{aligned}\{\boldsymbol{p_0}, \boldsymbol{p_1} :& \theta_{00} < p_{00} < 1, 0 < p_{02} < \theta_{02}, \\ & 0 < p_{10} < \theta_{10}, \theta_{12} < p_{12} < 1, \\ & p_{00} = ap_{02} + b, p_{10} = ap_{12} + b, \\ & p_{00} > p_{01}, p_{01} > p_{02}, \\ & p_{10} < p_{11}, p_{11} < p_{12}\}.\end{aligned}$$

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

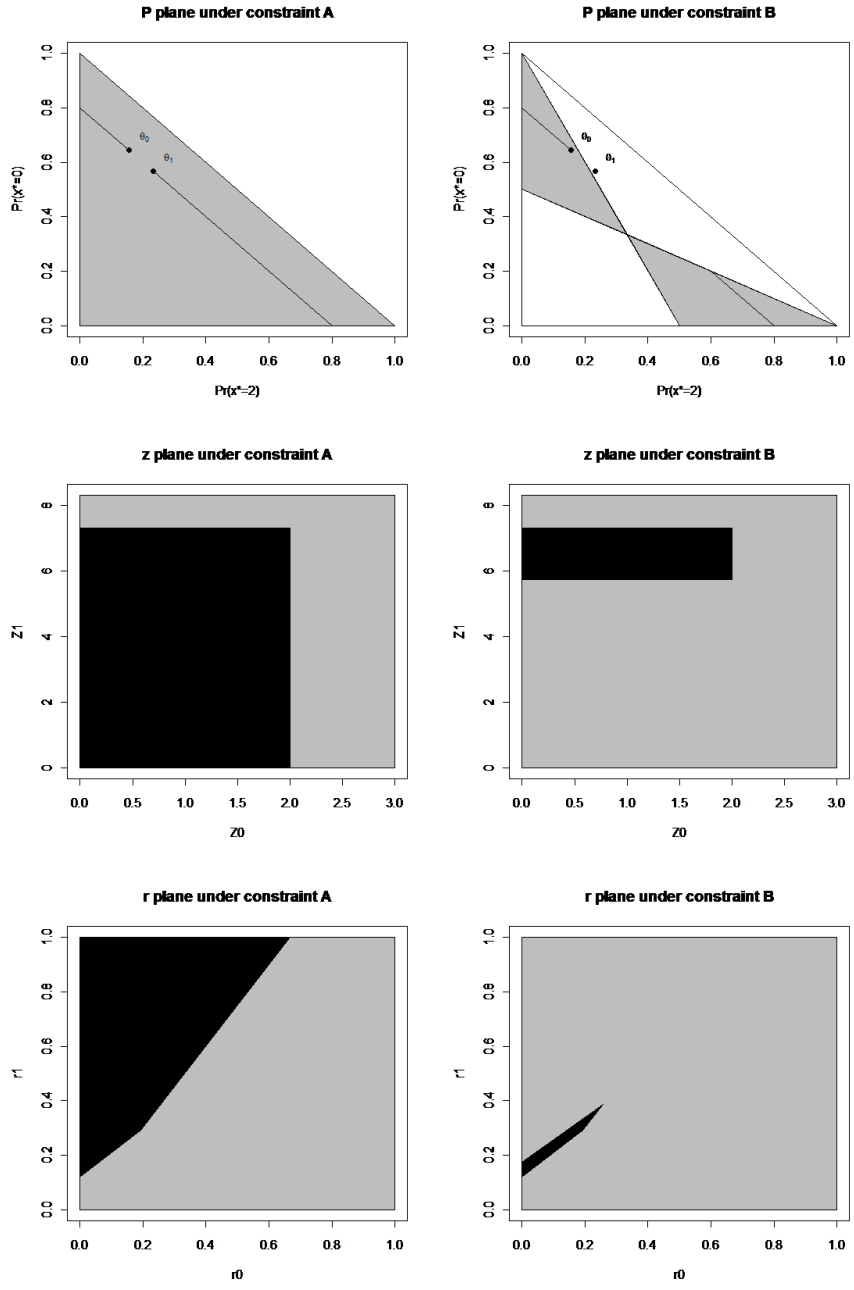

Figure 3.1: Prior and identification regions. The three plots on the left are under constraint A, and the three plots on the right are under constraint B. The shaded areas with gray are the prior regions, and the shaded areas with black are the identification regions.

respectively, and therefore be the same upper bounds (3.3) and (3.4) that apply under constraint A. Note that if $\boldsymbol{\theta}$ is compatible with constraint B, then $\bar{z}_0(\boldsymbol{\theta})$ and $\bar{z}_1(\boldsymbol{\theta})$ can only take values of $\theta_{02}/(\theta_{12}-\theta_{02})$ and $\theta_{10}/(\theta_{00}-\theta_{10})$. It is also clear from the geometric view (upper-right panel of Figure 3.1 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 $i = 0$ or 1, where

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

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 3.1. Formally, the identification region for (r_0, r_1) can be expressed as:

$$\begin{aligned} \mathbb{Q}_B(\boldsymbol{\theta}) = \Big\{ (r_0, r_1) : \quad & r_1 > \tfrac{\bar{z}_0(\boldsymbol{\theta})+1}{\bar{z}_0(\boldsymbol{\theta})} r_0, r_1 > \tfrac{\bar{z}_1(\boldsymbol{\theta})}{\bar{z}_1(\boldsymbol{\theta})+1} r_0 + \tfrac{1}{\bar{z}_1(\boldsymbol{\theta})+1}, \\ & r_1 < I(\underline{z}_0(\boldsymbol{\theta}) \neq 0) \tfrac{\underline{z}_0(\boldsymbol{\theta}))+1}{\underline{z}_0(\boldsymbol{\theta})} r_0, r_1 < \tfrac{\underline{z}_1(\boldsymbol{\theta})}{\underline{z}_1(\boldsymbol{\theta})+1} r_0 + \tfrac{1}{\underline{z}_1(\boldsymbol{\theta})+1}, \\ & 0 < r_0, r_1 < 1 \Big\}. \end{aligned}$$

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 northeast 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 $\mathbb{Q}_A(\boldsymbol{\theta})$.

3.1.3 Comparison Between Constraint A and B

Based on the identification regions for for $\boldsymbol{p}$, for $\boldsymbol{z}$, and for $\boldsymbol{r}$ described in Section 3.1.1 and 3.1.2, we can summarize some salient features of the identification regions under constraints A and B in the following theorems.

Theorem 3.1.1 *Say that $\boldsymbol{\theta}$ is compatible with constraint A. Assume without loss of generality that $\theta_{00} > \theta_{10}$ and $\theta_{02} < \theta_{12}$ (as must arise if $r_0 < r_1$). 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* $\boldsymbol{\theta_0}$ *is outside the prior region for* $\boldsymbol{p_0}$ and $\boldsymbol{\theta_1}$ *is outside the prior region for* $\boldsymbol{p_1}$*.*
(iv) Constraint A yields a lower bound on the odds ratio achieved at $r_0 = \bar{z}_0(\boldsymbol{\theta})/(\bar{z}_0(\boldsymbol{\theta}) + \bar{z}_0(\boldsymbol{\theta}) + 1)$ *and* $r_1 = (\bar{z}_0(\boldsymbol{\theta}) + 1)/(\bar{z}_0(\boldsymbol{\theta}) + \bar{z}_0(\boldsymbol{\theta}) + 1)$*. 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 connectging line. By inspection, it is obvious that the prior region under constraint A never precludes either possibility (upper-left panel of Figure 3.1). Both are precluded under constraint B, however, if and only if both $\boldsymbol{\theta_0}$ and $\boldsymbol{\theta_1}$ lie outside the respective components of $\mathbb{P}_B$ (upper-right panel of Figure 3.1).

(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 $\mathbb{Q}_B(\boldsymbol{\theta})$. ∎

3.1.4 Collapsing Exposure to Two Categories

As alluded to in Section 1.3, 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 exposed' and 'maybe exposed' categories could be merged. Then the binary apparent exposure would be

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

Given the setting

$$\begin{aligned} p_0^* &= Pr\{X^{**} = 2 \mid X = 0\} = p_{02}, \\ p_1^* &= Pr\{X^{**} = 1 \mid X = 1\} = p_{12}, \end{aligned}$$

the quality of classification can be described by specificity $1-p_0^*$ and sensitivity p_1^*. 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 I 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:

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

The identification region for this problem is determined by Gustafson (2001) [12]. However, we express the results in a form more amenable for comparison with the results in Sections 3.1.1 and 3.1.2. Assume without loss of generality that $\theta_0^* < \theta_1^*$ (as must arise if $r_0 < r_1$). As per (3.1) and (3.2), 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 $i = 0$ or 1, such that:

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

Compared with the value of $\boldsymbol{z}$ under constraint A and B, $\theta_0^* = \theta_{12}$ and $\theta_1^* = \theta_{12}$ in this collapsed case.

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 $\boldsymbol{r} \in \mathbb{Q}^*(\boldsymbol{\theta^*})$. The minimum odds ratio occurs when $r_0 = \theta_0^*$ and $r_1 = \theta_1^*$.

It is very easy to compare the effect of collapsing to the use of three categories under constraint A. The conclusions can be generated as theorems in Theorem 3.1.2.

Theorem 3.1.2 *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:*

- $\mathbb{Q}_A(\boldsymbol{\theta}) \subset \mathbb{Q}^*(\boldsymbol{\theta^*})$;
- $\mathbb{Q}^*(\boldsymbol{\theta^*})$ *cannot produce a finite upper bound on the odds ratio, which is the same as* $\mathbb{Q}_A(\boldsymbol{\theta})$;

- *the lower bound on the odds ratio of* $\mathbb{Q}^*(\boldsymbol{\theta^*})$ *cannot exceed that of* $\mathbb{Q}_A(\boldsymbol{\theta})$.

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 $i = 0$ or 1). By mapping the identification region for (z_0, z_1) to (r_0, r_1), the conclusion can be got directly 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 ratio for the collapsed case will always smaller or equal to the lower bound under constraint A. ∎

3.1.5 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 by both the values of r_0 and OR. The symbols '-' and '+' are used to label the first and second settings for each of these three factors. 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 $OR = 2.0$. The first setting for the classification probabilities is a 'symmetric' situation with $\boldsymbol{p}_0 = (0.750, 0.200, 0.050)$ and $\boldsymbol{p}_1 = (0.050, 0.200, 0.750)$. The second setting corresponds to exposure being hard to detect, with $\boldsymbol{p}_0 = (0.900, 0.075, 0.025)$ and $\boldsymbol{p}_1 = (0.200, 0.300, 0.500)$. In total, there are $2^3 = 8$ values of $\boldsymbol{\theta}$ arising from all combinations of these three factors. The identification regions for these eight scenarios are displayed in Figures 3.2 through 3.9.

From these 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 on 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 $\boldsymbol{p_0}$,

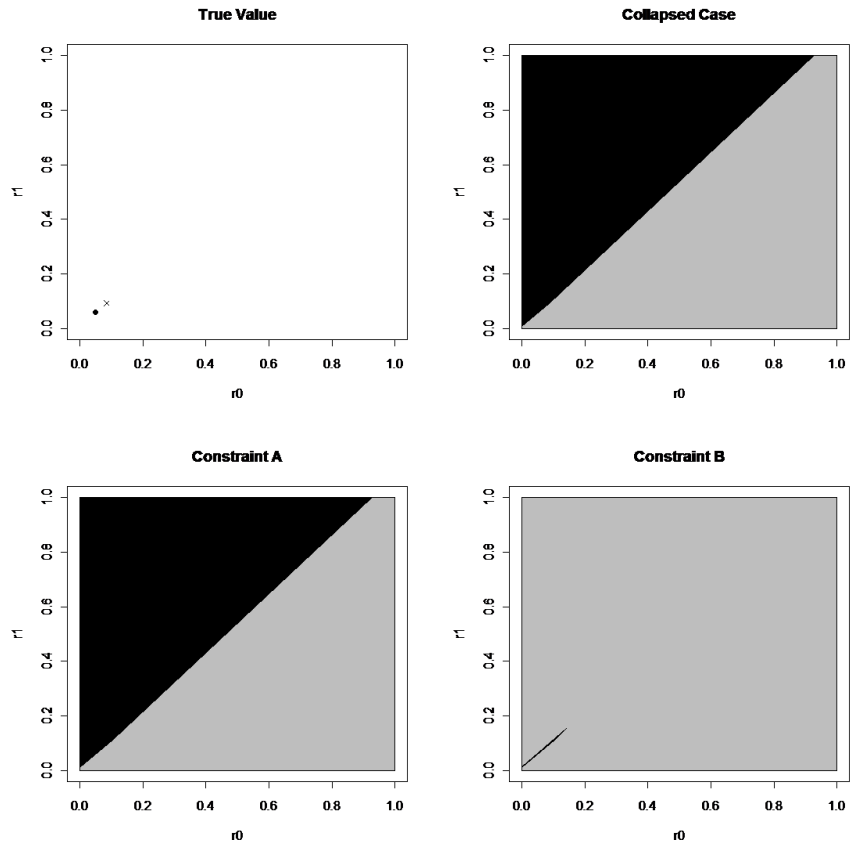

Figure 3.2: Identification regions for the combination $(---)$. Based on the given values of r_0 and OR, r_1 can be calculated to be 0.0594. In the upper-left panel, the dot indicates the true exposure prevalences (r_0, r_1), while the cross indicates apparent exposure prevalences upon collapsing to 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).

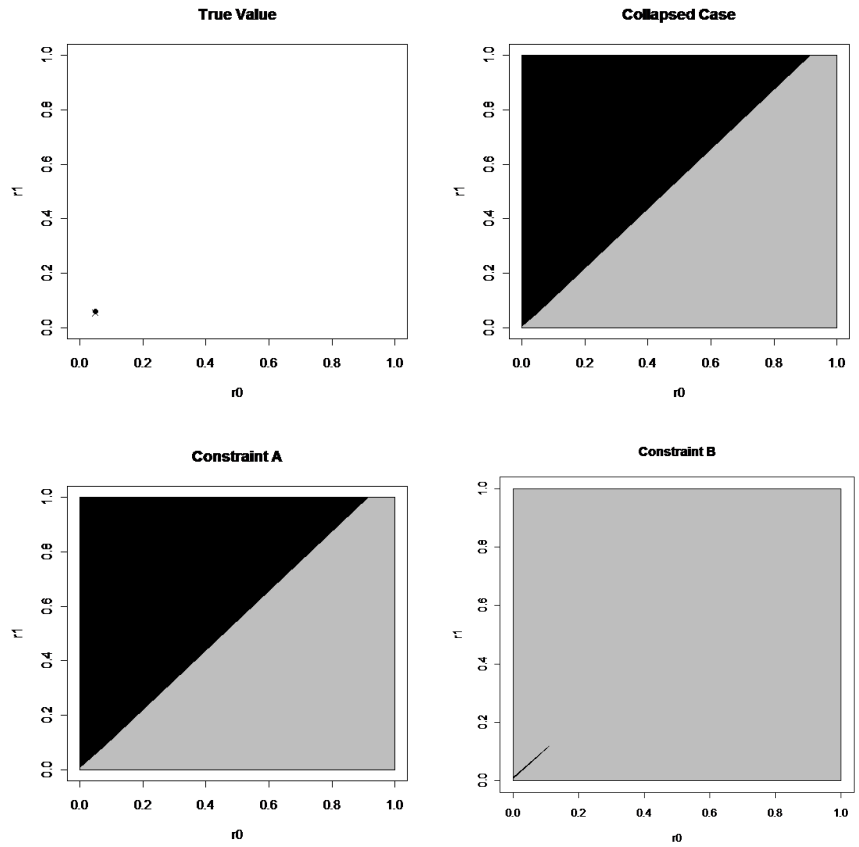

Figure 3.3: Identification regions for the combination $(--+)$. The layout is the same as Figure 3.2. Based on the given values of r_0 and OR, r_1 can be calculated to be 0.0594. In the collapsed case, θ_0^*=0.0488 and θ_1^*=0.0532. Under constraint A and B, $\boldsymbol{\theta_0}$=(0.865, 0.0863, 0.0488) and $\boldsymbol{\theta_1}$=(0.8584, 0.0884, 0.0532).

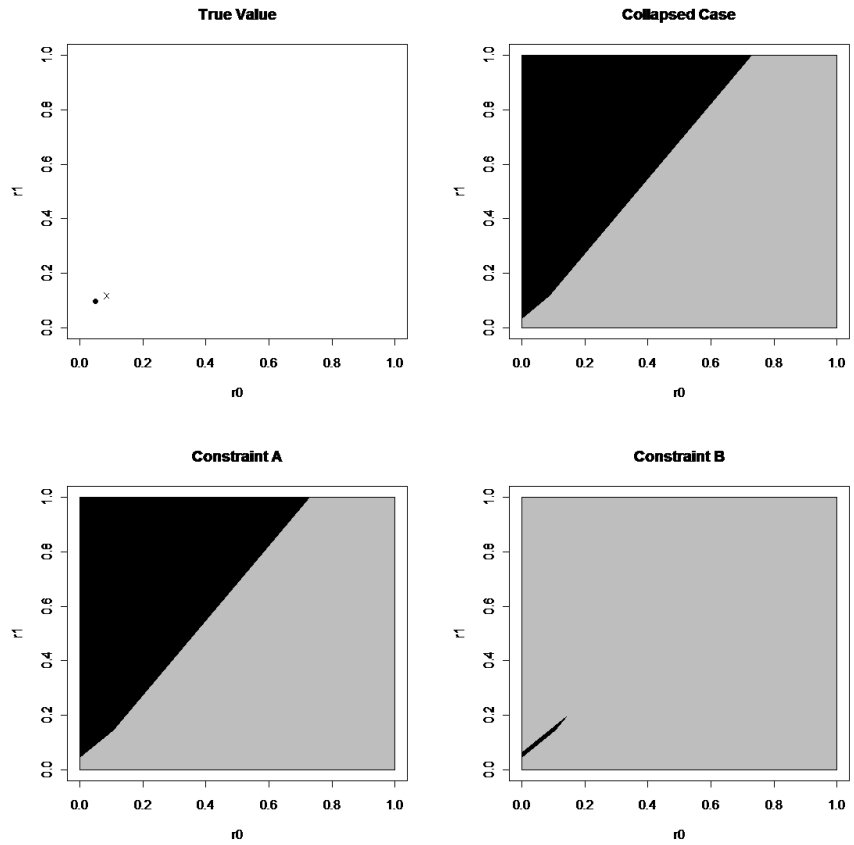

Figure 3.4: Identification regions for the combination $(-+-)$. The layout is the same as Figure 3.2. Based on the given values of r_0 and OR, r_1 can be calculated to be 0.0952. In the collapsed case, θ_0^*=0.0850 and θ_1^*=0.1167. Under constraint A and B, $\boldsymbol{\theta_0}$=(0.7150, 0.2000, 0.0850) and $\boldsymbol{\theta_1}$=(0.6833, 0.2000, 0.1167).

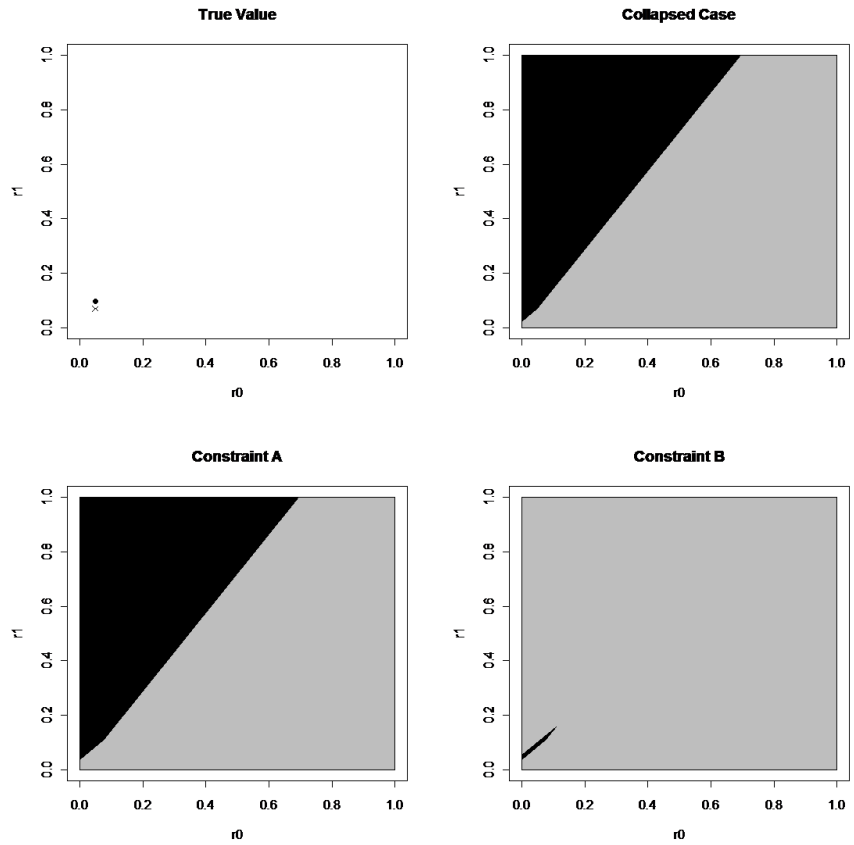

Figure 3.5: Identification regions for the combination $(- + +)$. The layout is the same as Figure 3.2. Based on the given values of r_0 and OR, r_1 can be calculated to be 0.0952. In the collapsed case, θ_0^*=0.0488 and θ_1^*=0.0702. Under constraint A and B, $\boldsymbol{\theta_0}$=(0.8650, 0.0863, 0.0488) and $\boldsymbol{\theta_1}$=(0.8333, 0.0964, 0.0702).

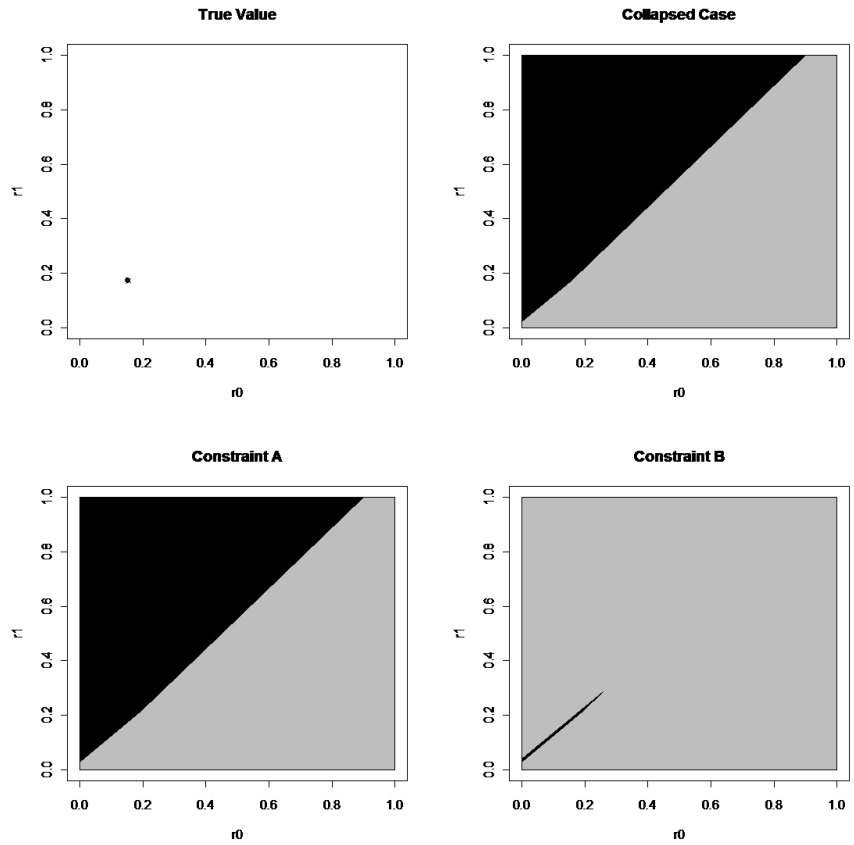

Figure 3.6: Identification regions for the combination $(+--)$. The layout is the same as Figure 3.2. Based on the given values of r_0 and OR, r_1 can be calculated to be 0.1748. In the collapsed case, θ_0^*=0.1550 and θ_1^*=0.1723. Under constraint A and B, $\boldsymbol{\theta_0}$=(0.6450, 0.2000, 0.1550) and $\boldsymbol{\theta_1}$=(0.6277, 0.2000, 0.1723).

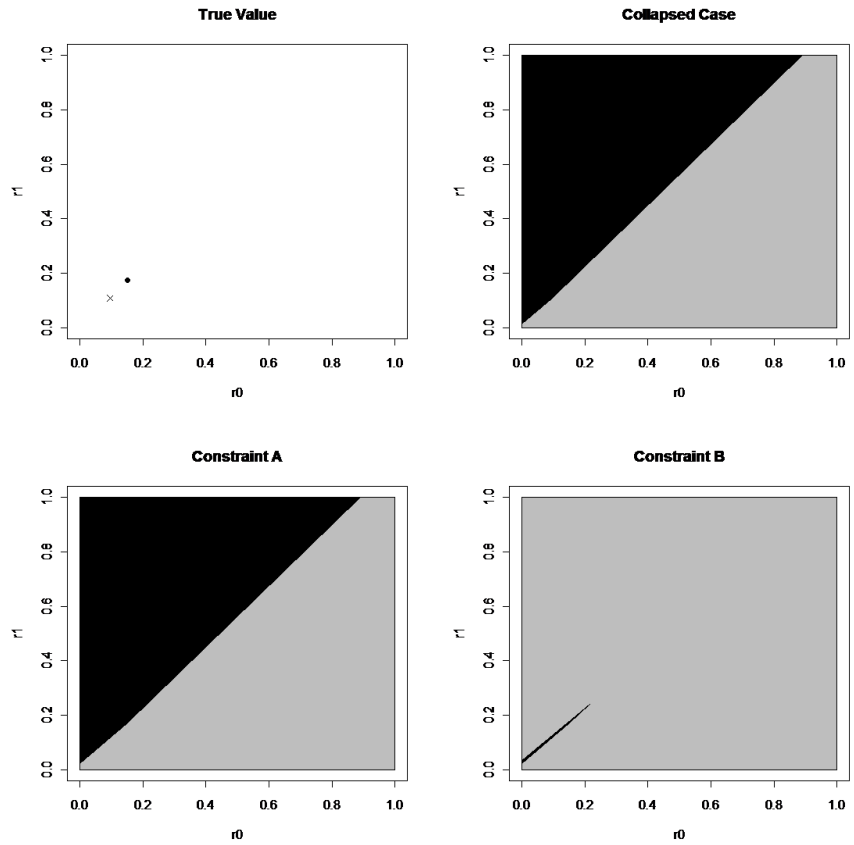

Figure 3.7: Identification regions for the combination $(+-+)$. The layout is the same as Figure 3.2. Based on the given values of r_0 and OR, r_1 can be calculated to be 0.1748. In the collapsed case, θ_0^*=0.0963 and θ_1^*=0.1080. Under constraint A and B, $\boldsymbol{\theta_0}$=(0.7950, 0.1088, 0.0963) and $\boldsymbol{\theta_1}$=(0.7777, 0.1143, 0.1080).

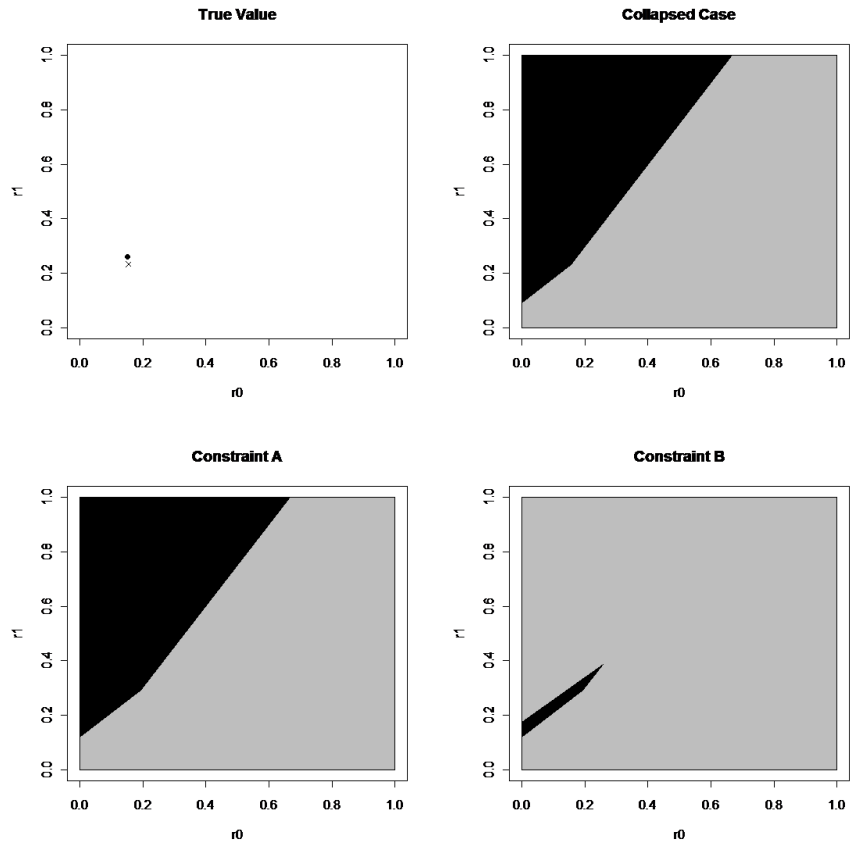

Figure 3.8: Identification regions for the combination $(++-)$. The layout is the same as Figure 3.2. Based on the given values of r_0 and OR, r_1 can be calculated to be 0.2609. In the collapsed case, θ_0^*=0.1550 and θ_1^*=0.2326. Under constraint A and B, $\boldsymbol{\theta_0}$=(0.6450, 0.2000, 0.1550) and $\boldsymbol{\theta_1}$=(0.5674, 0.2000, 0.2326).

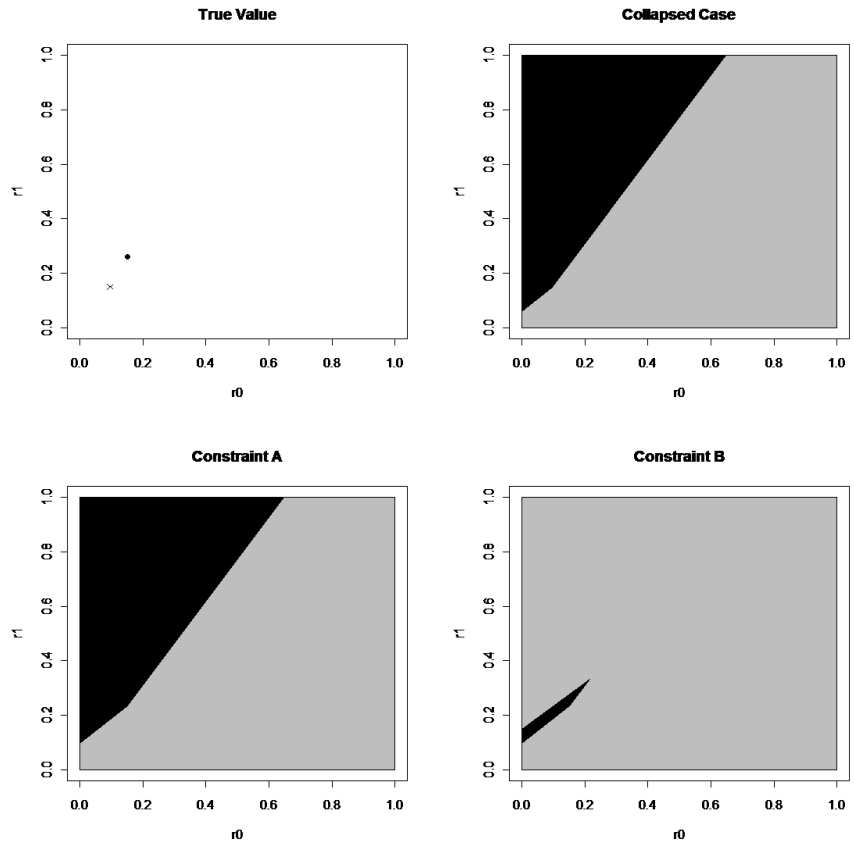

Figure 3.9: Identification regions for the combination of (+++). The layout is the same as Figure 3.2. Based on the given values of r_0 and OR, r_1 can be calculated to be 0.2609. In the collapsed case, θ_0^*=0.0963 and θ_1^*=0.1489. Under constraint A and B, $\boldsymbol{\theta_0}$=(0.7950, 0.1088, 0.0963) and $\boldsymbol{\theta_1}$=(0.7174, 0.1337, 0.1489).

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 on the odds ratio will not commonly arise. More specifically, for given classification probabilities $\boldsymbol{p}$ it is a simple matter to characterize how large r_0 must be (and how small r_1 must be) in order 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 $i = 0$ or 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 on the odds ratio in these eight scenarios are given in Table 3.1. As guaranteed by Theorem 3.1.1, the lower bound is always the same under both constraints A and B, but smaller in the collapsed case. In most scenarios, the lower bound for collapsed case 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 remembering 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 exposed' and 'maybe exposed' categories together and treating this is the unexposed category.

3.2 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 some 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.

Table 3.1: The lower bound on the odds ratio for collapsed case and for constraint A and B.

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

3.2.1 Principle

Suppose r_0, r_1, $\boldsymbol{p_0}$, and $\boldsymbol{p_1}$ are independent of each other a priori. Assume that r_0 and r_1 are both uniformly distributed such that $r_0 \sim U(0,1)$ and $r_1 \sim U(0,1)$. Also, assume that $\boldsymbol{p_0}$ and $\boldsymbol{p_1}$ follow Dirichlet distribution, written as $\boldsymbol{p_0} \sim \text{Dirichlet}(c_{00}, c_{01}, c_{02})$ and $\boldsymbol{p_1} \sim \text{Dirichlet}(c_{10}, c_{11}, c_{12})$, with the additional truncation of $\boldsymbol{p}_0$ and $\boldsymbol{p}_1$ to the assumed prior region $\mathbb{P}$. Under these assumptions, the joint prior density can be written as:

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

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 written 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, the limiting posterior density of the log odds ratio, say $s = \text{logit } r_1 - \text{logit } r_0$, can be expressed as:

$$f(s|\boldsymbol{\theta_0}, \boldsymbol{\theta_1}) \;=\; \int g(r_0; s) f(r_0, \text{expit}(s + \text{logit } r_0)|\boldsymbol{\theta_0}, \boldsymbol{\theta_1}) \, dr_0, \quad (3.5)$$

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 (3.5) 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 3.1), for given $(s, \boldsymbol{\theta})$ this could be either an interval of r_0 values or a pair of disjoint intervals. Particularly, 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 $\log OR$, since the support of the integrand in (3.5) 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 (3.5) using one-dimensional numerical integration.

3.2.2 Examples

The eight scenarios (values of $\boldsymbol{\theta}$) from Section 3.1.5 are revisited, in combination with two different settings of the prior distribution according to hyperparameters $\boldsymbol{c_0}$ and $\boldsymbol{c_1}$. 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. The second setting is $\boldsymbol{c_0^+} = (6, 4, 2)$, $\boldsymbol{c_1^+} = (2, 4, 6)$, which assigns more prior weight to better classifications (henceforth we refer to this setting as the 'weighted' prior). It is also possible to 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 in the collapsed case, $\boldsymbol{c_{0*}^-} = \boldsymbol{c_{1*}^-} = (1, 1)$ is taken as an instance of uniform priors. In light of the collapsibility property of Dirichlet distributions, the analogous 'weighted prior' setting when the 'maybe exposed' and 'unlikely exposed' categories are combined is $\boldsymbol{c_{0*}^+} = (10, 2)$, $\boldsymbol{c_{1*}^+} = (6, 6)$.

The limiting posterior distributions of $\log OR$ for these eight scenarios appear in Figures 3.10 through 3.17. 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 interval of $\log OR$ 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 interval of $\log OR$ are only very marginally bigger for the collapsed case analysis. Thus there is 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 starting 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.

3.3 Finite-Sample Posteriors

Until now, only the limiting behavior in the infinite sample size limit is under consideration. 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 $\log OR$ moves toward the limiting posterior distribution when the sample size increases, by simulating data under several of the previous scenarios.

The prior distributions are taken as $\boldsymbol{p}_0 \sim \mathit{Dirichlet}(6, 4, 2)$ and $\boldsymbol{p}_1 \sim \mathit{Dirichlet}(2, 4, 6)$, truncated with constraint A. We simulate five independent data sequences with equal numbers of controls and cases ($n_i = n$, for $i = 0$ or 1), and then determine the posterior distribution of $\log OR$ after $n = 100$, $n = 1000$, and $n = 5000$ observations, using WinBUGS [17]. We generically write $\boldsymbol{D_n}$ for the observed data. Posterior densities arising under the $(---)$ and $(+++)$ scenarios appear in Figures 3.18 and 3.19, with the limiting posterior densities also given.

In both scenarios, we see the sampling variation in the posterior distribution diminish with sample size. However, 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. Denote the posterior variance as

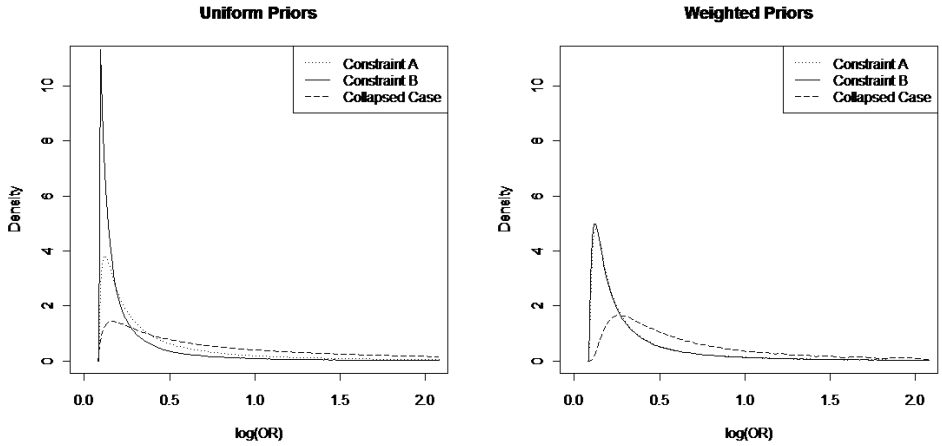

Figure 3.10: Limiting posterior distributions under the combination $(---)$. The left panel are the limiting posterior distributions of $\log OR$ for constraint A, constraint B, and collapsed case when $\boldsymbol{p_0}$ and $\boldsymbol{p_1}$ have noninformative uniform priors. The right panel are the limiting posterior distributions of $\log OR$ for constraint A, constraint B, and collapsed case when the prior distribution gives more weight to better classifications. In this scenario, the true $\log OR$ is 0.1823. The lower bound of $\log OR$ is 0.0839 under both constraint A and B, and 0.0818 under collapsed case.

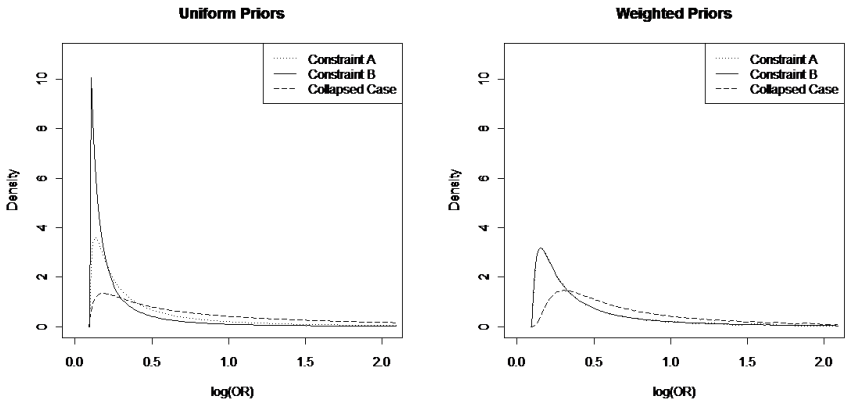

Figure 3.11: Limiting posterior distributions under the combination $(--+)$. The layout is the same as Figure 3.10. In this scenario, the true log odds ratio is 0.1823. The lower bound of $\log OR$ is 0.0953 under both constraint A and B, and 0.0934 under collapsed case.

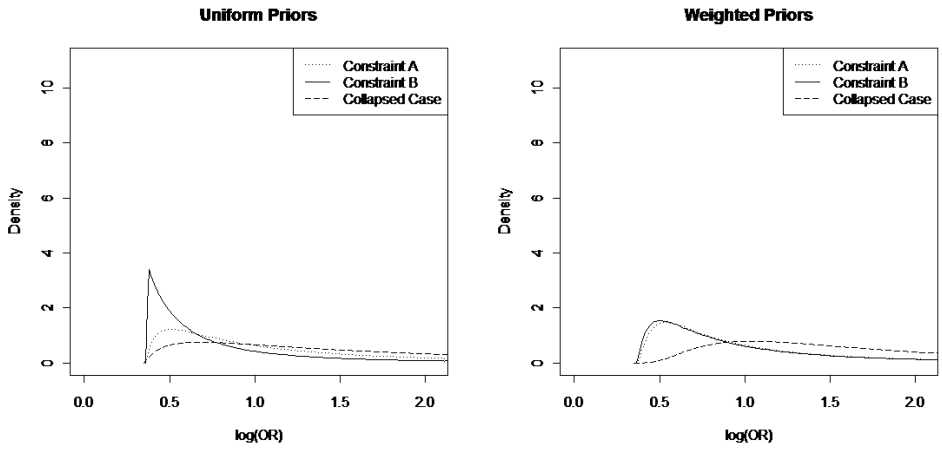

Figure 3.12: Limiting posterior distributions under the combination $(-+-)$. The layout is the same as Figure 3.10. In this scenario, the true log odds ratio is 0.6932. The lower bound of $\log OR$ is 0.3620 under both constraint A and B, and 0.3519 under collapsed case.

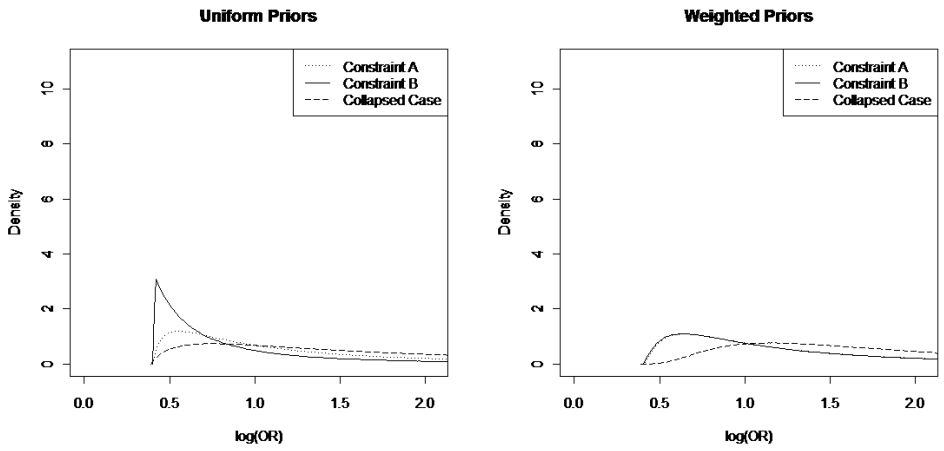

Figure 3.13: Limiting posterior distributions under the combination $(-++)$. The layout is the same as Figure 3.10. In this scenario, the true log odds ratio is 0.6932. The lower bound of $\log OR$ is 0.4025 under both constraint A and B, and 0.3880 under collapsed case.

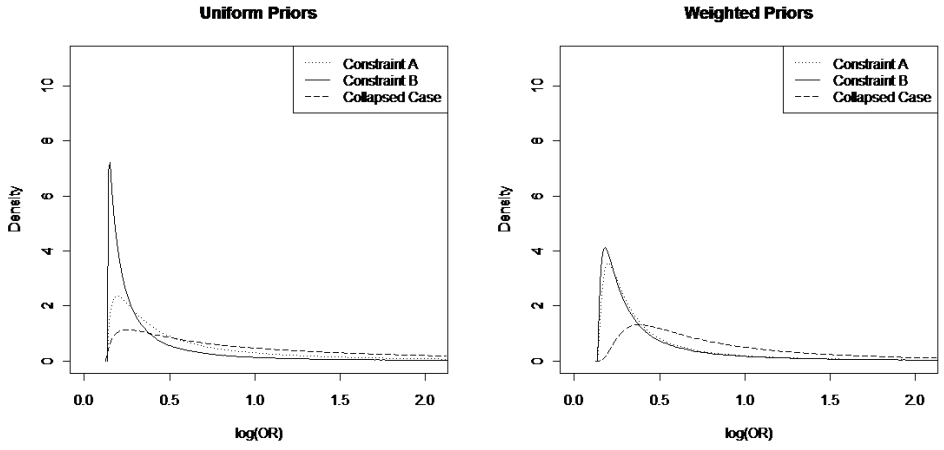

Figure 3.14: Limiting posterior distributions under the combination $(+--)$. The layout is the same as Figure 3.10. In this scenario, the true log odds ratio is 0.1823. The lower bound of $\log OR$ is 0.1332 under both constraint A and B, and 0.1267 under collapsed case.

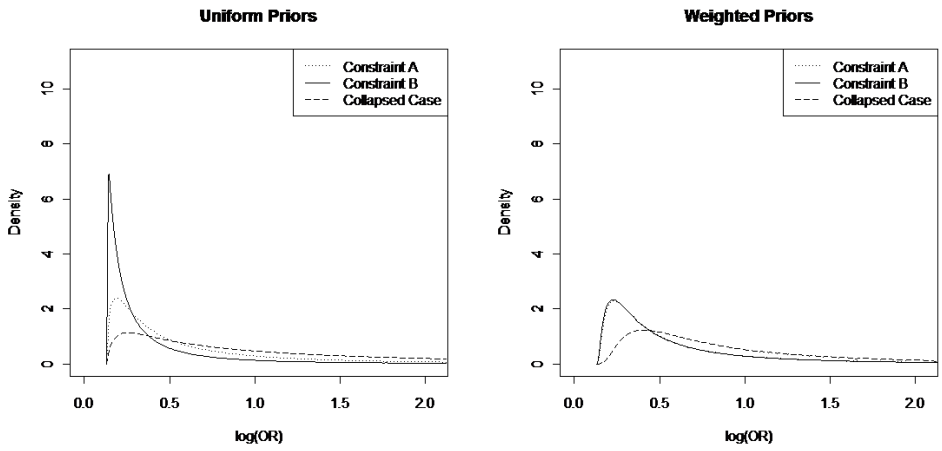

Figure 3.15: Limiting posterior distributions under the combination $(+-+)$. The layout is the same as Figure 3.10. In this scenario, the true log odds ratio is 0.1823. The lower bound of $\log OR$ is 0.1373 under both constraint A and B, and 0.1284 under collapsed case.

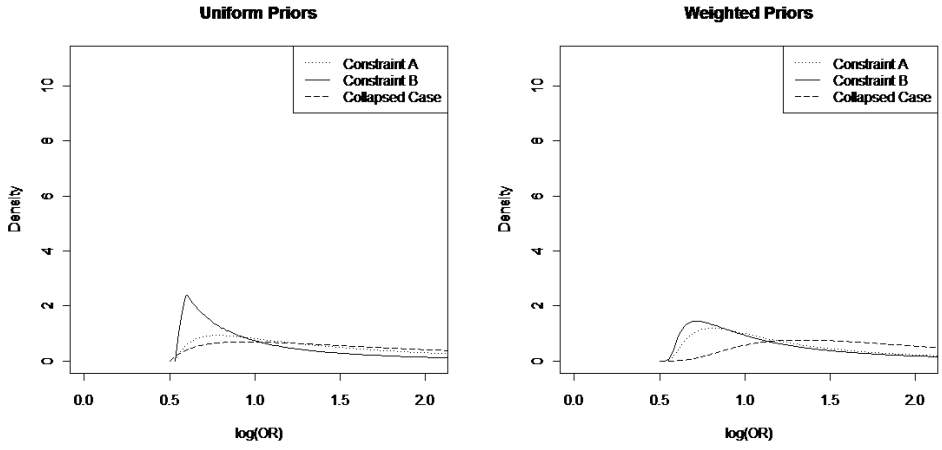

Figure 3.16: Limiting posterior distributions under the combination $(++-)$. The layout is the same as Figure 3.10. In this scenario, the true log odds ratio is 0.6932. The lower bound of $\log OR$ is 0.5341 under both constraint A and B, and 0.5023 under collapsed case.

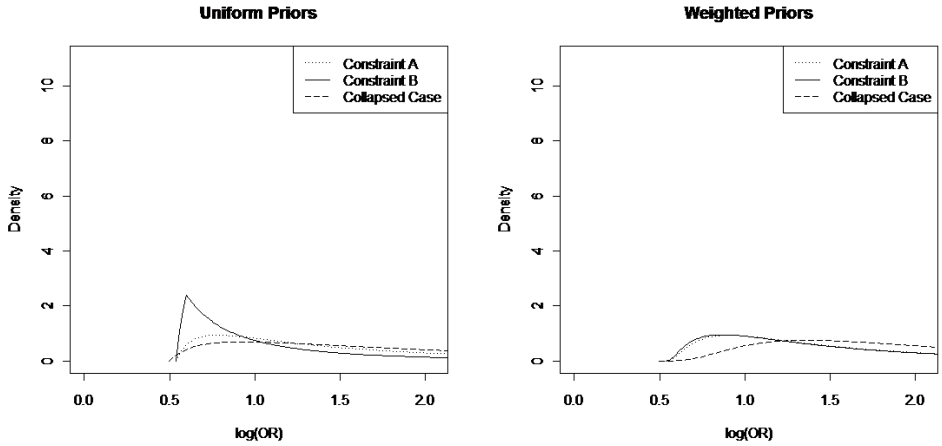

Figure 3.17: Limiting posterior distributions under the combination (+++). The layout is the same as Figure 3.10. In this scenario, the true log odds ratio is 0.6932. The lower bound of $\log OR$ is 0.5391 under both constraint A and B, and 0.4965 under collapsed case.

$\mathrm{Var}\{s(\boldsymbol{r})|\boldsymbol{D_n}\}$, where $s(\boldsymbol{r}) = \mathrm{logit}r_1 - \mathrm{logit}r_0$ is the log odds ratio. Note that

$$\mathrm{Var}\{g(\boldsymbol{r})|D_n\} \quad = \quad E\left[\mathrm{Var}\{g(\boldsymbol{r})|\boldsymbol{\theta}\}|\boldsymbol{D_n}\right] + \mathrm{Var}\left[E\{g(\boldsymbol{r})|\boldsymbol{\theta}\}|\boldsymbol{D_n}\right], \quad (3.6)$$

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 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 (3.6) in partially identified models are studied at length by Gustafson (2006) [10].

The simulated data sets are also analyzed via the informal method alluded to in Section 3.1.4. That is, 'unlikely exposed' and 'maybe exposed' subjects are merged and taken as 'unexposed', while the 'likely exposed'

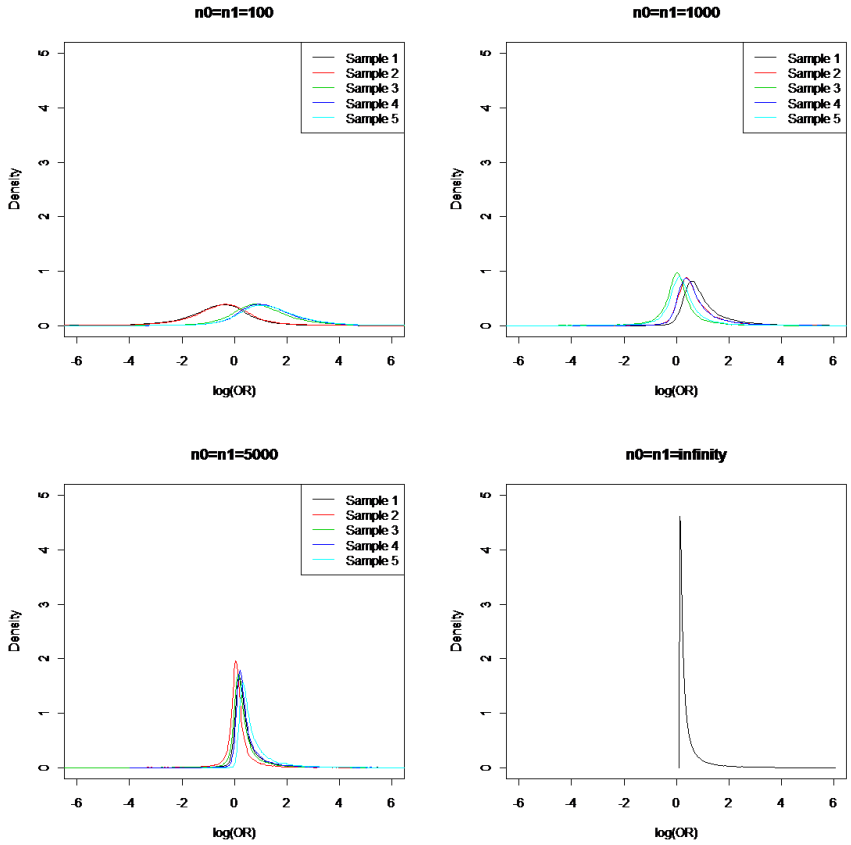

Figure 3.18: Posterior distributions under the combination $(- - -)$. From the upper-left panel to the lower-right panel, the posterior distributions of $\log OR$ for the sample sizes 100, 1000, 5000, and the limiting posterior distribution are displayed.

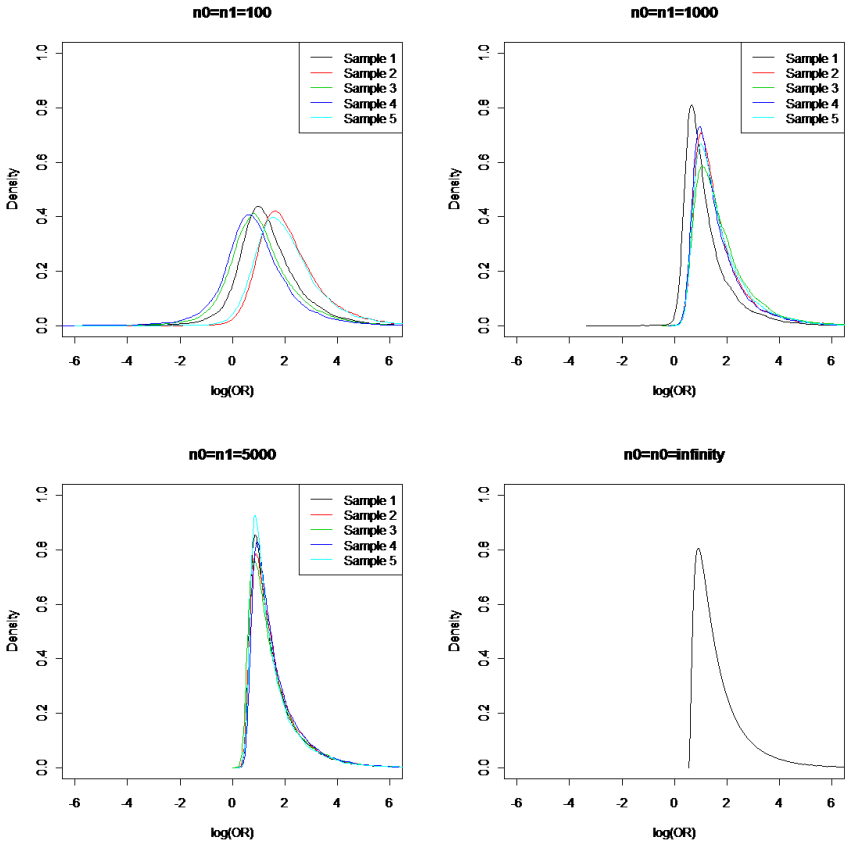

Figure 3.19: Posterior distributions under the combination $(+ + +)$. The layout is the same as Figure 3.18.

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 $\log OR$ appear in Figures 3.20 and 3.21. 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 $(---)$ setting or $OR = 2.0$ in the $(+++)$ setting. Thus the informal strategy of choosing to treat 'maybe exposed' subjects as being unexposed in light of low exposure prevalence proves to be useful. Of course with enough data people would eventually be lead astray. That is, from Table 3.1 it shows 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.

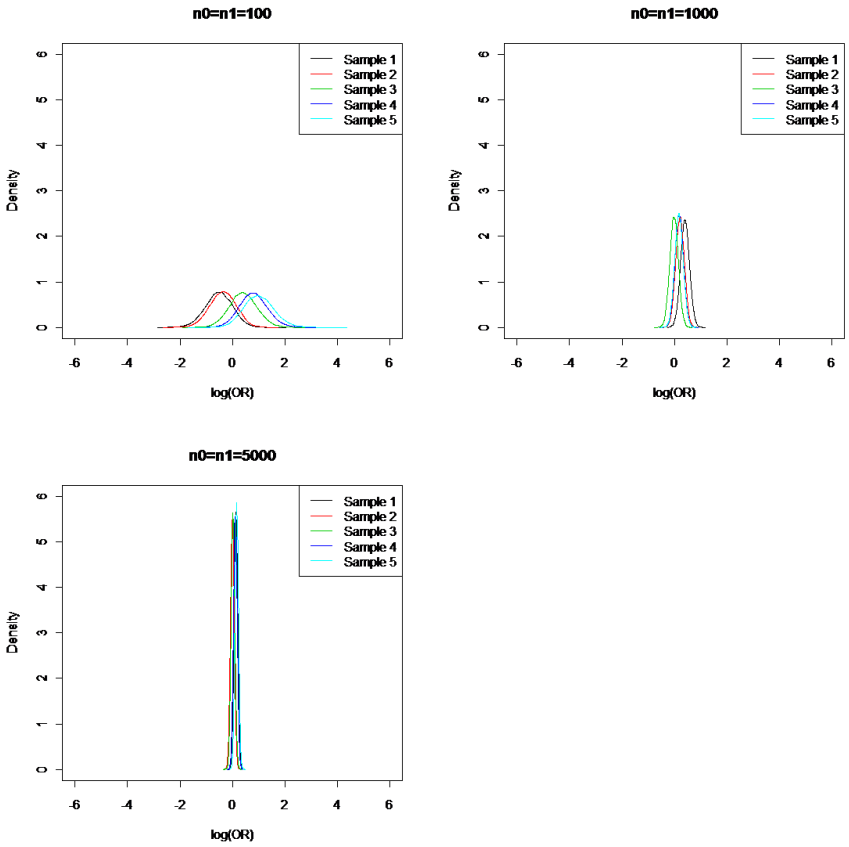

Figure 3.20: Posterior distributions via informal analysis under the combination $(- - -)$. From the upper-left panel to the lower panel, the posterior distributions of $\log OR$ for the sample sizes 100, 1000, and 5000.

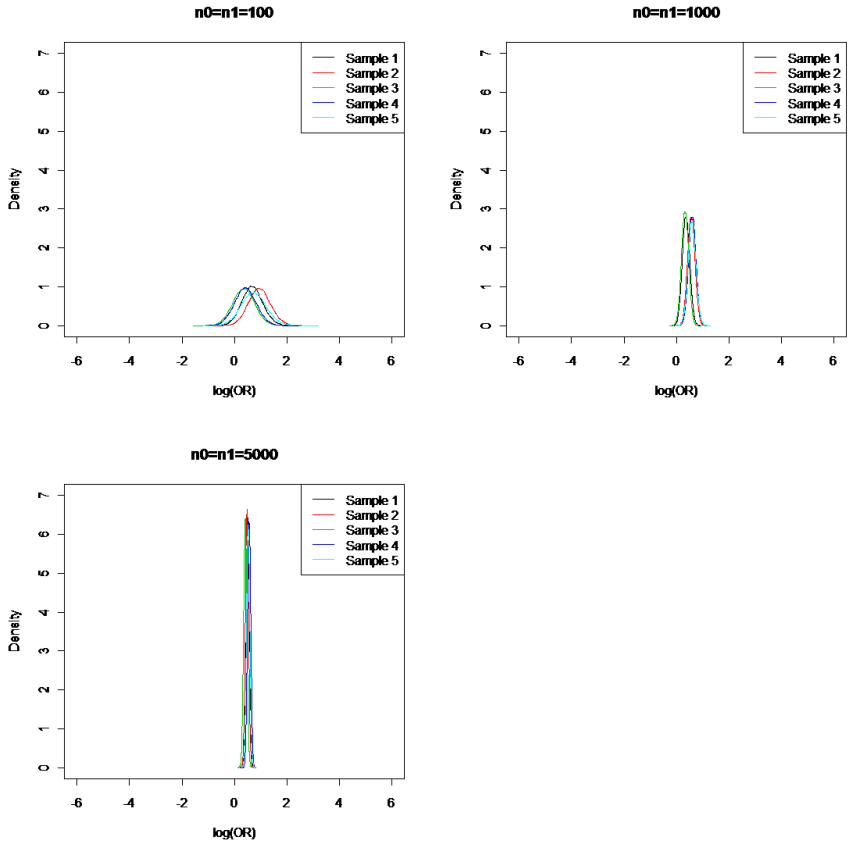

Figure 3.21: Posterior distributions via informal analysis under the combination $(+++)$. The layout is the same as Figure 3.20.

Chapter 4

Conclusion

In Chapter 2, we investigated the impact of misclassification for polychotomous exposure. The relationship between the group mean outcomes of apparent exposure and actual exposure can be calculated using Bayes' theorem. We first summarize that only under the least severe misclassification, where subjects cannot be misclassified more than one category away from the true exposure levels, monotone group mean outcomes of actual exposure will also lead to monotone group mean outcomes of apparent exposure. Whenever the classification is worse, the conclusion does not hold anymore. Then, we focus on the effect of the exposure distribution under the least severe misclassification. For a given exposure level, it is possible to compare the effect of misclassification at this level between uniformly distributed and non-uniformly distributed actual exposure. Moreover, as the goal of the study is to analyze the effect of misclassification for polychotomous exposure, we performed a trend test to investigate the overall effect of misclassification. By comparing the power of the trend test, we can find a counterexample in which misclassification strengthens the exposure-disease association. Therefore, we conclude that misclassification does not always attenuate the exposure-disease association for polychotomous exposure. It means that the effect of misclassification for polychotomous exposure does not always the same as for continuous exposure or binary exposure.

In Chapter 3, 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 (1996) [3]. 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 modeled 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. 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. Section 3.1 illustrates how an infinite amount of data can rule out near-null values of the exposure-disease association. 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 3.2, is the extent to which the posterior distribution, in the large-sample limit, is flat or concentrated across the identification region. It shows that in many circumstances the limiting posterior distribution of $\log OR$ is indeed quite peaked. In Section 3.3, 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.

Bibliography

[1] J. Buonaccorsi. *Measurement Error: Models, Methods, and Applications*. Chapman and Hall, CRC Press, 2010.

[2] R. Chu. Bayisan adjustment for exposure misclassifiction in case-contral studies. *Statistics in Medicine*, 29:994–1003, 2010.

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

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

[5] G. Espino-Hernandez, P. Gustafson, and I. Burstyn. Bayesian adjustment for measurement error in continuous exposures in an individually matched case-contral study. *BMC Medical Research Methodology*, 11:67, 2011.

[6] M. Garcia-Zattera, T. Mutsvari, A. Jara, D. Declerck, and E. Lesaffre. Correcting for Mislcassification for a Monotone Disease Process with an Application in Dental Research. *Wiley Online Libary*, 29:3103–3117, 2010.

[7] L. Gordis. *Epidemiology*. SAUNDERS, 1996.

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

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

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

[11] P. Gustafson. Bayesian inference for partially identified models. *International Journal of Biostatistics*, 6:issue 2 article 17, 2010.

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

[13] D. Kenkel, D. Lillard, and A. Mathios. Accounting for Misclassification Error in Retrospective Smoking Data. *HEALTH ECONOMICS*, 13:1031–1044, 2004.

[14] H. Kraemer, J. Measelle, J. Ablow, M. Essex, W. Boyce, and D. Kupfer. A New Approach to integrating Data From Multiple Informants in Psychiatric Assessment and Research: Mixing and Matching Contexts and Perspectives. *American Journal of Psychiatry*, 160:9:1566–1577, 2003.

[15] S. Krisbnaiab, K. Vilas, B. Shamanna, G. Rao, R. Thomas, and D. Balasubramanian. Local sensitivity of inferences to prior marginals. *IOVS*, 46:58–65, 2005.

[16] B. Lindblad, N. Hakansson, H. Svensson, B. Philipson, and A. Wolk. Intensity of Smoking and Smoking Cessation in Relation to Risk of Cataract Extaction: A Prospective Study of Women. *American Journal of Epidemiology*, 162:73–79, 2005.

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

[18] C. F. Manski. *Partial Identification of Probability Distributions*. Springer, 2003.

[19] E. Savoca. Accounting for Misclassification Bias in Binary Outcome Measures of Illness: The Case Of Post-Traumatic Stress Disorder in Male Veterans. *SOCIOLOGICAL METHODOLOGY*, 41:49–76, 2011.

[20] T. Shen. Formal and Informal Approaches to Adjusting for Exposure Misclassification. *Thesis, Department of Statistics, UBC*, 2009.

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