Costs and Benefits of Environmental Data in Investigations of Gene-Disease Associations

by

Hao Luo

B.Sc., Nanjing University, 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)

August 2012

Abstract

The inclusion of environmental exposure data may be beneficial, in terms of statistical power, to investigation of gene-disease association when it exists. However, resources invested in obtaining exposure data could instead be applied to measure disease status and genotype on more subjects. In a cohort study setting, we consider the tradeoff between measuring only disease status and genotype for a larger study sample and measuring disease status, genotype, and environmental exposure for a smaller study sample, under the 'Mendelian randomization' assumption that the environmental exposure is independent of genotype in the study population. We focus on the power of tests for gene-disease association, applied in situations where a gene modifies risk of disease due to particular exposure without a main effect of gene on disease. Our results are equally applicable to exploratory genome-wide association studies and more hypothesis-driven candidate gene investigations. We further consider the impact of misclassification for environmental exposures. We find that under a wide range of circumstances research resources should be allocated to genotyping larger groups of individuals, to achieve a higher power for detecting presence of gene-environment interactions by studying gene-disease association.

Table of Contents

List of Tables

List of Figures

Acknowledgments

First, I really want to give thanks to my supervisor, Professor Paul Gustafson, who helped me a lot with the completion of my thesis. It was my great honor to work with him. I would give thanks to Professor Igor Burstyn, who provided support and guidance as an epidemiological expertise. I would also like to thank Professor Gabriela Cohen Freue for agreeing to be my second reader.

I would express my gratitude to Professor John Pekau, Ruben Zamar, Jennifer Bryan, Harry Joe, William Welch, Lang Wu and Eugenia Yu for their constant support and excellent teaching. I am also grateful to Peggy Ng, Elaine Salameh and Andrea Sollberger for their hard work and kind help. Thanks are given to everyone in our department for making the department such a good place.

Finally, I owe special thanks to my parents for their support and understanding of my study.

Chapter 1

Introduction

In recent decades, advances in genotyping technology and reductions in associated cost have made it feasible to conduct large-scale genome-wide association studies to locate disease susceptibility loci among thousands or millions of screened markers. However, such studies usually ignore the joint effects of genetic and environmental exposures, which may result in a loss in statistical power, as it is widely accepted that complex diseases are likely to be caused by the interplay of both genetic and environmental factors. Thus, it may be beneficial to collect concurrent environmental exposure data to conduct a study taking into account the gene-environment interaction [Kraft et al., 2007, Williamson et al., 2010].

It can be challenging to measure environmental exposure well. In the context of a binary exposure, however, [Kraft et al., 2007] found that the benefit of having environmental data is seen to be maintained in the face of misclassification levels with both sensitivity and specificity of 80%, with such levels seen commonly (e.g, [England et al., 2007, Pickett et al., 2009]). In occupational and environmental epidemiology, however, exposure misclassification can occur at a much higher rate, with sensitivity often around 50% or less (e.g. [Burstyn et al., 2009, Teschke et al., 2002]). Therefore the misclassification rates studied in [Kraft et al., 2007] are neither extreme nor typical of epidemiology in general. Furthermore, it is typically costly to obtain exposure data, and this cost could instead be applied to measure disease status and genotype on more subjects. Thus, from a fixed resource per-

spective, the additional cost of exposure assessment may be so high that measuring only disease status and genotype for a larger study sample may yield more power than measuring disease status, genotype, and environmental exposure for a smaller study sample. It has been shown before that there is a balance between costly exposure estimates that are perfect and cheaper error-prone methods in achieving optimal study power on a fixed budget [Armstrong, 1996]. It has also been argued that to detect gene-environment interaction when exposure measures are assessed with error, it is beneficial to fit a marginal (gene-only) model to the data, when it can be assumed that gene is associated with outcome only in the presence of exposure, and the acquisition of environmental exposure is independent of genotype. The latter assumption is usually referred to as the 'Mendelian randomization' assumption [Smith, 2004], permitting the detection and quantification of environmental effects in the presence of latent confounding and heterogeneity in genetic susceptibility to environmental exposure.

We compare the power of two cohort study designs aimed at detecting conditional dependence between disease and a genetic locus given environmental exposure, with and without assessment of this exposure, in terms of cost-effectiveness, i.e., which one achieves higher power on a fixed-budget basis. We identify the situations in which the resources should be allocated to enlarging the sample size of the study instead of assessing environmental exposure. We focus on scenarios where Mendelian randomization can be assumed; a test for marginal gene-disease association is intuitively sensible under this assumption. For the joint test which incorporates environmental data, a null of no gene effect is tested against an alternative that there is a main effect and/or an interaction effect of gene. However, we study the power of this test in the setting that there is no main gene effect, i.e., the gene effect is only evident in the presence of exposure. This is referred to as a 'qualitative interaction' [Williamson et al., 2010], and was also stressed in earlier work [Burstyn et al., 2009]. We also discuss under what conditions our findings apply to case-control studies, and contrast our approach with the case-only study design. Our analysis is equally applicable to investigations that in genotyping rely on candidate genes selected for their involvement in affecting toxicity of specific exposure and genome-wide association studies.

Chapter 2

Study Designs

Let Y be the binary disease status, X the environmental exposure, and G one of possibly very many ascertained genetic markers. We assume G is binary, as would result if we distinguished only between the homozygous dominant versus other genotypes. We assume X is binary for the sake of simplicity and ease of illustration, with the view that X could in fact be a dichotomized version of a continuous exposure variable. The assumption of a binary X is commonly made in investigating methodologies for gene-environment interaction studies (see, for instance, [Kraft et al., 2007, Li and Conti, 2009, Umbach and Weinberg, 1997, Williamson et al., 2010]), since even if the underlying exposure is continuous, interpretation of effect above a certain threshold is often desirable in development of policy for interventions.

2.1 (Y,X,G) Design

When (Y,G,X) are all observed, a saturated logistic regression model, allowing a gene-environment interaction, is commonly fit:

$$\text{logit}Pr(Y=1|X,G)=\beta_0+\beta_1X+\beta_2G+\beta_3XG.$$

Within this model, the null and alternative hypotheses are:

$$\mathrm{H}_0 : \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \text{vs} \qquad \mathrm{H}_1 : \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

This null hypothesis states that the gene is not associated with disease, given the environmental exposure status.

We assume that being exposed is independent of having a specific gene in study population. Further, we denote $\pi_G = Pr(G = 1)$ the genotype prevalence, and $\pi_X = Pr(X = 1)$ the environmental exposure prevalence. Then the log-likelihood implied by the full model is:

$$\begin{aligned} \ell = \; & Y \log \left[\frac{\exp[\beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG]}{1 + \exp[\beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG]} \right] \\ & + (1 - Y) \log \left[\frac{1}{1 + \exp[\beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG]} \right] \\ & + G \log \pi_G + (1 - G) \log(1 - \pi_G) + X \log \pi_X + (1 - X) \log(1 - \pi_X). \end{aligned}$$

By standard large-sample theory, the asymptotic distribution of $\hat{\beta}$ is a multivariate normal distribution with mean its true value and variance the inverse of the expected Fisher information matrix. The expected Fisher information, $\mathbf{I}$, is the negative of the expectations of the second derivatives of the log-likelihood. As an example, we show the algebra for the $[\beta_3, \beta_3]$ entry, $\mathbf{I}_{44}$:

- The first derivative

$$\frac{\partial \ell}{\partial \beta_3} = YXG - XG \times \frac{\exp[\beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG]}{1 + \exp[\beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG]}.$$

- The second derivative

$$\frac{\partial^2 \ell}{\partial \beta_3^2} = -X^2 G^2 \times \frac{\exp[\beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG}{(1 + \exp[\beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG])^2}.$$

- The negative of the expectation

$$\begin{aligned}
\mathbf{I}_{44} &= -\mathbf{E}\left\{\frac{\partial^2 \ell}{\partial \beta_3^2}\right\} \\
&= \mathbf{E}\left\{X^2G^2 \times \frac{\exp[\beta_0+\beta_1X+\beta_2G+\beta_3XG}{(1+\exp[\beta_0+\beta_1X+\beta_2G+\beta_3XG])^2}\right\} \\
&= Pr(G=1,X=1) \times \frac{\exp[\beta_0+\beta_1+\beta_2+\beta_3}{(1+\exp[\beta_0+\beta_1+\beta_2+\beta_3])^2} \\
&= \pi_G\pi_X \frac{\exp[\beta_0+\beta_1+\beta_2+\beta_3}{(1+\exp[\beta_0+\beta_1+\beta_2+\beta_3])^2}.
\end{aligned}$$

Similarly, we can calculate the rest elements of the information matrix. After some algebra, it turns out that

$$\mathbf{I} = \begin{pmatrix}
B_0+B_1+B_2+B_3 & B_1+B_3 & B_2+B_3 & B_3 \\
B_1+B_3 & B_1+B_3 & B_3 & B_3 \\
B_2+B_3 & B_3 & B_2+B_3 & B_3 \\
B_3 & B_3 & B_3 & B_3
\end{pmatrix},$$

where

$$\begin{aligned}
B_0 &= (1-\pi_G)(1-\pi_X)\exp(\beta_0)(1+\exp(\beta_0))^{-2}, \\
B_1 &= (1-\pi_G)\pi_X\exp(\beta_0+\beta_1)(1+\exp(\beta_0+\beta_1))^{-2}, \\
B_2 &= \pi_G(1-\pi_X)\exp(\beta_0+\beta_2)(1+\exp(\beta_0+\beta_2))^{-2}, \\
B_3 &= \pi_G\pi_X\exp(\beta_0+\beta_1+\beta_2+\beta_3)(1+\exp(\beta_0+\beta_1+\beta_2+\beta_3))^{-2}.
\end{aligned}$$

The power calculation is based on the Wald test. Under the alternative hypothesis, the Wald statistic follows a non-central χ^2 distribution, $\chi^2_2(\lambda)$, with 2 degrees of freedom and non-centrality parameter:

$$\lambda = n\begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix}^T \times \left(\left[\mathbf{I}^{-1}\right]_{(3:4,3:4)}\right)^{-1} \times \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix}$$

$$= n \left[\beta_2^2 \frac{B_0^{-1} + B_1^{-1} + B_2^{-1} + B_3^{-1}}{B_0^{-1} + B_2^{-1}} + 2\beta_2\beta_3 + \beta_3^2 \right] (B_1^{-1} + B_3^{-1}).$$

Then, the power of this joint test is calculated as:

$$\text{Power} = Pr(\chi_2^2(\lambda) > \chi_{1-\alpha,2}^2),$$

where $\chi_{1-\alpha,2}^2$ is defined as the $1-\alpha$ percentile of the χ^2 distribution with 2 degrees of freedom.

2.2 (Y, G) Design

On the other hand, without X data the (Y, G) association can be represented by a saturated logistic regression, or a 'reduced form' of the above model:

$$\text{logit}Pr(Y = 1|G) = \alpha_0 + \alpha_1 G.$$

In this reduced model, the parameter of interest is the marginal gene-disease odds ratio, α_1, which is equal to 0 if there is no gene-disease association. Correspondingly, the null and alternative hypotheses for this reduced model are:

$$\text{H}_0 : \alpha_1 = 0 \qquad \text{vs} \qquad \text{H}_1 : \alpha_1 \neq 0.$$

Further, we have

$$Pr(Y = 1|G) = \sum_X Pr(Y = 1|X, G) Pr(X|G).$$

Thus, the parameter α_1 can be obtained by substituting the probabilities from the full model.

Under the assumption that being exposed is independent of having a specific

gene in study population, $(\beta_2, \beta_3) = (0,0)$ implies that

$$\begin{aligned}
Pr(Y=1|G=1) =& Pr(Y=1|X=1,G=1)Pr(X=1) \\
&+ Pr(Y=1|X=0,G=1)Pr(X=0) \\
=& \pi_X \frac{\exp[\beta_0+\beta_1]}{1+\exp[\beta_0+\beta_1]} + (1-\pi_X)\frac{\exp[\beta_0]}{1+\exp[\beta_0]} \\
=& Pr(Y=1|X=1,G=0)Pr(X=1) \\
&+ Pr(Y=1|X=0,G=0)Pr(X=0) \\
=& Pr(Y=1|G=0).
\end{aligned}$$

This further implies that gene, marginally, does not affect the risk of disease and hence $\alpha_1 = 0$. On the other hand, $\alpha_1 = 0$ implies that

$$\begin{aligned}
&\frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}\pi_X + \frac{\exp(\beta_0)}{1+\exp(\beta_0)}(1-\pi_X) = \\
&\frac{\exp(\beta_0+\beta_1+\beta_2+\beta_3)}{1+\exp(\beta_0+\beta_1+\beta_2+\beta_3)}\pi_X + \frac{\exp(\beta_0+\beta_2)}{1+\exp(\beta_0+\beta_2)}(1-\pi_X).
\end{aligned}$$

Hence, $\alpha_1 = 0$ corresponds to a single curve in the (β_2, β_3) parameter space that goes through the origin and depends upon $(\pi_X, \beta_0, \beta_1)$. Further, we notice that the above equation holds only when $(\beta_2, \beta_3) = 0$ or $\beta_2(\beta_2+\beta_3) < 0$. Thus, the null hypothesis of the marginal model is nearly equivalent to the null hypothesis of the full model. Then, (Y, G) design can be alternatively used to test for the null hypothesis of conditional independence between Y and G given X. Without the assumption of gene-environment independence, a non-zero value of α_1 can arise when $(\beta_2, \beta_3) = (0,0)$. In such instances then, evidence of a non-null (Y, G) association cannot be taken as evidence that Y and G are conditionally dependent given X.

The log-likelihood implied by the reduced model is:

$$\ell = \; Y \log\left[\frac{\exp[\alpha_0+\alpha_1 G]}{1+\exp[\alpha_0+\alpha_1 G]}\right]$$

$$+ (1-Y)\log\left[\frac{1}{1+\exp[\alpha_0+\alpha_1 G]}\right]$$
$$+ G\log\pi_G + (1-G)\log(1-\pi_G).$$

Similar to the calculation shown in Section 2.1, after some algebra, we have the expected Fisher information matrix for this marginal model as:

$$\mathbf{I} = \begin{pmatrix} A_0+A_1 & A_1 \\ A_1 & A_1 \end{pmatrix},$$

where

$$A_0 = (1-\pi_G)\exp(\alpha_0)(1+\exp(\alpha_0))^{-2},$$
$$A_1 = \pi_G\exp(\alpha_0+\alpha_1)(1+\exp(\alpha_0+\alpha_1))^{-2}.$$

Therefore, the power of the marginal test is calculated as:

$$\text{Power} = Pr(\chi_1^2(\lambda) > \chi^2_{1-\alpha,1}),$$

where $\chi_2^2(\lambda)$ is a non-central χ^2 distribution with 1 degrees of freedom and non-centrality parameter

$$\lambda = n\alpha^2\ (A_0^{-1}+A_1^{-1}).$$

2.3 Mixed Design

There is another possible sampling scheme, involving (Y,X,G) measurements for some subjects and (Y,G) measurements for others. In this case, we can still fit a full model and the corresponding null and alternative hypotheses are:

$$\mathrm{H}_0 : \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \text{vs} \qquad \mathrm{H}_1 : \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

.

Suppose our sample consists of (Y,X,G) measurements on N_1 subjects and (Y,G) measurements on N_2 subjects. Then the expected Fisher information matrix is:

$$\mathbf{I} = \frac{N_1}{N_1+N_2}\mathbf{I_1} + \frac{N_2}{N_1+N_2}\mathbf{I_2},$$

where $\mathbf{I_1}$ and $\mathbf{I_2}$ correspond to (Y,X,G) data and (Y,G) data respectively. For (Y,X,G) data, we have already derived the expected Fisher information matrix under the full model. Therefore, $\mathbf{I_1}$ takes the form as given in Section 2.1:

$$\mathbf{I_1} = \begin{pmatrix} B_0+B_1+B_2+B_3 & B_1+B_3 & B_2+B_3 & B_3 \\ B_1+B_3 & B_1+B_3 & B_3 & B_3 \\ B_2+B_3 & B_3 & B_2+B_3 & B_3 \\ B_3 & B_3 & B_3 & B_3 \end{pmatrix}.$$

On the other hand, applying the full model to (Y,G) data implies that the marginal gene-disease association should be expressed in terms of β through:

$$\begin{aligned} Pr(Y=1|G) =& Pr(Y=1|X=1,G)Pr(X=1) + Pr(Y=1|X=0,G)Pr(X=0) \\ =& \frac{\exp[\beta_0+\beta_1+(\beta_2+\beta_3)G]}{1+\exp[\beta_0+\beta_1+(\beta_2+\beta_3)G]} \times \pi_X + \frac{\exp[\beta_0+\beta_2 G]}{1+\exp[\beta_0+\beta_2 G]} \times (1-\pi_X). \end{aligned}$$

The log-likelihood for (Y,G) data with full model is

$$\begin{aligned} \ell = \;& Y \log\left[\frac{\exp[\beta_0+\beta_1+(\beta_2+\beta_3)G]}{1+\exp[\beta_0+\beta_1+(\beta_2+\beta_3)G]}\pi_X + \frac{\exp[\beta_0+\beta_2 G]}{1+\exp[\beta_0+\beta_2 G]}(1-\pi_X)\right] \\ &+ (1-Y)\log\left[\frac{1}{1+\exp[\beta_0+\beta_1+(\beta_2+\beta_3)G]}\pi_X + \frac{1}{1+\exp[\beta_0+\beta_2 G]}(1-\pi_X)\right] \\ &+ G\log\pi_G + (1-G)\log(1-\pi_G). \end{aligned}$$

Then, $\mathbf{I_2}$ can be derived based on this log-likelihood. After some algebra, it turns

out that:

$$\mathbf{I_2} = \begin{pmatrix} \frac{(B_0+B_1)^2}{A_0} + \frac{(B_2+B_3)^2}{A_1} & \frac{B_1(B_0+B_1)}{A_0} + \frac{B_3(B_2+B_3)}{A_1} & \frac{(B_2+B_3)^2}{A_1} & \frac{B_3(B_2+B_3)}{A_1} \\ \frac{B_1(B_0+B_1)}{A_0} + \frac{B_3(B_2+B_3)}{A_1} & \frac{B_1{}^2}{A_0} + \frac{B_3{}^2}{A_1} & \frac{B_3(B_2+B_3)}{A_1} & \frac{B_3{}^2}{A_1} \\ \frac{(B_2+B_3)^2}{A_1} & \frac{B_3(B_2+B_3)}{A_1} & \frac{(B_2+B_3)^2}{A_1} & \frac{B_3(B_2+B_3)}{A_1} \\ \frac{B_3(B_2+B_3)}{A_1} & \frac{B_3{}^2}{A_1} & \frac{B_3(B_2+B_3)}{A_1} & \frac{B_3{}^2}{A_1} \end{pmatrix},$$

where A_0, A_1, and B_0 to B_3 are given in previous two sections.

Having determined the Fisher information matrix, the power calculation is carried out again by:

$$\text{Power} = Pr(\chi_2^2(\lambda) > \chi_{1-\alpha,2}^2).$$

where the non-centrality parameter is

$$\lambda = n \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix}^T \times \left(\left[\mathbf{I}^{-1} \right]_{(3:4,3:4)} \right)^{-1} \times \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix}.$$

Chapter 3

Cost Effectiveness

Many studies have compared the (Y,G) design with the (Y,X,G) design, in terms of statistical power. [Kraft et al., 2007] found that collecting concurrent environmental data within large cohort studies could be beneficial for investigating gene-disease associations in situations where the gene effect is only evident in the presence of exposure.

In practice, however, this kind of comparison may be 'unfair' since they ignored the fact that a (Y,X,G) design typically costs more money or resources than a (Y,G) design with the same sample size. It is clear that resources invested on obtaining exposure data could instead be applied to measure disease status and genotype on more subjects. Therefore, it might be more appropriate to compare the power of different cohort study designs in terms of cost-effectiveness, i.e., which one achieves higher power on a fixed-budget basis.

We presume the cost of measuring Y, X and all the genetic markers on a subject to be c times the cost of measuring Y and the markers alone, referring to $c > 1$ as the cost-ratio. For the purpose of illustration, we focus on only one genetic marker, denoted by G as mentioned in Chapter 2. A pilot example is shown in Figure 3.1 to display the power of the (Y,X,G) design as the cost ratio varies. We assume that $(\pi_G, \pi_X, \beta_0, \beta_1, \beta_3) = (0.19, 0.4, \text{logit}0.05, \log 1.5, \log 1.5)$, and the study budget can afford 80% power for the (Y,G) design. From Figure 3.1, we can see that

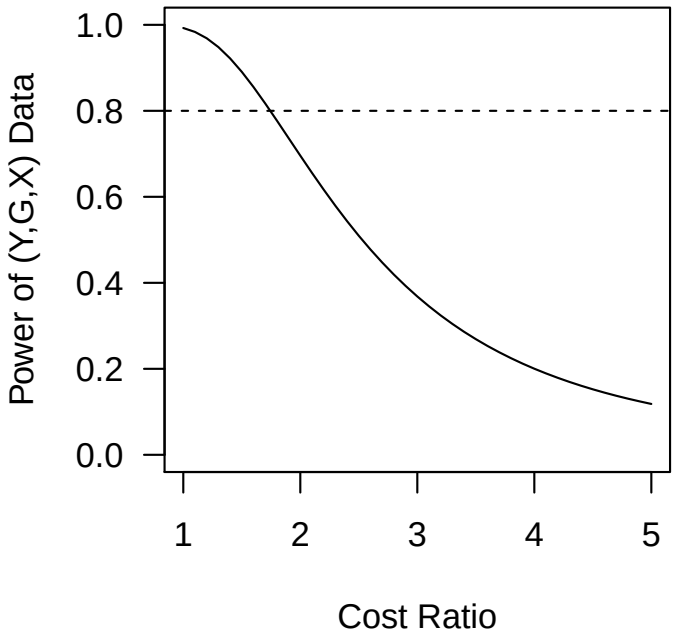


Figure 3.1: Effect of cost ratio on relative performance of (Y,X,G) design.

the change in power is quite sensitive to cost ratio. Particularly, collecting (Y,G,X) data will only yield power of around 50% when the cost ratio is at the value of 2.5.

To help our comparison, we introduce the break-even cost c^*, which is the value of c for which the same total cost spent on either a smaller (Y,X,G) sample or a larger (by a factor of c^*) (Y,G) sample will yield the same power to detect a gene effect. If the actual cost ratio c exceeds c^*, then collecting only (Y,G) data and fitting the reduced model is a better use of resources than collecting (Y,X,G) data and fitting the full model.

3.1 Performance of the Mixed Design

We begin by investigating the performance of the mixed design. Suppose our sample consists of (Y,X,G) measurements on a proportion w $(0 < w < 1)$ of the subjects and (Y,G) measurements only on the remaining $(1-w)\times 100\%$ subjects. Then the budget used to obtain (Y,G) data for m subjects, or (Y,X,G) data for m/c subjects, can also be used to obtain this kind of mixed data type for $m/(cw+1-w)$ subjects. Then the power of this mixed design can be calculated following the discussion in Section 2.3. Under the same setting of the pilot example, we examine the power for different values of w, as shown in Figure 3.2, with the cost ratio being 1.5 (top panel) and 2 (bottom panel).

We note that the power calculation for the mixed data ($w \in (0,1)$) is still valid when we have (Y,X,G) data only ($w = 1$), but not applicable when we have (Y,G) data only ($w = 0$) since applying the full model to (Y,G) data would lead to a non-identifiability problem. Thus, the power as a function of w is not continuous at $w = 0$ (as evident in Figure 3.2). Also, the performance of the mixed design depends on the value of the cost ratio. When the cost ratio is small, (Y,X,G) data are preferred. Thus, the more weight on (Y,X,G), the higher power. On contrary, when the cost ratio is large, (Y,G) data are preferred and increasing the proportion of (Y,X,G) decreases the power. But the model will become nearly non-identifiable if too few (Y,X,G) are collected. Therefore, in this case, the power of the mixed design is maximized at some point between 0 and 1. However, (Y,G) alone can achieve even higher power.

To make the comparison more tractable, we consider the situation where the cost ratio is the break-even cost. Then, the power at two endpoints, $w = 0$ and $w = 1$, are the same, so our problem is simplified as maximizing the power with $w \in (0,1]$. We conduct a factorial experiment, as shown in Table 3.1, to investigate the optimal w. In all settings, the shape of the power function is similar to that in the top panel of Figure 3.2, and the maximum power is always reached when $w = 1$. Then, we can conclude that this mixed data type is not a good choice compared to

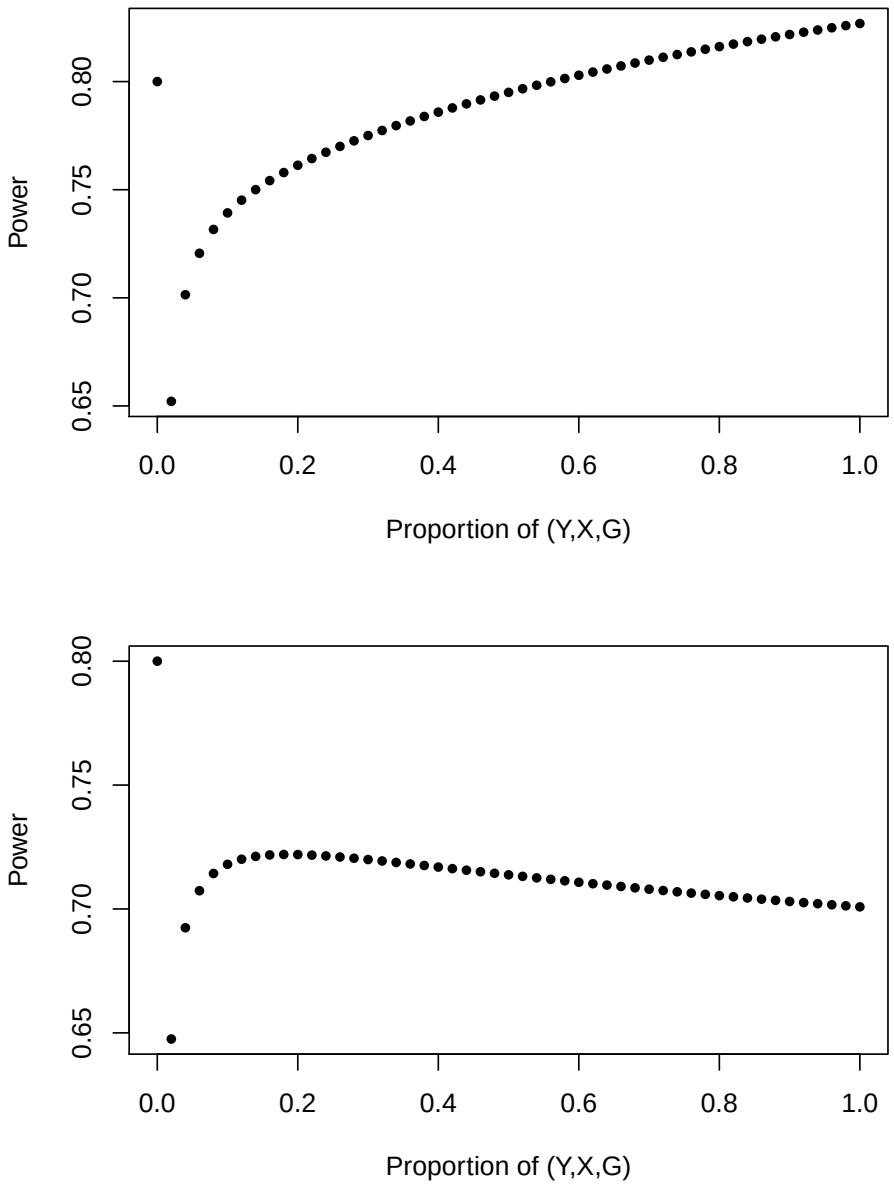


Figure 3.2: Power of the mixed design as proportion of (Y,X,G) data varies.

the other two sampling schemes with one single data type. This also matches our intuition that we should invest all resources on the most cost-effective data type.

3.2 (Y,X,G) Design *vs.* (Y,G) Design

We have shown that the 'mixed type' sampling scheme is always less cost-effective than the better of the two single type schemes, so we can focus on the comparison between (Y,X,G) data alone and (Y,G) data alone. To match the realistic situation that gene alone confers no additional disease risk in the absence of exposure, we consider scenarios where $\beta_2 = 0$ and $\beta_3 \neq 0$, which is termed as a 'qualitative' gene-environment interaction by [Williamson et al., 2010].

Let $F_q(\cdot,k)$ denote the cumulative distribution function for the noncentral χ^2 distribution with degree-of-freedom q and non-centrality parameter k. Let r_i denote the solutions for equation $1 - F_i(F_i^{-1}(1-s,0),x) = \text{Power},\ i = 1,2$, where s is the pre-specified significance level. Based on the power calculation described in Chapter 2, the sample sizes required for two study designs to achieve a certain power are:

$$N_{(Y,G)} = \frac{r_1}{\alpha_1^2} \times (\frac{1}{A_0} + \frac{1}{A_1}),$$
$$N_{(Y,X,G)} = \frac{r_2}{\beta_3^2 + 2\beta_2\beta_3 + \beta_2^2 Q/P} \times (Q - P),$$

where $P = 1/B_0 + 1/B_2$ and $Q = P + 1/B_1 + 1/B_3$. The break-even cost is just the ratio of sample size of (Y,G) design to sample size of (Y,X,G) design. Particularly, for the scenarios considered with a qualitative interaction, the break-even cost takes the following form:

$$c^* = \frac{r_1}{r_2} \times \frac{\beta_3^2}{\alpha_1^2} \times \frac{\frac{1}{A_0} + \frac{1}{A_1}}{\frac{1}{B_1} + \frac{1}{B_3}}.$$

As a technical point, the first term $\frac{r_1}{r_2}$ is only a function of the significance level and the desired power. Its value increases as the magnitude of desired power in-

creases, but decreases as the significance level increase. Once the type I error and the desired power are specified, it becomes a constant. Given that large numbers of markers may be screened, so that some form of multiple comparison adjustment or false discovery rate control will be required, we report power when the significance level is 10^{-4}.

We start with the same parameter setting as given in the pilot example. Figure 3.3 shows the break-even cost c^* as a function of the desired power. We can see that the break-even cost increases modestly with the desired power. Its shape also supports the fact that the c^* can vary according to the magnitude of the desired power since we are comparing the power of a two degree of freedom test to a one degree of freedom test. Particularly, we can read from Figure 3.3 that the break-even cost is around 1.7 when 80% power is desired. This implies that (Y,G) data are more cost effect than (Y,X,G) data when the per-subject cost of obtaining X is more than 70% of the cost of obtaining (Y,G).

Factorial Experiment

Next, we investigate the break-even cost under different parameter settings through a factorial experiment. All $440,000$ possible combinations of the parameter values listed in Table 3.1 have been considered.

	π_G	π_X	β_0	β_1	β_3
From	0.05	0.05	-5	0.1	0.1
Step	0.05	0.05	0.5	0.1	0.1
To	0.50	0.50	0	2	2

Table 3.1: Parameter settings of the factorial experiment.

We find that among the 5 parameters, prevalence of exposure (π_X) and background rate of the health outcome (β_0) have consistent effects, while the impacts

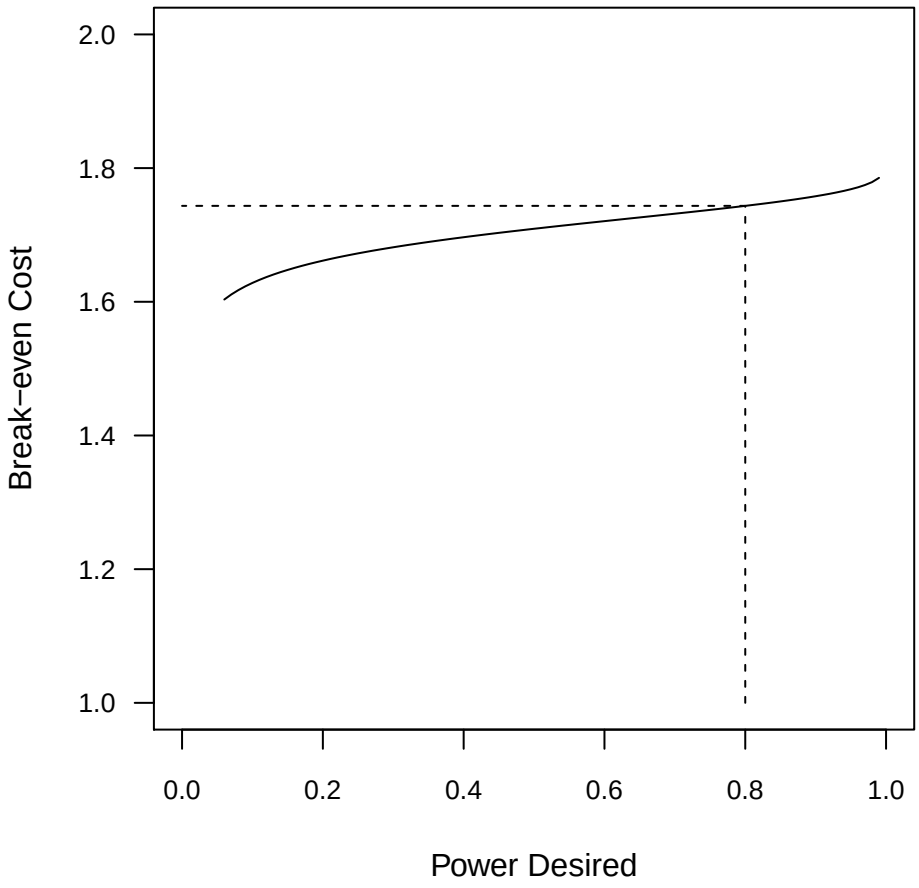


Figure 3.3: Break-even cost as a function of desired power.

of the other three parameters involve interactions with the other parameters. Figure 3.4 shows the joint effect of (π_X, β_0) and the corresponding contour plot, with all other parameter values set as in Figure 3.3. We can see that increasing the value of β_0 (i.e., studying the outcome more prevalent in the population) will increase the break-even cost, while increasing the value of π_X (i.e., studying a population with higher prevalence of environmental exposure) has the opposite effect.

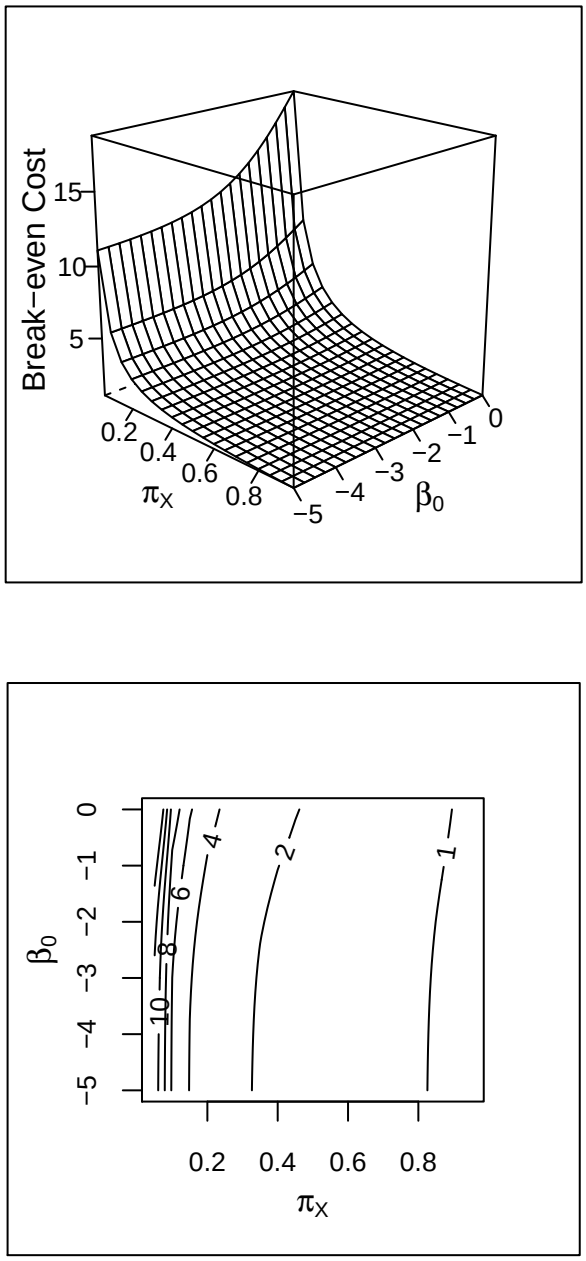


Figure 3.4: Joint effect of (π_X, β_0) on break-even cost.

Particularly, we notice that when the exposure is rare, the break-even cost will become extremely large, which is a strong signal for the necessity of assessing environmental exposure (collecting data on X) as opposed to relying on the marginal model to detect qualitative interaction. In our current comparison setting, gene alone confers no additional risk, so the difference in disease risk is only evident among exposed subjects. Therefore, when the prevalence of exposure is very rare, the sub-groups based on genotype are dominated by unexposed subjects, and hence exhibit very little difference in terms of disease risk between the two groups. That is why we need X data to identify the exposed subjects in two groups and focus on understanding differences in risk mainly in those two sub-groups: susceptible exposed and resistant exposed.

On the other hand, when the exposure is common, (Y,G) data are more likely to be preferred. Of all the settings with $\pi_X > 0.3$, about 77% have a break-even cost below 2, and 89% have a break-even cost below 2.5. So we can generalize that for common exposure, when the cost ratio is greater than 2, collection of (Y,G) data is typically more efficient. This may represent realities of studies of highly exposed groups such as industry-based cohorts or for contaminants that are widespread at 'toxic' levels in the general environment.

Collecting X Data can be Harmful

Interestingly, if we focus on a very prevalent environmental exposure, say with $\pi_X > 0.7$, we find that the break-even cost can even be below 1, which means that using X data would decrease power even if they could be obtained for free! (See Figure 3.5.) We have chosen several sets of parameter values for which the break-even cost is below one, and conducted simulation to verify these theoretical results. Although the empirical results may not be consistent with the theoretical results since the power calculation is only asymptotically true, there do exists situations where the (Y,G) study outperforms the (Y,X,G) study even with the same sample size, as reported in Table 3.2.

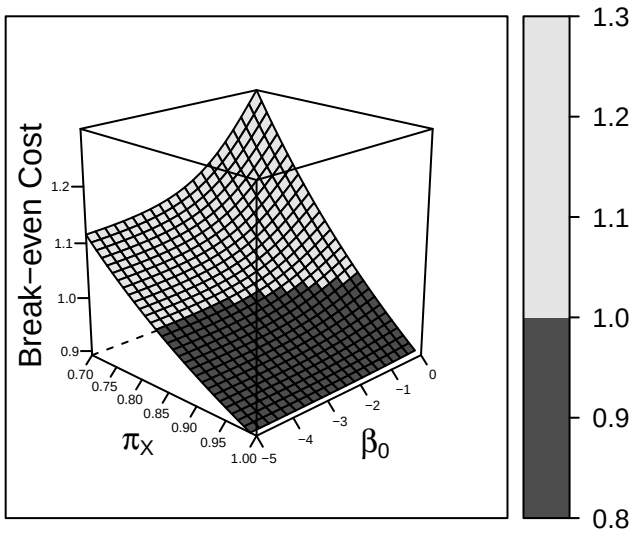


Figure 3.5: Situations where break-even cost is below 1.

Thus, when the exposure is very common, even perfectly measured X data are likely to be harmful, which seems counter-intuitive. To understand why this is the case, let us consider the extreme situation where everyone is exposed. If this is the case, binary X data clearly are useless, conveying no additional information beyond (Y,G). However, when we fit the full model to (Y,G,X) data, we are attempting to 'parse' any gene effect into the main effect β_2 and the interaction effect β_3. This is impossible without variation in X, and by extension inefficient when the X prevalence is very high. This inefficiency is also manifested when the X prevalence is very low, however the joint test still beats the marginal test in this case for the reasons described above.

π_G	π_X	β_0	β_1	β_3	Sample Size	Power of (Y,G)	Power of (Y,X,G)
0.50	0.40	-5.0	1.4	2.0	1400	0.899	0.890
0.35	0.50	-5.0	0.9	1.9	1700	0.831	0.820
0.30	0.50	-5.0	1.7	1.3	2100	0.774	0.766
0.45	0.50	-5.0	1.9	1.3	1800	0.829	0.811
0.45	0.50	-5.0	1.6	1.2	2900	0.835	0.834
0.35	0.45	-5.0	1.5	1.9	1100	0.824	0.808
0.20	0.50	-5.0	1.9	1.2	2600	0.766	0.761
0.50	0.50	-5.0	1.9	1.2	2300	0.850	0.838
0.40	0.45	-5.0	0.8	2.0	1900	0.858	0.852
0.15	0.50	-5.0	1.9	2.0	800	0.771	0.767
0.50	0.45	-5.0	2.0	1.6	1200	0.849	0.839

$\star$ 100,000 data sets were simulated under each condition.

Table 3.2: Situations where collecting X data can be harmful.

Chapter 4

Misclassification

4.1 Misclassification

In Chapter 3, we have compared the (Y,X,G) design with the (Y,G) design in terms of cost effectiveness, assuming all data are measured without error. However, this assumption is unrealistic. Whereas genetic information is relatively stable throughout life and can be measured nearly perfectly, exposure assessment is generally considered to be almost always error-prone. [England et al., 2007] found that reliance on self-reported smoking status among pregnant women can result in exposure misclassification, where 21.6% of self-reported quitters had evidence of active smoking. In some studies in occupation and epidemiology, exposure misclassification can occur at a much higher rate, with sensitivity often around 50% or less [Burstyn et al., 2009, Teschke et al., 2002].

In this chapter, we consider the situations with the presence of exposure misclassification. We denote X^* the imperfect environmental exposure, to distinguish it from the true environmental exposure X. Again, we assume X^* and G are independent given X and Y. Since the environmental exposure is assumed to be binary, the magnitude of misclassification can be described by sensitivity ($SN = P(X^* = 1|X = 1)$) and specificity ($SP = P(X^* = 0|X = 0)$). When we treat

X^* as if it were X, we are actually working under a true relationship of the form:

$$\begin{aligned} Pr(Y=1|X^*,G) &= \sum_X \{Pr(Y=1|X,X^*,G)Pr(X|X^*)\} \\ &= \sum_X \left\{ \frac{Pr(Y=1|X,G)Pr(X^*|Y=1,X,G)}{Pr(X^*|X,G)} \times \frac{Pr(X^*|X)Pr(X)}{Pr(X^*)} \right\} \\ &= \sum_X \left\{ Pr(Y=1|X,G)Pr(X^*|X,Y=1)\frac{Pr(X)}{Pr(X^*)} \right\}. \end{aligned}$$

Thus, we need a new set of parameters $({\pi_G}^*, {\pi_X}^*, {\beta_0}^*, {\beta_1}^*, {\beta_2}^*, {\beta_3}^*)$ rather than the true parameter setting $(\pi_G, \pi_X, \beta_0, \beta_1, \beta_2, \beta_3)$ to capture the true nature of (Y, X^*, G) data. The prevalence of genotype remain unchanged, indicating that $\pi_G^* = \pi_G$. The probability of being classified as exposed is

$$\pi_X^* = \sum_Y \sum_X \{Pr(X^*=1|X,Y)Pr(Y|X)Pr(X)\}.$$

Finally, β^* can be derived from the original parameters by solving:

$$\begin{aligned} Pr(Y=1|X^*,G) &= \text{expit}(\beta_0^* + \beta_1^* X^* + \beta_2^* G + \beta_3^* X^* G) \\ &= \sum_X \left[\text{expit}(\beta_0 + \beta_1 X + \beta_2 G + \beta_3 X G) \frac{Pr(X^*|X,Y=1)Pr(X)}{Pr(X^*)} \right], \end{aligned}$$

and it turns out that

$$\begin{aligned} \beta_0^* =& \text{logit} \left\{ \sum_X \left[\text{expit}(\beta_0 + \beta_1 X) \frac{Pr(X^*=0|X,Y=1)Pr(X)}{Pr(X^*=0)} \right] \right\}, \\ \beta_1^* =& \text{logit} \left\{ \sum_X \left[\text{expit}(\beta_0 + \beta_1 X) \frac{Pr(X^*=1|X,Y=1)Pr(X)}{Pr(X^*=1)} \right] \right\} \\ & - \beta_0^*, \\ \beta_2^* =& \text{logit} \left\{ \sum_X \left[\text{expit}(\beta_0 + \beta_2 + (\beta_1 + \beta_3) X) \frac{Pr(X^*=0|X,Y=1)Pr(X)}{Pr(X^*=0)} \right] \right\} \\ & - \beta_0^*, \end{aligned}$$

$$\beta_3^* = \text{logit}\left\{\sum_X \left[\text{expit}(\beta_0+\beta_2+(\beta_1+\beta_3)X)\frac{Pr(X^*=1|X,Y=1)Pr(X)}{Pr(X^*=1)}\right]\right\} - \beta_0^* - \beta_1^* - \beta_2^*.$$

We can see from the above expressions that fitting the full model to (Y, X^*, G) data generally gives biased point estimates of β. When $(\beta_2, \beta_3) = (0,0)$, however, it can be easily verified that the following two equations should both be satisfied:

$$\beta_2^* + \beta_0^* = \beta_0^*,$$
$$\beta_3^* + \beta_2^* + \beta_1^* + \beta_0^* = \beta_1^* + \beta_0^*.$$

This implies that $(\beta_2^*, \beta_3^*) = (0,0)$. Hence, $(\beta_2, \beta_3) = (0,0)$ in the $(Y|X,G)$ relationship implies zero coefficients for G and X^*G in the $(Y|X^*,G)$ relationship, so that fitting the full model to (Y,G,X^*) data still yields a valid test of the null hypothesis that Y and G are conditionally independent given X. However, the use of X^* rather than X will reduce power (e.g. [Burstyn et al., 2009, Rothman et al., 1999, Vineis, 2004, Wong et al., 2003]), and the power calculation should be adjusted for the presence of misclassification. We should plug the adjusted parameters given above into the power calculation shown in section 2.1 to obtain the power for misclassified (Y, X^*, G) data.

4.2 (Y, X^*, G) Design *vs.* (Y, G) Design.

In this section, we compare (Y, X^*, G) study to (Y, G) study in terms of cost effectiveness. Two types of misclassification are considered:

Non differential misclassification occurs when the probability of being misclassified are the same for all study subjects.

Differential misclassification occurs when the probability of being misclassified differs across groups of study subjects.

4.2.1 Non-differential Misclassification

We first focus on non-differential misclassification models. Figure 4.1 and Figure 4.2 show the effect of non-differential misclassification under different parameter settings:

- Figure 4.1 – rare exposure:

$$(\pi_G, \pi_X, \beta_0, \beta_1, \beta_3) = (0.19, 0.2, \text{logit}0.05, \log 1.5, \log 1.5),$$

- Figure 4.2 – common exposure:

$$(\pi_G, \pi_X, \beta_0, \beta_1, \beta_3) = (0.19, 0.7, \text{logit}0.05, \log 1.5, \log 1.5).$$

In both figures, we evaluate the break-even cost under four scenarios:

(i) *top-left:* $SN = 0.75, SP = 0.75$;

(ii) *top-right:* $SN = 0.95, SP = 0.95$;

(iii) *bottom-left:* $SN = 0.75, SP = 0.95$;

(iv) *bottom-right:* $SN = 0.95, SP = 0.75$.

The solid curve represents the break-even cost of the misclassified data, while the dash curve represents the break-even cost of the perfect data. The bigger gap between these two curves, the more the break-even cost is influenced by misclassification.

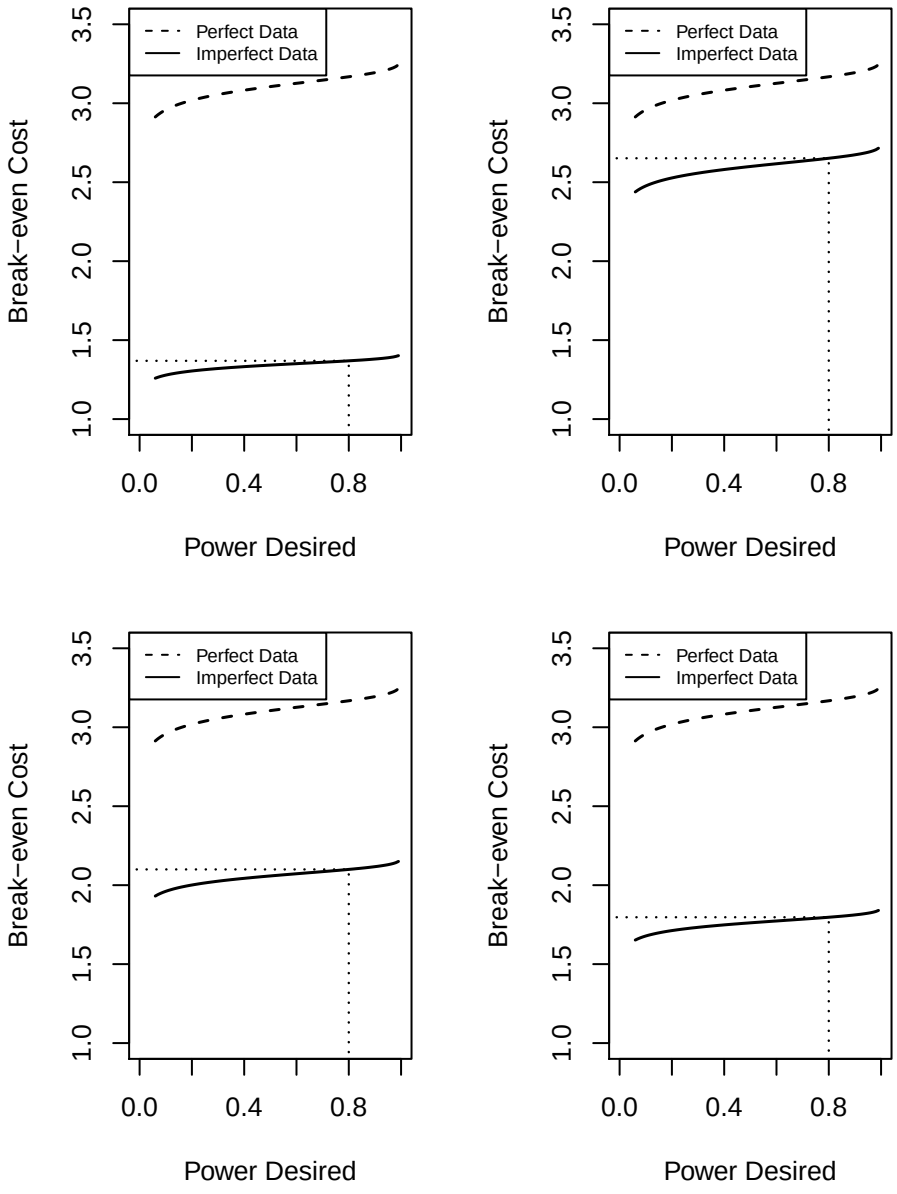


Figure 4.1: Effect of non-differential misclassification: rare exposure.

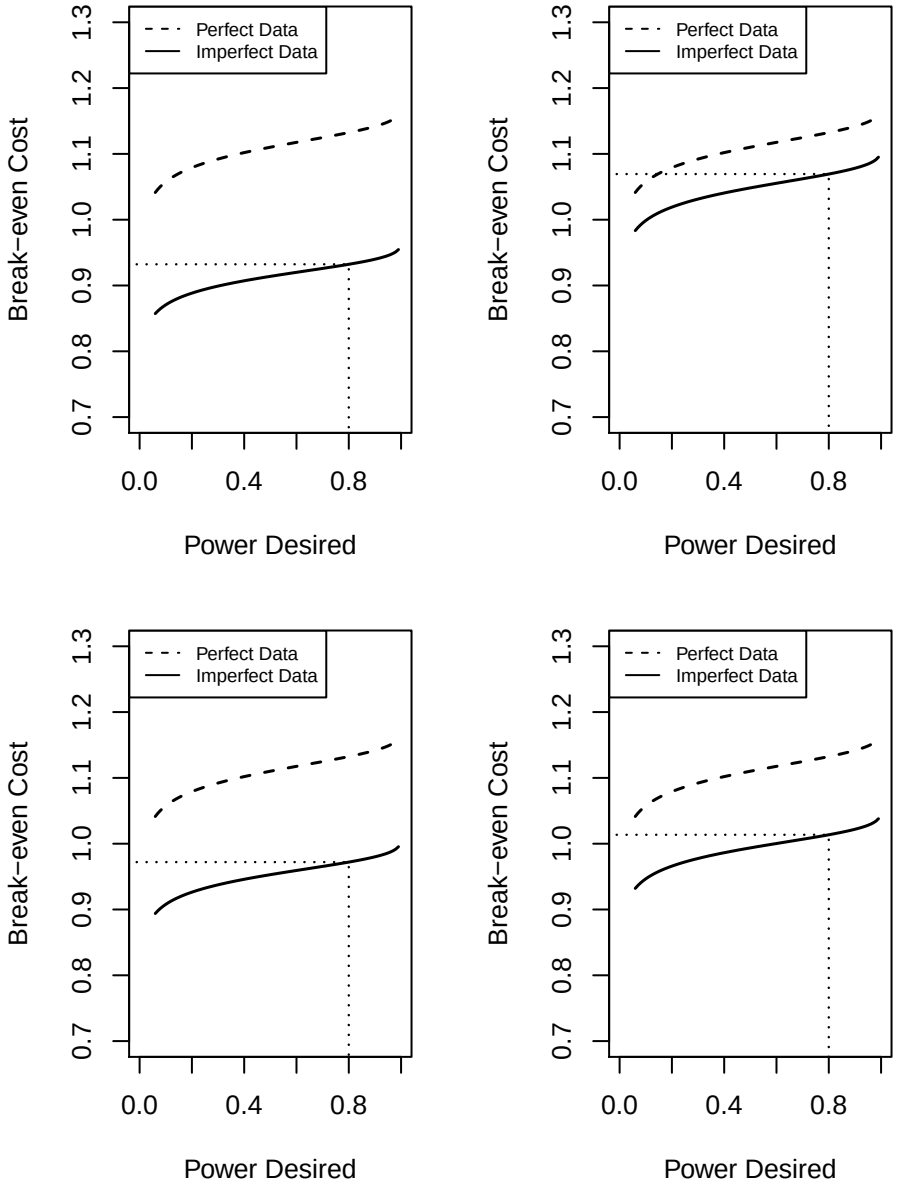


Figure 4.2: Effect of non-differential misclassification: common exposure.

From these two figures, we can see that similar to the impact of misclassification on estimation [Gustafson, 2004], power is more influenced by sensitivity (SN) when π_X is large (common exposure) and more by specificity (SP) when π_X is small (rare exposure). Also as expected, the break-even cost decreases as the misclassification becomes more severe, because more data have to be collected to compensate for lower power due to the use of an imperfect surrogate for exposure. In other words, misclassified exposure has to be rather cheap compared to collection of health outcome data and genotyping to be worthwhile, whereas greater expense can be justified for perfect (or near-perfect) exposure assessment.

Particularly, when the quality of the X^* data is very poor, the use of X^* data can be harmful. Therefore, it is important to determine at what values of sensitivity and specificity the break-even cost dips below 1. We already know from Chapter 3 that when the environmental exposure is very common, the break-even cost is likely to be below 1 even if X data are perfectly classified. So we investigate only situations with moderate prevalence of exposure. All combinations described in Table 3.1 were investigated again, under various values for the quality of exposure classification. We find that for a moderate prevalence of exposure, say that $0.2 < \pi_X < 0.5$, when $SN = SP = 0.6$, all combinations have a break-even cost below 1. Thus, if we cannot guarantee the quality of our environmental exposure data, it may be better to avoid assessing exposure. While it is well known that estimation bias can be removed by appropriate statistical adjustment for exposure measurement error, there is no way to recover the power lost by having X^* measurements rather than X measurements [Greenland and Gustafson, 2006].

4.2.2 Differential Misclassification

Next, let us turn our attention to the situation where differential misclassification occurs, which are more likely in cohort studies. We study three differential misclassification models that were also considered by [Williamson et al., 2010]:

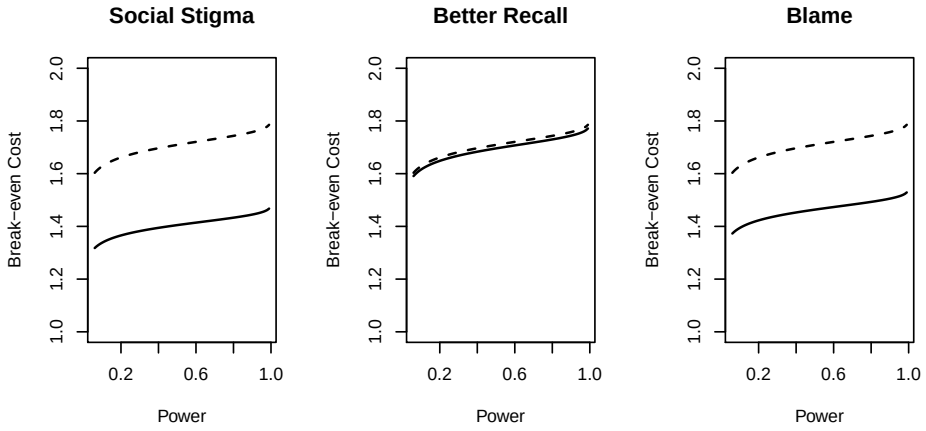


Figure 4.3: Effect of differential misclassification

(i) the "social stigma" model, where diseased subjects are unwilling to report their exposure status (sensitivity given disease = 80%, otherwise perfect classification); (ii) the "better recall" model, where those with disease keep an eye on the exposure and thus can give perfect classification, but those without disease are not able to recall the exposure history well (sensitivity and specificity for undiseased = 80%, perfect classification for diseased); (iii) the "blame" model, where subjects with disease blame their disease on an exposure and hence report it more often (specificity given disease = 80%, otherwise perfect classification). Figure 4.3 shows the break-even cost under these three scenarios, where the dash lines show the break-even cost without misclassification. We can see that there is a big drop under the "social stigma" model and the "blame" model, while no big change is seen under the "better recall" model. Hence, it seems that the loss in power is more driven by the misclassification among diseased subjects.

4.3 Extension of Comparison

Finally, we can extend our comparison to be among three data types: (Y,G,X), (Y,G,X^*), and (Y,G). For example, consider again the scenario under which Figure 3.3 is created: $(\pi_G,\pi_X,\beta_0,\beta_1,\beta_3) = (0.19,0.4,\text{logit}0.05,\log 1.5,\log 1.5)$. Let us say non-differential misclassification occurs with sensitivity and specificity both being 0.9. We aim to achieve 80% power, with 10^{-4} significance level. We can then determine two corresponding break-even costs relative to (Y,G) data: $c_1 = 1.7$ for (Y,G,X) data and $c_2 = 1.4$ for (Y,G,X^*) data. The break-even cost between (Y,G,X) and (Y,G,X^*) is simply the ratio $c_1/c_2 = 1.22$. That is, when a decision should be made between the (Y,X,G) design and the (Y,X^*,G) design, the former is less cost-effective if its per-subject cost of data acquisition is 23% more than the latter. Hence, Figure 4.4 can be created.

Presuming that (Y,G,X) data are indeed more costly than (Y,G,X^*) data which are in turn more costly than (Y,G) data, the shaded area in this plot corresponds to unrealistic cost ratios. The region of plausible cost values is divided into 3 subregions: area (i) is where (Y,G) data are the most cost-effective, area (ii) is where (Y,G,X^*) data should be collected; and area (iii) is where (Y,G,X) data are preferred. Thus, according to this, we can 'see' the best data type to be collected as a function of the two actual per subject cost ratios: for (Y,G,X) versus (Y,G) and (Y,G,X^*) versus (Y,G).

Plots of this form could be used for study-planning purposes, with pilot values of the parameters used to create a customized version of the plot. Anticipated sampling costs can then be located on the plot to visualize which of the three study designs is most cost effective. In some practical situations though, the choice in study design may be between collecting X^* versus no exposure assessment at all, since exposure status cannot be ascertained without error at any price. Moreover, with low SN and SP the comparison may lead investigators to favour inferring the presence of a gene effect from (Y,G) data. Such situations are common in practice and have profound implications for rational allocation of research resources and

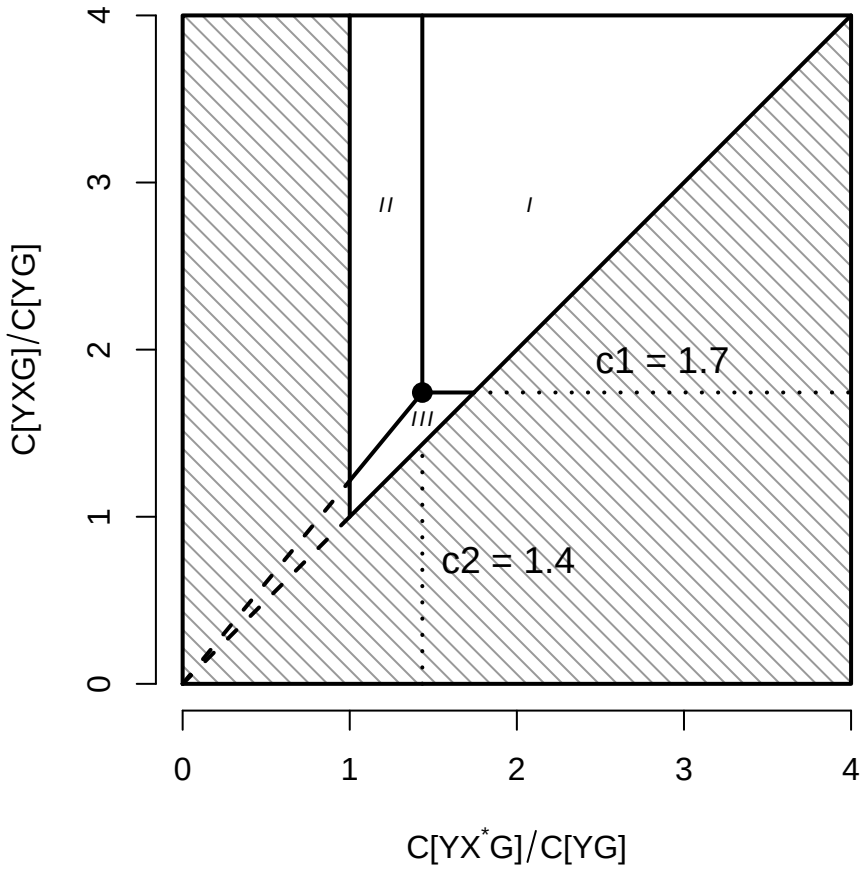


Figure 4.4: Comparison among three data types.

choice of research questions. Of course, one can usually allocate more resources to upgrade the instruments and improve the quality of data. Therefore, further extensions can be made to compare two environmental exposure surrogates with different qualities.

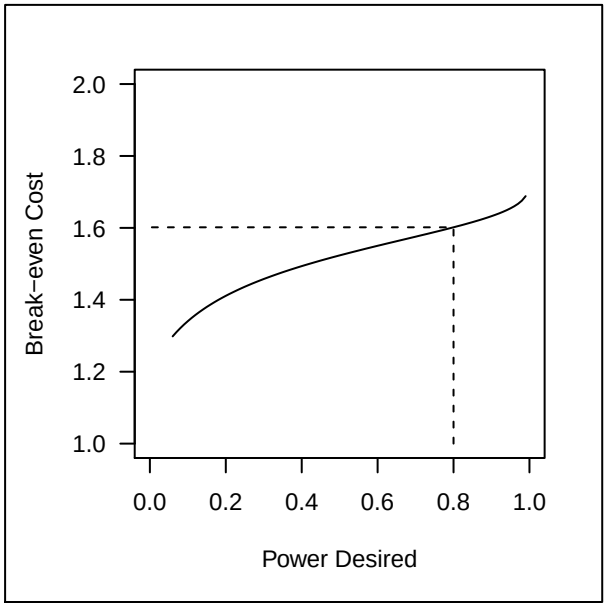
Chapter 5

Other Issues

5.1 Significance Level

In the previous two chapters, our results are presented in the context of a 10^{-4} significance level. Our use of the 10^{-4} significance level is a compromise for illustrative purposes. In the context of a genome-wide association study, a much more stringent level, such as 5×10^{-8}, would be employed (or, perhaps more likely, false discovery rate control would be implemented). Conversely, for some environmental exposures the number of candidate genes is very limited. For example, certain common and important exposures are associated with only one or two single nucleotide polymorphisms, e.g., SNPs in paraoxonase (PON1) gene and organophosphates [Burstyn et al., 2009]. In contexts with very limited numbers of candidate genes, much more liberal significance levels than 10^{-4} would be applied.

Figure 5.1 and Figure 5.2 reproduce some figures in Chapter 2 with a 0.05 significance level and a 5×10^{-8} significance level, respectively. The break-even cost for achieving 80% power is 1.6 when the significance level is 0.05, and 1.8 when the significance level is 5×10^{-8}. Thus, the relative performance of the (Y,X,G) study is improved with a more stringent significance level, although the circumstances under which the (Y,G) only design outperforms the (Y,X,G) design are insensitive to this choice.



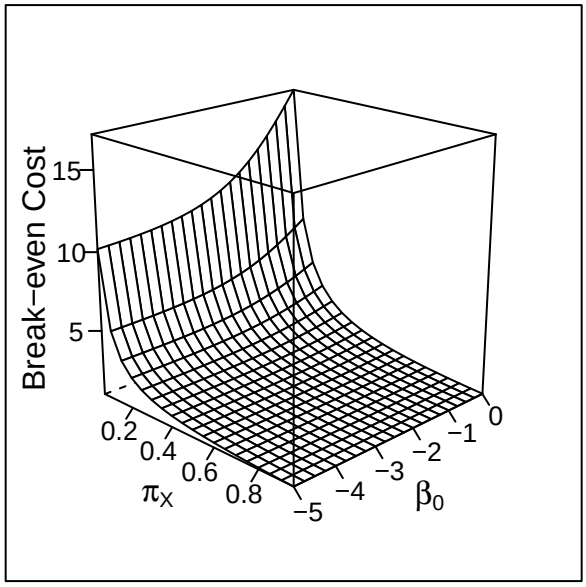
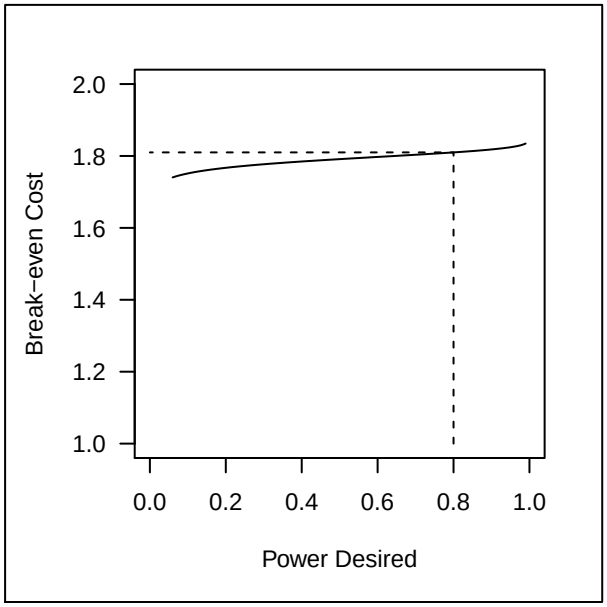


Figure 5.1: Results with a liberal significance level, 0.05.



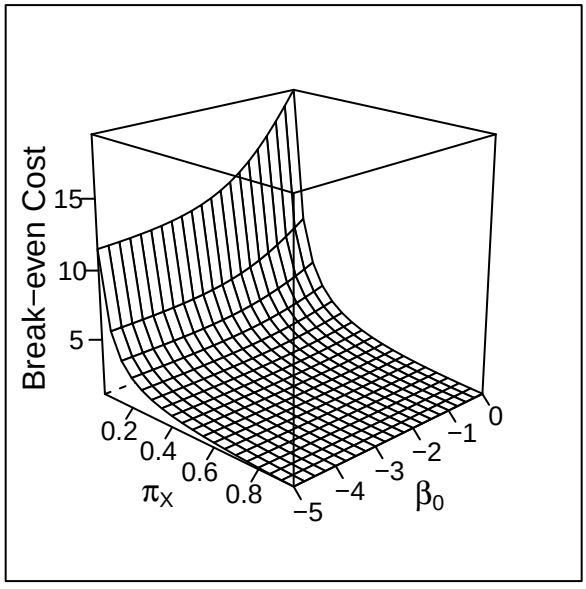


Figure 5.2: Results with a stringent significance level, 5×10^{-8}.

5.2 Presence of Main Gene Effect

Our reported power evaluations have been in the qualitative interaction setting ($\beta_2 = 0, \beta_3 \neq 0$), as we believe this may be a typical circumstance. However, when we evaluate power in the presence of a main effect of gene ($\beta_2 \neq 0$), we find that the break-even cost decreases with the magnitude of the main effect, whilst other parameters remain fixed at $(\pi_G, \pi_X, \beta_0, \beta_1, \beta_3) = (0.19, 0.4, \text{logit}0.05, \log 1.5, \log 1.5)$, as shown in Figure 5.3. In fact, as the main gene effect comes to dominate the

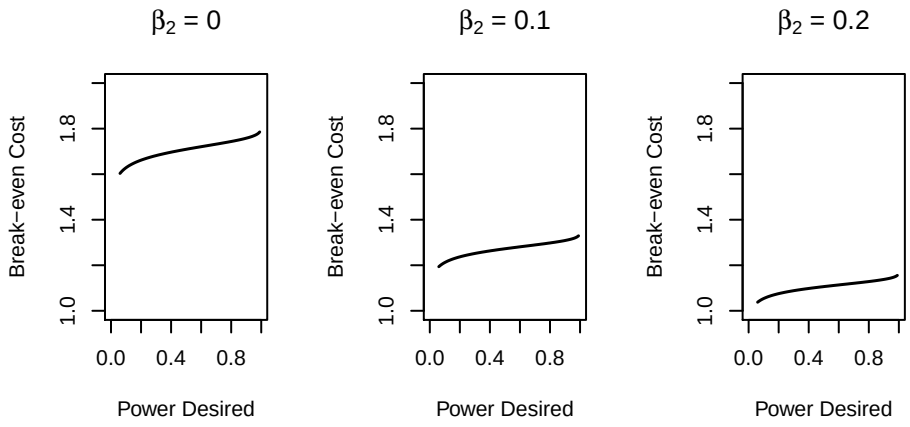


Figure 5.3: Break-even cost with the presence of main gene effect.

other effects in the model, the marginal gene effect may be large enough to be easily detected, and the involvement of exposure data may not help much. Thus, by evaluating power when in fact the gene-environment interaction is qualitative, we are considering a 'least favorable' setting for the (Y, G) only design, yet it often outperforms the (Y, X, G) design nonetheless.

5.3 Case-Control & Case-Only

While our results are presented in the cohort study setting, the power calculations are valid for case-control studies, with the proviso that the relevant intercept β_0 would be that induced by the case-control sampling scheme rather than that describing disease prevalence in the target population:

$$\beta_0^* = \beta_0 + \log \frac{N_{\text{Case}}}{N_{\text{Control}}} - \log \frac{Pr(Y=1)}{Pr(Y=0)}.$$

The other caveat is that our results are developed under sampling from a distribution in which X and G are independent. In the cohort study setting, this naturally corresponds to independence in the study population. In the case-control setting the situation is less clear, as the (X,G) distribution induced by case-control sampling is not identical to the (X,G) distribution in the target population.

We have addressed the value of X data (or X^* data) by comparing tests for a main effect and/or interaction effect of G obtained from equally costly samples with and without X (or X^*). In either case, the null hypothesis $\beta_2 = \beta_3 = 0$ is considered. However, if X data are to be collected, it may not be worth collecting any information on controls. [Piegorsch et al., 1994] showed that a gene-environment interaction can be estimated more efficiently with a case-only design than with either a cohort or a case-control study, under the assumption that the environmental exposure and gene are independent among controls. Let's first have a brief review of the case-only approach.

The justification of the case-only design can be shown by expressing β_3 in the following form:

$$\exp(\beta_3) = \frac{\text{Odds}(G=1|X=1,Y=1)/\text{Odds}(G=1|X=0,Y=1)}{\text{Odds}(G=1|X=1,Y=0)/\text{Odds}(G=1|X=0,Y=0)}.$$

Note that the denominator is equal to 1 when X and G are independent among controls. This is approximately true under the assumption of gene-environment in-

dependence (on population level) and under a rare-disease assumption. Therefore, β_3 can be estimated without any controls. Let N_{ijk} be the number of subjects with $Y = i, X = j$ and $G = k$ $(i, j, k = 0, 1)$ for a case-control study, as shown in Table 5.1.

	Y=1		Y=0	
	G=0	G=1	G=0	G=1
X=0	N_{100}	N_{101}	N_{000}	N_{001}
X=1	N_{110}	N_{111}	N_{010}	N_{011}

Table 5.1: Notations for the data of a case-control study.

In a case-control study, β_3 can be estimated by

$$\hat{\beta}_3 = \log \frac{N_{111}N_{100}N_{010}N_{001}}{N_{011}N_{000}N_{110}N_{101}},$$

and the estimated variance of $\hat{\beta}_3$ is

$$\widehat{\mathrm{Var}(\hat{\beta}_3)} = \frac{1}{N_{111}} + \frac{1}{N_{110}} + \frac{1}{N_{101}} + \frac{1}{N_{100}} + \frac{1}{N_{011}} + \frac{1}{N_{010}} + \frac{1}{N_{001}} + \frac{1}{N_{000}}.$$

On the other hand, by collecting case only, β_3 can also be estimated through

$$\hat{\beta}_3 = \log \frac{N_{111}N_{100}}{N_{110}N_{101}},$$

and the corresponding estimated variance is

$$\widehat{\mathrm{Var}(\hat{\beta}_3)} = \frac{1}{N_{111}} + \frac{1}{N_{110}} + \frac{1}{N_{101}} + \frac{1}{N_{100}}.$$

Thus, the case-only design provides a more efficient way for estimating gene-environment interaction.

However, we have deliberately not compared the case–only design to the (Y,G)

only design, since this would be a 'category mistake' (or an 'apples and oranges' comparison): the case-only design can only test the null $\beta_3 = 0$ (under the rare-disease assumption), whereas the (Y,G) only design can only test the null $\beta_2 = \beta_3 = 0$ (without invocation of the rare disease assumption).

Furthermore, in thinking about how the case-only design relates to the present discussion, exposure misclassification must be considered in deriving a comparison of study designs that are applicable to epidemiologic practice. We have already mentioned misclassification of environmental exposure as a point in favor of the (Y,G) only design compared to the (Y,X^*,G) design. In fact, the case-only design is even more susceptible to such misclassification than the (Y,X^*,G) design. Consider the worst-case of a "useless" exposure classification having sensitivity equal to $1-$specificity. When $(\beta_2,\beta_3) \neq (0,0)$, according to the reparameterization given in Section 4.1, this will induce $\beta_1^* = \beta_3^* = 0$ but $\beta_2^* \neq 0$, hence the (Y,X^*,G) data still have some power to detect the gene effect. Conversely, X^* and G will be conditionally independent given $Y = 1$, which will render the case-only design completely powerless. This gives a sense in which this design is an order-of-magnitude more susceptible to exposure misclassification than the 'full data' design.

5.4 3-Category Genotype

Finally, we can extend the discussion to the situation where the genotype has three categories. In cases where a gene exists in two allelic forms (designated A and a), three combinations of alleles (genotypes) are possible:

Homozygous-dominant Genotype when both alleles are dominant, i.e., AA;

Homozygous-recessive Genotype when both alleles are recessive, i.e., aa;

Heterozygous when two alleles are different, i.e., Aa.

Let H denote the 3-category genetic factor, which can take values in $\{0,1,2\}$. The power calculations for the (Y,X,H) design and the (Y,H) design are analogous to those shown in Chapter 2. In what follows, we only show some important pieces for the power calculation.

(Y,X,H) Design

The model applied to (Y,X,H) data is

$$\begin{aligned}\text{logit}Pr(Y=1|X,H) =& \tilde{\beta}_0+\tilde{\beta}_1 X+\tilde{\beta}_{21}\mathrm{I}(H=1)+\tilde{\beta}_{22}\mathrm{I}(H=2)\\ &+\tilde{\beta}_{31}X\mathrm{I}(H=1)+\tilde{\beta}_{32}X\mathrm{I}(H=1),\end{aligned}$$

where $\mathrm{I}(\cdot)$ is an indicator function. Correspondingly, the null and alternative hypotheses are

$$\mathrm{H}_0: \begin{pmatrix}\tilde{\beta}_{21}\\ \tilde{\beta}_{22}\\ \tilde{\beta}_{31}\\ \tilde{\beta}_{32}\end{pmatrix} = \begin{pmatrix}0\\0\\0\\0\end{pmatrix} \qquad \text{vs} \qquad \mathrm{H}_a: \begin{pmatrix}\tilde{\beta}_{21}\\ \tilde{\beta}_{22}\\ \tilde{\beta}_{31}\\ \tilde{\beta}_{32}\end{pmatrix} \neq \begin{pmatrix}0\\0\\0\\0\end{pmatrix}.$$

Finally, we have the expected Fisher information matrix as

$$\mathbf{I}=\begin{pmatrix}
D_0+D_1+D_{21}+D_{22}+D_{31}+D_{32} & D_1+D_{31}+D_{32} & D_{21}+D_{31} & D_{22}+B_{32} & D_{31} & D_{32}\\
D_1+D_{31}+D_{32} & D_1+D_{31}+D_{32} & D_{31} & D_{32} & D_{31} & D_{32}\\
D_{21}+D_{31} & D_{31} & D_{21}+D_{31} & 0 & D_{31} & 0\\
D_{22}+D_{32} & D_{32} & 0 & D_{22}+D_{32} & 0 & D_{32}\\
D_{31} & D_{31} & D_{31} & 0 & D_{31} & 0\\
D_{32} & D_{32} & 0 & D_{32} & 0 & D_{32}
\end{pmatrix},$$

where

$$\begin{aligned}D_0 &= Pr(H=0,X=0)\exp(\tilde{\beta}_0)(1+\exp(\tilde{\beta}_0))^{-2},\\ D_1 &= Pr(H=0,X=1)\exp(\tilde{\beta}_0+\tilde{\beta}_1)(1+\exp(\tilde{\beta}_0+\tilde{\beta}_1))^{-2},\end{aligned}$$

$$D_{21} = Pr(H=1,X=0)\exp(\tilde{\beta}_0+\tilde{\beta}_{21})(1+\exp(\tilde{\beta}_0+\tilde{\beta}_{21}))^{-2},$$
$$D_{22} = Pr(H=2,X=0)\exp(\tilde{\beta}_0+\tilde{\beta}_{22})(1+\exp(\tilde{\beta}_0+\tilde{\beta}_{22}))^{-2},$$
$$D_{31} = Pr(H=1,X=1)\exp(\tilde{\beta}_0+\tilde{\beta}_1+\tilde{\beta}_{21}+\tilde{\beta}_{31})(1+\exp(\tilde{\beta}_0+\tilde{\beta}_1+\tilde{\beta}_{21}+\tilde{\beta}_{31}))^{-2}.$$
$$D_{32} = Pr(H=2,X=1)\exp(\tilde{\beta}_0+\tilde{\beta}_1+\tilde{\beta}_{22}+\tilde{\beta}_{32})(1+\exp(\tilde{\beta}_0+\tilde{\beta}_1+\tilde{\beta}_{22}+\tilde{\beta}_{32}))^{-2}.$$

(Y,H) Design

The model applied to (Y,H) data is

$$\text{logit}Pr(Y=1|H) = \tilde{\alpha}_0 + \tilde{\alpha}_{11}\text{I}(H=1) + \tilde{\alpha}_{12}\text{I}(H=2).$$

The corresponding null and alternative hypotheses are:

$$\text{H}_0 : \begin{pmatrix} \tilde{\alpha}_{11} \\ \tilde{\alpha}_{12} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \text{vs} \qquad \text{H}_a \begin{pmatrix} \tilde{\alpha}_{11} \\ \tilde{\alpha}_{12} \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Finally, the expected Fisher information matrix is

$$\mathbf{I} = \begin{pmatrix} C_0+C_{11}+C_{12} & C_{11} & C_{12} \\ C_{11} & C_{11} & 0 \\ C_{12} & 0 & C_{12} \end{pmatrix},$$

where

$$C_0 = Pr(G=0)\exp(\tilde{\alpha}_0)(1+\exp(\tilde{\alpha}_0))^{-2},$$
$$C_{11} = Pr(G=1)\exp(\tilde{\alpha}_0+\tilde{\alpha}_{11})(1+\exp(\tilde{\alpha}_0+\tilde{\alpha}_{11}))^{-2},$$
$$C_{12} = Pr(G=2)\exp(\tilde{\alpha}_0+\tilde{\alpha}_{12})(1+\exp(\tilde{\alpha}_0+\tilde{\alpha}_{12}))^{-2}.$$

Chapter 6

Conclusion & Discussion

Our main finding is that under a wide range of circumstances research resources aimed at identification of association between genes and diseases can be more efficiently (in a sense of study power) allocated to genotyping larger groups of individuals rather than investing in exposure assessment, when exposure and genes interact. Likewise, efficient study design to detect qualitative gene-environment interactions can typically omit exposure assessment if there is convincing evidence that the gene only influences risk of disease by modifying exposure. (The evidence for mode of action of gene in conferring risk of disease would have to arise from studies outside of realm of epidemiology.) These conclusions do not negate the need for exposure assessment in quantifying gene-disease and gene-environment interactions, but do suggest the claim in [Williamson et al., 2010] that it is always desirable to assess exposures in such studies does not hold when resource constraints are considered. Of course there may well be circumstances where neither analytical approach will yield satisfactory power, but our results support the claim made in [Burstyn et al., 2009] that test for qualitative interaction typically requires smaller sample size to achieve the same power as study that collects error-prone exposure data and estimates interaction directly. It should be noted that when prior information is available on the magnitude of gene-environment interaction, data on gene and health outcome alone, under the Mendelian randomization assumption, can be used to estimate the magnitude of interaction through a Bayesian procedure [Gustafson and Burstyn, 2011].

It must be recognized that even in situations where (Y,G) data yields more power than (Y,X,G) data, the latter data structure does permit partitioning of the gene effect into main and interaction components. If (Y,G) data alone are collected and indicate a gene effect, then the question arises of whether this might arise via a main effect of G alone, an interaction effect of G alone (a qualitative interaction), or a combination of main and interaction effects. In some contexts a main effect is implausible a priori, so the results can be interpreted as evidence for a qualitative interaction. In other contexts, it may make sense to seek additional resources, in order to obtain exposure or surrogate exposure measurements for a subsample, permitting estimation of the coefficients in the full model. The question of resource use is now more complex, since the sub-sample cost of exposure assessment is only incurred if the initial (Y,G) sample indicates association.

It is also paramount to consider that the number of environmental exposures of potential interest is large. For example, a conservative list used by the U.S. National Health and Nutrition Examination Survey consists of at least 266 'core' exposures [Patel et al., 2010], and it is believed that the environmental exposures epidemiologists ought to be considering in exploratory studies number in the thousands, at least [Wild, 2005]. It is also clear that the cost of exposure assessment that meets the needs of epidemiology by providing both accurate and biologically meaningful measures will continue to escalate in near term, given the experimental nature of approaches that are being proposed [Rappaport and Smith, 2010]. There is hope that the costs of exposure assessment will decline in time, as they have done for genotyping, but exposure assessment is a much more complex technical challenge than genotyping. For the foreseeable future, scientists must contend with exposure assessment costs that can be on the order of hundreds of dollars per subject per exposure. Under these conditions, selecting appropriate exposures and genes to study in addressing important questions in public health will remain central to designing feasible and cost-effective investigations.

Overall, we conclude that in many situations not collecting environmental exposure to boost sample size is an efficient approach to assessing qualitative gene-

environment interactions when the disease is known not be causes by gene alone. The approach may also prove to be valuable as an efficient first stage of identifying role of gene (via interaction or main effect) in causing a disease.

Bibliography

B. Armstrong. Optimizing power in allocating resources to exposure assessment in an epidemiologic study. *American Journal of Epidemiology*, 144(2): 192–197, 1996.

I. Burstyn, H. Kim, Y. Yasui, and C. N.M. The virtues of a deliberately mis-specified disease model in demonstrating a gene-environment interaction. *Occupational and Environmental Medicine*, 66(6):374–380, 2009.

L. England, A. Grauman, C. Qian, D. Wilkins, E. Schisterman, F. Kai, and R. Levine. Misclassification of maternal smoking status and its effects on an epidemiologic study of pregnancy outcomes. *Nicotine & Tobacco Research*, 9 (10):1005–1013, 2007.

S. Greenland and P. Gustafson. Accounting for independent nondifferential misclassification does not increase certainty that an observed association is in the correct direction. *American Journal of Epidemiology*, 164(1):63–68, 2006.

P. Gustafson. *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. Boca Raton: Chapman and Hall/CRC, 2004.

P. Gustafson and I. Burstyn. Bayesian inference of geneCenvironment interaction from incomplete data: What happens when information on environment is disjoint from data on gene and disease? *Statistics in Medicine*, 30(8):877–889, 2011.

P. Kraft, Y. Yen, D. Stram, J. Morrison, and W. Gauderman. Exploiting gene-environment interaction to detect genetic associations. *Human Heredity*, 63(2):111–119, 2007.

D. Li and D. Conti. Detecting gene-environment interactions using a combined case-only and case-control approach. *American Journal of Epidemiology*, 169 (4):497–504, 2009.

C. Patel, J. Bhattacharya, and A. Butte. An environment-wide association study (EWAS) on type 2 diabetes mellitus. *PLoS One*, 5(5):e10746, 2010.

K. Pickett, K. Kasza, G. Biesecker, R. Wright, and L. Wakschlag. Women who remember, women who do not: A methodological study of maternal recall of smoking in pregnancy. *Nicotine & Tobacco Research*, 11(10):1166–1174, 2009.

W. Piegorsch, C. Weinberg, and J. Taylor. Non–hierarchical logistic models and case–only designs for assessing susceptibility in population–based case–control studies. *Statistics in Medicine*, 13(2):153–162, 1994.

S. Rappaport and M. Smith. Environment and disease risks. *Science*, 330(6003): 460–461, 2010.

N. Rothman, M. Garcia-Closas, W. Stewart, and J. Lubin. The impact of misclassification in case-control studies of gene-environment interactions. *IARC Scientific Publications*, 148:89–96, 1999.

S. Smith, GD nad Ebrahim. Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology*, 33(1):30–42, 2004.

K. Teschke, A. Olshan, J. Daniels, A. De Roos, C. Parks, M. Schulz, and T. Vaughan. Occupational exposure assessment in caseCcontrol studies: opportunities for improvement. *Occupational and Environmental Medicine*, 59 (9):575–594, 2002.

D. Umbach and C. Weinberg. Designing and analysing case–control studies to exploit independence of genotype and exposure. *Statistics in Medicine*, 16(15): 1731–1743, 1997.

P. Vineis. A self-fulfilling prophecy: are we underestimating the role of the environment in geneCenvironment interaction research? *International Journal of Epidemiology*, 33(5):945–946, 2004.

C. Wild. Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiology Biomarkers & Prevention*, 14(8):1847–1850, 2005.

E. Williamson, A. Ponsonby, J. Carlin, and T. Dwyer. Effect of including environmental data in investigations of gene–disease associations in the presence of qualitative interactions. *Genetic Epidemiology*, 34(6):552–560, 2010.

M. Wong, N. Day, J. Luan, K. Chan, and N. Wareham. The detection of geneCenvironment interaction for continuous traits: should we deal with measurement error by bigger studies or better measurement? *International Journal of Epidemiology*, 32(1):51–57, 2003.