INFERRING GENE-ENVIRONMENT INTERACTION FROM CASE-PARENT TRIO DATA: EVALUATION OF AND ADJUSTMENT FOR SPURIOUS $G\times E$ AND DEVELOPMENT OF A DATA-SMOOTHING METHOD TO UNCOVER TRUE $G\times E$

by

Ji-Hyung Shin

M.Sc., SIMON FRASER UNIVERSITY, 2004

B.Sc., SIMON FRASER UNIVERSITY, 2002

Thesis submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

in the

Department of Statistics and Actuarial Science

Faculty of Science

SIMON FRASER UNIVERSITY

Summer 2012

APPROVAL

Name: Ji-Hyung Shin

Degree: Doctor of Philosophy

Title of Thesis: Inferring gene-environment interaction from case-parent trio data: Evaluation of and adjustment for spurious $G\times E$ and development of a data-smoothing method to uncover true $G\times E$

Examining Committee:

Dr. Tim Swartz
Chair, Professor

Dr. Jinko Graham
Senior Supervisor, Associate Professor

Dr. Brad McNeney
Senior Supervisor, Associate Professor

Dr. Joan Hu
Supervisory Committee, Professor

Dr. Jiguo Cao
Internal Examiner, Assistant Professor

Dr. J. Concepción Loredo-Osti,External Examiner, Associate Professor, Department of Mathematics and Statistics, Memorial University

Date Approved: April 16, 2012

Partial Copyright Licence

Abstract

Most complex diseases are influenced jointly by genes (G) and environmental or non-genetic attributes (E). Gene-environment interaction ($G\times E$) is measured by statistical interaction between G and E, which occurs when genotype relative risks (GRRs) vary with E. In this thesis, we explore the sources of spurious $G\times E$ and propose a data-smoothing approach to $G\times E$ for case-parent trio data.

In the first project, we address the problem of making inference about $G\times E$ based on the transmission rates of alleles from parents to affected offspring. Since GRRs that vary with E lead to transmission rates that do too, transmission rates have been used to make inference about $G\times E$. However transmission-based tests of $G\times E$ are found to be invalid in general. To understand the bias of the transmission-based test, we derive theoretical transmission rates and compare their variation with E to that in the GRRs. Through simulation, we investigate the practical implication of the bias.

Valid approaches that are not based on transmission rates require specifying or are designed to work well under a parametric form for $G\times E$. In the second project, we develop a data-smoothing method to explore $G\times E$ that does not require model specification for the interaction component when we work with genotypes for a causal marker. The data-driven method produces graphical displays of $G\times E$ that suggest its form. For testing significance of $G\times E$, we take a permutation approach to account for the additional uncertainty introduced by the smoothing process.

For many approaches to inference of $G\times E$ with case-parent trio data, including our own, a key assumption is that the test marker is causal; however, in reality, it may not be causal but in linkage disequilibrium with a causal locus. In this case, the approaches can give a false impression of $G\times E$ due to a form of population

stratification that has not been appreciated well. In the final project, we investigate, through simulation, the source of the spurious $G \times E$ and propose an adjustment that uses additional unlinked markers genotyped in the affected offspring.

Keywords: Case-parent trios; gene-environment interaction; genotype relative risk; population stratification; generalized additive model; penalized maximum likelihood estimation

Acknowledgments

I would like to thank my supervisors Jinko Graham and Brad McNeney for their continuous help, support, guidance and encouragement during my time at SFU. I especially thank them for teaching me patiently how to be a good researcher.

I would like to thank my thesis committee members: Drs. J. Concepción Loredo-Osti, Jiguo Cao and Joan Hu, for their encouragement, insightful comments and questions. I also thank thank all the other professors, in the Department of Statistics and Actuarial Science, for not only helping me to learn the discipline well during my school years but also giving me valuable advice that will help me throughout my career life.

I would like to thank Sadika, Kelly and Charlene in the department for their help and support that made my student life much easier. I grateful for Pam, Brian, Kelly, Andy, Alyssa, Kody and all the other previous IRMACS staff for their help and support.

I thank my many friends at SFU who have helped me in many ways throughout the years, and with whom I've shared not only statistical knowledge but also many wonderful memories together outside the school: Aruni, Barbara, Caroll, Carolyn, Chunfang, Cindy, Dilnur, Donghong, Elizabeth, Flora, Huanhuan, Huijing, Jeong Eun, Jorge, Joslin, Kelly, Lihui, Lilian, Linda, Linnea, Matt, Maria, Pritam, Ruth, Ryan, Saman, Shrin, Simon, Soyean, Suli, Vivien Wilson, Yingying and Yuanzhen. I also thank all my other friends for all their prayers and encouraging words.

I would to thank my family for their endless love, care, understanding and prayers throughout my whole life and especially the past several years of my student life; for all these years, my dad has given me a ride to and picked up from the skytrain

station every morning and night, and my mom has packed me lunch/dinner that makes everyone jealous of me.

And, I thank God for giving me the opportunity to be here and to meet all these wonderful people through my PhD program!

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Complex diseases, such as diabetes or cancer, are thought to result from an interplay between genes G and environmental or non-genetic attributes E. The case-parent trio study design is often used for estimation and testing genetic effects and gene-by-environment interactions for such diseases. The design collects genotypes from unrelated children affected with a disease and also from their parents. Information may also be collected on environmental factors in the children. Genetic effects can be measured by genotype relative risk (GRR) in individuals with one genotype compared to those with some reference genotype. Under a log-additive penetrance model, statistical interaction between G and E, or $G\times E$, occurs when GRRs vary with the levels or the values of E. In this thesis, we explore the sources of spurious $G\times E$ and propose methods to uncover true $G\times E$ using data from a case-parent trio study.

The thesis consists of three projects. The work in Chapters 2 and 4 has been published. As a result, some introductory material and the description of the simulation settings are repeated in more than one chapter.

1.1 Overview of the thesis

Allelic transmission rates from parents to cases are frequently stratified by an environmental risk factor E and compared, with heterogeneity interpreted as $G\times E$. Although such transmission-based approaches to $G\times E$ are found to be invalid in

general under population stratification (Umbach and Weinberg, 2000), such analyses continue to appear. In Chapter 2, we revisit why heterogeneity is not equivalent to $G\times E$ in a range of settings not considered previously. The objective is a fuller understanding of the bias in transmission rates and what is driving it. Extending previously published findings of Umbach and Weinberg (2000), we derive parental mating-type probabilities in cases and use them to obtain transmission rates, which we then compare to $G\times E$. Through simulation, we investigate the practical implications of the bias for a transmission-based test of $G\times E$. For exploring $G\times E$, we suggest graphical displays of the transmission rates within parental mating types, as they are robust to population stratification and the penetrance model. This work has been published in Shin *et al.* (2010).

Numerous approaches have been proposed to assess $G\times E$ using data from case-parent trios (e.g., Schaid, 1999; Umbach and Weinberg, 2000; Lake and Laird, 2004). Many of these approaches require specifying a parametric regression model for $G\times E$, such as linearity, or are designed to work well under a specific form of $G\times E$. When the form of the underlying $G\times E$ differs from that specified by the regression model, or from the form for which an approach was designed, it can lead to bias and loss of statistical power. To address this issue, in Chapter 3, we develop a penalized maximum likelihood method to graphically explore the form of $G\times E$, under a generalized additive modelling framework (e.g., Wood, 2006). This data-smoothing approach offers the advantage of allowing the data to suggest the functional form of $G\times E$, rather than specifying it in advance. For testing $G\times E$, we adopt a permutation-based approach in order to account for the additional uncertainty introduced by the smoothing process. We investigate the statistical properties of the proposed permutation test through simulation. We also illustrate the use of the method with a simulated data set.

Many approaches to inference of $G\times E$ with data from case-parent trios, including that of the previous chapter, rely on genotypes G being measured at a causal locus and G being independent of E within families. Then, under the log-additive penetrance model, dependence of G and E within *affected* families is equivalent to

$G\times E$ (Umbach and Weinberg, 2000). At a causal locus, one may therefore, infer $G\times E$ from association between G and E within affected families. However, a test locus may in fact be in linkage disequilibrium with a causal locus. As noted by Shi *et al.* (2011), when genotypes G' are measured at a non-causal test locus, population stratification can create association between G' and E within affected families in the absence of $G\times E$. In Chapter 4, we describe this apparent interaction as a consequence of mis-specification of the penetrance model and population stratification. A log-additive penetrance model for the causal locus does not apply to the test locus. The mis-specification of the penetrance model for the test locus, together with population stratification, gives rise to G'-E dependence within affected families in the absence of $G\times E$. One design-based solution to avoid incorrectly inferring interaction involves collecting data on the environmental variable in an unaffected sibling of the affected child (Shi *et al.*, 2011). We propose an analysis-based solution that uses genotypes for random or ancestral informative markers in the affected child to adjust the penetrance model. Our approach does not require data on unaffected siblings and has been published in Shin *et al.* (2012).

In the last chapter, we make concluding remarks. Some of theoretical and simulation details for Chapters 1, 2 and 3 are provided in Appendices A, B and C, respectively.

Chapter 2

On the use of allelic transmission rates for assessing $G \times E$ in case-parent trios

2.1 Introduction

For many complex diseases, both genes (G) and environmental exposure or non-genetic attributes (E) act jointly to increase risk. For example, even though cigarette smoking is one of the most important risk factors for chronic obstructive pulmonary disease, only 10-20% of chronic smokers develop the disease, indicating the possible contribution of genetic factors (e.g., glutathione S-transferase gene family; Cheng *et al.*, 2004). For such diseases, failure to account for the interplay between G and E may lead to incorrect conclusions about their etiological roles. One way to measure the interplay between G and E is through variation with E in genotype relative risks (GRRs), which are ratios of disease risks compared between individuals with a genotype of interest and those with some reference genotype. We refer to this variation in GRRs as statistical interaction between G and E, or $G \times E$.

For data from case-parent trios, it is natural to work with allelic transmission

rates from parents to cases, as in the transmission disequilibrium test (TDT; Spielman *et al.*, 1993). It is then tempting to stratify transmission analyses by E and interpret heterogeneity with E in transmission rates as $G{\times}E$, assuming the robustness of the TDT to population stratification carries over. However, as noted by Umbach and Weinberg (2000), heterogeneity in transmission rates does not necessarily reflect heterogeneity in GRRs. These authors illustrated the point with a counter-example involving a recessive penetrance model and no $G{\times}E$, but transmission rates that vary with E because of population stratification. Their purpose was to motivate an alternate likelihood-based approach to inference of $G{\times}E$, which they subsequently discussed. The current investigation revisits their initial point about the non-equivalence of heterogeneity in transmission rates and GRRs. We feel that this non-equivalence is not widely appreciated because transmission analyses stratified by E continue to appear (e.g., Wang *et al.*, 2006; Bellgrove *et al.*, 2006; Brookes *et al.*, 2008; Du *et al.*, 2008; Ma *et al.*, 2009). Hence, there is a need to expand on some of the ideas touched on by Umbach and Weinberg. The current investigation aims to fulfil this need, and to gain further insight into how and why the bias in transmission rates arises when they are used as a proxy for GRRs.

In this work, we continue the line of investigation started by Umbach and Weinberg, and derive general expressions for the mating-type probabilities in the parents of cases under population stratification. We then use these expressions to compare the variation in transmission rates to that in the GRRs, under different penetrance models and levels of G-E dependence induced by population stratification. The comparison gives a fuller understanding of the bias in transmission rates and what is driving it. Along the way, we also clarify how to derive Umbach and Weinberg's expressions for the mating-type specific transmission rates. The practical implications of the bias in transmission rates are explored through a simulation study comparing the error rates of a transmission-based test for $G{\times}E$ to those of a likelihood-based test. We conclude by suggesting descriptive summaries for exploring $G{\times}E$. Unlike stratified transmission rates, these summaries are not biased by population stratification or by non-multiplicative penetrance models.

2.2 Models and Methods

Consider a single nucleotide polymorphism (SNP). Let G, G_M and G_F denote the number of copies of the putative risk allele carried by a child, his/her mother and father, respectively. Each of G, G_M and G_F can take a value from 0, 1 or 2. We assume symmetry of mating such that, for example, $(G_M, G_F) = (0, 1)$ has the same probability of occurring as $(G_M, G_F) = (1, 0)$. Under this assumption there are six distinctive parental mating types G_p as described in Appendix A.1. Let D denote the event that a child develops disease and E denote his/her continuously varying non-genetic attribute. Let $R_g(e)$ be the attribute-specific GRR of an individual with g copies of the risk allele compared to an individual with no copies. The details of the notation and model are given in Appendix A.1.

The transmission rate is defined to be the probability that a heterozygous parent transmits the risk allele to his/her affected child. The attribute-specific transmission rate $\tau(e)$ can be written as

$$\tau(e) \quad = \quad \sum_{m \in \mathcal{I}} \tau_m(e) w_m(e),$$

where $\mathcal{I}$ is the set of mating types with at least one heterozygous parent (reviewed in Appendix A.2). This is a weighted average of the mating-type-specific transmission rates $\tau_m(e)$, with the weight $w_m(e)$ being the proportion of heterozygous parents of cases with $E = e$ that come from mating type m (Umbach and Weinberg, 2000). In Table 1 of Umbach and Weinberg (2000), reproduced as our Table 2.1, $\tau_m(e)$ and $w_m(e)$ are written, respectively, in terms of the $R_g(e)$ and the proportion $\pi_m(e) \equiv \Pr(G_p = m \mid D, E = e)$ of cases with $E = e$ that come from mating type $G_p = m$. In Appendix A.3, we derive the expressions for $\tau_m(e)$.

By way of a numerical example, Umbach and Weinberg showed that, when GRRs do not vary with E (i.e. there is no $G \times E$), transmission rates can still vary with E because the weights $w_m(e)$ vary under G-E dependence. From the form of $w_m(e)$ in Table 2.1, we can see that they vary with E because of $\pi_m(e)$. Thus, to understand

Table 2.1: Components of the transmission rate of a risk allele from heterozygous parents to affected children with $E = e$ (Reproduced from Umbach and Weinberg, 2000.)

Informative mating type	Proportion of parents of cases that come from given mating type*	Proportion of heterozygous parents of cases that come from given mating type†	Mating-type-specific transmission rate
$(0,1)$	$\pi_{01}(e)$	$w_{01}(e) = \frac{\pi_{01}(e)}{d(e)}$	$\tau_{01}(e) = \frac{R_1(e)}{1+R_1(e)}$
$(1,2)$	$\pi_{12}(e)$	$w_{12}(e) = \frac{\pi_{12}(e)}{d(e)}$	$\tau_{12}(e) = \frac{R_2(e)}{R_1(e)+R_2(e)}$
$(1,1)$	$\pi_{11}(e)$	$w_{11}(e) = \frac{2\pi_{11}(e)}{d(e)}$	$\tau_{11}(e) = \frac{R_1(e)+R_2(e)}{1+2R_1(e)+R_2(e)}$

* Expressions for $\pi_m(e) \equiv \Pr(G_p = m \mid D, E = e)$ are given in equation (2.1).
† $d(e) \equiv \pi_{01}(e) + \pi_{12}(e) + 2\pi_{11}(e)$.

how transmission rates $\tau(e)$ vary with E, we need to understand how $\pi_m(e)$ does. Towards this goal, we derive expressions for $\pi_m(e)$ and use these to clarify how G-E dependence impacts $\tau(e)$. The model of G-E dependence that we use is the one considered by Umbach and Weinberg, in which dependence is induced by population stratification.

2.2.1 Example settings

In their example illustrating the bias in transmission rates for assessing $G\times E$, Umbach and Weinberg considered a structured population consisting of two equal-sized subpopulations, assuming E had no effect on disease risk and that a recessive gene affected the disease penetrance with a relative risk of size 3. For our investigation, we considered settings with or without $G\times E$ for dominant and recessive penetrance models, as defined in Appendix A.1, under both G-E dependence and independence in the general population. We did not consider multiplicative penetrance models because the variation with E in transmission rates is equivalent to that in GRRs, as reviewed in Appendix A.2 (see equation (A4)). The specific settings considered for our investigation are no $G\times E$ and G-E dependence, $G\times E$ and G-E dependence, and $G\times E$ and G-E independence. The setting with no $G\times E$ and G-E independence is

omitted because both the transmission rates and GRRs are constant in E in this case; that is, $\tau(e)$ reflects the (lack of) $G{\times}E$.

We discuss the details of the model and related notation in Appendix A.1. In the risk model, we set $\beta = \log(3)$ for GRRs in (A1), $f(e) \equiv 0$ under no $G{\times}E$ and $f(e) = -0.25e$ under $G{\times}E$. To induce G-E dependence due to population stratification (e.g., Figure 2.1) we considered a general population with two hidden subpopulations, denoted by $S = 0$ or 1, of equal sizes in which genotype frequencies follow Hardy-Weinberg proportions, and the risk allele frequencies are $q_0 = 0.1$ and $q_1 = 0.9$, as in the example of Umbach and Weinberg (2000). The general population is subdivided, with all subpopulations in Hardy-Weinberg equilibrium. The general population is therefore, subject to the well-known Wahlund effect (Li, 1955; Wahlund, 1928) in which the number of heterozygotes tends to be less than expected under Hardy-Weinberg equilibrium. For the non-genetic attribute E, we let the general population have a mean of 0 and variance of 1 and the subpopulations have a common variance σ^2. The conditional expected value $\mathbb{E}(E \mid S)$ is linear in the binary variable S, implying $\mathrm{V}(\mathbb{E}(E \mid S))/\mathrm{V}(E) = \rho^2_{ES}$, where ρ_{ES} is the correlation between E and S (Hogg *et al.*, 2005). Using this identity, the subpopulation-specific means are $\mathbb{E}(E \mid S = 0) \equiv \mu_0 = -\rho_{ES}$ and $\mathbb{E}(E \mid S = 1) \equiv \mu_1 = \rho_{ES}$, and their variances are $\sigma^2 = 1 - \rho^2_{ES}$. Within each subpopulation, we let E be normally distributed and independent of G. With fixed subpopulation-specific allele frequencies $q_0 = 0.1$ and $q_1 = 0.9$, one can show that $Cov(G, E) = 0.8 \times \rho_{ES}$, by first expressing $Cov(G, E) = \mathbb{E}(Cov(G, E \mid S)) + Cov(\mathbb{E}(G \mid S), \mathbb{E}(E \mid S))$. A similar conditioning calculation yields $\mathrm{V}(G) = 0.82$ for this population in which $\mathrm{V}(E) = 1$. Thus, $\rho_{GE} = \dfrac{0.8}{\sqrt{0.82}} \times \rho_{ES}$. Using this identity, we controlled the G-E correlation by varying ρ_{ES}. The values of ρ_{ES} considered were 0.2, 0.5 and 0.8, which correspond to ρ_{GE} values of 0.18, 0.44 and 0.71, respectively. Under G-E independence, we let $q_0 = q_1 = q$ with $q = 0.1, 0.5$ and 0.9, and $\rho_{ES} = 0$ (i.e., $\mu_0 = \mu_1 = 0$). The G-E dependence can arise from the same (hidden) population stratification responsible for the Wahlund effect; or it can arise by chance through genetic sampling of a random population. In this work, we view this dependence ρ_{GE} as a population-level parameter that is invariant to the sampling design or artefacts

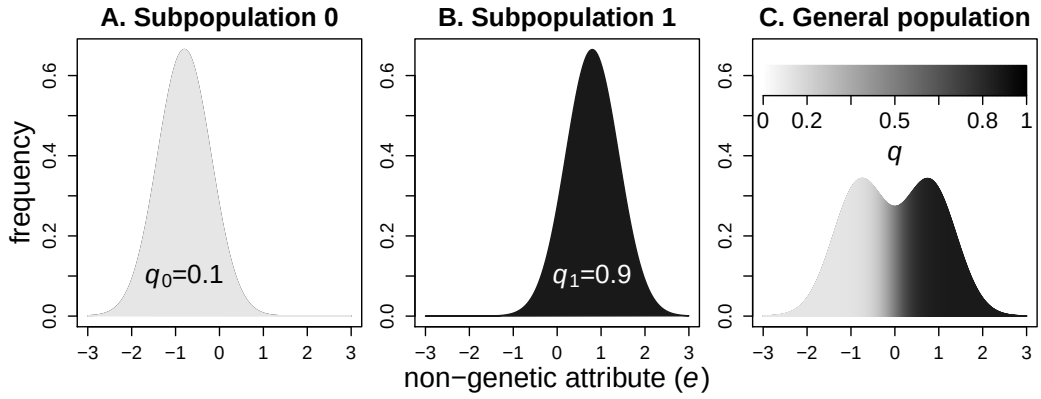

Figure 2.1: Inducing G-E dependence through population structure. The subpopulation-specific and population allele frequencies (q_0, q_1 and q, respectively) for the SNP G are indicated by the grey scale in panel C: higher allele frequencies are indicated by darker shadings and lower frequencies, by lighter shadings. The E-S correlation is $\rho_{ES} = 0.8$. As indicated in panels A and B, within each subpopulation, G and E occur independently. However, as shown in panel C, in the (combined) general population, G and E become dependent in the sense that, for the individuals with lower values of E, the risk-allele frequency tends to be lower, whereas for those with higher values of E, the risk-allele frequency tends to be higher. The resulting G-E correlation is $\rho_{GE} = 0.71$. Individuals with lower values of E are more likely to be from subpopulation 0 which has a lower risk-allele frequency, whereas those with higher values of E are more likely to be from subpopulation 1 which has a higher risk-allele frequency.

of statistical sampling from a fixed population.

2.2.2 Simulation study

The transmission rates can give a biased assessment of GRRs. To assess the practical implications of such bias, we evaluated the power of a transmission-based test of $G\times E$ by simulation. False-positive rates correspond to the power of tests under the no-$G\times E$ null hypothesis. False-negative rates correspond to one minus the power under the $G\times E$ alternative hypothesis. As previously noted (Umbach and Weinberg, 2000), these error rates are also influenced by incorrectly assuming independence of transmission events. The transmissions from two heterozygous parents to their affected child are independent only under a multiplicative penetrance model

or when the variant allele has no effect on disease risk. The transmission-based test was compared to a likelihood-based benchmark. Following Umbach and Weinberg (2000), we simulated 1000 affected trios with informative mating types according to the settings described in the previous subsection. The risk allele frequency in the general population was $q = 0.5$. Estimates of power were based on 10,000 simulation replicates, giving simulation errors of ≤ 0.01. Simulations were programmed in R (R Development Core Team, 2011).

For the transmission-based approach, the log-odds of transmitting the risk allele to an affected child were modelled as a linear function of E and the slope term was assessed via a likelihood-ratio test assuming independence of transmissions. For the likelihood-based approach, a conditional logistic regression was used to model the conditional probability of the affected child's genotype given E and the parental genotypes G_p (Schaid, 1999). A likelihood-ratio test was applied, based on a co-dominant penetrance model, as defined in Appendix A.1. For categorical E, this approach is equivalent to the log-linear modelling approach of Umbach and Weinberg (2000).

2.3 Results

The probabilities $\pi_m(e)$ of mating-types in parents of cases with $E = e$ can be written as

$$
\begin{aligned}
&\pi_m(e) \equiv \Pr(G_p = m \mid D, E = e) \\
&= \frac{\sum_{g \in \mathcal{G}_m} \Pr(G_p = m, D, G = g, E = e)}{\sum_{m' \in \mathcal{M}} \sum_{g' \in \mathcal{G}_{m'}} \Pr(G_p = m', D, G = g', E = e)} \\
&= \frac{\sum_{g \in \mathcal{G}_m} \Pr(D \mid G = g, E = e) \Pr(G = g \mid G_p = m) \Pr(G_p = m \mid E = e) \Pr(E = e)}{\sum_{m' \in \mathcal{M}} \sum_{g' \in \mathcal{G}_{m'}} \Pr(D \mid G = g', E = e) \Pr(G = g' \mid G_p = m') \Pr(G_p = m' \mid E = e) \Pr(E = e)} \\
&= \frac{\Pr(G_p = m \mid E = e) \sum_{g \in \mathcal{G}_m} \Pr(G = g \mid G_p = m) R_g(e)}{\sum_{m' \in \mathcal{M}} \Pr(G_p = m' \mid E = e) \sum_{g' \in \mathcal{G}_{m'}} \Pr(G = g' \mid G_p = m') R_{g'}(e)},
\end{aligned}
\tag{2.1}
$$

where $\mathcal{M}$ is the set of distinct parental mating types (see Appendix A.1). The first line of the equation is the definition of $\pi_m(e)$. The third line follows under conditional G-E independence given parental genotypes and no parent-of-origin effects, so that

$$\begin{aligned}&\Pr(G_p = m, D, G = g, E = e) = \\ &\quad \Pr(D \mid G = g, E = e) \Pr(G = g \mid G_p = m) \Pr(G_p = m \mid E = e) \Pr(E = e).\end{aligned}$$

The final line follows from dividing through by $\Pr(D \mid G = 0, E = e)$ in the numerator and denominator and factoring out the terms that do not depend on g. Even with GRRs that are constant in E, equation (2.1) shows that $\pi_m(e)$ can vary with E through the stratum-specific mating-type probabilities $\Pr(G_p = m \mid E = e)$.

Through equation (2.1), we see that $\Pr(G_p = m \mid E = e)$ can be as important as the GRRs in determining whether $\pi_m(e)$, and hence the overall transmission rates $\tau(e)$, vary with e. Under G-E independence within subpopulations,

$$\begin{aligned}&\Pr(G_p = m \mid E = e) \\ &\quad = \sum_s \Pr(G_p = m \mid S = s) \Pr(E = e \mid S = s) \Pr(S = s) / \Pr(E = e) \\ &\quad = \sum_s \Pr(G_p = m \mid S = s) \Pr(S = s \mid E = e).\end{aligned}$$

Thus, $\Pr(G_p = m \mid E = e)$ vary with e when both G_p and E depend on S; i.e., when there is G-E dependence in the overall population.

2.3.1 G-E dependence

In the hypothetical population of Figure 2.1, the population stratification induces G-E dependence. The resulting probabilities $\Pr(G_p = m \mid E = e)$ are shown in Figure 2.2. At lower values of E, parents are more likely to come from subpopulation 0 ($q_0 = 0.1$), which has a higher frequency of mating type $(0, 1)$ than mating types $(1, 1)$ and $(1, 2)$. At higher values of E, parents are more likely to come from subpopulation 1 ($q_1 = 0.9$), which has a higher frequency of mating type $(1, 2)$ than the other

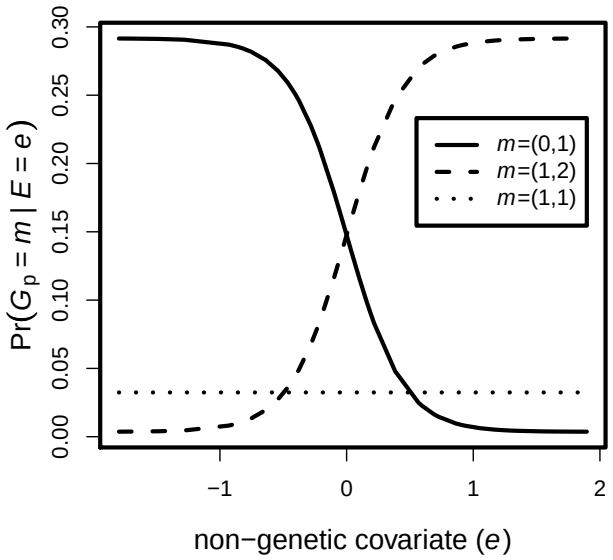

Figure 2.2: The mating-type probabilities $\Pr(G_p = m \mid E = e)$ in the parents of children with $E = e$ under G-E dependence with $\rho_{GE} = 0.71$.

informative mating types.

The impact of the stratified mating-type probabilities, $\Pr(G_p = m \mid E = e)$, on the weights, $w_m(e)$, in Table 2.1 is shown in Figure 2.3 for the setting with $G{\times}E$ and G-E correlation of $\rho_{GE} = 0.71$ under the dominant penetrance model. The weights under the recessive penetrance model are similar. The pattern in the weights closely mirrors that in the conditional mating-type probabilities. For example, at low values of E, parental mating type $(0, 1)$ is weighted heavily compared to the other two informative mating types. By contrast, at high values of E, parental mating type $(1, 2)$ is weighted heavily compared to the other two informative mating types.

Figures 2.4 and 2.5 illustrate the impact of $\Pr(G_p = m \mid E = e)$ on the transmission rates under $\rho_{GE} = 0.18, 0.44$ and 0.71. Define $R(e) = R_1(e) = R_2(e)$ under dominant penetrance and $R(e) = R_2(e)$ under recessive penetrance. Throughout, variation with E in $\text{logit}(\tau(e))$ is compared with that in $\log(R(e))$, as would be done under multiplicative penetrance based on the relationship (A4).

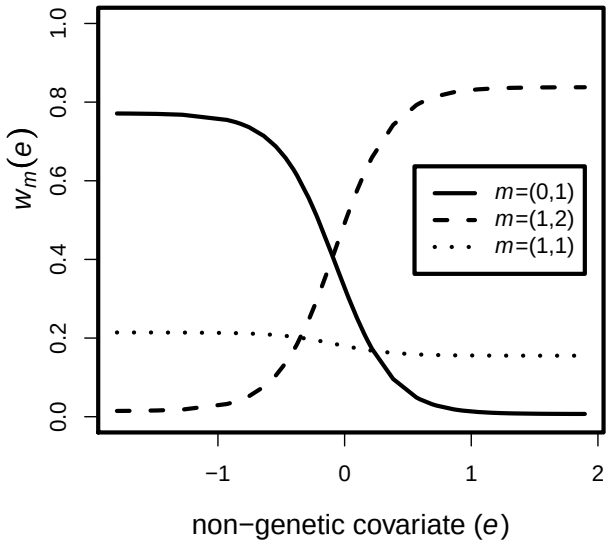

Figure 2.3: The weights $w_m(e)$ under no $G\times E$ and G-E correlation of $\rho_{GE} = 0.71$ for the dominant penetrance model.

Figure 2.4 plots logit$(\tau(e))$ versus e calculated under the setting where G and E are dependent in the absence of $G\times E$, for dominant (panel A) and recessive (panel B) penetrance models. From the figure, we can see logit$(\tau(e))$ varies with E while $\log(R(e))$ does not, as expected based on the form of $\pi_m(e)$ in equation (2.1). The figure also shows that logit$(\tau(e))$ varies more as G-E dependence increases. Hence, inferring $G\times E$ based on variation with E in transmission rates may lead to false-positive results.

Figure 2.5 plots logit$(\tau(e))$ versus e calculated under the setting where G and E are dependent in the presence of $G\times E$. As shown in the figure, the form of variation with E in logit$(\tau(e))$ differs from that in $\log(R(e))$ at any level of G-E dependence. The figure suggests that inferring $G\times E$ based on variation in transmission rates may lead to false-negative results in some cases; for example, logit$(\tau(e))$ curves under recessive penetrance with G-E correlations of 0.18 or 0.44 are relatively close to a horizontal line, which represents no variation in $\tau(e)$.

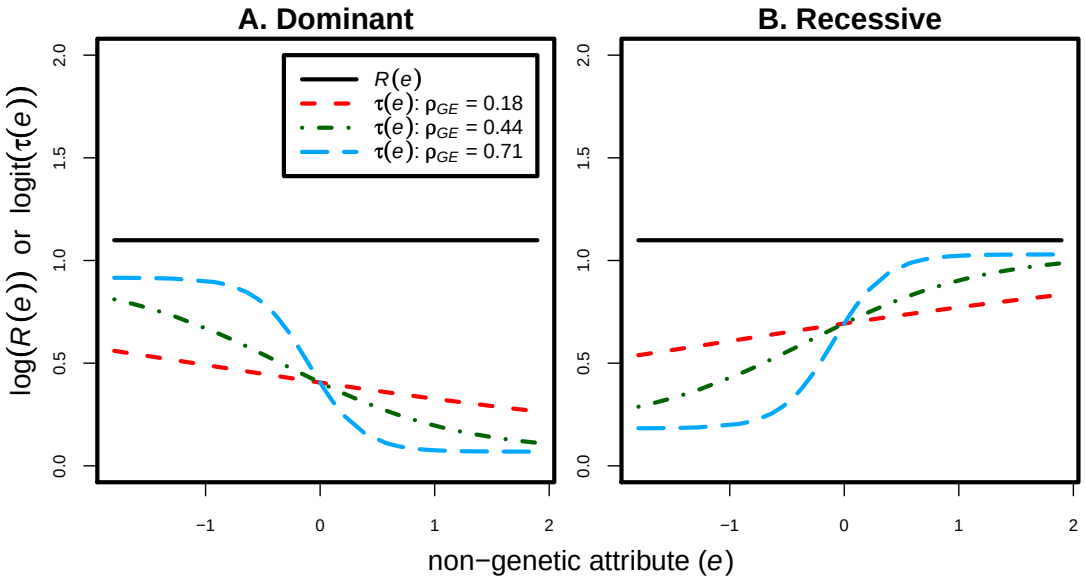

Figure 2.4: Variation in GRRs and transmission rates under no $G{\times}E$ and G-E dependence. Solid lines indicate $\log(R(e))$, which represents the true $G{\times}E$ measure. Broken lines indicate $\text{logit}(\tau(e))$ for different levels of G-E dependence, ρ_{GE}, induced by ρ_{ES}. Curves in the left column represent the variation for a dominant penetrance model, and those in the right column, the variation for a recessive model.

In the presence of $G{\times}E$, the weights $w_m(e)$ are very similar to those in the absence of $G{\times}E$ (results not shown), indicating that $\Pr(G_p = m \mid E = e)$ continues to drive their behaviour. Moreover, the patterns of variation in transmission rates with (Figure 2.5) and without (Figure 2.4) $G{\times}E$ are also similar. These similar patterns indicate that, even in the presence of $G{\times}E$, the behaviour of $\tau(e)$ can be determined by $\Pr(G_p = m \mid E = e)$ rather than by $G{\times}E$.

In summary, for the population stratification that we have considered, variation in the stratified mating-type probabilities, $\Pr(G_p = m \mid E = e)$, drives variation in the weights $w_m(e)$. In turn, the weights drive variation in the transmission rates $\tau(e)$. For example, in the dominant penetrance model, $\tau_{01}(e) > \tau_{11}(e) > \tau_{12}(e) \equiv 1/2$. From Figure 2.3, we see that, for the overall transmission rate, the large mating-type specific transmission rate $\tau_{01}(e)$ gets most of the weight at low values of e, whereas the small mating-type specific transmission rate $\tau_{12}(e)$ gets most of the weight at high

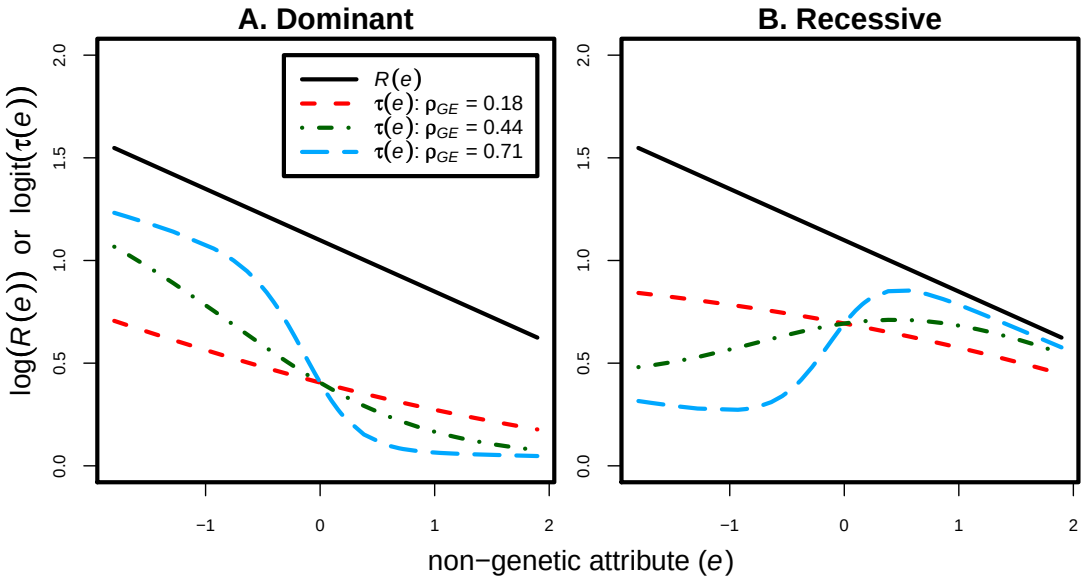

Figure 2.5: Variation in GRRs and transmission rates under $G{\times}E$ and G-E dependence. Solid lines indicate $\log(R(e))$, the true $G{\times}E$ measure. Broken lines indicate $\text{logit}(\tau(e))$ for different levels of G-E dependence, ρ_{GE}, induced by ρ_{ES}.

values. The influence of the weights and/or the probabilities $\Pr(G_p = m \mid E = e)$ can be seen in panel A of Figures 2.4 and 2.5. Similar arguments apply in the case of the recessive penetrance model.

2.3.2 G-E independence

Even when G and E are independent, transmission rates still may not reflect GRRs, as indicated by the odds of transmission under no population stratification:

$$\frac{\tau(e)}{1-\tau(e)} = \begin{cases} \dfrac{R(e)}{(1-q)+R(e)\cdot q} & \text{if dominant} \\ (1-q)+R(e)\cdot q & \text{if recessive,} \end{cases} \tag{2.2}$$

where q is the relative frequency of the risk allele in the population. The details of the derivations are provided in Appendix A.4. Hence, under the dominant (recessive) penetrance model, $\text{logit}(\tau(e))$ reflects $\log(R(e))$ correctly only as $q \to 0$ ($q \to 1$) in the presence of $G{\times}E$. To illustrate, Figure 2.6 shows $\text{logit}(\tau(e))$ for a high ($q = 0.9$) and

a low ($q = 0.1$) risk-allele frequency. In general, for this setting under $G{\times}E$ and G-E

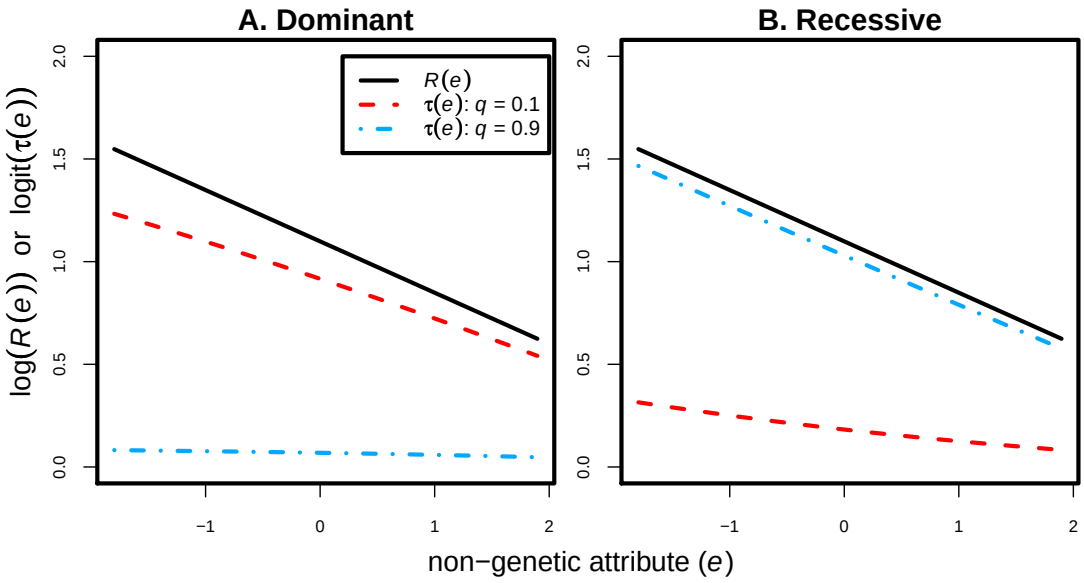

Figure 2.6: Variation in GRRs and transmission rates under $G{\times}E$ and G-E independence. Solid lines indicate $\log(R(e))$, the true $G{\times}E$ measure. Dashed lines indicate $\text{logit}(\tau(e))$ for a low population risk-allele frequency $q = 0.1$, and dotted-dashed lines indicate $\text{logit}(\tau(e))$ for a high risk-allele frequency $q = 0.9$.

independence, $\text{logit}(\tau(e))$ varies less than $\log(R(e))$. Hence, inferring $G{\times}E$ based on variation in $\text{logit}(\tau(e))$ may lead to false-negative results.

2.3.3 Simulation results

Table 2.2 summarizes the results of the simulation study. Under no $G{\times}E$ and G-E correlation, the false-positive error rate of the transmission-based test is inflated above the nominal 5% level. These results are consistent with the bias in transmission rates shown in Figure 2.4. By contrast, the error rates of the likelihood-based test is within simulation error of the nominal 5% level.

Under $G{\times}E$ and G-E correlation, we expect the transmission-based test to have greater power than the likelihood-based test given its grossly inflated false-positive error rates. In those cases when its power is less, the transmission rates vary noticeably less with E than the GRRs. For example, under recessive penetrance and $\rho_{GE} = 0.44$

Table 2.2: Estimated power of the transmission and likelihood tests under the example settings, with q =0.5. For each configuration, the column "error type" indicates whether errors under that configuration are false-positive (fp) or false-negative (fn) results. Under no $G\times E$, power estimates are false-positive (type 1 error) rates; under $G\times E$, they are one minus false-negative rates. Power estimates are based on 10,000 simulation replicates and the simulation error is ≤ 0.01. The nominal level of all tests is 5%.

					Power	
Setting	$G\times E$	penetrance	ρ_{GE}	error type	transmission	likelihood
1	no	dom	0.71	fp	1.000	0.047
			0.44	fp	0.909	0.047
			0.18	fp	0.251	0.050
		rec	0.71	fp	0.999	0.049
			0.44	fp	0.872	0.051
			0.18	fp	0.246	0.050
2	yes	dom	0.71	fn	1.000	0.284
			0.44	fn	0.997	0.414
			0.18	fn	0.670	0.474
		rec	0.71	fn	0.794	0.539
			0.44	fn	0.092	0.684
			0.18	fn	0.368	0.736
3	yes	dom	0	fn	0.177	0.487
		rec	0	fn	0.884	0.830

or 0.18 the curves for transmission rates in Figure 2.5, panel B, are relatively flat. The transmission-based test relies on a linear approximation to the log-odds of transmission; its power depends on the slope of this approximation. Accordingly, we see that, under the recessive penetrance model, power of the transmission-based approach is lowest for $\rho_{GE} = 0.44$, the configuration in which a linear approximation to the transmission rates has slope closest to zero. It is interesting to note that, when the penetrance model is recessive and $\rho_{GE} = 0.44$ or 0.71, the transmission-based test displays the perverse behaviour of rejecting more often under the null hypothesis than under an alternative hypothesis. Specifically, the transmission-based test has substantially larger type 1 error rates (0.872 and 0.999, respectively) than its power under the specified alternative (0.092 and 0.794, respectively).

Under no $G\times E$ and no G-E correlation (results not shown), the false-positive error rates for the dominant penetrance model are 0.034 and 0.048, for the transmission-based and likelihood tests, respectively; for the recessive penetrance model, the error rates are 0.067 and 0.052. In this setting, the variation in transmission rates is an unbiased reflection of that in the GRRs. In this case, the bias in the transmission-based test arises from incorrectly assuming that the transmission events of parents are independent. By contrast, in the previous two example settings, the transmission-based test was biased under population stratification.

Under $G\times E$ and no G-E correlation, the power of the transmission-based test is lower than that of the likelihood-based test for the dominant penetrance model. This lower power is consistent with the conservative type 1 error rate of this test and the conservative bias in the logit($\tau(e)$) curves under under dominant penetrance (see Figure 2.6, panel A). By conservative bias in the logit($\tau(e)$) curves, we mean curves that are closer to horizontal than the log($R(e)$) curve. For the recessive penetrance model, a slight conservative bias in the transmission rates is countered by the anti-conservative nature of the transmission-based test.

2.4 Discussion

With the case-parent trio design and no $G\times E$, the TDT is an attractive test for genetic association, as it is robust to population structure, regardless of the penetrance mode of the underlying disease (Spielman *et al.*, 1993). However, as noted by Umbach and Weinberg (2000), extensions which detect $G\times E$ based on variation with E in allelic transmission rates are not robust to population stratification unless the disease risk follows a multiplicative penetrance model. To illustrate this point, Umbach and Weinberg provided a counter-example involving a recessive penetrance model and no $G\times E$. They expressed the transmission rate in terms of the GRRs and the mating-type probabilities $\pi_m(e)$ of parents of cases. In this paper, we have investigated this point more extensively, using a wider variety of settings, for a continuously-varying E. To do so, we derived expressions for $\pi_m(e)$ and used them to obtain theoretical

transmission rates. These rates were then compared to GRRs for various penetrance models and *G*-*E* correlations induced by population stratification. This comparison enabled a fuller understanding of the bias in transmission rates and what is driving it.

We showed that, when G and E are dependent, the stratum-specific mating-type probabilities $\Pr(G_p = m \mid E = e)$ can drive the weighting of the mating-type-specific transmission rates $\tau_m(e)$ when determining the overall transmission rate $\tau(e)$. As a result, $\tau(e)$ varies with e in the absence of $G{\times}E$ (Figure 2.4) and varies with e to a greater or less extent than GRRs in the presence of $G{\times}E$ (Figure 2.5). When G and E are independent, $\tau(e)$ depends on the GRRs and the variant allele frequency and varies less with e than the GRRs under $G{\times}E$ (Figure 2.6).

The practical implications of such bias were investigated through a simulation study. We have reported results for simulation configurations with $\rho_{GE} > 0$ and decreasing interaction parameter $f(e)$; similar conclusions (results not shown) are obtained for $\rho_{GE} < 0$ and $f(e)$ increasing. For simulation settings with notable bias in transmission rates, the error rates of the transmission-based test were inflated relative to those of the likelihood-based test. For example, the false-positive error rates of the transmission-based test were grossly inflated above the nominal 5% level for higher levels of *G*-*E* dependence, whereas those of the likelihood-based test matched the nominal level. As another example, the false-negative error rates of the transmission-based test were inflated relative to those of the likelihood-based test in a recessive penetrance model, under moderate levels of *G*-*E* dependence.

Our results reinforce the message that transmission-based analyses of $G{\times}E$ can be misleading. This message applies not only to tests but also to descriptive summaries. For example, a common descriptive summary involves pooling the transmissions from informative parental mating types and graphically comparing the observed transmission rates across strata for E. Heterogeneity in the stratified transmission rates is taken to be suggestive of $G{\times}E$. Figure 2.7A gives an example of such a graphical display for a simulated data set under no $G{\times}E$ with $\rho_{GE} = 0.71$ and a dominant penetrance model. There is a striking but erroneous impression of $G{\times}E$ due to the

population stratification. To avoid such bias, we suggest comparing the transmission rates within parental mating types instead, as they depend on the GRRs only (Table 2.1). Within a mating type $G_p = m$, let the observed transmission rate for the stratum defined by $E = e$ be $\hat{\tau}_m(e)$ and let $\hat{V}_m(e)$ be an estimate of its variance given by equation (A8) in Appendix A.5. Then the suggested display is of $\hat{\tau}_m(e) \pm 2\sqrt{\hat{V}_m(e)}$ across the strata, within a parental mating type m. Figure 2.7B illustrates this display using the same simulated data set shown in Figure 2.7A. The display of transmission rates within mating types in Figure 2.7B is robust to the population stratification, whereas the display of pooled transmission rates in Figure 2.7A is not. We stress that the display of transmission rates within mating types is intended only as a descriptive summary of the data. Inference of $G{\times}E$ should be based on valid statistical approaches developed for this purpose (e.g., Lake and Laird, 2004; Cordell *et al.*, 2004; Lim *et al.*, 2005).

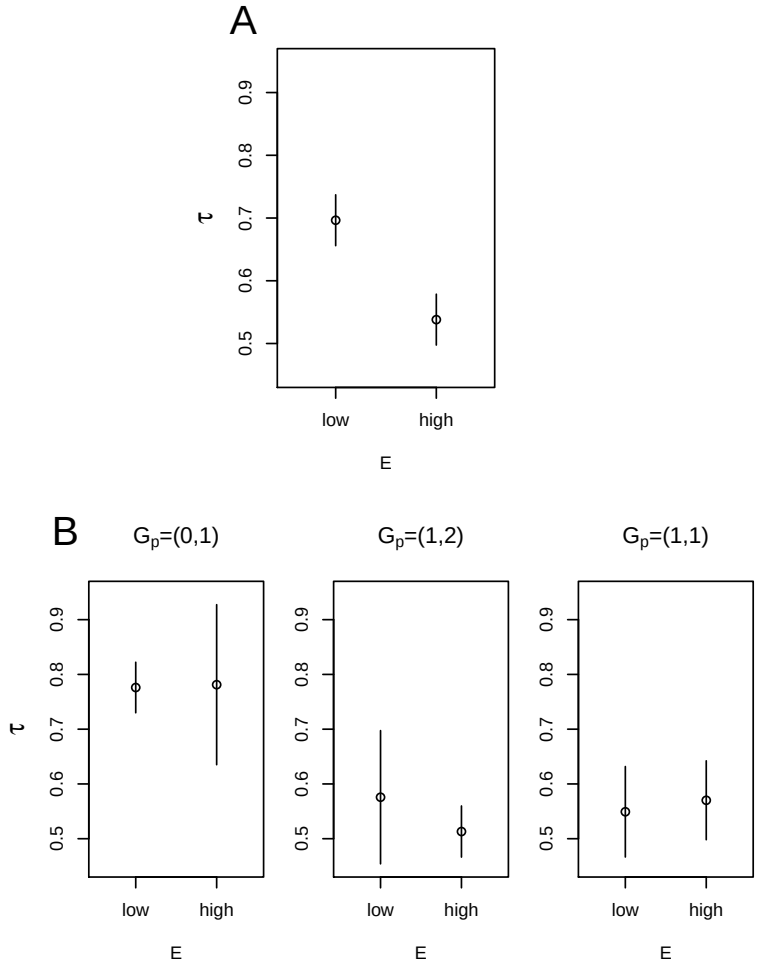

Figure 2.7: Observed transmission rates stratified by low (< 0) and high (≥ 0) values of E, for simulated data from 1000 case-parent trios under no $G{\times}E$ with $\rho_{GE} = 0.71$ and a dominant penetrance model. Panel A shows the transmission rates for pooled data, and Panel B, those within each informative parental-mating-type. Lines represent approximate 95% confidence intervals calculated under a multiplicative penetrance model.

Chapter 3

A data smoothing method to uncover gene-environment interaction using data from case-parent trios

3.1 Introduction

A case-parent trio study collects the genotypes of unrelated affected children and their parents. Information on cases' non-genetic covariates can also be collected. The design allows conditioning on parent genotypes, which has the effect of creating family-based controls matched to the case for ancestry; the inference of genetic effects is robust to population stratification.

The joint effect of genetic and non-genetic factors, that is gene-by-environment interaction ($G\times E$), is often of interest. As for the genetic effects, conditioning on parent genotypes G_p provides robust inference of $G\times E$ from case-parent trio data against population stratification when the test marker is causal (e.g., Umbach and Weinberg, 2000). Various approaches have been developed to examine $G\times E$ using data from case-parent trios. Such approaches include the log-linear modelling method

(Umbach and Weinberg, 2000), the conditional logistic regression (e.g., Schaid, 1999) and the family-based association test of interaction (FBAT-I; Lake and Laird, 2004). These methods condition on parent genotypes for assessing $G\times E$ in data from case-parent trios.

However, there are issues with these approaches. For example, the log-linear modelling approach can only handle a categorical non-genetic covariate. Therefore, if E is a continuous covariate, it needs to be categorized, which results in a loss of information. Conditional logistic regression can handle a continuous E, but it needs to assume a parametric form (e.g., linear) for the $G\times E$ model. When the interaction model is mis-specified, it can lead to invalid conclusions about $G\times E$. FBAT-I does not assume any parametric model for $G\times E$; however, it uses a test statistic that works best when G and E are linearly associated. The test also needs to specify the mode of inheritance, which can also lead to mis-leading inference about $G\times E$ when the mode is mis-specified.

To address such issues, we develop a penalized maximum likelihood method to assess $G\times E$, using a generalized additive modelling framework (e.g., Wood, 2006). The proposed method does not require specification of either the $G\times E$ model or the mode of inheritance. The resulting point and interval estimates may be used to displayed to graphically explore the form of $G\times E$ and the mode of inheritance. For assessing the significance of $G\times E$, we adopt a permutation-based approach that takes into account of the additional uncertainty introduced by the smoothing process. A simulation study is conducted to evaluate the type 1 error rates and the statistical power of the proposed test under various scenarios. We also compare the power of the proposed test to that of the other available methods mentioned above. For the simulation study, we generate and use datasets with a large sample size; the power of our permutation test is expected to be low since it is difficult to detect $G\times E$ in general (Smith and Day, 1984; Dempfle *et al.*, 2008), and on top of that, we only make minimal assumptions about the $G\times E$ model.

3.2 Model

Let G denote the number of copies of the index allele for a SNP carried by an individual, which can take a value from 0, 1, or 2; E, his/her continuously varying non-genetic covariate; and D denote the binary indicator of his/her disease status ($D = 1$ if affected). Let G_M and G_F denote the numbers of the copies of the index allele carried by the mother and the father of the individual.

We assume mating symmetry and Mendelian segregation. Under these assumptions, the *informative* mating type G_p can take a value from 1, 2 or 3, indicating $(G_M, G_F) = \{(0,1) \text{ or } (1,0)\}$, $\{(1,2) \text{ or } (2,1)\}$ and $\{(1,1)\}$, respectively. G and E are assumed to be conditionally independent given G_p. For disease risk probability, we assume the following log-additive model (Shin *et al.*, 2010):

$$P(D = 1 \mid G = g, E = e) = \exp\{k + \boldsymbol{z}(g)\boldsymbol{\gamma} + \xi(e) + \boldsymbol{z}(g)\boldsymbol{f}(e)\}, \tag{3.1}$$

where k is the baseline disease probability, $\boldsymbol{z}(g) = (z_1(g), z_2(g))$ where $z_1(g)$ and $z_2(g)$ are indicator variables for $g > 0$ and $g = 2$, representing the co-dominant genetic coding; $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)^{\mathsf{T}}$, where γ_1 and γ_2 represent genetic main effect; $\xi(e)$, an unspecified smooth function of E representing the non-genetic main effect; and $\boldsymbol{f}(e) = (f_1(e), f_2(e))^{\mathsf{T}}$ where $f_1(e)$ and $f_2(e)$ are unspecified smooth functions of E. $G\times E$ occurs when genotype relative risks (GRRs) vary with values of non-genetic covariates. The parameterization in model (3.1) focuses on the idea of differential genetic effects on the disease risk due to differential gene dose effects, which is a natural interpretation in our context. Other parameterizations that allow for two GRRs are also possible, and one such example is presented in Appendix B.1.

Under the risk model (3.1), we have

$$\begin{aligned} \log\left(\mathrm{GRR}_1(e)\right) &\equiv \log\left\{\frac{P(D = 1 \mid G = 1, E = e)}{P(D = 1 \mid G = 0, E = e)}\right\} = \gamma_1 + f_1(e); \text{ and} \\ \log\left(\mathrm{GRR}_2(e)\right) &\equiv \log\left\{\frac{P(D = 1 \mid G = 2, E = e)}{P(D = 1 \mid G = 1, E = e)}\right\} = \gamma_2 + f_2(e). \end{aligned} \tag{3.2}$$

The smooth functions $f_1(e)$ and $f_2(e)$ model $G\times E$ since GRR's depend on E through them. When $f_1(e) = f_2(e) = 0$ for all values $E = e$, it indicates there is no $G\times E$. When $f_1(e)$ and/or $f_2(e)$ vary with e, it indicates that there is $G\times E$.

Under different genetic inheritance modes, $f_1(e)$ and $f_2(e)$ behave differently. Under dominant models, $f_1(e)$ varies with E, but $f_2(e) = 0$ since

$$\frac{P(D=1 \mid G=1, E=e)}{P(D=0 \mid G=1, E=e)} = \frac{P(D=2 \mid G=1, E=e)}{P(D=0 \mid G=1, E=e)} \neq 1 \tag{3.3}$$

$$\Updownarrow$$

$$\log(\mathrm{GRR}_1(e)) \neq 0, \quad \log(\mathrm{GRR}_2(e)) = 0.$$

Under multiplicative or log-additive models, both functions vary with E in the same way (i.e., $f_1(e) = f_2(e)$) since

$$\frac{P(D=1 \mid G=1, E=e)}{P(D=0 \mid G=1, E=e)} = \frac{P(D=2 \mid G=1, E=e)}{P(D=1 \mid G=1, E=e)} \neq 1 \tag{3.4}$$

$$\Updownarrow$$

$$\log(\mathrm{GRR}_1(e)) = \log(\mathrm{GRR}_2(e)) \neq 0.$$

Under recessive models, $f_1(e) = 0$, but $f_2(e)$ varies with E since

$$\frac{P(D=1 \mid G=1, E=e)}{P(D=0 \mid G=1, E=e)} = 1, \frac{P(D=2 \mid G=1, E=e)}{P(D=0 \mid G=1, E=e)} \neq 1 \tag{3.5}$$

$$\Updownarrow$$

$$\log(\mathrm{GRR}_1(e)) = 0, \log(\mathrm{GRR}_2(e)) \neq 0.$$

For the case-parent trio design, the data are ascertained conditional on $D = 1$, and hence the likelihood for the observed data for a single family is

$$\begin{aligned} & P(G=g, E=e, G_p=m \mid D=1) \\ = \;& P(G=g \mid E=e, G_p=m, D=1) \cdot P(E=e, G_p=m \mid D=1). \end{aligned} \tag{3.6}$$

However, unconditional inference based on the joint probability distribution $P(G =$

$g, E = e, G_p = m \mid D = 1)$ requires knowledge about the joint distribution of E and G_p, which is not available from case-parent trios. An alternative way is to make conditional inference by conditioning on E and G_p (e.g., Schaid, 1999), such that the likelihood is based only on the first factor in equation (3.6). Conditioning on E and G_p would result in loss of information (e.g., Liang, 1983) due to ignoring the second factor in equation (3.6), which also contains information on $G{\times}E$. However, Moerkerke *et al.* (2010) have shown that the conditional inference is asymptotically efficient under linear $G{\times}E$. Hence, we expect that the loss of information about $G{\times}E$ from conditioning on G_p and E would be minimal, provided that the sample size is big enough.

For the purpose of writing the likelihood based on $P(G = g \mid E = e, G_p = m, D = 1)$, it is convenient to introduce a binary variable Y_{mjg} coding the genotype of the affected child in j^{th} trio from m^{th} mating type, such that

$$Y_{mjg} = \begin{cases} 1 & \text{if the child has } G = g, \\ 0 & \text{otherwise} \end{cases}.$$

Y_{mjg} are mutually independent, and if two affected children within a mating type have the same value of $E = e$, their responses are identically distributed.

The mean responses $\mu_{mg}(e)$ are

$$\mu_{mg}(e) \equiv \mathrm{E}(Y_{mgj} \mid E = e) = P(G = g \mid E = e, G_p = m, D = 1),$$

for which the expressions under the considered assumptions are provided in Table 3.1. Note that the baseline parameter k and the non-genetic main effect parameter $\xi(e)$ of the disease risk model (3.1) are not estimable since, as shown in Table 3.1, they are cancelled out in the calculation of $P(G = g \mid E = e, G_p = m, D = 1)$ on which we base our conditional likelihood inference. However, we are not concerned with this because k and $\xi(e)$ are nuisance parameters in an analysis of $G{\times}E$.

Assuming no mutation, $G = 2$ is impossible for trios from $G_p = 1$, and $G = 0$ is impossible for those from $G_p = 2$. Therefore, $\mu_{12}(e)$ and $\mu_{20}(e)$ are not defined, as

Table 3.1: Expressions for $\mu_{mg}(e) \equiv P(G = g \mid E = e, G_p = m, D = 1)$ under the assumptions.

Mating type (m)	Genotype (g)		
	0	1	2
1	$\frac{1}{1+\exp(\gamma_1 + f_1(e))}$	$\frac{\exp(\gamma_1 + f_1(e))}{1+\exp(\gamma_1 + f_1(e))}$	–
2	–	$\frac{1}{1+\exp(\gamma_2 + f_2(e))}$	$\frac{\exp(\gamma_2 + f_2(e))}{1+\exp(\gamma_2 + f_2(e))}$
3[a]	$\frac{1}{d(\gamma_1,\gamma_2,f_1(e),f_2(e))}$	$\frac{2\exp(\gamma_1 + f_1(e))}{d(\gamma_1,\gamma_2,f_1(e),f_2(e))}$	$\frac{\exp(\gamma_1 + f_1(e) + \gamma_2 + f_2(e))}{d(\gamma_1,\gamma_2,f_1(e),f_2(e))}$

[a] $d(\gamma_1,\gamma_2,f_1(e),f_2(e)) \equiv 1 + 2\exp(\gamma_1 + f_1(e)) + \exp(\gamma_1 + f_1(e) + \gamma_2 + f_2(e))$

indicated in Table 3.1. Letting $\boldsymbol{\mu}_1(e) \equiv (\mu_{10}(e), \mu_{11}(e))^{\mathsf{T}}$, $\boldsymbol{\mu}_2(e) \equiv (\mu_{21}(e), \mu_{22}(e))^{\mathsf{T}}$ and $\boldsymbol{\mu}_3(e) \equiv (\mu_{30}(e), \mu_{31}(e), \mu_{32}(e))^{\mathsf{T}}$, the log-likelihood function can be expressed as a sum of three components:

$$\begin{aligned} l(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3) &= \sum_{j=1}^{n_1} [y_{1j0}\log(\mu_{10}(e_{1j})) + y_{1j1}\log(\mu_{11}(e_{1j}))] \\ &\quad + \sum_{j=1}^{n_2} [y_{2j1}\log(\mu_{21}(e_{2j})) + y_{2j2}\log(\mu_{22}(e_{2j}))] \\ &\quad + \sum_{j=1}^{n_3} [y_{3j0}\log(\mu_{30}(e_{3j})) + y_{3j1}\log(\mu_{31}(e_{3j})) + y_{3j2}\log(\mu_{32}(e_{3j}))] \\ &\equiv \sum_{j=1}^{n_1} l_{1j}(\boldsymbol{\mu}_1(e_{1j})) + \sum_{j=1}^{n_2} l_{2j}(\boldsymbol{\mu}_2(e_{2j})) + \sum_{j=1}^{n_3} l_{3j}(\boldsymbol{\mu}_3(e_{3j})). \end{aligned}$$

where n_m is the number of affected trios from $G_p = m$, and $l_{mj}(\cdot)$ is a log-likelihood contribution for a trio from $G_p = m$. The first two components of the log-likelihood function are binomial log-likelihoods, and the last one is a trinomial log-likelihood.

From Table 3.1, we can see that $\boldsymbol{\mu}_1(e)$ involves γ_1 and $f_1(e)$, and $\boldsymbol{\mu}_2(e)$, γ_2 and $f_2(e)$, while $\boldsymbol{\mu}_3(e)$ involves γ_1, γ_2, $f_1(e)$ and $f_2(e)$. Consequently, the log-likelihood

function can be re-expressed in terms of GRR-model parameters as

$$l(\gamma_1, \gamma_2, f_1(e), f_2(e)) = \sum_{j=1}^{n_1} l_{1j}\left(\gamma_1, f_1(e_{1j})\right) + \sum_{j=1}^{n_2} l_{2j}\left(\gamma_2, f_2(e_{2j})\right) + \sum_{j=1}^{n_3} l_{3j}\left(\gamma_1, \gamma_2, f_1(e_{3j}), f_2(e_{3j})\right), \tag{3.7}$$

which indicates that the trios from $G_p = 1$ are used for estimating γ_1 and $f_1(e)$, those from $G_p = 2$, for γ_2 and $f_2(e)$; and those from $G_p = 3$ are used for all the parameters $\gamma_1, \gamma_2, f_1(e)$ and $f_2(e)$.

3.3 Methods

3.3.1 Penalized likelihood setup

For modelling the smooth $G\times E$ functions $f_1(e)$ and $f_2(e)$, we consider natural cubic splines, respectively, with K_1 and K_2 knots. The K_1 and K_2 knots are selected based on the observed E in the trios from mating types $G_p = 1, 2$ and those from $G_p = 2, 3$, respectively. Under a penalized estimation framework with a fixed basis dimension, the exact number and positions of the knots do not contribute much impact on the resulting fit, as long as the basis dimension is large so that there are enough degrees of freedom for representing the true function (Wood, 2006).

Since $f_1(e)$ and $f_2(e)$ are not expected be complex, by default, we assume $K_1 = K_2 = 5$ are enough to represent the smooth functions. In order to make good use of the data, we let them be distributed evenly through out the data by placing them at sample quantiles. For example, under the default numbers of knots, the we place three interior knots at the 25^{th}, 50^{th} and 75^{th} quantiles and two boundary knots, at the endpoints of the data.

With the chosen K_h knots, $f_h(e)$ can be expressed as

$$f_h(e) = \sum_{k=1}^{K_h} b_{hk}(e) c^*_{hk} = X^*_h(e)\boldsymbol{c}^*_h, \qquad \text{for } h = 1, 2, \tag{3.8}$$

where the k^{th} basis function $b_{hk}(e)$is evaluated at $E = e$ based on the K_h knots, c^*_{hk} is the corresponding coefficient, $X^*_h(e) = [b_{h1}(e), ..., b_{hk}(e)]$, and $\boldsymbol{c}^*_h = (c^*_{h1}, ..., c^*_{hK_h})^\mathsf{T}$. There are several ways to define a natural cubic spline function and hence the basis function vector $X^*_h(e)$ (e.g., Wood, 2006). We show one definition of $X^*_h(e)$ in equation (B2) in Appendix B.2.

The roughness penalties associated with f_h, for $h = 1, 2$ are measured by the integrated squared second derivative

$$\int \{f''_h(e)\}^2 de \;=\; \boldsymbol{c}^{*\mathsf{T}}_h \mathbf{S}^*_h \boldsymbol{c}^*_h, \tag{3.9}$$

where $f''_h(e)$ is the second derivative function of $f_h(e)$, $\mathbf{S}^*_h$ is the $K_h \times K_h$ penalty matrix. The $(i, j)^{th}$ elements of $\mathbf{S}^*_h$ are $\{s^*_{h,ij}\} = \int b''_{hi}(e) b''_{hj}(e) de$, for which $b''_{hi}(e)$ is the second derivative function of the i^{th} basis function evaluated at $E = e$ (Wood and Augustin, 2002).

To identify the genetic main effect terms γ_1 and γ_2 from log-GRR$_h(e)$ in equations (3.2), we impose the following two constraints for $h = 1, 2$ that the sums of $f_h(e_{mj})$ over all observed covariate values of cases in trios from mating type $m = h$ or 3 are zero:

$$\sum_{m\in\{h,3\}} \sum_j f_h(e_{mj}) = \sum_k \sum_{m\in\{h,3\}} \sum_j b_{hk}(e_{mj}) c^*_{hk} = \mathbf{C}_h \boldsymbol{c}^*_h = 0, \tag{3.10}$$

where $\mathbf{C}_h$ is the $1\times K_h$ matrix with k^{th} element $C_{hk} = \sum_{m\in\{h,3\}} \sum_j b_{hk}(e_{mj})$. The fitting problem may be reparameterized in terms of a new basis coefficient vector $\boldsymbol{c}_h$ of length $(K_h - 1)$ induced by the constraints (3.10), by letting

$$\boldsymbol{c}^*_h = \mathbf{A}_h \boldsymbol{c}_h.$$

Thus, the constraints (3.10) will be automatically satisfied if $\mathbf{A}_h$ can be chosen such that

$$\mathbf{C}_h \mathbf{A}_h = \mathbf{0}. \tag{3.11}$$

One way to find an $\mathbf{A}_h$ is to use the QR-decomposition on $\mathbf{C}^\mathsf{T}_h$ (Section 1.8.1 Wood,

2006); that is, we write

$$\mathbf{C}_h^{\mathsf{T}} = \mathbf{Q}_h \mathbf{R}_h = \Big[\mathbf{Q}_{1h} : \mathbf{Q}_{2h}\Big] \mathbf{R}_h,$$

where $\mathbf{Q}_h$ is a $K_h \times K_h$ orthogonal matrix, and $\mathbf{R}_h$ is a $K_h \times 1$ (upper triangular) matrix. Let $\mathbf{Q}_h$ be partitioned into two parts: $\mathbf{Q}_{1h}$, a matrix containing the first column of $\mathbf{Q}_h$ and $\mathbf{Q}_{2h}$, a matrix containing the last $(K_h - 1)$ columns. Then, setting $\mathbf{A}_h = \mathbf{Q}_{2h}$ will lead to equation (3.11) since the columns of $\mathbf{Q}_{2h}$ are in the null space of $\mathbf{C}_h$ (Fundamental Theorem of Linear Algebra).

Hence, letting $X_h(e) = X_h^*(e)\mathbf{A}_h$, we obtain

$$f_h(e) = X_h(e)\boldsymbol{c}_h,$$

while satisfying the constraints on $f_h(e)$ in equation (3.10). Consequently, the log-likelihood function in (3.7) can be re-written in terms of the parameter vector $\boldsymbol{\beta} = (\gamma_1, \boldsymbol{c}_1^{\mathsf{T}}, \gamma_2, \boldsymbol{c}_2^{\mathsf{T}})^{\mathsf{T}}$. Similarly, letting $S_h = \mathbf{A}_h^{\mathsf{T}} \mathbf{S}_h^* \mathbf{A}_h$, we obtain the corresponding roughness penalty

$$\int \{f_h''(e)\}^2 de = \boldsymbol{c}_h^{\mathsf{T}} S_h \boldsymbol{c}_h.$$

The roughness penalty can be further re-expressed in terms of the parameter vector $\boldsymbol{\beta}$ as

$$\int \{f_h''(e)\}^2 de = \boldsymbol{\beta}^{\mathsf{T}} \mathbf{S}_h \boldsymbol{\beta},$$

by letting the penalty matrix $\mathbf{S}_h$ be $(K_1 + K_2)\times(K_1 + K_2)$ square matrices

$$\mathbf{S}_1 = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & s_{1,11} & \cdots & s_{1,1(K_1-1)} & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & s_{1,(K_1-1)1} & \cdots & s_{1,(K_1-1)(K_1-1)} & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \end{bmatrix},$$

and

$$\mathbf{S}_2 = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & s_{2,11} & \cdots & s_{2,1(K_2-1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & s_{2,(K_2-1)1} & \cdots & s_{2,(K_2-1)(K_2-1)} \end{bmatrix},$$

where $\{s_{h,ij}\}$ represent the elements of S_h. Consequently, the penalized log-likelihood function can be written in terms of $\boldsymbol{\beta}$ as

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2}\sum_{h=1}^{2} \lambda_h \boldsymbol{\beta}^{\mathsf{T}} \mathbf{S}_h \boldsymbol{\beta},$$

where λ_h represents the smoothing parameter that controls the trade-off between the fit to the observed data (i.e., bias) and smoothness (i.e., variance) for the estimator of $f_h(e)$.

3.3.2 Penalized maximum likelihood estimation

To write the additive predictors for estimating the GRR_h-parameters $\boldsymbol{\beta}_h = (\gamma_h, \boldsymbol{c}_h^{\mathsf{T}})^{\mathsf{T}}$ for $h = 1, 2$, we let $\mathbf{X}_h(e) = [1, X_h(e)]$. Then, we define the mating-type-specific additive predictors $\eta_m(e)$ for $m = 1, 2, 3$ to be

$$\begin{aligned}
\eta_1(e) &= \log\left[\frac{\mu_{11}(e)}{\mu_{10}(e)}\right] = \mathbf{X}_1(e)\boldsymbol{\beta}_1; \\
\eta_2(e) &= \log\left[\frac{\mu_{22}(e)}{\mu_{21}(e)}\right] = \mathbf{X}_2(e)\boldsymbol{\beta}_2; \text{ and} \\
\eta_3(e) &\equiv \begin{pmatrix}\eta_{31}(e)\\ \eta_{32}(e)\end{pmatrix} = \begin{pmatrix}\log\{\mu_{31}(e)/\mu_{30}(e)\}\\ \log\{\mu_{32}(e)/\mu_{31}(e)\}\end{pmatrix}^{\mathsf{T}} = \begin{pmatrix}\mathbf{X}_1(e)\boldsymbol{\beta}_1 + \log(2)\\ \mathbf{X}_2(e)\boldsymbol{\beta}_2 - \log(2)\end{pmatrix}^{\mathsf{T}},
\end{aligned}$$

where $\mu_{mg}(e)$ are as defined in Table 3.1. These additive predictors indicate that the likelihood contribution from the j^{th} trio from m^{th} mating type is

$$l_{mj} \equiv \begin{cases}
y_{1j1}\eta_1(e_{1j}) - \log(1 + \mathrm{e}^{\eta_1(e_{1j})}) & \text{if } m = 1 \\
y_{2j2}\eta_2(e_{2j}) - \log(1 + \mathrm{e}^{\eta_2(e_{2j})}) & \text{if } m = 2 \\
\begin{array}{c} y_{3j1}\eta_{31}(e_{3j}) + y_{3j2}(\eta_{31}(e_{3j}) + \eta_{32}(e_{3j})) \\ -\log(1 + \mathrm{e}^{\eta_{31}(e_{3j})} + \mathrm{e}^{\eta_{31}(e_{3j})+\eta_{32}(e_{3j})}) \end{array} & \text{if } m = 3.
\end{cases}$$

From the forms of $\eta_m(e)$ above, we can see that $\eta_{31}(e) = \eta_1(e) + \log(2)$ and $\eta_{32}(e) = \eta_2(e) - \log(2)$, and hence the penalized log-likelihood is a function of $\eta_1(e)$ and $\eta_2(e)$ only, which can be expressed as

$$l_p(\boldsymbol{\beta}) = \sum_{m=1}^{3}\sum_{j=1}^{n} l_{mj}(\eta_1(e_{mj}), \eta_2(e_{mj})) - \frac{1}{2}\sum_{h=1}^{2}\lambda_h \boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}_h\boldsymbol{\beta}. \tag{3.12}$$

For given smoothing parameters λ_1 and λ_2, we can find the penalized maximum likelihood estimate (PMLE) $\hat{\boldsymbol{\beta}}$ numerically, using the Newton-Raphson method. The Newton-Raphson update for $\boldsymbol{\beta}$ can be derived using the fact that the penalized log-likelihood function in (3.12) has a similar form to that of a size-2 vector GAM (Yee

and Wild, 1996). It can be shown that the update is equivalent to the penalized iteratively re-weighted least squares (P-IRLS) solution in a matrix form

$$\boldsymbol{\beta}^{new} = \Big(\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X} + \sum_{h=1}^{2} \lambda_h \mathbf{S}_h\Big)^{-1}\Big(\mathbf{X}^{\mathsf{T}}\mathbf{W}Z\Big),$$

where the descriptions for $\mathbf{X}$, $\mathbf{W}$ and Z are as follows.

The model matrix $\mathbf{X}$ is obtained by stacking the trio-specific matrices $\mathbf{X}_{mj}$ corresponding to the j^{th} trio in mating type m. Respectively for $m = 1, 2$ and 3, these matrices are

$$\mathbf{X}_{1j} = \Big[\mathbf{X}_1(e_{1j}) \;\; : \;\; \mathbf{0}\Big]_{1\times(K_1+K_2),}$$

$$\mathbf{X}_{2j} = \Big[\mathbf{0} \;\; : \;\; \mathbf{X}_2(e_{2j})\Big]_{1\times(K_1+K_2),}$$

and

$$\mathbf{X}_{3j} = \begin{bmatrix} \mathbf{X}_1(e_{3j}) & : & \mathbf{0} \\ \\ \mathbf{0} & : & \mathbf{X}_2(e_{3j}) \end{bmatrix}_{2\times(K_1+K_2).}$$

The weight matrix $\mathbf{W}$ is a block-diagonal matrix having the trio-specific diagonal blocks. Setting $\eta_{mj} \equiv \eta_m(e_{mj})$, we can express these blocks as

$$\mathbf{W}_{1j} = -\frac{\partial^2 l_{1j}}{\partial \eta_{1j}^2} = \frac{\mathrm{e}^{\eta_{1j}}}{(1+\mathrm{e}^{\eta_{1j}})^2},$$

$$\mathbf{W}_{2j} = -\frac{\partial^2 l_{2j}}{\partial \eta_{2j}^2} = \frac{\mathrm{e}^{\eta_{2j}}}{(1+\mathrm{e}^{\eta_{2j}})^2}$$

and

$$\mathbf{W}_{3j} = -\frac{\partial^2 l_{3j}}{\partial \eta_{3j}\partial \eta_{3j}^{\mathsf{T}}} = \begin{bmatrix} \dfrac{\mathrm{e}^{\eta_{31j}} + \mathrm{e}^{\eta_{31j}+\eta_{32j}}}{(1+\mathrm{e}^{\eta_{31j}} + \mathrm{e}^{\eta_{31j}+\eta_{32j}})^2} & \dfrac{\mathrm{e}^{\eta_{31j}+\eta_{32j}}}{(1+\mathrm{e}^{\eta_{31j}} + \mathrm{e}^{\eta_{31j}+\eta_{32j}})^2} \\ \dfrac{\mathrm{e}^{\eta_{31j}+\eta_{32j}}}{(1+\mathrm{e}^{\eta_{31j}} + \mathrm{e}^{\eta_{31j}+\eta_{32j}})^2} & \dfrac{\mathrm{e}^{\eta_{31j}+\eta_{32j}}\,(1+\mathrm{e}^{\eta_{31j}})}{(1+\mathrm{e}^{\eta_{31j}} + \mathrm{e}^{\eta_{31j}+\eta_{32j}})^2} \end{bmatrix}.$$

Furthermore, the pseudo-response vector Z can be obtained by stacking the trio-specific pseudo-responses

$$Z_{mj} = \mathbf{X}_{mj}\boldsymbol{\beta} + W_{mj}^{-1} d_{mj},$$

where

$$d_{1j} = \frac{\partial l_{1j}}{\partial \eta_{1j}} = y_{1j1} - \frac{\mathrm{e}^{\eta_{1j}}}{1+\mathrm{e}^{\eta_{1j}}},$$

$$d_{2j} = \frac{\partial l_{2j}}{\partial \eta_{2j}} = y_{2j2} - \frac{\mathrm{e}^{\eta_{2j}}}{1+\mathrm{e}^{\eta_{2j}}},$$

and

$$d_{3j} = \begin{pmatrix} \dfrac{\partial l_{3j}}{\partial \eta_{31j}} \\ \dfrac{\partial l_{3j}}{\partial \eta_{32j}} \end{pmatrix} = \begin{pmatrix} y_{3j1} + y_{3j2} - \dfrac{\mathrm{e}^{\eta_{31j}} + \mathrm{e}^{\eta_{31j}+\eta_{32j}}}{1+\mathrm{e}^{\eta_{31j}} + \mathrm{e}^{\eta_{31j}+\eta_{32j}}} \\ y_{3j2} - \dfrac{\mathrm{e}^{\eta_{31j}+\eta_{32j}}}{1+\mathrm{e}^{\eta_{31j}} + \mathrm{e}^{\eta_{31j}+\eta_{32j}}} \end{pmatrix}.$$

3.3.3 Smoothing parameter estimation and confidence intervals

Smoothing parameter estimation is done by using two one-dimensional grid searches to find the values of λ_1 and λ_2 that minimize the generalized AIC function:

$$\mathcal{V}(\lambda_1, \lambda_2) = \frac{1}{n} D(\hat{\boldsymbol{\beta}}) - \phi + \frac{2}{n}\mathrm{tr}(\mathbf{A})\phi, \tag{3.13}$$

where $D(\hat{\boldsymbol{\beta}})$ is the model deviance, which is negative twice the unpenalized log-likelihood, ϕ is the scale parameter, which is 1 in this context, and $\mathbf{A} = \mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X} + \sum_{h=1}^{2} \lambda_h \mathbf{S}_h)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{W}$ is the hat matrix through which the objective function depends on the smoothing parameters (Wood, 2006). The computational details showing how to estimate the smoothing parameters are presented in Appendix B.3.

Naive confidence bands for $f_h(e)$ based on the asymptotic normal distribution of $\hat{\boldsymbol{c}}_h$ have coverage probabilities less than the nominal confidence level due to the intentional bias introduced by penalized estimation (e.g., Wood, 2006). As an alternative, we consider the Bayesian intervals that have been shown to have good frequentist coverage probabilities (e.g., Nychka, 1988; Marra and Wood, 2012). To construct the Bayesian intervals for $f_h(e)$, we use the approximate normality of the posterior distribution of $\boldsymbol{c}_h$, so that

$$f_h(e) \dot{\sim} N\Big(X_h(e)\hat{\boldsymbol{c}}_h,\ X_h(e)\mathrm{V}_{\boldsymbol{c}_h}X_h^{\mathsf{T}}(e)\Big),$$

where the variance-covariance matrix $\mathrm{V}_{\boldsymbol{c}_h}$ is obtained by extracting appropriate rows and columns from the Bayesian posterior variance-covariance matrix for the full parameter vector $\boldsymbol{\beta}$ (Wood, 2006)

$$\mathrm{V}_{\boldsymbol{\beta}} = (\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X} + \sum_{h=1}^{2} \lambda_h \mathbf{S}_h)^{-1}\phi.$$

For displaying the smoothed fits for $G\times E$ functions, we plot the fitted curves $\hat{f}_1$ and $\hat{f}_2$ along with their $(1-\alpha)100\%$ Bayesian confidence bands for a given confidence level α. Different genetic inheritance modes will yield different patterns in $\hat{f}_h$: under the dominant mode, non-horizontal $\hat{f}_1$ and horizontal $\hat{f}_2$ would be produced; under the recessive mode, horizontal $\hat{f}_1$ and non-horizontal $\hat{f}_2$ would be produced; and under the multiplicative or log-additive mode, non-horizontal $\hat{f}_1$ and $\hat{f}_2$ having equivalent forms would be produced.

3.3.4 Permutation test of $G\times E$

To account for the extra uncertainty introduced by the estimation of the smoothing parameters, we take a permutation-based approach to testing $G\times E$. We define the test statistic T as

$$T = \hat{\boldsymbol{c}}^{\mathsf{T}}\{V_{\boldsymbol{c}}\}^{-1}\hat{\boldsymbol{c}},$$

where $\boldsymbol{c} = (\boldsymbol{c}_1^{\mathsf{T}}, \boldsymbol{c}_2^{\mathsf{T}})^{\mathsf{T}}$, and $V_{\boldsymbol{c}}$ is the $(K_1 + K_2 - 2) \times (K_1 + K_2 - 2)$ matrix formed by extracting, from $V_{\boldsymbol{\beta}}$, the appropriate columns and rows corresponding to $\boldsymbol{c}$. As the analysis is conditional on parental genotypes, we estimate the distribution of T under the hypothesis of no $G\times E$ by shuffling E within mating types. Under no $G\times E$, G and E are independent within a random affected trio when they are independent within a trio from the general population (Umbach and Weinberg, 2000). The p-value is obtained by computing the proportion of test statistics that are more extreme than or as extreme as the observed test statistic. For our analysis conditional on parental mating types, an alternative to permutation is bootstrap re-sampling of E within parental mating types. The advantage of a bootstrap-based approach is that its

statistical properties are better understood. However, when the number of affected trios within each parental mating types is large, both approaches should reach similar conclusions.

3.4 Simulation

We conducted a simulation study to evaluate type 1 error rate and power of the proposed permutation test of $G\times E$ under various scenarios. The test statistic T was computed based $K_1 = K_2 = 5$ knots placed at the quartiles of the observed distribution of E in the appropriate subsets of case-parent trios (e.g., trios in mating types $m = 1, 3$ for estimating $f_1(e)$). The p-values were obtained based on 1000 permutations.

When assessing the size of the test, we considered both a homogeneous (or unstratified) and a stratified population to verify unbiasedness of the test regardless of population stratification. However, when assessing the power, we did not consider population stratification. For comparison, we also evaluated three other tests, which will be discussed below: (i) a likelihood ratio test based on a conditional logistic regression model (e.g., Schaid, 1999), (ii) a likelihood ratio test based on a log-linear model (Umbach and Weinberg, 2000), and (iii) a family-based association test of gene-environment interaction (FBAT-I; Lake and Laird, 2004).

In the conditional logistic regression, we used linear $G\times E$ via $f_1(e) = \beta_{ge1}e$ and $f_2(e) = \beta_{ge2}e$ in model (3.1). For the log-linear modelling approach, we dichotomized E based on its sample median $\tilde{\mu}$ in the affected trios and set $f_1(e) = \beta_{ge1}\mathrm{I}\{e > \tilde{\mu}\}$ and $f_2(e) = \beta_{ge2}\mathrm{I}\{e > \tilde{\mu}\}$ in model (3.1). For FBAT-I, we set $z_1(g) = z_2(g) = g$, $\gamma_1 = \gamma_2$ and $f_1(e) = f_2(e)$ in model (3.1) and calculated the p-values based on 10,000 permutations. The type 1 error rates for these three tests were not examined because it is well established that they maintain the nominal level of significance when the test marker is causal (e.g., Lake and Laird, 2004; Shin *et al.*, 2010).

All computation was done by R (R Development Core Team, 2011). The proposed method was implemented in R and is soon to be available on CRAN. For FBAT-I,

we used the 'fbati' package (Hoffmann, 2009). We simulated 1000 data sets in the absence and 500 in the presence of $G\times E$ under dominant, log-additive and recessive penetrance models, where each data set consisted of (G, E, G_p) of informative case-parent trios. We chose to use a large sample size of 3000 in order to get enough resolution for comparing power since it is well known that the power to detect $G\times E$ is low (Smith and Day, 1984; Dempfle *et al.*, 2008).

3.4.1 Simulation setting

Under population stratification, we considered a stratified population with two equal-sized subpopulations $S = 0$ and 1, assuming random-mating within but not between subpopulations. Within subpopulations, the index allele frequencies were chosen as $q_0 = 0.1$ and $q_1 = 0.9$, and the means and common variance of E, as $\mu_0 = 0.8$, $\mu_1 = -0.8$ and $\sigma^2 = 0.36$, respectively. Under no population stratification, the two subpopulations were set to have the same allele frequency $q_0 \equiv q_1 \equiv q = 0.1$ and the same mean and variance of E so that $\mu_0 \equiv \mu_1 \equiv \mu = 0$ and $\sigma^2 = 1$. In each subpopulation, G_p were simulated under Hardy-Weinberg proportions (HWP) and mating symmetry; G were simulated under Mendelian segregation with no mutation; and E were simulated independently of G_p and G under normal distributions.

Parameters in the disease risk model in equation (3.1) were chosen as follows. Since the baseline disease probability and the non-genetic main effect are not estimable from case-parent trio data, we let $k = 0$ and $\xi(e) = 0$ for all values of e, for convenience. Under a dominant penetrance model, we took $\gamma_1 = \log(3)$ and $\gamma_2 = 0$, giving GRR of 3 between the individuals with one or two copies and those with zero copies of the index allele. Under the log-additive penetrance model, $\gamma_1 = \gamma_2 = \log(\sqrt{3})$, giving GRR of 3 between the individuals with two copies and those with zero copies of the index allele and GRR of 1.5 between those with one copy and those with zero copies of the index allele. Under a recessive penetrance model, we took $\gamma_1 = 0$ and $\gamma_2 = \log(3)$, giving GRR of 3 between the individuals with two copies and those with one or zero copies of the index allele.

Under no $G\times E$, we let $f_1(e) = f_2(e) = 0$ for all $E = e$ both in the absence (setting

H_{0U}) and in the presence (setting H_{0S}) of population stratification. Under $G\times E$, we let $f_1(e) \equiv f(e)$ and $f_2(e) = 0$ under a dominant penetrance model, $f_1(e) \equiv f_2(e) = \frac{1}{2}\cdot f(e)$ under a multiplicative penetrance model, and let $f_1(e) = 0$ and $f_2 \equiv f(e)$ under a recessive penetrance model, to have the equivalent GRR between the individuals with two copies of the index allele and those with zero copies under both penetrances.

For $f(e)$, we considered linear (setting H_{1L}), piecewise linear (setting H_{1P}) and quadratic (setting H_{1Q}) models in the absence of population stratification. Table 3.2 summarizes the scenarios we considered for the simulation study. Under setting H_{1L},

Table 3.2: Simulation scenarios

Setting	Population Stratification	$G\times E$
H_{0S}	Yes	No
H_{0U}	No	No
H^*_{1L}	No	Yes
H_{1P}	No	Yes
H_{1Q}	No	Yes

we let $f(e)$ be a linear function with slope β_{ge}. Under setting H_{1P}, we let $f(e)$ be a piecewise linear function created by joining one horizontal line and one straight line having a slope of β_{ge} together at a point z_p, which represents the p^{th}-quantile of the standard normal distribution of E in general population. Although using a piecewise linear function violates the assumption that $f_1(e)$ and $f_2(e)$ are smooth functions, we chose to use it since it is easier to control the shape of the function and hence the effect size of $G\times E$ than a smooth function with a similar form (e.g., exponential). Under setting H_{1Q}, we let $f(e)$ be a quadratic function with coefficient β_{ge} and axis of symmetry z_p placed at p^{th} quantile of $N(0,1)$. The specific forms for $f(e)$ and the ranges of β_{ge} and p under different models of $G\times E$ are presented in Table 3.3.

Table 3.3: Parameterizations for $f(e)$ under $G\times E$

Setting	$f(e)$	β_{ge}	p[a]
H_{1L}	$\beta_{ge}e$	[-0.24, -0.10]	–
H_{1P}[b]	$\beta_{ge}I\{e < z_p\} \cdot (e - z_p)$	[0.20, 1.00]	[0.10, 0.50]
H_{1Q}	$\beta_{ge}(e - z_p)^2$	[-0.20, -0.04]	[0.10, 0.50]

[a] z_p indicates the p^{th} quantile of the standard normal distribution of E in the general population.

[b] $f(e)$ is not smooth, but chosen for convenience.

3.4.2 Simulation results

Under no $G\times E$, the proposed test maintained the nominal significance level of 0.05 within simulation error both in the absence (setting H_{0U}) and in the presence (setting H_{0S}) of population stratification. Under the dominant penetrance models, the empirical type 1 error rates were 0.053 under H_{0U} (SE = 0.007) and 0.060 under H_{0S} (SE=0.008). The rates were similar under the log-additive and the recessive models (results not shown).

Figures 3.1 – 3.3 show the empirical power results for different penetrance models under various simulation configurations. The simulation results can be summarized as follows: i) when the underlying $G\times E$ is non-linear, and there is little or no linear association between G and E, the proposed test has the highest power among the four tests; ii) when $G\times E$ is non-linear, but there is some linear association between G and E, the proposed test has comparable power to that of conditional logistic regression approach and/or FBAT-I but higher power than the log-linear approach; iii) when $G\times E$ is linear, the proposed test has lower power than conditional logistic regression and can have lower or higher power than FBAT-I depending on whether the FABT-I mode of inheritance is incorrectly specified as additive rather than recessive.

For the other tests, conditional logistic regression had more power than the log-linear approach. FBAT-I performed as well as conditional logistic regression under the dominant penetrance model, better than conditional logistic regression under the

log-additive model and worse than conditional logistic regression under the recessive penetrance. The results indicated that FBAT-I can lose power under recessive penetrance when the mode of inheritance is mis-specified as additive.

Under H_{1L}, the proposed test had lower power than the conditional logistic regression approach with linear $G\times E$ (Figure 3.1). Under the dominant and the log-additive models, it also had lower power than FBAT-I since FBAT-I's sample covariance test statistic measures the strength of the linear association between G and E (Figure 3.1 panels A and B). However, under the recessive penetrance models, the proposed test performed better than FBAT-I, which has a huge power loss due to mis-specification of the penetrance mode (Figure 3.1C). The proposed test had comparable power to that of the log-linear approach under all penetrance models.

Under H_{1P}, the power of the proposed and the other tests increased with the effect size $|\beta_{ge}|$ (e.g., Figure 3.2, panels A and B) and with the joining point p (e.g., Figure 3.2, panels C and D). Under the dominant models, the proposed test had more power than the other tests when the joining point was at a lower quantile (Figure 3.2A). This is because when p is low, the linear association between G and E is weak, leading to the loss of power for the conditional logistic regression approach and FBAT-I. When $|\beta_{ge}|$ is low, the proposed test had similar power to conditional logistic regression and FBAT-I (e.g., Figure 3.2C) since it tends to fit the $G\times E$ curves as linear functions (results not shown). Similar but weaker patterns were observed under the log-additive and the recessive models (results not shown).

Under H_{1Q}, the proposed test had comparable or superior power to that of the other competing tests (e.g., Figure 3.3). The power of the proposed test increased with the effect size $|\beta_{ge}|$, while the power of the other tests did not always increase with $|\beta_{ge}|$ (e.g., Figure 3.3, panels A and B). The power of the other tests increased with $|\beta_{ge}|$ when the axis of symmetry z_p is far from the median (e.g., Figure 3.3A) but not when z_p was at the median (e.g., Figure 3.3B). The power of both the proposed and the other tests decreased as z_p became closer to the median of E; however, the relative power for the proposed test increased as z_p became closer to the median (e.g., Figure 3.3 panels C and D). When z_p was far from the median, both the proposed and

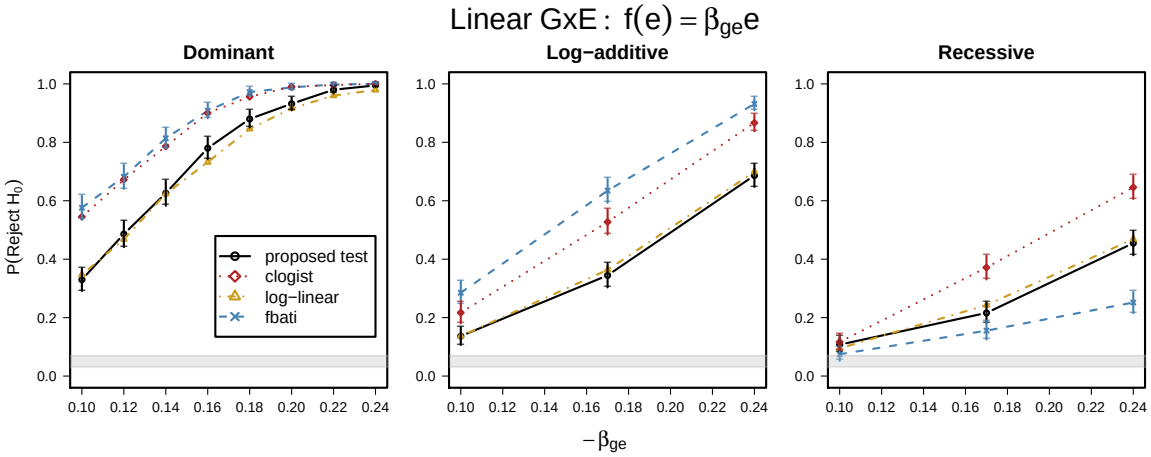

Figure 3.1: Empirical power results for gene-environment tests under linear $G{\times}E$ (Setting H_{1L} in Table 3.3). The black circles with solid lines represent power for the proposed test (triogam); the blue cross-marks with dashed lines represent power for FBAT-I (fbati); the red diamonds with dotted lines represent power for the conditional logistic regression approach (clogist); and the yellow triangles with dot-dashed lines represent power for the log-linear modelling approach (log-linear). The shaded area represents simulation error about the nominal 5% level: a test with the correct size would have estimated type 1 error within the shaded region 19 times out of 20. The vertical bars represent ± 2 simulation errors, which we present for the proposed test, FBAT-I and/or conditional logistic regression since the latter two tests can be more powerful than the proposed test under linear $G{\times}E$. Results are based on 500 simulation replicates of 3000 informative case-parent trios generated under the dominant (panel A), the log-additive (panel B) and the recessive (panel C) penetrance models.

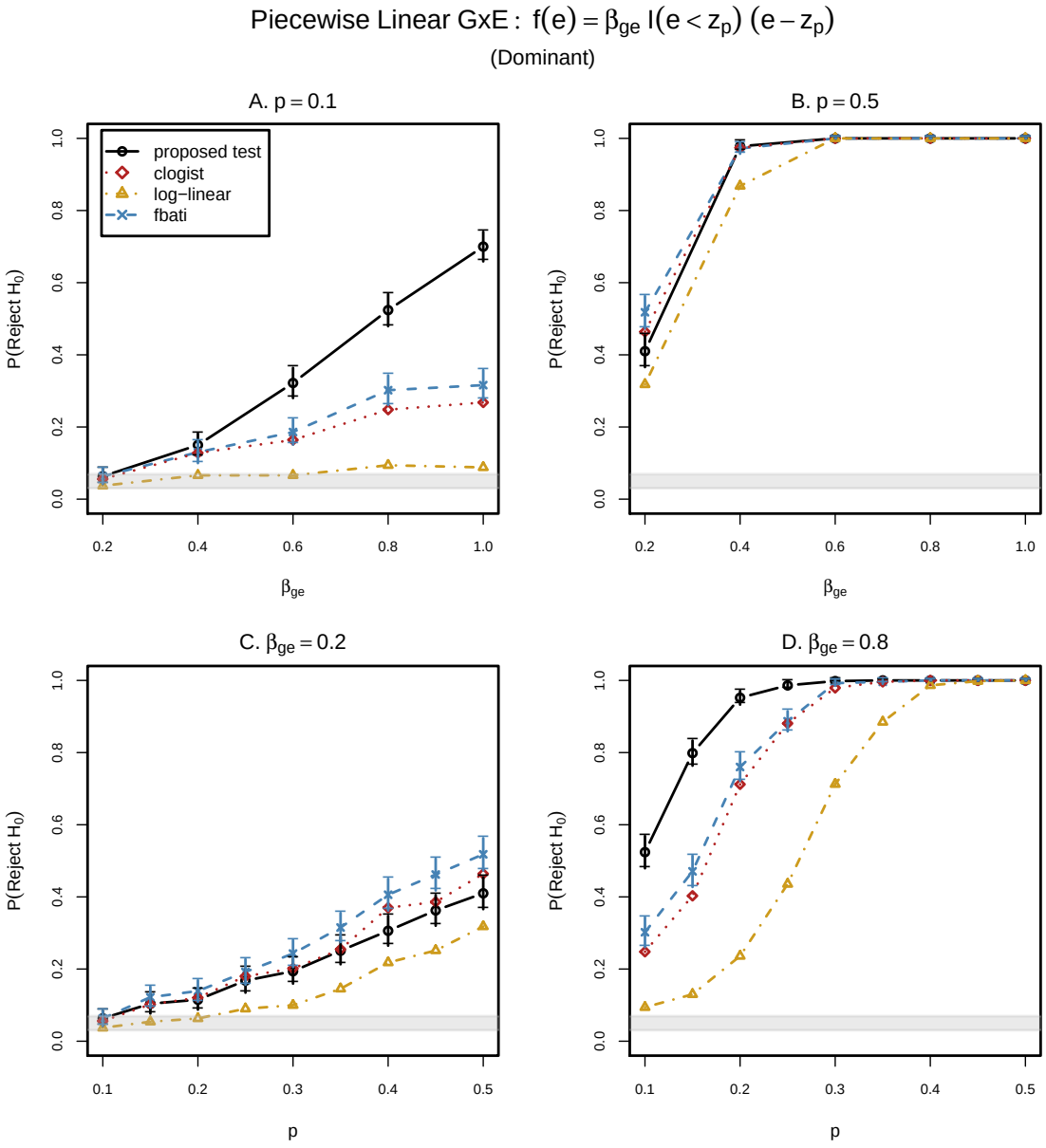

Figure 3.2: Empirical power results for gene-environment tests under piecewise linear $G\times E$ (Setting H_{1P} in Table 3.3). The black circles with solid lines represent power for the proposed test (triogam); the blue cross-marks with dashed lines represent power for FBAT-I (fbati); the red diamonds with dotted lines represent power for the conditional logistic regression approach (clogist); and the yellow triangles with dot-dashed lines represent power for the log-linear modelling approach (log-linear). The shaded area represents simulation error about the nominal 5% level - a test with the correct size would have estimated type 1 error within the shaded region 19 times out of 20. The vertical bars represent ± 2 simulation errors, which we present only for the proposed test and FBAT-I since FBAT-I is uniformly more powerful than the other two competing tests. Results are based on 500 simulation replicates of 3000 informative trios generated under the dominant penetrance models.

the competing tests performed well and had comparable power. When z_p was close to the median, both the proposed and the other tests did not perform as well as when z_p was far from the median; however, the power of the proposed test was greater than that of the other tests, and it decreased at a slower rate than that of the other tests. The reason why the power of all the tests decreased as z_p was closer to the median of E is that the $G\times E$ function varies with E more rapidly in the boundary areas where there is little information available when its axis of symmetry is closer to the median, while it varies more rapidly in the regions where there is a lot of information (e.g., near median) when its z_p is far from the median. The reason why the other tests performed worse when z_p was close to the median is that in addition to the previous reason, the linear association between G and E also becomes weaker as the axis of symmetry of a quadratic $G\times E$ curve is closer to the median.

3.5 Illustration: Application to acute lymphoblastic leukemia simulated data

Acute lymphoblastic leukemia (ALL) is the most common type of leukemia in children aged 1 – 19 years old. ALL can occur at any age, but the age-adjusted incidence rates are highest during childhood between age 2 and 6 years, decrease during young-adulthood, and then increase again at older ages around > 50 years (Ries *et al.*, 1999) (e.g., Figure 3.4). We examined the C609T polymorphism in NAD(P)H:quinone oxidoreductase 1 (NQO1), which plays a role in detoxification of carcinogenic byproducts. Homozygous individuals for the variant allele (T/T) are deficient in NQO1 activity, and lower activity of the NQO1 has been shown to be associated with infant ALL (Wiemels *et al.*, 1999). The bimodal distribution of incidence with age is consistent with different disease mechanisms for younger- and older- patients. For example, younger cases could have a genetic basis whereas older cases could be sporadic. This motivates us to search for age-dependent NQO1 genotype relative risks for ALL patients.

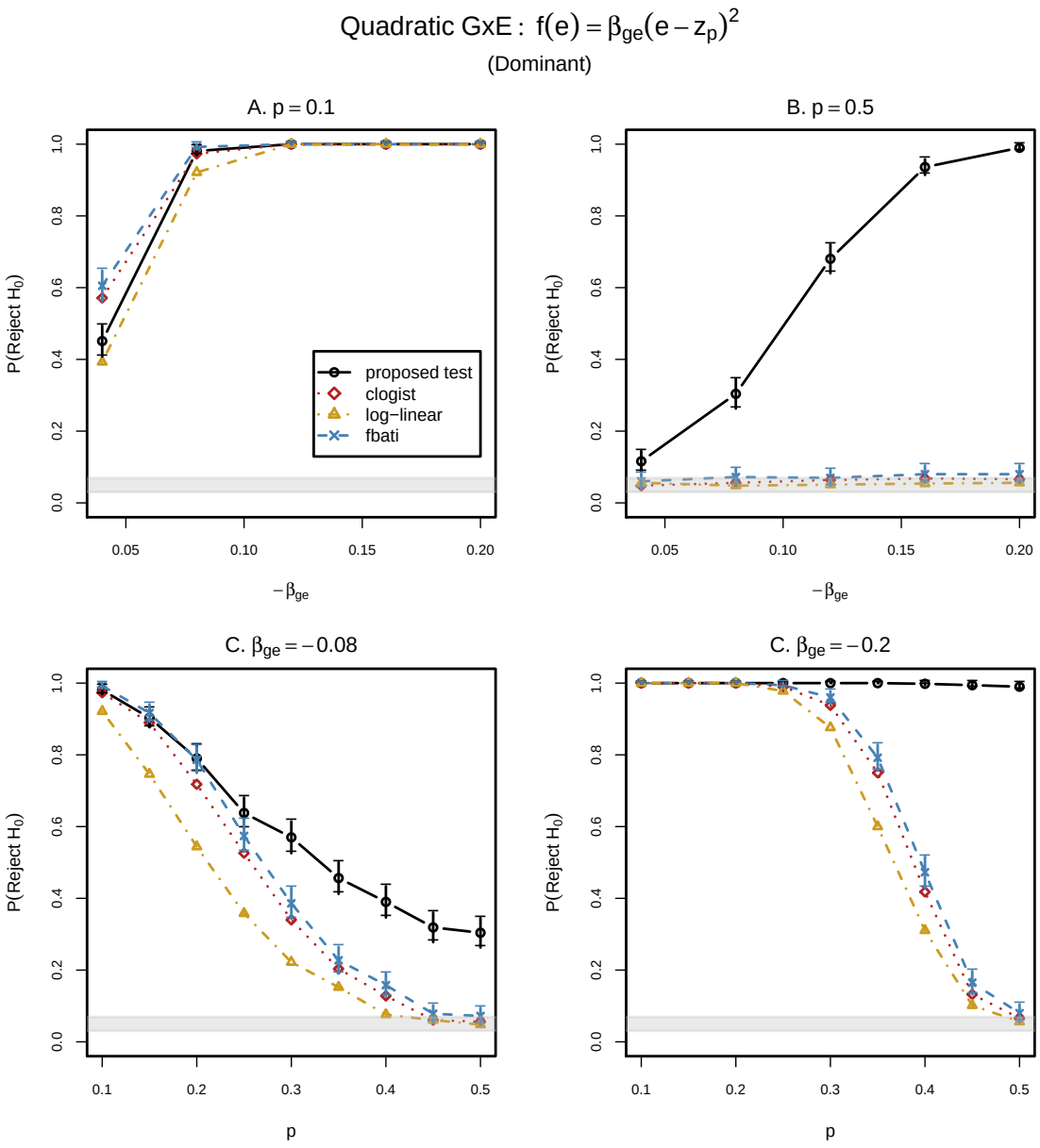

Figure 3.3: Empirical power results for gene-environment tests under quadratic $G\times E$ (Setting H_{1Q} in Table 3.3). The black circles with solid lines represent power for the proposed test (triogam); the blue cross-marks with dashed lines represent power for FBAT-I (fbati); the red diamonds with dotted lines represent power for the conditional logistic regression approach (clogist); and the yellow triangles with dot-dashed lines represent power for the log-linear modelling approach (log-linear). The shaded area represents simulation error about the nominal 5% level - a test with the correct size would have estimated type 1 error within the shaded region 19 times out of 20. The vertical bars represent ± 2 simulation errors, which we present only for the proposed test and FBAT-I since FBAT-I is uniformly more powerful than the other two competing tests. Results are based on 500 simulation replicates of 3000 informative trios generated under the dominant penetrance models.

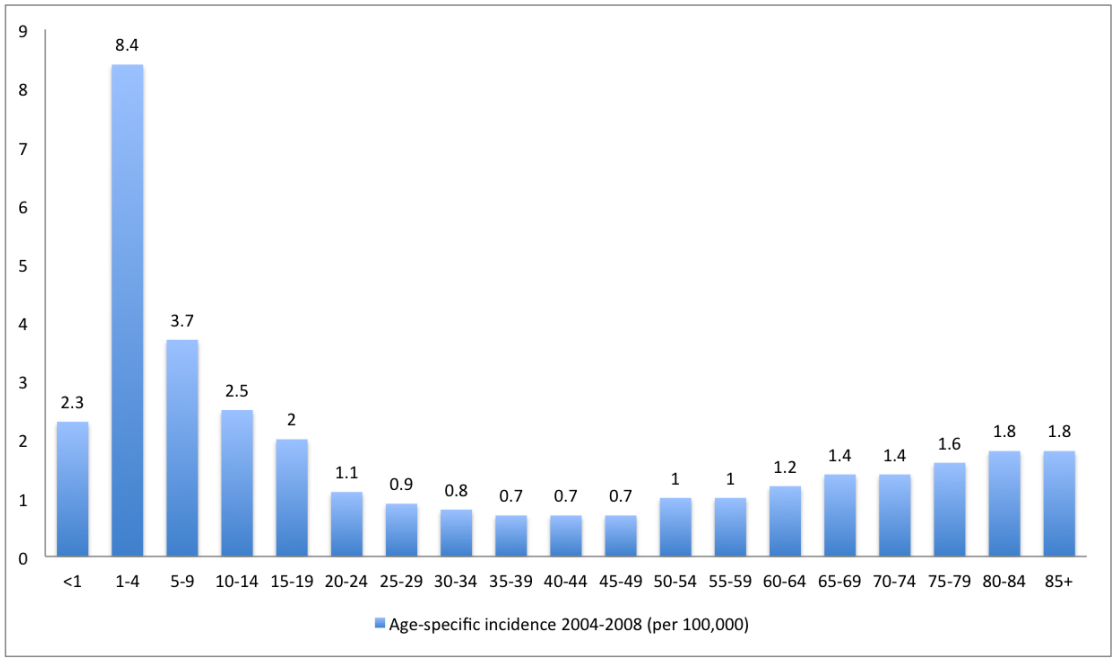

Figure 3.4: Age-specific incidence of ALL in white patients in the US during 2004–2008. The horizontal axis shows five-year age intervals. The vertical axis shows the frequency of new cases of ALL per 100,000 in a given age-group. (source: Surveillance, Epidemiology and End Results [SEER] Program, 2004-2008, National Cancer Institute, 2011).

We illustrate our method by using a simulated data set that mimics case-parent trio data obtained from a genetic association study of childhood ALL. The real data arise from two family studies from Québec (Infante-Rivard *et al.*, 2000; Infante-Rivard, 2003; Infante-Rivard *et al.*, 2007) and France. The French data are from a case-control study (Perrillat *et al.*, 2001; Clavel *et al.*, 2005), for which parental genotype information was later collected. There were 1031 case-parent trios, of which 288 were informative for the polymorphism of interest. It is well known that the sample size requirements to detect $G\times E$ are much larger than those to detect the main effects of G or E (Smith and Day, 1984; Dempfle *et al.*, 2008). A typical rule of thumb is that, for a given power, the sample size for detecting $G\times E$ should be at least four times that for detecting a marginal effect with the same power. To increase the power to detect interaction, we used the original data to simulate 1000 informative case-parent trios. Simulated data were generated based on the characteristics of the real data. Note that the data were simulated in the presence of $G\times E$ (i.e., $f_1(e) \neq 0$

and $f_2(e) \neq 0$) as indicated by the theoretical log-GRR curves shown in Figure 3.5. Other details describing how the data were simulated are presented in Appendix B.4.

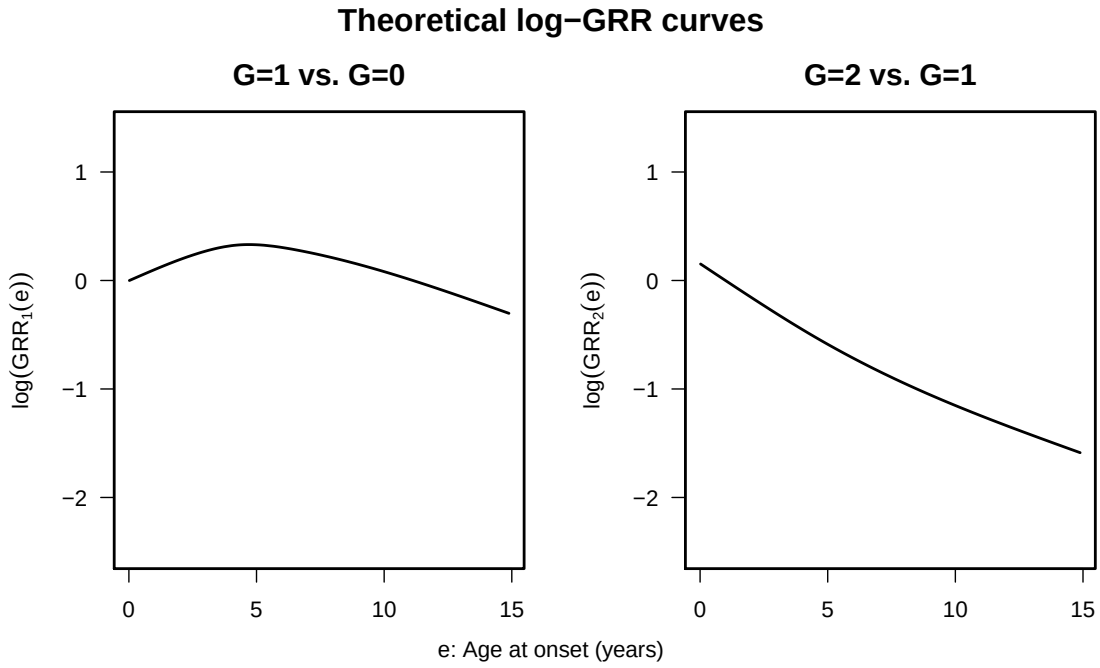

Figure 3.5: Theoretical log-GRR curves used for simulating the ALL data set. The curves were constructed based on the $G\times E$ curves estimated from fitting the 288 informative case-parent trios in the original data set of ALL.

The mating-type-specific distribution of genotype frequencies and the histogram of age-at-diagnosis among the cases from the resulting 1000 simulated informative trios are shown in Table 3.4 and Figure 3.6, respectively. According to Table 3.4, heterozygous parents transmit the variant allele to the cases $634/1175 = 54.0\%$ of the time. The transmission disequilibrium test (TDT; Spielman *et al.*, 1993) confirms that the variant allele was transmitted slightly more frequently than expected ($p = 0.007$). A similar trend was observed in the original data, for which the observed proportion was $181/348 = 52\%$ ($p = 0.49$).

Figure 3.7 shows the fitted curves of $G\times E$ and their corresponding Bayesian 95% confidence intervals. The confidence intervals are suggestive for $G\times E$ between *NQO1 C609T* and age-at-onset of ALL. To test for $G\times E$, we applied the proposed permutation test to the simulated data set using 1000 replications. The resulting p-value

Table 3.4: Mating-type-specific frequencies (%) of case-genotypes in 1000 simulated informative trios

	Informative mating type (m^*)		
Number of copies of *NQO1 C609T* variant	1	2	3
0	343 (44)	–	32 (18)
1	435 (57)	27 (57)	107 (61)
2	–	20 (43)	36 (21)
n_m	778 (100)	47 (100)	175 (100)

$^*m = 1, 2, 3$ corresponds to parental genotype pairs $(G_M, G_F) = \{(1,0) \text{ or } (0,1)\}$, $\{(1,2) \text{ or } (2,1)\}$ and $\{(1,1)\}$, respectively.

for our approach indicated there was $G{\times}E$ ($p = 0.03$), while the p-values of the conditional logistic regression, FBAT-I and the log-linear modelling approach did not ($p = 0.12, 0.19$ and 0.34, respectively).

3.6 Discussion

Complex diseases, such as diabetes and cancer, are the result of both genetic and non-genetic factors acting jointly on the disease risk. For example, the effects of multiple genes in the HLA region on the risk of type 1 diabetes vary with age-at-onset (e.g., Caillat-Zucman *et al.*, 1992). The more we learn about gene-environment interactions, the better insights we can get into the disease aetiology.

In this work, we proposed a smoothing approach to exploring $G{\times}E$ using data from case-parent trios. The method provides a flexible way of modelling $G{\times}E$ via spline functions, which are estimated under a penalized maximum likelihood framework. Rather than making assumptions about the parametric form and the inheritance mode of $G{\times}E$, the proposed approach lets the data determine them. The revealed patterns are displayed graphically, which can provide new or better insights into the biological mechanisms for G and E under the study. For testing $G{\times}E$,

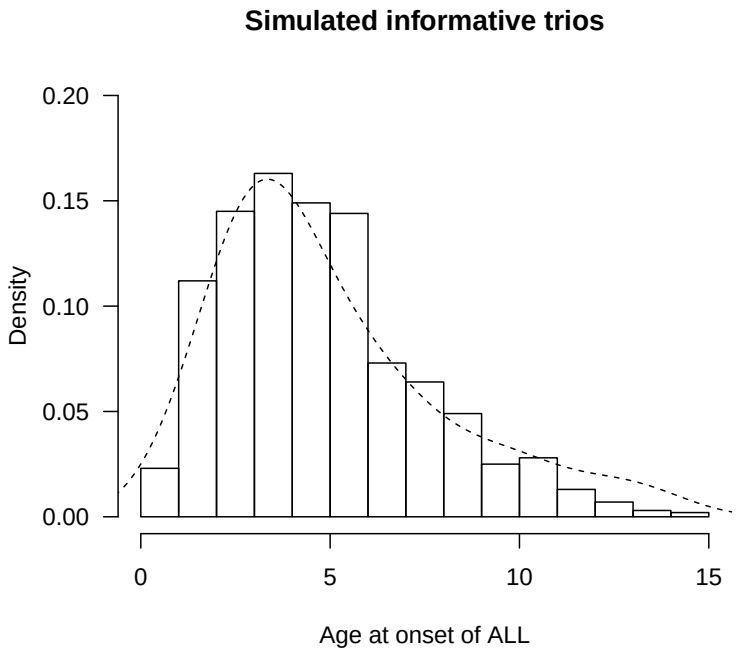

Figure 3.6: Histogram of age-at-diagnosis for the cases from 1000 simulated informative trios with complete information on NQO1 genotype and age-at-onset. The dashed-line curve represents the density curve estimated from the observed E of the informative affected trios in the real-data.

we adopted a permutation-based test that takes into account the extra uncertainty arising from the estimation of the smoothing parameters.

The simulation study results demonstrate that the proposed test can have much greater power to detect non linear $G\times E$, compared to the other available tests we considered (e.g., Figure 3.2A and 3.3B). The power of the permutation test can be low since we make minimal assumptions about the parametric form of $G\times E$ model. However, the proposed test can be useful in a unique way. For example, when considering $G\times E$ with continuous E, an analyst can fit a conditional logistic regression model with linear $G\times E$; if such inference suggests that $G\times E$ is not significant, she/he could look at the form for $G\times E$ by applying our method and, if the estimated curve is not linear, our permutation test can be applied to get the p-value.

One advantage of the case-parent trio design is that it can allow for the genetic

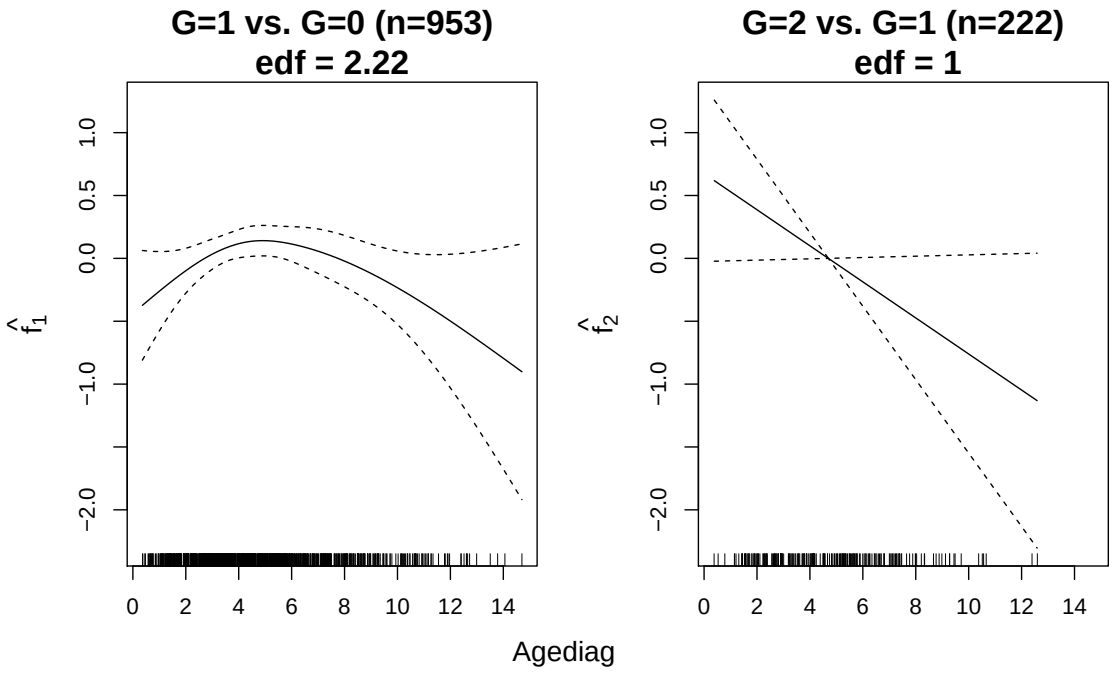

Figure 3.7: Fitted $G\times E$ curves for the simulated ALL data set. The left panel shows the fitted $G\times E$ for GRR between the individuals with 1 and 0 copies of the *NQO1 C609T* variant, and the right one, the fitted $G\times E$ for GRR between those with 2 and 1 copy of the variant allele. The dashed lines indicate the 95% pointwise Bayesian confidence limits.

effects such as parent-of-origin effects. A parent-of-origin effect exists when the disease risk of an individual is affected by whether the allele responsible for the disease was transmitted from the mother or the father. Numerous association and linkage studies of complex disorders have suggested existence of parent-of-origin effects, and hence incorporating parent-of-origin effects will improve our understanding of disease aetiology (Guilmatre and Sharp, 2012). One possible way to extend the current method to incorporate parent-of-origin effects is outlined in Chapter 5.

Chapter 4

Adjusting for spurious gene-by-environment interaction using case-parent triads.

4.1 Introduction

In the case-parent design, unrelated children affected with a disease are genotyped along with their parents. The requirement of parental genotypes makes this design most practical for early-onset diseases. Information on the cases' non-genetic covariates, such as age at onset, may also be collected. Throughout we use the notation G to denote a genotype at a causal SNP, with possible values 0, 1 or 2 for number of copies of an index allele. We let E denote the exposure to an environmental factor, which we assume is continuous.

If G is associated with the disease, it may then be of interest to ask whether the association is modified by E. Alternately, if E is associated with disease, it may be of interest to ask whether G modifies this association. In either case, interest is in gene-by-environment interaction which exists when genotype relative risks (GRRs)

depend on E:

$$GRR_1(e) = \frac{P(D = 1 \mid G = 1, E = e)}{P(D = 1 \mid G = 0, E = e)} \quad \text{or} \quad GRR_2(e) = \frac{P(D = 1 \mid G = 2, E = e)}{P(D = 1 \mid G = 1, E = e)}.$$

Here $D = 1$ is the event that the child in the trio is affected. The choice of parametrization of the two GRR functions is arbitrary. In the above parametrization, each GRR function is the factor by which risk increases for an additional copy of the index allele in G, for a fixed value of E. Throughout we define models with gene-by-environment interaction ($G \times E$) to be those that imply E-dependent GRRs.

As recently noted by Shi *et al.* (2011), inference of interaction based on a non-causal genotype G' at a test locus that is in linkage disequilibrium (LD) with the causal locus can be misleading under population stratification. They show that GRRs at G' can vary with E without $G \times E$ when both the distribution of E and the GG' haplotypes vary by sub-population. The interpretation of $G \times E$ in such a case is spurious and may be regarded as a bias due to population stratification. We refer to this situation as spurious interaction.

Whether or not such spurious interaction is a concern depends on how plausible it is to have haplotype and E distributions that vary by sub-population. Differences in haplotype distributions between populations may result from genetic drift or positive selection (Bersaglieri *et al.*, 2004). Examples of genomic regions under positive selection include the lactase gene and the Duffy blood group locus. The lactase gene is thought to be under positive selection in Europeans and other dairy-dependent societies for a variant that confers the ability to digest milk into adulthood (Bersaglieri *et al.*, 2004; Tishkoff *et al.*, 2007). The Duffy blood group locus is thought to be under positive selection in African populations for a variant that confers resistance to malaria (Hamblin *et al.*, 2002). In both cases, haplotype frequencies in the population under selection differ from those of other populations (Teo *et al.*, 2009). Thus, when studying non-causal loci, differences in E between genetic subgroups in the population could lead to inference of spurious interaction.

Shi et al. discuss a robust approach to spurious interaction that requires E to also

be available on an unaffected sibling of the affected child. In this paper we further explore the source of spurious interaction and suggest an alternate approach that mitigates its effects using the existing case-parent triads. In contrast to Shi et al., our approach does not require data on unaffected siblings. Simulations of case-parent trio data used an approach that is implemented in a freely-available R package soon to be released on the Comprehensive R Archive Network (CRAN).

A summary of the remainder of the chapter is as follows. In the next section we present data simulated under population stratification that lead to inference of spurious interaction. This is followed by a brief description of the risk model used for inference. We then propose an adjustment to the risk model that uses information on unlinked markers to provide robustness to spurious interaction. We also comment on risk model adjustments that can preserve power to detect interaction. Simulations show that, with enough independent marker information, the proposed adjustments achieve the nominal type 1 error rate and reasonable power. We conclude with a summary and comparison of approaches to avoiding spurious interaction and ideas for future work.

4.2 Example of spurious interaction

For illustration, we consider a population comprised of two equi-sized and non-mixing sub-populations in which both the distribution of E and the GG' haplotypes vary by sub-population. The subpopulation-specific haplotype distributions are given in Table 4.1. In this example, alleles in G are denoted by R (risk) and N (non-risk), while alleles in G' are denoted by 1 and 0. Here, and in what follows, it will be convenient to summarize haplotype distributions by the implied allelic correlations between the index alleles R and 1. Under the GG' haplotype frequencies given in the table, these correlations are $r_0 = -1$ in sub-population $S = 0$ and $r_1 = 1$ in sub-population $S = 1$. Consequently, $G' = 1$ is associated with low disease risk in $S = 0$ and high disease risk in $S = 1$. The subpopulation-specific E-distributions are given in Figure 4.1. Low values of E tag individuals as being likely from $S = 0$ while

high values tag them as likely from $S = 1$.

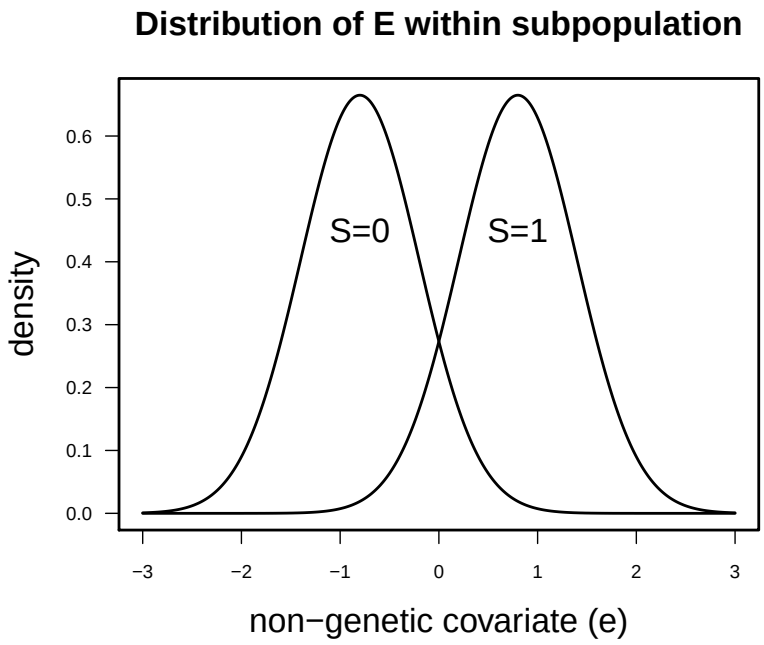

Figure 4.1: Example distributions of E in two sub-populations

When data are simulated from these settings under no $G\times E$, the non-zero slopes of the fitted log-GRR curves for G' indicate spurious interaction, as shown in Figure 4.2. The curves were obtained by applying likelihood methods to fit a model described in the next section. The source of the spurious interaction is illustrated in the schematic of Figure 4.3. Though GRRs do not vary within either sub-population, low E tags sub-population 0 in which $G' = 1$ is low risk, while high E tags sub-population 1 in which $G' = 1$ is high risk. As a result, GRRs for G' vary with E; i.e., there is $G' \times E$. However, the interaction is spurious because it is the allelic correlation between G and G' that modifies the GRRs, not E.

Table 4.1: Example haplotype frequencies for haplotypes comprised of causal and non-causal loci in two sub-populations. Alleles in G (causal) are denoted by R (risk) and N (non-risk), and alleles in G' (non-causal) are denoted by 1 and 0. Table entries are GG' haplotype frequencies in the two sub-populations, denoted by $S = 0$ and $S = 1$.

	Sub-population	
	$S = 0$	$S = 1$
R1	0	0.5
R0	0.5	0
N1	0.5	0
N0	0	0.5

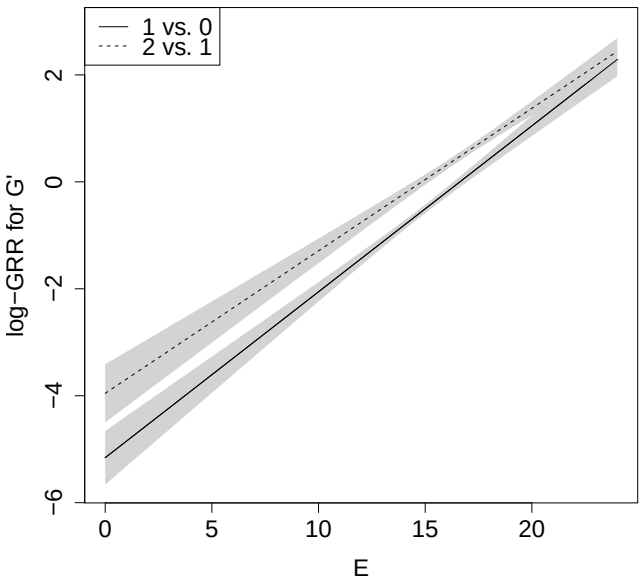

Figure 4.2: Plot of fitted log-GRRs for G' versus E. The non-zero slopes of the fitted log-GRRs for $G' = 1$ versus $G' = 0$ (solid) and $G' = 2$ versus $G' = 1$ (dashed) suggest interaction. Shaded regions represent approximate point-wise 95% confidence intervals.

4.3 Methods

4.3.1 Model

Throughout, we use the following log-additive model:

$$P(D = 1 \mid G = g, E = e) \propto \exp\left\{z_1(g)\beta_1 + z_2(g)\beta_2 + \eta(e) + z_1(g)f_1(e) + z_2(g)f_2(e)\right\}, \tag{4.1}$$

where $z_1(g)$ and $z_2(g)$ are indicator variables for $g > 0$ and $g = 2$, respectively (Shin *et al.*, 2010). To emphasize the dependence of the resulting GRRs on E, we write them as a function of e:

$$\mathrm{GRR}_1(e) = \frac{P(\ D = 1 \mid G = 1,\ E = e\ )}{P(\ D = 1 \mid G = 0,\ E = e\ )} = \exp\big(\beta_1 + f_1(e)\big)$$

and

$$\mathrm{GRR}_2(e) = \frac{P(\ D = 1 \mid G = 2,\ E = e\ )}{P(\ D = 1 \mid G = 1,\ E = e\ )} = \exp\big(\beta_2 + f_2(e)\big).$$

The model allows for $G \times E$ because GRRs can vary with E through the functions f_1 and f_2. If $f_1 = f_2 \equiv 0$ then there is no interaction.

Assuming G and E are conditionally independent given parental genotypes G_p, and conditioning on G_p and E we can obtain a likelihood based on

$$P(G = g \mid D = 1, E = e, G_p) = \frac{\exp\left(z_1(g)\beta_1 + z_2(g)\beta_2 + z_1(g)f_1(e) + z_2(g)f_2(e)\right) P(G = g \mid G_p)}{\sum_{g^*} \exp\left(z_1(g^*)\beta_1 + z_2(g^*)\beta_2 + z_1(g^*)f_1(e) + z_2(g^*)f_2(e)\right) P(G = g^* \mid G_p)},$$

where the sum in the denominator is over genotypes g^* consistent with parental genotypes G_p and the probabilities $P(G = g \mid G_p)$ are given by Mendel's laws.

4.3.2 Avoiding spurious interaction

Recall the diagram in Figure 4.3 that illustrates the source of spurious $G' \times E$ described by Shi *et al.* (2011). The different log-GRRs for G' reflect different patterns of G'-G LD in the different sub-populations. A correctly specified penetrance model

with separate genetic effects for each sub-population would eliminate the spurious interaction. As illustrated in Figure 4.4, spurious interaction is also avoided if we allow separate genetic effects for *groups* of sub-populations with different E distributions. The variable X distinguishes groups, taking value 0 for the group comprised of sub-populations $S = 0$ and $S = 1$, and value 1 for the group comprised of sub-populations $S = 2$ and $S = 3$. In the spirit of Shi *et al.* (2011), we could adjust the penetrance model my X to account for different E-distributions. Although adjusting for X avoids spurious interaction, the E-specific GRRs for G' are mis-specified for each sub-population. In future work we plan to correct for this mis-specification of the penetrance model by allowing each sub-population to have separate genetic effects.

With X we might adjust the risk model to, e.g.,

$$\mathrm{GRR}_i(e) = \frac{P(D=1|G'=i, E=e, X=x)}{P(D=1|G'=i-1, E=e, X=x)} = \exp\big(\beta_i + f_i(e) + \beta_{iX}x\big);\ i = 1, 2 \tag{4.2}$$

so that subjects from group $X = 0$ have genetic main effects β_i and subjects from group $X = 1$ have genetic main effects $\beta_i + \beta_{iX}$. Appendix C provides details on how such adjustments avoid inference of spurious interaction.

In practice, sub-populations and their groupings X are not known. Shi *et al.* (2011) collect dichotomous E on an unaffected sibling and allow separate genetic effects for different values of sibship-averaged E. In effect, the sibship-average is used to tag the distribution of E from which the siblings are sampled. Here we take the alternative approach of allowing separate genetic effects for different values of $\mathrm{E}(E \mid \text{sub-population})$. The idea is to use the sub-population-specific average of E to tag the distribution from which the affected child is sampled. We use $\mathrm{E}(E \mid M) \equiv \mu(M)$ as a proxy for $\mathrm{E}(E \mid \text{sub-population})$ where M is a set of ancestry-informative markers (AIMs) or random SNPs collected on the affected child. We allow GRRs for G' to vary with $\mu(M)$ through the adjusted risk model

$$\frac{P(D=1|G'=i, E=e, M)}{P(D=1|G'=i-1, E=e, M)} = \exp\big(\beta_i + f_i(e) + h_i(\mu(M))\big);\ i = 1, 2. \tag{4.3}$$

That is, we replace $\beta_{iX}x$ in equation (4.2) with $h_i(\mu(M))$.

To model $\mu(M)$ there are several possibilities. We opted to regress E on "important" principal components (PCs), where the number of important PCs is determined by the selection procedure of Zhu and Ghodsi (2006). This selection procedure is an automated version of the standard approach based on inspection of a scree plot (e.g. Jolliffe, 2002, Chapter 6), which graphs the proportion of variance in the data explained by each PC against its rank. Given a number k of important PCs, $PC_1, \ldots, PC_k$, the estimates of $\mu(M)$ for each child are taken to be the child's fitted value from an ordinary least-squares regression of E on $PC_1, \ldots, PC_k$. We took $h_i(\mu(M)) = \gamma_i\mu(M)$, $i = 1, 2$, since in our simulations there are always two modes in the distribution of $\mu(M)$.

4.3.3 Power

Under $G \times E$, patterns of $G' \times E$ can vary with GG' haplotype distributions in different groups of sub-populations. Hence, enforcing a common interaction across sub-populations may yield low power. For example, an allelic correlation flip in the two sub-populations can flip the sub-population-specific effects of $G' \times E$, as illustrated in Figure 4.5. The figure is a schematic of the log-GRRs for G' under negative linear interactions for the causal G such that $f_1(e) = f_2(e) = \beta_{GE} \times e$ with $\beta_{GE} < 0$ (as in our simulation study), and allelic correlations $r_0 = -1$ in $S = 0$ and $r_1 = 1$ in $S = 1$. In $S = 1$, the log-GRR curve for G' is the same as for G. In $S = 0$, the log-GRR curves for G' are the negative of those for G. The result is log-GRR curves for G' with opposite-signed slopes in two sub-populations. Enforcing a common interaction in the two sub-populations misses the signal.

To maximize power in this example, we would like to allow the $G' \times E$ effect to vary by haplotype distribution. For example, letting Y distinguish the two groups of sub-populations with different haplotype distributions, we could modify the risk

model for G' to

$$\begin{aligned} \mathrm{GRR}_i(e) &= \frac{P(D=1|G'=i, E=e, X=x, Y=y)}{P(D=1|G'=i-1, E=e, X=x, Y=y)} \\ &= \exp\big(\beta_i + f_i(e) + \beta_{iX}x + f_{iY}(e)y\big);\ i=1,2 \end{aligned}$$

so that subjects from group $Y=0$ have interaction effects $f_i(e)$ and subjects from group $Y=1$ have interaction effects $f_i(e)+f_{iY}(e)$. However, sub-populations, their haplotype distributions and hence Y are not known.

Assuming different E distributions correspond to different haplotype distributions, as in our example, one could use $\mu(M)$ in the model

$$\frac{P(D=1|G'=i, E=e, M)}{P(D=1|G'=i-1, E=e, M)} =$$

$$\exp\big(\beta_i + f_i(e) + \gamma_i\mu(M) + f_{iY}(e)\mu(M)\big);\ i=1,2.$$

In simple cases of two sub-populations this adjustment would be expected to perform well. However, power is lost if haplotype distributions vary within the distribution of E, as illustrated in Figure 4.6. We emphasize though that the primary motivation for modifying the risk model is to prevent detection of spurious interaction, and that preserving power when interaction is present is a secondary consideration.

4.4 Simulation Study

4.4.1 Simulation settings

We generated 5000 data sets comprised of 3000 case-parent trios with at least one heterozygous parent and information on (G'_p, G', S, E). To induce spurious interaction we simulated a stratified population with different distributions of E and GG' haplotypes in each of two sub-populations. The two sub-populations were of equal size, were randomly mating, and did not mix.

The sub-population distributions of E were chosen to be normal with common

variance $\sigma^2 = 0.36$, and means $\mu_0 = -0.8$ and $\mu_1 = 0.8$ in $S = 0$ and $S = 1$, respectively (Figure 4.1). The resulting population distribution of E has mean zero and variance one. Given sub-population status, E values were simulated independently of all genetic data on the trio.

The GG' haplotype distributions in the two sub-populations were chosen to have common marginal allele frequencies of 0.5 for the index alleles R and 1 of the causal and test locus, respectively. Haplotype distributions were then determined by the allelic correlation between R and 1. We considered two scenarios for the allelic correlations r_0 and r_1 in sub-populations $S = 0$ and $S = 1$: (i) $r_0 = -r$ and $r_1 = r$ and (ii) $r_0 = 0$ and $r_1 = r$, for a grid of r values from zero to one. The first scenario may occur when, for example, an advantageous variant arises independently on different backgrounds in different sub-populations, as in the case of the lactase persistence trait (Tishkoff *et al.*, 2007). The second scenario is plausible in the case of an older sub-population $S = 0$, in which recombination has broken up haplotype structure in the neighborhood of the test locus, and a younger sub-population $S = 1$, in which LD is present due to a founder effect. The haplotype distributions were used to simulate parental haplotypes under Hardy-Weinberg proportions (HWP). Children's haplotypes were then sampled according to Mendel's laws, assuming no recombination between the causal and test loci during parental meioses.

Parameters in the disease risk model (equation 4.1) were chosen as follows. We took $\beta_1 = \beta_2 = \log\sqrt{3}$, giving multiplicative inheritance with GRRs of $\sqrt{3}$ under no $G \times E$. For interaction we set $f_1(e) = f_2(e) = 0$ under no $G{\times}E$ and $f_1(e) = f_2(e) = -0.25e$ under $G{\times}E$. Since the main effect of E is not estimable from case-parent trio data, we set $\eta(e) = 0$ for simplicity. We simulated informative trios according to their population distribution, keeping only those with an affected child. An R function `trioGESim()` written to perform the simulations will be made available in a forthcoming R package.

In the simulation configurations described so far, the allele frequencies at both causal and test loci were always the same in the two sub-populations. However, we also conducted limited simulations in which both allele frequencies at the two loci and

GG' haplotype frequencies vary among sub-populations. The simulation configuration was intended to mimic a situation in which the "risk" allele at the causal locus arose on a common haplotypic background and was under positive selection in one of the sub-populations. Results were qualitatively similar to those obtained in the other configurations, and are not shown.

4.4.2 Unadjusted tests of interaction

We evaluated two tests that did not adjust for population structure: (i) a likelihood ratio test based on a conditional logistic regression (Schaid, 1999) and (ii) FBAT-I (Lake and Laird, 2004). For the conditional logistic regression approach, we fit a model with separate linear interactions for each GRR and tested the significance of these linear interactions with a likelihood ratio test. For the FBAT-I, we computed the p-value based on 10000 Monte-Carlo iterations in the R package fbati (Hoffmann, 2009). The nominal significance level was 0.05.

4.4.3 Adjusted test of interaction

For the adjusted approach, we generated panels of various numbers of substructure-informative markers. The genotypes of these markers were independent of G and G' and each other, and followed HWP (Hardy-Weinberg proportions) within sub-populations. The subpopulation-specific index allele frequencies for these SNPs were generated independently from a uniform distribution $U(l, u)$, with $l < u$. We considered two types of markers: ancestry informative markers (AIMs) and random SNPs. AIMs were chosen such that the subpopulation-specific allele frequencies differed by at least $\delta = 0.4$. To be conservative in evaluating our procedure, our threshold for classifying a SNP as ancestry informative was more inclusive than the $\delta = 0.5$ value that has been recommended (Shriver *et al.*, 1997). Random SNPs included all markers, without any restriction on allele frequency differences. We considered panels of 50 and 100 AIMs and 100 and 250 random SNPs. Models were fit by conditional logistic regression and hypotheses were tested by likelihood ratio tests.

4.4.4 Type 1 error results

Simulation results under no $G \times E$ are given in Figures 4.7 (for $r_0 = -r$, $r_1 = r$) and 4.8 (for $r_0 = 0$, $r_1 = r$). In both cases, the type 1 error rate of the unadjusted methods quickly increases beyond the nominal 5% with increasing r. However, with enough independent marker data, the type 1 error rate of the adjusted approach stays near the nominal level. In our simulation setting, about 250 random SNPs or 100 AIMs appears to be enough to reliably identify the E distribution of each subject.

4.4.5 Power results

Results under $G \times E$ are summarized in Figures 4.9 (for $r_0 = -r$, $r_1 = r$) and 4.10 (for $r_0 = 0$, $r_1 = r$). Each figure displays the relative power of the test for interaction, defined as the power based on the test SNP divided by power based on the causal SNP. The first simulation setting, with allelic correlations of $r_0 = -r$ and $r_1 = r$, is the one in which interaction curves flip, so that enforcing a common interaction curve would miss the signal. Reassuringly, the simulation results from this configuration (Figure 4.9) show that, with the proposed adjustment, relative power increases in r and approaches one as r approaches one. In the second configuration with $r_0 = 0$ and $r_1 = 1$, the test SNP provides no information about $G \times E$ in sub-population $S = 0$. This is reflected by lower powers in the second configuration (Figure 4.10), relative to the first, for the same value of r.

4.5 Discussion

$G \times E$ describes statistical interaction; i.e., departures from additive effects on some scale defined by a linear model (Cordell, 2002). When this linear model is mis-specified, the interpretation of statistical interaction is problematic. The general issue is well known in both plant genetics (Gauch, 2006) and epidemiology (Clayton, 2009). The issue has also been recognized in the context of case-parent trio studies (Umbach and Weinberg, 2000; Lake and Laird, 2004; Cordell, 2009). More recently, Shi *et al.*

(2011) have shown how misspecification of a log-additive model of disease risk can arise from LD between a test locus G' and causal locus G. They demonstrate spurious interaction when GG' haplotype frequencies and the distribution of the environmental variable E vary by sub-population. Such variation in haplotype distributions may occur due to various forces such as migration, genetic drift, selection, etc. Shi et al. proposed to avoid this LD-based source of spurious interaction via a design-based approach requiring case-parent trios and measurements of E on an unaffected sibling of the affected child. We provide further investigation of this LD-based source of spurious interaction and propose an alternate approach to inference that relies only on data from the case-parent triads.

Both the design-based approach and our adjusted approach require extra data compared to a traditional case-parent design. The design-based approach requires E on unaffected siblings, while the adjusted approach requires independent marker genotypes on affected children. For studies in the planning stages, either requirement would be straightforward to incorporate. For example, including AIMs on custom genotyping chips is common practice, while in genome-wide studies there is an abundance of genotypes available and no extra genotyping is necessary. For studies in which genotyping has been completed and does not include independent markers, the extra genotyping is an extra cost, but is at least logistically straightforward, relying only on the availability of DNA samples for study subjects. By contrast, adding unaffected siblings to a study that has completed contact with subjects would likely be difficult. We therefore feel that our approach can be useful in studies which have completed contact with subjects but have DNA samples available.

Our approach is based on an adjustment to the risk model that uses independent marker information to identify each subject's E distribution. In our simulations with two underlying subpopulations, this adjustment maintained the nominal type 1 error rate when sufficient independent marker information was available to reliably identify the E distribution of each subject. In our setting, about 250 random SNPs or 100 AIMs was sufficient. Under our simulation setting, a simple adjustment to the risk model with $h_i(\mu(M)) = \gamma_i \mu(M)$ (see equation 4.3) worked well, but it is less clear

how to proceed with more than two sub-populations. One possibility that we are currently investigating will be discussed in Chapter 5.

We have also discussed adjustments to the risk model aimed at preserving power when $G \times E$ exists in the special case when the distributions of E and haplotypes coincide. These adjustments performed well in our simulations under the special case but would not be expected to work well in general. Development of a more general-purpose adjustment to preserve power is therefore of interest.

The method we used for simulating genotypes and environmental covariates in affected, informative case-parent trios has been implemented in a freely-available R package that will soon be released to the Comprehensive R Archive Network (CRAN).

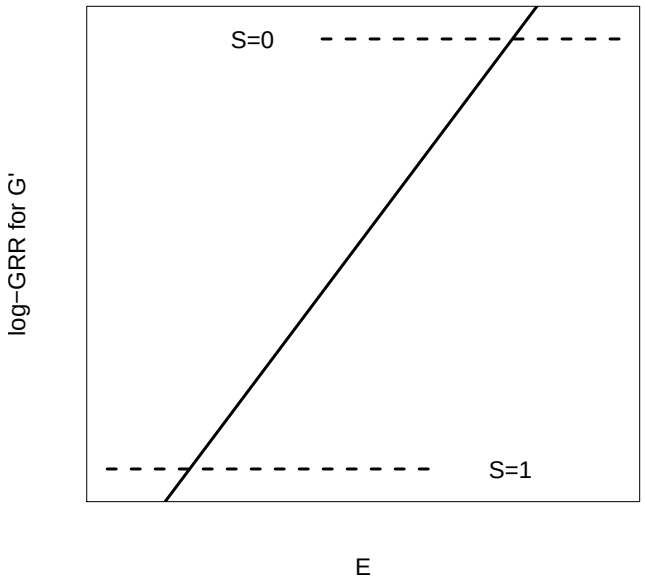

Figure 4.3: Schematic of log-GRRs for G' versus E in a structured population with two sub-populations $S = 0$ and $S = 1$. Dashed curves represent log-GRRs within each sub-population. Solid curve represents a linear log-GRR curve fit to data from both sub-populations combined.

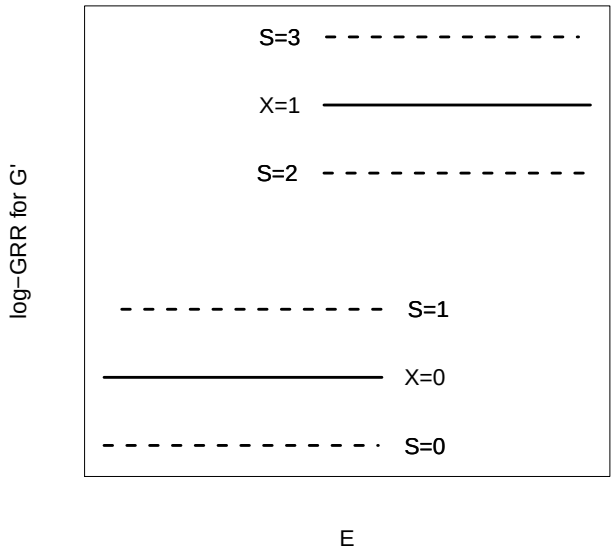

Figure 4.4: Schematic of log-GRRs for G' versus E in a structured population with four sub-populations $S = 0, 1, 2, 3$. Dashed lines represent log-GRRs in the different sub-populations. Spurious interaction is avoided if we allow separate genetic effects (solid lines) for the group $X = 0$ comprised of sub-populations $S = 0$ and $S = 1$ and the group $X = 1$ comprised of sub-populations $S = 1$ and $S = 2$.

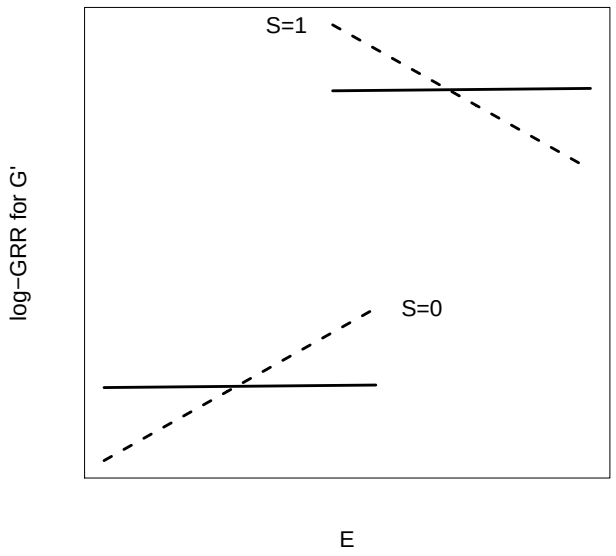

Figure 4.5: Schematic to illustrate how a flip in allelic correlations flips $G' \times E$. Dashed lines indicate log-GRR curves for $G' = 1$ versus 0 in sub-populations $S = 0$ and $S = 1$. The solid line is the fitted log-GRR that would result from enforcing a common interaction in the two sub-populations.

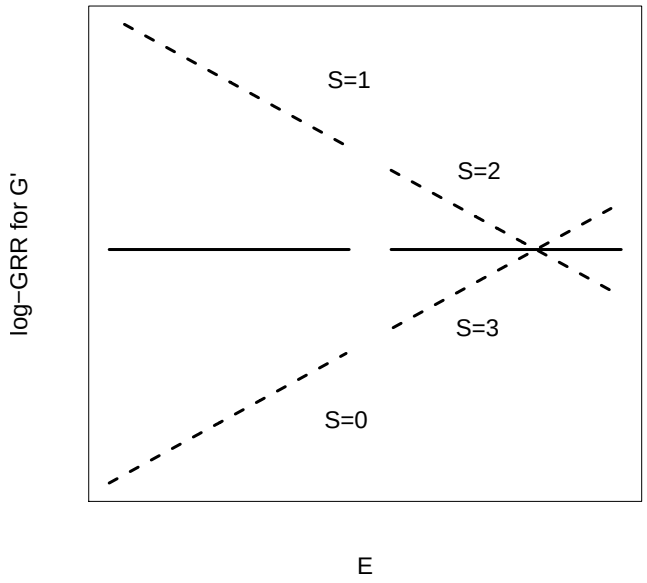

Figure 4.6: The proposed approach may miss interaction signal when haplotype distributions vary within E distributions. Dashed lines indicate log-GRR curves for $G' = 1$ versus 0 in sub-populations $S = 0$ through $S = 3$. The solid line is the fitted log-GRR that would result from allowing interaction to vary according to the distribution of E rather than the distribution of haplotypes.

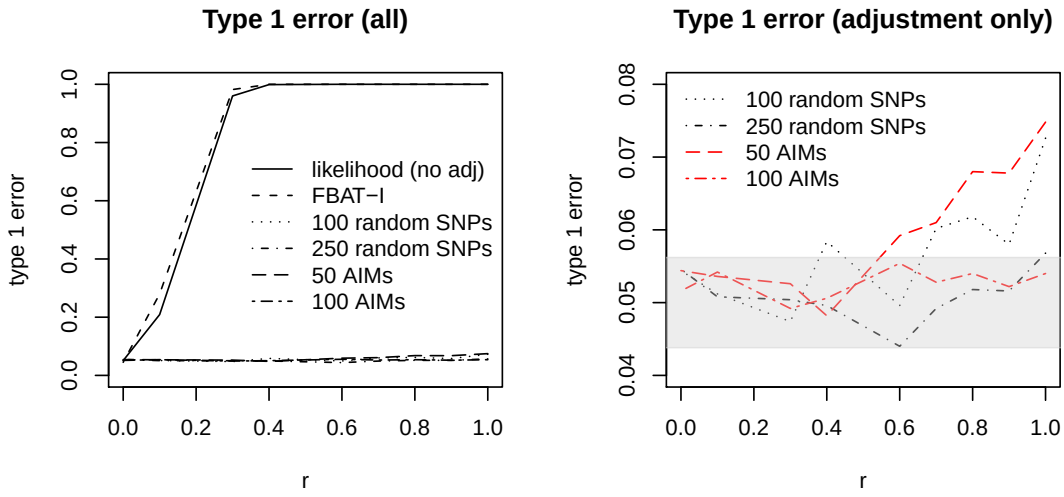

Figure 4.7: Type 1 error as a function of the allelic correlation r, with correlation between R at G and 1 at G' being $-r$ in $S = 0$ and r in $S = 1$. The left panel includes all methods considered, while the right panel includes only those that make an adjustment to avoid spurious interaction. The shaded area in the right panel represents simulation error about the nominal 5% level – a test with the correct size would have estimated type 1 error within the shaded region 19 times out of 20.

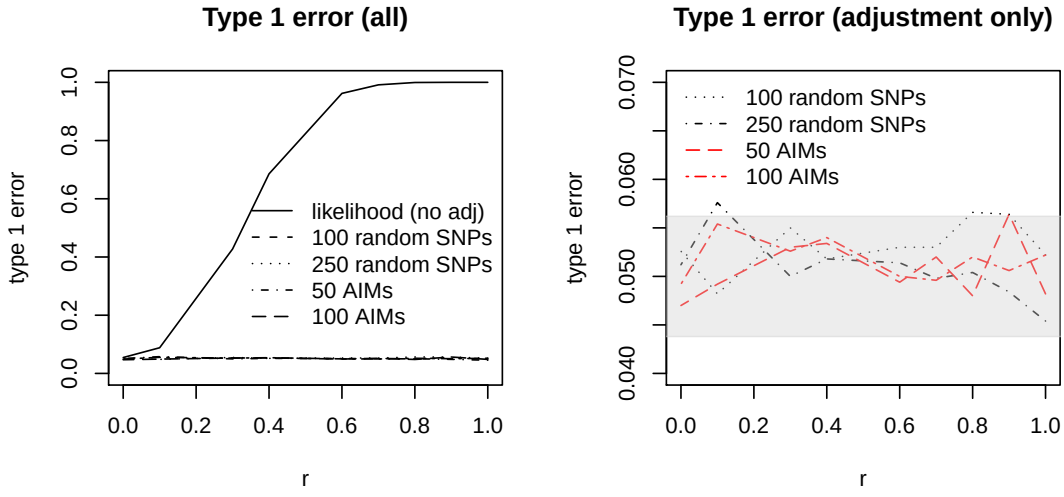

Figure 4.8: Type 1 error as a function of the allelic correlation r, with correlation between R at G and 1 at G' being 0 in $S = 0$ and r in $S = 1$. The left panel includes all methods considered, while the right panel includes only those that make an adjustment to avoid spurious interaction. The shaded area in the right panel represents simulation error about the nominal 5% level – a test with the correct size would have estimated type 1 error within the shaded region 19 times out of 20. Results based on 5000 simulation replicates of 3000 informative trios.

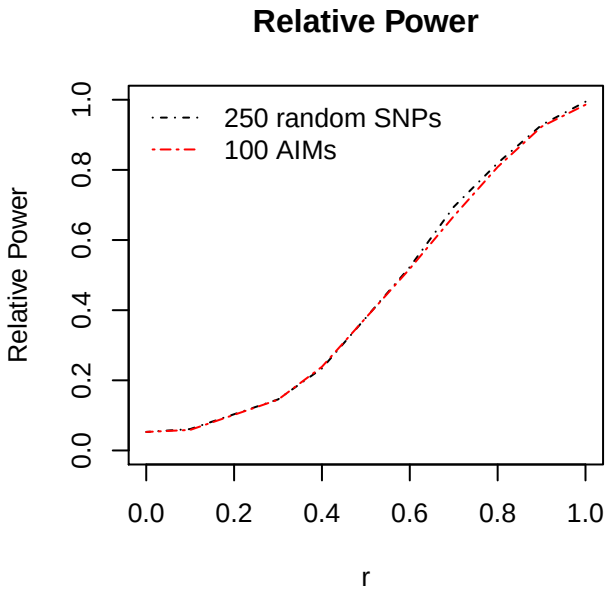

Figure 4.9: Relative power of the non-causal SNP to detect $G \times E$. Sub-population-specific haplotype distributions are summarized by GG' allelic correlations of $r_0 = -r$ and $r_1 = r$. Results based on 5000 simulation replicates of 3000 informative trios. Interaction effects were $f_1(e) = f_2(e) = -0.25e$.

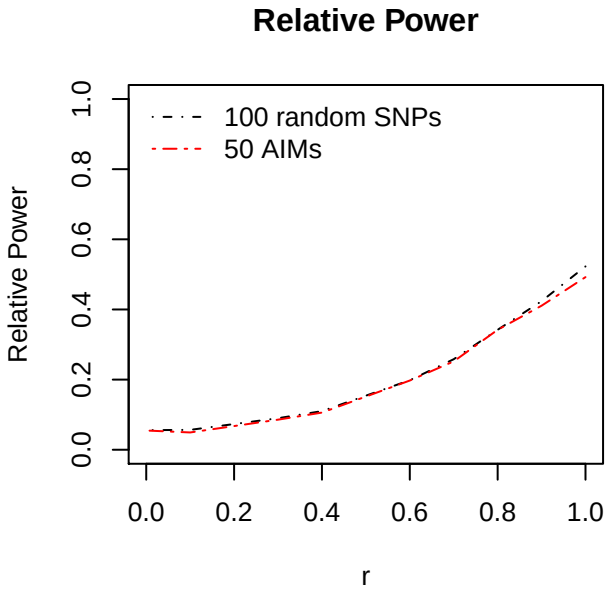

Figure 4.10: Relative power of the non-causal SNP to detect $G \times E$. Sub-population-specific haplotype distributions are summarized by GG' allelic correlations of $r_0 = 0$ and $r_1 = r$. Results based on 5000 simulation replicates of 3000 informative trios. Interaction effects were $f_1(e) = f_2(e) = -0.25e$.

Chapter 5

Conclusions

The three projects in this thesis consider the sources and remedies of bias that can arise in the analysis of $G{\times}E$ in data from case-parent trios.

In Chapter 2, we revisited the problem of inferring $G{\times}E$ with a transmission-based approach. We discussed how, in the absence of $G{\times}E$, transmission rates can vary with E because of, for example, correlation between G and E due to population stratification. A simulation study demonstrated that the transmission-based test can have inflated type 1 error rate and lower power than a likelihood-based test. To guard against false-positive $G{\times}E$ due to population stratification, we suggest exploratory graphical displays of transmission-rates within parent mating types.

In Chapter 3, we developed a data-smoothing method to explore $G{\times}E$, using data from case-parent trios. The method protects against the bias caused by mis-specification of the parametric form or the mode of inheritance of $G{\times}E$. In particular, it models the $G{\times}E$ functions using spline functions and allows for separate genotype relative risks depending on the number of copies of the variant allele. A permutation-based test was adopted to test $G{\times}E$, taking into account the additional uncertainty due to the smoothing process. A simulation study demonstrated that the proposed test can detect non-linear $G{\times}E$ better than other available approaches.

One concern with our proposed smoothing approach is that the $G{\times}E$ curve estimates and the performance of the associated confidence intervals are sensitive to how data are distributed, which is a general issue in smoothing approaches (e.g., Nychka,

1988). In general, when the observations are distributed evenly throughout the range of the data, smoothing methods perform well; the estimated curves have low bias, and their associated confidence intervals have good coverage probabilities. However, when the data are sparsely distributed, the approaches can miss a more complicated feature of the underlying smooth function (e.g., curvature in quadratic $G\times E$). Sparse data regions can have estimates with large bias and hence confidence intervals with coverage probabilities that are lower than the nominal confidence level.

Poor performance in sparse data regions was observed for our smoothing approach in a preliminary simulation study. In the simulation study, the performance was evaluated based on the coverage probabilities for the 95% Bayesian confidence intervals of $G\times E$ functions under the no-$G\times E$ null hypothesis, linear $G\times E$ and quadratic $G\times E$. The confidence intervals have good coverage probabilities when $G\times E$ is a simple linear or horizontal (i.e., no $G\times E$) function. However, when the $G\times E$ function is quadratic, the proposed approach can perform poorly. Under no population stratification, the confidence intervals have poor coverage probabilities when the variant allele frequency is either high or low. Under population stratification, the confidence intervals behave poorly when the degree of population stratification is high. In both situations, a large bias at sparse boundary regions seemed to be responsible for the poor performance (e.g., Figure 5.1). One solution to this problem is to increase the sample size to obtain more points across the range of E. However, collecting more data may be hard in practice. Thus, analysts should be aware of the limitations of sparse data before using this or any data smoothing approach.

In the third project, we investigated how approaches to $G\times E$, including our own smoothing approach, are not robust to population stratification when the test marker is not causal but linked to an unobserved causal gene. When we have such a test marker, population stratification arises when there exist hidden subpopulations that have different distributions of the non-genetic covariate and different distributions for the frequency of the haplotype comprised of the test and the causal marker. The project provided further insights into how this type of population stratification can lead to both false-positive and false-negative $G\times E$. To protect against population

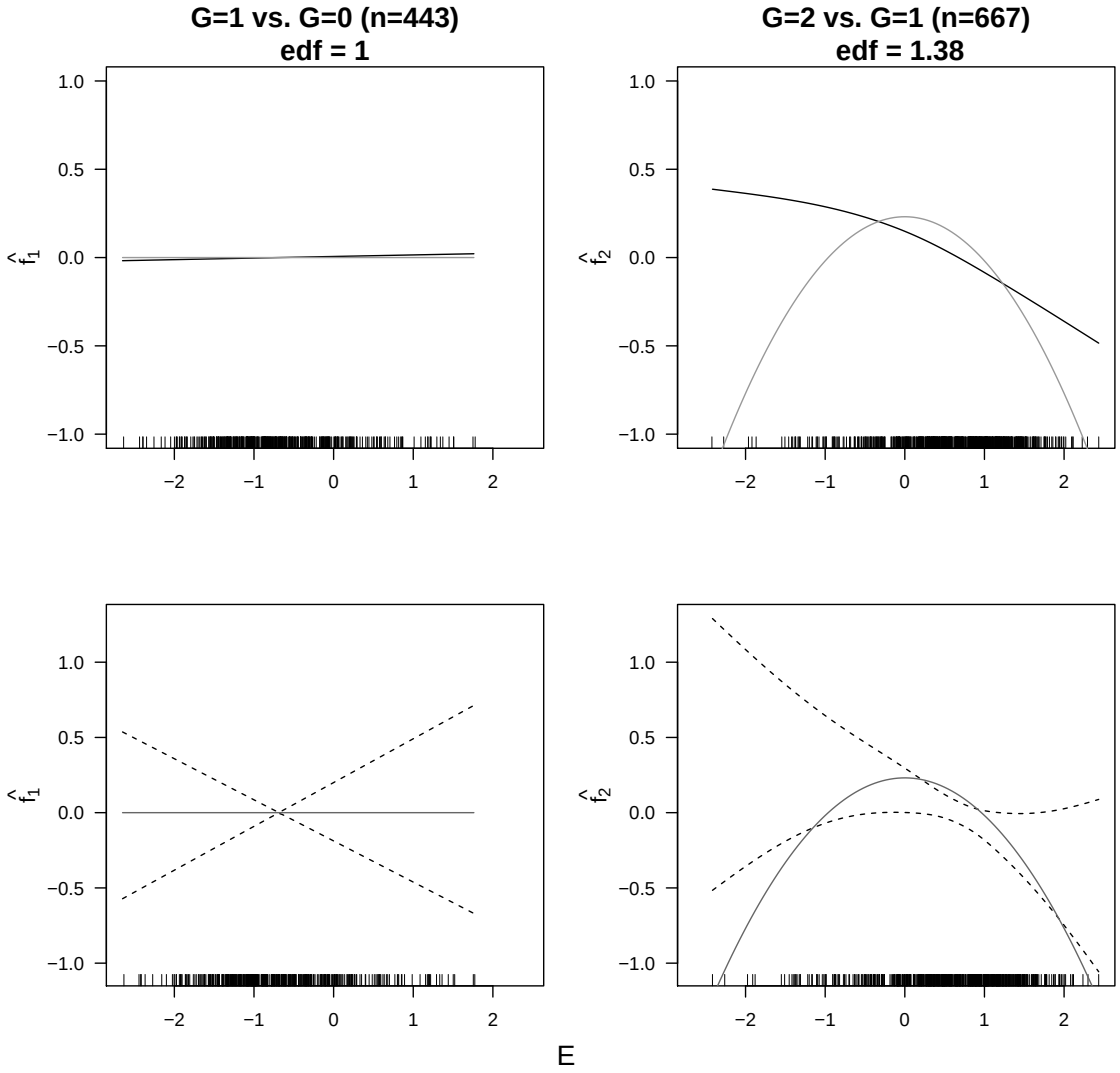

Figure 5.1: Estimated $G\times E$ functions for a dataset based on 1000 case-parent trios under quadratic $G\times E$ with $f_1(e) = 0, f_2(e) = 0.25e^2$ (i.e., recessive penetrance). The data were generated under population stratification with two subpopulations $S = 0$ and 1. Subpopulation $S = 0$ ($S = 1$) has a variant allele frequency of $q_0 = 0.1$ ($q_1 = 0.9$) and non-genetic covariate distributed as $E \sim N(-0.8, 0.36)$ ($E \sim N(0.8, 0.36)$). The black solid lines represent the estimated $\hat{f}_1(e)$ and $\hat{f}_2(e)$; the grey solid lines, the true $f_1(e)$ and $f_2(e)$; and the black-dashed lines indicate the 95% Bayesian confidence intervals.

stratification, we proposed an approach to adjusting the disease risk model, based on ancestry informative markers or random markers measured on the affected children. The simulation results demonstrated that when sufficient independent marker information is available, the adjustments maintain the nominal level while maintaining power.

The approaches developed in Chapters 3 and 4 can be extended in several directions to incorporate other genetic effects and/or a more general population structure. In what follows, we will discuss some possibilities for future research.

For the smoothing approach developed in Chapter 3, one possible direction for future research is to extend the current approach to incorporate parent-of-origin effects, as follows. For this, we consider the ordered child-genotypes $G = 0, 1_M, 1_F, 2$, where 1_M (1_F) is a heterozygote with index allele inherited from the mother (father). The informative mating types, $G_p = (G_M, G_F) \in \{(1,0),(0,1),(1,2),(2,1),(1,1)\}$, now have five possibilities rather than three. The disease risk model can be written similarly to the one in (3.1), such that

$$P(D = 1 \mid G = g, E = e) = \exp(k + \xi(e) + \boldsymbol{\gamma}\boldsymbol{z}(g) + \boldsymbol{f}(e)\boldsymbol{z}(g)),$$

where $\boldsymbol{\gamma}, \boldsymbol{z}(g)$ and $\boldsymbol{f}(e)$ are length-three vectors rather than two; the genetic coding vector $\boldsymbol{z}(g) = (z_{1M}(g), z_{2M}(g), z_2(e))^{\mathsf{T}}$, the main genetic effect parameters $\boldsymbol{\gamma} = (\gamma_{1M}, \gamma_{1F}, \gamma_2)$, and the $G{\times}E$ functions $\boldsymbol{f}(e) = (f_{1M}(e), f_{1F}(e), f_2(e))$. The elements of $\boldsymbol{z}(g)$ are defined as $z_{1M}(g) = \mathrm{I}\{g = 1_M\}$, $z_{1F}(g) = \mathrm{I}\{g = 1_F\}$ and $z_2(g) = \mathrm{I}\{g = 2\}$.

Under this model, genotype relative risks can be represented by

$$\begin{aligned} \mathrm{GRR}_{1M}(e) &= \gamma_{1M} + f_{1M}(e) \\ \mathrm{GRR}_{1F}(e) &= \gamma_{1F} + f_{1F}(e) \\ \mathrm{GRR}_2(e) &= \gamma_2 + f_2(e), \end{aligned}$$

where

$$\mathrm{GRR}_g(e) \equiv P(D = 1 \mid G = g, E = e)/P(D = 1 \mid G = 0, E = e),$$

indicating the genotype relative risk between individuals with $G = g$ ($g \neq 0$) and those with $G = 0$. The $G{\times}E$ functions $f_{1M}(e)$ and $f_{1F}(e)$ allow for separate interaction terms depending on whether the child inherits the variant allele from the mother or from the father. Consequently, under $G{\times}E$, $f_{1M}(e) \neq f_{1F}(e)$ indicates there is a parent-of-origin effect. When $f_{1M}(e) = f_{1F}(e) \neq 0$ or $f_{1M}(e) = f_{1F}(e) = 0$ and $f_2(e) \neq 0$, it indicates that there is $G{\times}E$ but no parent-of-origin effect. Under no $G{\times}E$, we have $f_{1M}(e) = f_{1F}(e) = f_2(e) \equiv 0$.

One important issue arising from incorporating parent-of-origin effects is that, for trios with all members heterozygous, one cannot observe whether $G = 1_M$ or $G = 1_F$, which leads to a missing-data problem. To handle the missing information, one can adopt an expectation-maximization (EM) approach (Dempster *et al.*, 1977).

It is anticipated that many case-parent trios would be required for this extended smoothing method since it estimates more parameters than the original smoothing approach, which already requires a large sample size due to its generality. One strategy for retaining power is to restrict the disease penetrance model (e.g., to dominant) in order to reduce the number of the parameters that must be estimated.

In Chapter 4, recall that M is a set of ancestry-informative markers (AIMs) or random SNPs collected on the affected child, and $\mathbb{E}(E|M) \equiv \mu(M)$. The proposed approach to adjusting the risk model using $h_i(\mu(M)) = \gamma_i \mu(M)$ in equation (4.3) was demonstrated to work well under two subpopulations, but it would not be expected to do well under three or more subpopulations. To incorporate a more general population structure, we are currently investigating one approach based on identifying clusters in the distribution of $\mu(M)$ among affected children. We use a clustering approach to determine the number of clusters k in the distribution of $\mu(M)$ and estimate probabilities $p(\mu(M)) \equiv (p_1, \ldots, p_k)^T$ of membership in each cluster for each affected child. These cluster membership probabilities are then used as predictors in the adjusted risk model; i.e., we take $h_i(\mu(M)) = \vec{\gamma}_i^T p(\mu(M))$ for equation (4.3), where $\vec{\gamma}_i$ is a vector of regression coefficients of length k. While this approach has shown promise in initial tests, more work is required to fully evaluate its properties.

Another possible direction for future research is to modify the smoothing approach

of Chapter 3 to protect against false $G\times E$ due to population stratification that occurs when the test marker is not causal. For example, to protect against false positive $G\times E$ under two subpopulations, we can modify the GRRs in equation (3.2), using the adjustment shown in equation (4.3), with $h_i(\mu(M)) = \delta_i\mu(M)$, $i = 1, 2$. This adjustment would lead to the estimation of two more parameters δ_1 and δ_2 in addition to the original genetic main effect terms γ_1 and γ_2 in equation (3.2).

Bibliography

Bellgrove, M. A., Mattingley, J. B., Hawi, Z., Mullins, C., Kirley, A., Gill, M., and Robertson, I. H. (2006). Impaired temporal resolution of visual attention and dopamine beta hydroxylase genotype in attention-deficit/hyperactivity disorder. *Biol Psychiatry*, **60**, 1039–1045.

Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., Rhodes, M., Reich, D. E., and Hirschhorn, J. N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet*, **74**(6), 1111–1120. PMID: 15114531 PMCID: 1182075.

Brookes, K. J., Neale, B., Xu, X., Thapar, A., Gill, M., Langley, K., Hawi, Z., Mill, J., Taylor, E., Franke, B., Chen, W., Ebstein, R., Buitelaar, J., Banaschewski, T., Sonuga-Barke, E., Eisenberg, J., Manor, I., Miranda, A., Oades, R. D., Roeyers, H., Rothenberger, A., Sergeant, J., Steinhausen, H. C., Faraone, S. V., and Asherson, P. (2008). Differential dopamine receptor D4 allele association with ADHD dependent of proband season of birth. *Am J Med Genet B Neuropsychiatr Genet*, **147B**, 94–99.

Caillat-Zucman, S., Garchon, H., Timsit, J., Assan, R., Boitard, C., Djilali-Saiah, I., Bougneres, P., and Bach, J. (1992). Age-dependent HLA genetic heterogeneity of type 1 insulin-dependent diabetes mellitus. *J of Clin Invest*, **90**(6), 2242–2250.

Cheng, S.-L., Yu, C.-J., Chen, C.-J., and Yang, P.-C. (2004). Genetic polymorphism of epoxide hydrolase and glutathione S-transferase in COPD. *Eur Respir J*, **23**, 818–824.

Clavel, J., Bellec, S., Rebouissou, S., Ménégaux, F., Feunteun, J., Bonaïti-Pellié, C., Baruchel, A., Kebaili, K., Lambilliotte, A., Leverger, G., *et al.* (2005). Childhood leukaemia, polymorphisms of metabolism enzyme genes, and interactions with maternal tobacco, coffee and alcohol consumption during pregnancy. *Eur J of Cancer Prev*, **14**(6), 531–540.

Clayton, D. G. (2009). Prediction and interaction in complex disease genetics: experience in type 1 diabetes. *PLoS Genet*, **5**, e1000540.

Cordell, H., Barratt, B., and Clayton, D. (2004). Case/pseudocontrol analysis in genetic association studies: A unified framework for detection of genotype and

haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects. *Genet Epidemiol*, **26**, 167–185.

Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet*, **11**, 2463–2468.

Cordell, H. J. (2009). Estimation and testing of gene-environment interactions in family-based association studies. *Genomics*, **93**, 5–9.

Database of Single Nucleotide Polymorphisms (dbSNP) (build 135). *National Center for Biotechnology Information, National Library of Medicine dbSNP accession: rs1800566*. Bethesda, MD. Avaliable from: http://www.ncbi.nlm.nih.gov/SNP/.

Dempfle, A., Scherag, A., Hein, R., Beckmann, L., Chang-Claude, J., and Schäfer, H. (2008). Gene–environment interactions for complex traits: definitions, methodological requirements and challenges. *Eur J Hum Genet*, **16**(10), 1164–1172.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *J Roy Stat Soc B Met*, **39**, 1–38.

Du, J., Duan, S., Wang, H., Chen, W., Zhao, X., Zhang, A., Wang, L., Xuan, J., Yu, L., Wu, S., Tang, W., Li, X., Li, H., Feng, G., Xing, Q., and He, L. (2008). Comprehensive analysis of polymorphisms throughout GAD1 gene: a family-based association study in schizophrenia. *J Neural Transm*, **115**, 513–519.

Duke, L. (2007). *A graphical tool for exploring SNP-by-environment interaction in case-parent trios.* Master's thesis, Statistics and Actuarial Science: Simon Fraser University.

Gauch, H. G. J. (2006). Statistical analysis of yield trials by AMMI and GCE. *Crop Sci*, **46**, 1488–1500.

Guilmatre, A. and Sharp, A. (2012). Parent of origin effects. *Clin Genet*, **81**(3), 201–209.

Hamblin, M. T., Thompson, E. E., and Rienzo, A. D. (2002). Complex signatures of natural selection at the Duffy blood group locus. *Am J Hum Genet*, **70**(2), 369–383.

Hoffmann, T. (2009). *fbati: Gene by Environment Interaction and Conditional Gene Tests for Nuclear Families*. R package version 0.7-1.

Hogg, R., McKean, J., and Craig, A. (2005). *Introduction to mathematical statistics*. Pearson Education, Upper Saddle River, N.J., 6th edition.

Infante-Rivard, C. (2003). Hospital or population controls for case-control studies of severe childhood diseases? *Am J Epidemiol*, **157**, 176–182.

Infante-Rivard, C., Fortier, I., and Olson, E. (2000). Markers of infection, breast-feeding and childhood acute lymphoblastic leukaemia. *Birt J Cancer*, **83**, 1559–1564.

Infante-Rivard, C., Vermunt, J., and Weinberg, C. (2007). Excess transmission of the NAD(P)H: Quinone Oxidoreductase 1 (NQO1) C609T polymorphism in families of children with acute lymphoblastic leukemia. *Am J Epidemiol*, **165**, 1248–1254.

Jolliffe, I. (2002). *Principal Component Analysis*. Springer, 2nd edition.

Lake, S. and Laird, N. (2004). Tests of gene-environment interaction for case-parent triads with general environmental exposures. *Ann Hum Genet*, **68**(1), 55–64.

Li, C. C. (1955). *Population Genetics*. University of Chicago Press.

Liang, K. (1983). On information and ancillarity in the presence of a nuisance parameter. *Biometrika*, **70**(3), 607–612.

Lim, S., Beyene, J., and Greenwood, C. (2005). Continuous covariates in genetic association studies of case-parent triads: Gene and gene-environment interaction effects, population stratification, and power analysis. *Stat Appl Genet Mol Biol*, **4**, Article 20. **Available at:** http://www.bepress.com/sagmb/vol4/iss1/art20.

Ma, J., Sun, J., Zhang, H., Zhang, R., Kang, W.-H., Gao, C.-G., Liu, H.-S., Ma, X.-H., Min, Z.-X., Zhao, W.-X., Ning, Q.-L., Wang, S.-H., Zhang, Y.-C., Guo, T.-W., and Lu, S.-M. (2009). Evidence for transmission disequilibrium at the daoa gene locus in a schizophrenia family sample. *Neurosci Lett*, **462**, 105–108.

Marra, G. and Wood, S. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scand J Stat*, **39**(1), 53–74.

Moerkerke, B., Vansteelandt, S., and Lange, C. (2010). A doubly robust test for gene-environment interaction in family-based studies of affected offspring. *Biostatistics*, **11**(2), 213–225.

Nychka, D. (1988). Bayesian confidence intervals for smoothing splines. *J Am Stat Assoc*, **83**(404), 1134–1143.

Perrillat, F., Clavel, J., Jaussent, I., Baruchel, A., Leverger, G., Nelken, B., Philippe, N., Schaison, G., Sommelet, D., Vilmer, E., Bonaïti-Pellié, C., and Hémon, D. (2001). Family cancer history and risk of childhood acute leukemia (france). *Cancer Causes Control*, **12**, 935–941.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ries, L., Smith, M., Gurney, J., Linet, M., Tamra, T., Young, J., and Bunin ,G. (eds) (1999). *Cancer Incidence and Survival among Children and Adolescents: United States SEER Program 1975-1995*. National Cancer Institute, SEER Program. NIH Pub. No. 99-4649, Bethesda, MD.

Schaid, D. J. (1999). Case-parents design for gene-environment interaction. *Genet Epidemiol*, **16**(3), 261–273.

Shi, M., Umbach, D., and Weinberg, C. (2011). Family based gene-by-environment interaction studies: Revelations and remedies. *Epidemiology*, **22**, 400–407.

Shin, J.-H., McNeney, B., and Graham, J. (2010). On the use of allelic transmission rates for assessing gene-by-environment interaction in case-parent trios. *Ann Hum Genet*, **74**(5), 439–451.

Shin, J.-H., Infante-Rivard, C., Graham, J., and McNeney, B. (2012). Adjusting for spurious gene-by-environment interaction using case-parent triads. *Stat Appl in Genet Molec Biol*, **11**, Article 7.

Shriver, M., Smith, M., Jin, L., Marcini, A., Akey, J., Deka, R., and Ferrell, R. (1997). Ethnic-affiliation estimation by use of population-specific DNA markers. *Am J Hum Genet*, **60**(4), 957–964.

Smith, P. and Day, N. (1984). The design of case-control studies: the influence of confounding and interaction effects. *Int J Epidemiol*, **13**(3), 356.

Spielman, R., McGinnis, R., and W., E. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Gen*, **52**, 506–516.

Teo, Y. Y., Fry, A. E., Bhattacharya, K., Small, K. S., Kwiatkowski, D. P., and Clark, T. G. (2009). Genome-wide comparisons of variation in linkage disequilibrium. *Genome Res*, **19**(10), 1849–1860. PMID: 19541915.

Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., Powell, K., Mortensen, H. M., Hirbo, J. B., Osman, M., Ibrahim, M., Omar, S. A., Lema, G., Nyambo, T. B., Ghori, J., Bumpstead, S., Pritchard, J. K., Wray, G. A., and Deloukas, P. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet*, **39**(1), 31–40.

Umbach, D. and Weinberg, C. (2000). The use of case-parent triads to study joint effects of genotype and exposure. *Am J Hum Gen*, **66**, 251–61.

Wahlund, S. (1928). Zuzammensetzung von populationen und korrelationserscheiunungen von standpunkt der vererbungslehreaus betrachte. *Hereditas*, **11**, 65–106.

Wang, T.-N., Chen, W.-Y., Huang, Y.-F., Shih, N.-H., Feng, W.-W., Tseng, H.-I., Lee, C.-H., and Ko, Y.-C. (2006). The synergistic effects of the IL-9 gene and environmental exposures on asthmatic Taiwanese families as determined by the transmission/disequilibrium test. *Int J Immunogenet*, **33**, 105–110.

Wiemels, J., Pagnamenta, A., Taylor, G., Eden, O., Alexander, F., Greaves, M., *et al.* (1999). A lack of a functional NAD (P) H: quinone oxidoreductase allele is selectively associated with pediatric leukemias that have MLL fusions. *Cancer Res*, **59**, 4095–4099.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC, Boca Ranton, FL.

Wood, S. N. and Augustin, N. H. (2002). GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecol Model*, **157**, 157 – 177.

Yee, T. and Wild, C. (1996). Vector generalized additive models. *J Roy Stat Soc B Met*, **58**, 481–493.

Zaykin, D. V. and Shibata, K. (2008). Genetic flip-flop without an accompanying change in linkage disequilibrium. *Am J Hum Genet*, **82**(3), 794–796.

Zhu, M. and Ghodsi, A. (2006). Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Comput Stat Data An*, **51**(2), 918–930.

Appendix A

A.1 Details of Model

Let $m = (m_1, m_2)$ index G_p with specific values in $\mathcal{M} = \{(0,0), (0,1), (0,2), (1,1), (1,2), (2,2)\}$, corresponding to $(G_M, G_F) \in \Big\{\{(0,0)\},\ \{(0,1),(1,0)\},\ \{(0,2),(2,0)\},\ \{(1,1)\},\ \{(2,1),(1,2)\}, \{(2,2)\}\Big\}$, respectively. The three mating types in the set $\mathcal{I} = \{(0,1), (1,2), (1,1)\}$ have at least one heterozygous parent and are *informative* because they yield variation in child genotypes under the Mendelian segregation law.

Throughout, we assume that G and E are conditionally independent given the parental genotypes; that both maternal and parent-of-origin effects on the disease risk are absent; and that the disease risk of a child with $G = g$ copies of the risk allele and non-genetic attribute $E = e$ follows a log-additive model: $\log\big[\Pr(D \mid G = g, E = e)\big] = k(e) + z(g)\beta + z(g)f(e)$, where $k(e)$ denotes the disease probability for those with the reference genotypes g_0 and value e of the non-genetic attribute; $z(g_0) = 0$ and $f(e)$ is a smooth function with $f(e_0) = 0$ for the reference value e_0 of the non-genetic attribute. This model generalizes the log-linear model of Umbach and Weinberg (2000). The term $z(g)f(e)$ represents $G \times E$. Throughout, we refer to the genetic coding $z(g) = g$ as the multiplicative penetrance model, to the coding $z(g) = I\{g > 0\}$ as the dominant penetrance model, to the coding $z(g) = I\{g = 2\}$ as the recessive penetrance model, and to the coding $z(g) = (I\{g = 1\}, I\{g = 2\})^T$ as the co-dominant penetrance model, where the indicator function $I\{\text{relation}\}$ takes the value 1 when the relation is true and 0 otherwise. Under the log-additive model,

$$R_g(e) = \frac{\Pr(D \mid G = g, E = e)}{\Pr(D \mid G = 0, E = e)} = \exp\left(z(g)\beta + z(g)f(e)\right), \quad \text{for } g = 1, 2, \qquad \text{(A1)}$$

and the hypothesis of no $G \times E$ is equivalent to the hypothesis that $f(e) = 0$ for every value of e. The genetic codings $z(g)$ above allow $R_1(e)$ and $R_2(e)$ to have the

appropriate relationships under each genetic model:

$$\begin{cases} R_1(e) = R_2(e) \equiv R(e) & \text{if dominant;} \\ R_1(e) \equiv R(e),\ R_2(e) = R^2(e) & \text{if multiplicative; and} \\ R_1(e) = 1,\ R_2(e) \equiv R(e) & \text{if recessive.} \end{cases} \tag{A2}$$

A.2 Transmission rates

Let T denote the event that a parent transmits the risk allele to his/her child, and H, the event that the parent is heterozygous. Then, the transmission rate may be written as

$$\begin{aligned} \tau(e) &= \Pr(T \mid H, D, E = e) \\ &= \sum_{m \in \mathcal{I}} \Pr(T \mid G_p = m, H, D, E = e) \Pr(G_p = m \mid H, D, E = e) \\ &\equiv \sum_{m \in \mathcal{I}} \tau_m(e) w_m(e). \end{aligned} \tag{A3}$$

Under multiplicative penetrance, the mating-type-specific transmission rate for any $m \in \mathcal{I}$, listed in Table 2.1, is $\tau_m(e) = R(e)/[1 + R(e)]$, where $R(e)$ is defined based on the appropriate relationship in (A2). Hence, the overall transmission rate $\tau(e)$ in equation (A3) becomes

$$\tau(e) = \sum_{m \in \mathcal{I}} \frac{R(e)}{1 + R(e)} w_m(e) = \frac{R(e)}{1 + R(e)}, \quad \text{because } \textstyle\sum_{m \in \mathcal{I}} w_m(e) = 1.$$

Rearranging the terms, we obtain

$$\text{logit}(\tau(e)) \equiv \log\left(\frac{\tau(e)}{1 - \tau(e)}\right) = \log\left(R(e)\right), \tag{A4}$$

demonstrating that variation in transmission rates with E is equivalent to variation in GRRs with E. Under non-multiplicative penetrance models, however, $\tau(e)$ does not reduce to a function of the GRRs only; transmission rates depend on both $R_g(e)$ and $w_m(e)$, except when $R_g(e) \equiv 1$ for all $g = 1, 2$ (i.e., no genetic effects).

A.3 Derivations of $\tau_m(e)$

The event that a heterozygous parent from mating type $(0,1)$ transmits the risk allele to his/her affected child is equivalent to the event that the child is heterozygous. Hence, the transmission rate for this mating type is

$$\begin{aligned}\tau_{01}(e) &\equiv \Pr(T \mid G_p = (0,1), H, D, E = e) \\ &= \Pr(G = 1 \mid G_p = (0,1), D, E = e). \end{aligned} \tag{A5}$$

Similarly, the transmission rate for mating type $(1,2)$ is

$$\begin{aligned}\tau_{12}(e) &\equiv \Pr(T \mid G_p = (1,2), H, D, E = e) \\ &= \Pr(G = 2 \mid G_p = (1,2), D, E = e). \end{aligned} \tag{A6}$$

The transmission rate for mating type $(1,1)$ is

$$\begin{aligned}\tau_{11}(e) &= \Pr(T \mid G_p = (1,1), D, E = e) \\ &= \sum_{g=0}^{2} \Pr(T \mid G = g, G_p = (1,1), D, E = e) \Pr(G = g \mid G_p = (1,1), D, E = e) \\ &= \frac{1}{2}\Pr(G = 1 \mid G_p = (1,1), D, E = e) + \Pr(G = 2 \mid G_p = (1,1), D, E = e). \end{aligned} \tag{A7}$$

The mating-type-specific genotype probabilities in (A5) – (A7) are

$$\Pr(G = g \mid G_p = m, D, E = e) = \frac{\Pr(G = g, G_p = m, D, E = e)}{\sum_{g' \in \mathcal{G}_m} \Pr(G = g', G_p = m, D, E = e)}.$$

Applying the identity,

$$\begin{aligned}&\Pr(G = g, G_p = m, D, E = e) = \\ &\quad \Pr(D \mid G = g, E = e) \Pr(G = g \mid G_p = m) \Pr(G_p = m \mid E = e) \Pr(E = e),\end{aligned}$$

which assumes G and E are conditionally independent given G_p, we obtain

$$\begin{aligned}\Pr(G = g \mid G_p = m, D, E = e) &= \frac{\Pr(D \mid G = g, E = e) \Pr(G = g \mid G_p = m)}{\sum_{g' \in \mathcal{G}_m} \Pr(D \mid G = g', E = e) \Pr(G = g' \mid G_p = m)} \\ &= \frac{R_g(e) \Pr(G = g \mid G_p = m)}{\sum_{g' \in \mathcal{G}_m} R_{g'}(e) \Pr(G = g' \mid G_p = m)}.\end{aligned}$$

Inserting the appropriate values for $\Pr(G_p = m \mid G = g)$ and substituting the result into (A5) – (A7), we obtain the final column of Table 2.1.

A.4 Derivation of odds of transmission under G-E independence

To derive the expressions in equation (2.2), consider the mating-type probabilities $\pi_m(e)$ in equation (2.1) for the parents of cases. Inserting the appropriate values for $\Pr(G = g \mid G_p = m)$, it can be shown that

$$\pi_m(e) \propto \begin{cases} \frac{1}{2} \cdot [1 + R_1(e)] \cdot \Pr(G_p = (0,1) \mid E = e) & \text{if } m = (0,1) \\ \frac{1}{2} \cdot [R_1(e) + R_2(e)] \cdot \Pr(G_p = (1,2) \mid E = e) & \text{if } m = (1,2) \\ \frac{1}{4} \cdot [1 + 2R_1(e) + R_2(e)] \cdot \Pr(G_p = (1,1) \mid E = e) & \text{if } m = (1,1). \end{cases}$$

Under G-E independence and Hardy-Weinberg genotype proportions in the population, this simplifies to

$$\pi_m(e) \propto \begin{cases} 2\,[1 + R_1(e)]\, q(1-q)^3 & \text{if } m = (0,1) \\ 2\,[R_1(e) + R_2(e)]\, q^3(1-q) & \text{if } m = (1,2) \\ [1 + 2R_1(e) + R_2(e)]\, q^2(1-q)^2 & \text{if } m = (1,1). \end{cases}$$

From the third column of Table 2.1, these $\pi_m(e)$ lead to

$$w_m(e) = \begin{cases} \dfrac{[1 + R_1(e)] \cdot (1-q)^2}{d(e)} & \text{if } m = (0,1) \\ \dfrac{[R_1(e) + R_2(e)] \cdot q^2}{d(e)} & \text{if } m = (1,2) \\ \dfrac{[1 + 2R_1(e) + R_2(e)] \cdot q(1-q)}{d(e)} & \text{if } m = (1,1), \end{cases}$$

where $d(e)$ is the sum of the numerators for all $m \in \mathcal{I}$. These $w_m(e)$ and the $\tau_m(e)$ in the last column of Table 2.1 can be substituted into equation (A3) to obtain the

overall transmission rate $\tau(e)$ as

$$\tau(e) = \frac{R_1(e) \cdot (1-q)^2 + R_2(e) \cdot q^2 + [R_1(e) + R_2(e)] \cdot q(1-q)}{d(e)},$$

from which the odds of transmission is

$$\frac{\tau(e)}{1-\tau(e)} = \frac{R_1(e) \cdot (1-q) + R_2(e) \cdot q}{(1-q) + R_1(e) \cdot q}.$$

Under the appropriate constraints in (A2), the expression above is reduced to the expressions in (2.2) for dominant and recessive penetrance models.

A.5 Variances of observed mating-type specific transmission rates

Initially, we suppress the conditioning on D and the stratum $E = e$ to simplify the notation. For a given stratum defined by $E = e$, let n_m be the number of case-parent trios with parents from mating type m. For the parental mating type $G_p = (0,1)$, the child's genotype can be $G = 0$ or 1. Thus $\hat{\tau}_{01} = \sum_{i=1}^{n_{01}} G_i/n_{01}$ and $V_{01} \equiv \mathrm{Var}(\hat{\tau}_{01}) = \mathrm{Var}(\, G \,|\, G_p = (0,1)\,)/n_{01}$. Similarly, for $G_p = (1,2)$, the child's genotype can be $G = 1$ or 2, and $\hat{\tau}_{12} = \sum_{i=1}^{n_{12}} (G_i - 1)/n_{12}$. It follows that $V_{12} = \mathrm{Var}(\, G \,|\, G_p = (1,2)\,)/n_{12}$. Finally, for $G_p = (1,1)$, the child's genotype can be $G = 0$ if neither parent transmitted the risk allele; $G = 1$ if one parent transmitted the risk allele and the other parent did not; and $G = 2$ if both parents transmitted the risk allele. Therefore, $\hat{\tau}_{11} = \sum_{i=1}^{n_{11}} G_i/(2n_{11})$ and $V_{11} = \mathrm{Var}(\, G \,|\, G_p = (1,1)\,)/(4n_{11})$.

To obtain the mating-type specific variances of the child's genotype G, we reason as follows. For $G_p = (0,1)$, the case genotype G is a Bernoulli random variable with parameter τ_{01}. For $G_p = (1,2)$, $G - 1$ is a Bernoulli random variable with parameter τ_{12}. Hence $\mathrm{Var}(\, G \,|\, G_p = (0,1)\,) = \tau_{01}(1-\tau_{01})$ and $\mathrm{Var}(\, G \,|\, G_p = (1,2)\,) = \tau_{12}(1-\tau_{12})$, so that $V_{01} = \tau_{01}(1-\tau_{01})/n_{01}$ and $V_{12} = \tau_{12}(1-\tau_{12})/n_{12}$. For $G_p = (1,1)$, we have

$$E(\, G \,|\, G_p = (1,1)\,) = \Pr(\, G = 1 \,|\, G_p = (1,1)\,) + 2\Pr(\, G = 2 \,|\, G_p = (1,1)\,) \overset{(A7)}{=} 2\,\tau_{11}$$

and

$$\begin{aligned} & E(\, G^2 \,|\, G_p = (1,1)\,) \\ &\quad = \quad \Pr(\, G = 1 \,|\, G_p = (1,1)\,) + 4\Pr(\, G = 2 \,|\, G_p = (1,1)\,) \\ &\quad \overset{(A7)}{=} \quad 4\,\tau_{11} - \Pr(\, G = 1 \,|\, G_p = (1,1)\,). \end{aligned}$$

Hence, $\mathrm{Var}(\, G \,|\, G_p = (1,1)\,) = 4\,\tau_{11}(1-\tau_{11}) - \Pr(\, G = 1 \,|\, G_p = (1,1)\,)$, and

$$V_{11} = \frac{1}{n_{11}}\left[\tau_{11}(1-\tau_{11}) - \frac{1}{4}\Pr(\, G = 1 \,|\, G_p = (1,1)\,)\right].$$

The variances V_{01}, V_{12} and V_{11}, of the observed mating-type-specific transmission rates may be estimated by substituting the observed mating-type-specific transmission rates $\hat{\tau}_{01}$, $\hat{\tau}_{12}$, $\hat{\tau}_{11}$, and the observed proportion, $\widehat{\Pr}(\, G = 1 \,|\, G_p = (1,1)\,)$, of trios from mating type $G_p = (1,1)$ with $G = 1$ into the preceding expressions. Restoring the conditioning on $E = e$ and D in the notation,

$$\hat{V}_m(e) = \begin{cases} \dfrac{1}{n_{01}(e)} \cdot \hat{\tau}_{01}(e)[1 - \hat{\tau}_{01}(e)] & \text{if } m = (0,1) \\ \dfrac{1}{n_{12}(e)} \cdot \hat{\tau}_{12}(e)[1 - \hat{\tau}_{12}(e)] & \text{if } m = (1,2) \\ \dfrac{1}{n_{11}(e)} \cdot \Big[\hat{\tau}_{11}(e)(1 - \hat{\tau}_{11}(e)) & \\ \quad -\dfrac{1}{4}\,\widehat{\Pr}(\, G = 1 \,|\, G_p = (1,1), D, E = e\,)\Big] & \text{if } \ m = (1,1), \end{cases} \tag{A8}$$

where $n_m(e)$ is the number of trios with parental mating type m and $E = e$.

Appendix B

B.1 Example of an alternate parameterization of the disease risk model

Disease risk can be modelled with a log-linear model similar to (4.1) but parameterized in a different way, such that

$$P(D = 1 \mid G = g, E = e) = \exp\{k + \boldsymbol{z}^*(g)\boldsymbol{\gamma}^* + \xi(e) + \boldsymbol{z}^*(g)\boldsymbol{f}^*(e)\},$$

where $\boldsymbol{z}^*(g) = (\mathrm{I}\{g = 1\}, \mathrm{I}\{g = 2\})$, $\boldsymbol{\gamma}^* = (\gamma_1^*, \gamma_2^*)^\mathsf{T}$, and $\boldsymbol{f}^*(e) = ({f^*}_1(e), {f^*}_2(e))^\mathsf{T}$. Then, under the disease risk model above, two GRRs can be expressed as

$$\mathrm{GRR}_1^*(e) \equiv \frac{P(D = 1 \mid G = 1, E = e)}{P(D = 1 \mid G = 0, E = e)} = \exp(\gamma_1^* + f_1^*(e)),$$

and

$$\mathrm{GRR}_2^*(e) \equiv \frac{P(D = 1 \mid G = 2, E = e)}{P(D = 1 \mid G = 0, E = e)} = \exp(\gamma_2^* + f_2^*(e)).$$

This parameterization allows for different genotype relative risks for the individuals with $G = 1$ and $G = 2$ compared to the individuals in the baseline group having $G = 0$. Comparing this parameterization to that of the risk model (4.1), we have $\gamma_1^* = \gamma_1$, $f_1^*(e) = f_1(e)$, $\gamma_2^* = \gamma_1 + \gamma_2$ and $f_2^*(e) = f_1(e) + f_2(e)$. As before, $f_1^*(e) = f_2^*(e) = 0$ indicates there is no $G{\times}E$. Under different modes of inheritance, $f_1^*(e)$ and $f_2^*(e)$ behave differently. Under the dominant mode shown in equation (3.3),

$$f_1^*(e) = f_2^*(e) \neq 0$$

since $\text{GRR}_1^*(e) = \text{GRR}_2^*(e) \neq 1$. Under the log-additive or multiplicative mode (3.4),

$$f_1^*(e) = \frac{1}{2} f_2^*(e) \neq 0$$

since $\left\{\text{GRR}_1^*(e)\right\}^2 = \text{GRR}_2^*(e) \neq 1$. Under the recessive mode (3.5),

$$f_1^*(e) = 0, \quad \text{and} \quad f_2^*(e) \neq 0$$

since $\text{GRR}_1^*(e) = 0$ and $\text{GRR}_2^*(e) \neq 1$.

In practice, under this parameterization, it may be more difficult to distinguish between the dominant and the log-additive inheritance models than under the original gene-dosage parameterization used in (4.1). Patterns in $G{\times}E$ curves can be similar under these two inheritance with the new parameterization as we have $f_1^*(e) \neq 0$ and $f_2^*(e) \neq 0$ under both modes. Patterns in the interaction curves are different with the original parameterization as we have $f_1(e) = 0$ and $f_2(e) \neq 0$ under the dominant mode and $f_1(e) = f_2(e) \neq 0$ under the log-additive mode.

B.2 Expression of $X^*(e)$

In this section, we will see how the natural cubic spline basis function vector $X_h^*(e) = [b_{h1}(e), \ldots, b_{hK_h}(e)]$ in expression (3.8) can be explicitly defined, using the representation of a natural cubic spline function used in Wood (2006), section 4.1.2. Since all the following arguments can be applied to any natural cubic spline function, we will suppress the dependence on h for $f_h(e)$ and its related terms, for notational simplicity.

Suppose $x_1 < \cdots < x_K$ are the knots chosen to represent $f(e)$. For this representation, the spline function is parameterized in terms of its values at the knots $c_k^* = f(x_k)$, $k = 1, ..., K$. Incorporating the conditions necessary for $f(e)$ to be a natural cubic spline function leads to

$$\begin{aligned} f(e) \;=\; & \frac{(x_{k+1} - e)}{t_k} c_k^* + \frac{(e - x_k)}{x_k} c_{k+1}^* + \\ & \frac{1}{6}\left[\frac{(x_{k+1} - e)^3}{t_k} - t_k(x_{k+1} - e)\right] F_k \boldsymbol{c}^* + \\ & \frac{1}{6}\left[\frac{(e - x_k)^3}{t_k} - t_k(e - x_k)\right] F_{k+1} \boldsymbol{c}^*, \qquad \text{if } x_k \leq e \leq x_{k+1} \end{aligned} \tag{B1}$$

where $\boldsymbol{c}^* = (c_1, \ldots, c_K)^\mathsf{T}$, and $F_k = [F_{k1}, \ldots, F_{kK}]$ is the k^{th} row of a $K \times K$ matrix $\mathbf{F}$ whose elements are either zero or in terms of the distance $t_k = x_{k+1} - x_k$ between

two adjacent knots. The matrix $\mathbf{F}$ will be defined at the bottom of the section.

Since

$$F_k \boldsymbol{c}^* = \sum_{k'=1}^{K} F_{kk'} c^*_{k'},$$

the representation in (B1) can be re-expressed as a linear combination in the form of

$$\begin{aligned} f(e) \;=\; & \frac{(x_{k+1}-e)}{t_k} c^*_k + \frac{(e-x_k)}{t_k} c^*_{k+1} + \\ & \frac{1}{6}\left[\frac{(x_{k+1}-e)^3}{t_k} - t_k(x_{k+1}-e)\right] \sum_{k'=1}^{K} F_{kk'} c^*_{k'} + \\ & \frac{1}{6}\left[\frac{(e-x_k)^3}{t_k} - t_k(e-x_k)\right] \sum_{k'=1}^{K} F_{k+1k'} c^*_{k'}, \qquad \text{if } x_k \le e \le x_{k+1}. \end{aligned}$$

This indicates that the row vector $X^*(e) = [b_1(e), \ldots, b_K(e)]$ of the basis functions in equation (3.8) is

$$X^*(e) = \begin{cases} \begin{bmatrix} a^-_1(e) + F_{11}c^-_1(e) + F_{21}c^+_1(e) \\ a^+_1(e) + F_{12}c^-_1(e) + F_{22}c^+_1(e) \\ F_{13}c^-_1(e) + F_{23}c^+_1(e) \\ \vdots \\ F_{1K}c^-_1(e) + F_{2K}c^+_1(e) \end{bmatrix}^{\mathsf{T}} & \text{if } x_1 \le e \le x_2, \\ \begin{bmatrix} F_{21}c^-_2(e) + F_{31}c^+_2(e) \\ a^-_2(e) + F_{22}c^-_2(e) + F_{32}c^+_2(e) \\ a^+_2(e) + F_{23}c^-_2(e) + F_{33}c^+_2(e) \\ F_{24}c^-_2(e) + F_{34}c^+_2(e) \\ \vdots \\ F_{2K}c^-_2(e) + F_{3K}c^+_2(e) \\ \vdots \end{bmatrix}^{\mathsf{T}} & \text{if } x_2 \le e \le x_3, \\ \begin{bmatrix} F_{(k-1)1}c^-_{K-1}(e) + F_{K1}c^+_{K-1}(e) \\ F_{(k-1)2}c^-_{K-1}(e) + F_{K2}c^+_{K-1}(e) \\ \vdots \\ a^-_{K-1}(e) + F_{K-1K-1}c^-_{K-1}(e) + F_{KK-1}c^+_{K-1}(e) \\ a^+_{K-1}(e) + F_{K-1K}c^-_{K-1}(e) + F_{KK}c^+_{K-1}(e) \end{bmatrix}^{\mathsf{T}} & \text{if } x_{K-1} \le e \le x_K, \end{cases} \tag{B2}$$

where

$$a_k^+(e) = \frac{(x_{k+1} - e)}{t_k},$$

$$a_k^-(e) = \frac{(e - x_k)}{t_k},$$

$$c_k^+(e) = \frac{1}{6}\left[\frac{(x_{k+1} - e)^3}{t_k} - t_k(x_{k+1} - e)\right],$$

and

$$c_k^-(e) = \frac{1}{6}\left[\frac{(e - x_k)^3}{t_k} - t_k(e - x_k)\right].$$

The $K \times K$ matrix $\mathbf{F}$ can be defined as

$$\mathbf{F} = \begin{bmatrix} \mathbf{0}_{1\times K} \\ \mathbf{B}^{-1}\mathbf{D} \\ \mathbf{0}_{1\times K} \end{bmatrix},$$

where the $(K-2) \times (K-2)$ band matrix $\mathbf{B}$ and the $(K-2) \times K$ matrix $\mathbf{D}$ are defined as, respectively,

$$\mathbf{B} = \begin{bmatrix} \frac{t_1+t_2}{3} & \frac{t_2}{6} & & & & & \\ \frac{t_2}{6} & \frac{t_2+t_3}{3} & \frac{t_3}{6} & & & & \\ \ddots & \ddots & \ddots & & & & \\ & \frac{t_{k-1}}{6} & \frac{t_{k-1}+t_k}{3} & \frac{t_k}{6} & & & \\ & \ddots & \ddots & \ddots & & & \\ & & \frac{t_{K-3}}{6} & \frac{t_{K-3}+t_{K-2}}{3} & \frac{t_{K-2}}{6} & & \\ & & & \frac{t_{K-2}}{6} & \frac{t_{K-2}+t_{K-1}}{3} & & \end{bmatrix},$$

and

$$\mathbf{D} = \begin{bmatrix} \frac{1}{t_1} & -\frac{1}{t_1} - \frac{1}{t_2} & \frac{1}{t_2} & & & & \\ \ddots & \ddots & \ddots & & & & \\ & \frac{1}{t_{k-1}} & -\frac{1}{t_{k-1}} - \frac{1}{t_k} & \frac{1}{t_k} & & & \\ & \ddots & \ddots & \ddots & & & \\ & & \frac{1}{t_{K-2}} & -\frac{1}{t_{K-2}} - \frac{1}{t_{K-1}} & \frac{1}{t_{K-1}} & & \end{bmatrix}.$$

B.3 Smoothing parameter estimation: computation details

For smoothing parameter estimation, we search for the values of λ_1 and λ_2 that minimize the generalized AIC function (3.13), using a grid search algorithm. Empirical results (not shown) suggested that minimizing the AIC function over one smoothing parameter is independent of minimizing the function over the other smoothing parameter. These results suggest that we can use an algorithm with two one-dimensional grid searches rather than one with a two-dimensional search and hence save computation time. To further save the computation time, we restrict the search for λ_h over a fixed number κ_h of grid points for $f_h(e)$, $h = 1, 2$.

In order to select 'good' κ_h grid points, we use the smoothing parameter estimates λ_h^* obtained by applying a likelihood approach that makes inference of $G \times E$ conditional on G_p, E and partial information on G (Duke, 2007). The estimates are computed by fitting two one-dimensional generalized additive models using the `gam()` function of the 'mgcv' package in R. Assuming that λ_h^* is close to the true value of the smoothing parameter λ_h, we choose grid points based on a truncated normal distribution for $\log(\lambda_h) \sim TN(\mu_h, a_h, b_h, \sigma_h)$, with mean μ_h, endpoints a_h and b_h ($a_h < b_h$) and standard deviation σ_h.

The mean μ_h of the truncated normal distribution is set to be $\log(\lambda_h^*)$, allowing for higher probability mass to the points in the neighbourhood of $\log(\lambda_h^*)$. The lower endpoints a_h, $h = 1, 2$ are set to be -20 since $\exp(-20)$ is effectively $-\infty$. Similarly, the upper endpoints b_h, $h = 1, 2$ are set to be 20 since $\exp(20)$ is effectively $+\infty$. The standard deviation is set to be $\sigma_h = \min\big(\log(\lambda_h^*) - a_h, b_h - \log(\lambda_h^*)\big)$. Then, the κ_h grid points for $\log(\lambda_h)$ can be chosen throughout the range $[-20, 20]$ using the quantiles of the truncated normal distribution. For example, as a default, we choose $\kappa_h = 6$ points that include $\log(\lambda_h^*)$, 25^{th}, 50^{th} and 75^{th} percentiles of the truncated normal distribution and the two endpoints a_h and b_h (i.e., the 0^{th} and 100^{th} percentiles, respectively).

With the chosen grid points, we obtain the optimal values of λ_1 and λ_2 by finding λ_{1j}, for $j = 1, 2, ..., \kappa_1$, that yields the minimum score of the generalized AIC function in (3.13) for a fixed value of λ_2 (e.g., λ_2^*) and λ_{2j} for $j = 1, 2, ..., \kappa_2$, that yields the minimum AIC score for the optimal value of λ_1 found in the first step.

Note that some power loss may occur when the proposed permutation test is carried out since we restrict each grid search to a small number (e.g., 6) of grid points. We could improve power by considering a finer grid for each smoothing parameter

but this would be computationally expensive.

B.4 ALL-mimicking data-simulation

In this section, we present the details showing how we generated the data used in Section 3.5 to illustrate the proposed smoothing approach.

We assumed a homogeneous population since about 91% of the 1030 cases in the original data set had both parents of European ethnicity. We set the allele frequency of the *NQO1 C609T* variant allele to be $q = 0.19$, which is the one estimated in the CEPH sample (Database of Single Nucleotide Polymorphisms (dbSNP), build 135). The parental mating types G_p were simulated assuming HWP and mating symmetry. Then, the child genotypes G were simulated given a pair of parents under Mendel's law. The ages E for children in the population meeting the age selection criteria of 0–15 years were simulated using a uniform distribution from ages 0–15.

For simulating disease status D, we used the disease risk model (4.1), whose parameter values were set as follows. For the baseline parameter k, we chose an arbitrary value since it does not affect the analysis of $G\times E$. For the main effect term $\xi(e)$ for the non-genetic factor, we let $\exp(\xi(e))$ be a function proportional to a gamma distribution estimated from the observed E in cases with reference genotype $G = 0$ since

$$\exp\big(\xi(e)\big) \propto P(D = 1 \mid E = e, G = 0) \propto P(E = e \mid D = 1, G = 0).$$

For the GRR-related parameters, their values were set to the estimated values obtained from fitting the original data set with 288 informative trios. Boundary knots were placed at $E = 0$ and 15 years of age in order to ensure that $f_1(e)$ and $f_2(e)$ were defined at any values of $E = e$ that were generated. Figure 3.5 shows the resulting theoretical log-GRR curves with parameterizations $\gamma_1 = 0.23, \gamma_2 = -0.56$, $f_1(e) = -0.23\, b_{11}(e) + 0.03\, b_{12}(e) + 0.10\, b_{13}(e) + 0.06\, b_{14}(e) - 0.54\, b_{15}(e)$, and $f_2(e) = 0.72\, b_{21}(e) + 0.30\, b_{22}(e) + 0.07\, b_{23}(e) - 0.22\, b_{24}(e) - 1.03\, b_{25}(e)$.

The natural cubic spline basis functions $b_{hk}(e)$ can be constructed using the definition (B2). Note that in this definition, all terms depend on the GRR-function index $h = 1, 2$. For example, for $f_1(e)$ with a covariate value located between the first and

the second knot (i.e., $x_{11} \leq e \leq x_{12}$), we can construct the basis function vector as

$$X_1(e) = [b_{11}(e), ..., b_{15}(e)] = \begin{bmatrix} a_{11}^-(e) + F_{1,11}c_{11}^-(e) + F_{1,21}c_{11}^+(e) \\ a_{11}^+(e) + F_{1,12}c_{11}^-(e) + F_{1,22}c_{11}^+(e) \\ F_{1,13}c_{11}^-(e) + F_{1,23}c_{11}^+(e) \\ F_{1,14}c_{11}^-(e) + F_{1,24}c_{11}^+(e) \\ F_{1,15}c_{11}^-(e) + F_{1,25}c_{11}^+(e) \end{bmatrix}^\mathsf{T},$$

where

$$a_{11}^+(e) = \frac{(x_{12} - e)}{t_{11}},$$

$$a_{11}^-(e) = \frac{(e - x_{11})}{t_{11}},$$

$$c_{11}^+(e) = \frac{1}{6}\left[\frac{(x_{12} - e)^3}{t_{11}} - t_{11}(x_{12} - e)\right],$$

and

$$c_{11}^-(e) = \frac{1}{6}\left[\frac{(e - x_{11})^3}{t_{11}} - t_{11}(e - x_{11})\right].$$

Appendix C

We first derive the GRRs for G' assuming no $G \times E$. As these depend on E, there is spurious interaction. We then show how spurious interaction can be avoided by adjusting the risk model involving G' to allow separate genetic main effects for each value of the grouping variable X. Recall that X distinguishes groups of sub-populations with different E distributions. Throughout this appendix the focus is on avoiding spurious interaction and so we are interested in the risk model without $G \times E$.

C.1 GRRs for G'

The no-interaction risk model specifies $P(D = 1 \mid G = g, E = e) \propto \psi(g) \exp(\eta(e))$, where $\psi(g) = \exp(z_1(g)\beta_1 + z_2(g)\beta_2)$. However, the model we fit is based on $P(D = 1 \mid G' = g, E = e)$. Assuming disease risk does not depend on G' once G is known, the law of total probability gives

$$\begin{aligned}
&P(D = 1 \mid G' = g, E = e) \\
&\quad = \sum_{g^*} P(D = 1 \mid G = g^*, G' = g, E = e) P(G = g^* \mid G' = g, E = e) \\
&\quad = \sum_{g^*} P(D = 1 \mid G = g^*, E = e) P(G = g^* \mid G' = g, E = e) \\
&\quad \propto \sum_{g^*} \psi(g^*) \exp(\eta(e)) P(G = g^* \mid G' = g, E = e) \\
&\quad = \exp(\eta(e)) \sum_{g^*} \psi(g^*) P(G = g^* \mid G' = g, E = e).
\end{aligned}$$

The GRRs for G' are thus

$$\frac{P(D = 1 \mid G' = 1, E = e)}{P(D = 1 \mid G' = 0, E = e)} = \frac{\sum_{g^*} \psi(g^*) P(G = g^* \mid G' = 1, E = e)}{\sum_{g^*} \psi(g^*) P(G = g^* \mid G' = 0, E = e)}$$

and

$$\frac{P(D=1 \mid G'=2, E=e)}{P(D=1 \mid G'=1, E=e)} = \frac{\sum_{g^*} \psi(g^*) P(G=g^* \mid G'=2, E=e)}{\sum_{g^*} \psi(g^*) P(G=g^* \mid G'=1, E=e)}$$

and depend on E only through $P(G=g^* \mid G'=g, E=e)$.

If $P(G=g^* \mid G'=g, E=e)$ doesn't depend on E then neither will the GRRs. However, this does not hold in our example. To see why, write

$$\begin{aligned}
P(G \mid G', E) &= \frac{P(G, E \mid G')}{P(E \mid G')} \\
&= \frac{\sum_S P(G, E \mid G', S) P(S \mid G')}{\sum_S P(E \mid G', S) P(S \mid G')} \\
&\overset{(i)}{=} \frac{\sum_S P(G \mid G', S) P(E \mid G', S) P(S \mid G')}{\sum_S P(E \mid G', S) P(S \mid G')} \\
&\overset{(ii)}{=} \frac{\sum_S P(G \mid G', S) P(E \mid S) P(S \mid G')}{\sum_S P(E \mid S) P(S \mid G')}
\end{aligned}$$

where the third and fourth lines of the equation use the example-specific identities (i) $P(G, E \mid G', S) = P(G \mid G', S) P(E \mid G', S)$, and (ii) $P(E \mid G', S) = P(E \mid S)$, respectively. These identities follow from E and (G, G') being conditionally independent given S. Now it can be seen that E will only cancel out of the final expression if either $P(E \mid S)$ or $P(G \mid G', S)$ can be factored out of the summations. That is, if either $P(E \mid S) = P(E)$ (the distribution of E does not depend on sub-population) or $P(G \mid G', S) = P(G \mid G')$ (the joint distribution of G and G' does not depend on sub-population). However, neither of these are true under the population stratification we consider. Therefore, $P(G \mid G', E)$ *does* depend on E. These calculations highlight that subpopulation-specific variation in GG' haplotype distributions is a driver of spurious interaction, not variation in patterns of linkage disequilibrium *per se* (Zaykin and Shibata, 2008).

C.2 Adjustment for X

Recall that X is a variable describing groups of sub-populations with different E distributions. If X were available for each subject we could adjust for it in analysis

of G', as follows.

$$\begin{aligned} P(D=1 \mid G'=g, E=e, X=x) \\ &= \sum_{g^*} P(D=1 \mid G=g^*, G'=g, E=e, X=x) P(G=g^* \mid G'=g, E=e, X=x) \\ &= \sum_{g^*} P(D=1 \mid G=g^*, E=e) P(G=g^* \mid G'=g, E=e, X=x). \end{aligned} \tag{C1}$$

Assuming that E is conditionally independent of (G, G') given X, we have

$$\begin{aligned} P(G=g^* \mid G'=g, E=e, X=x) \\ &= \frac{P(G=g^*, G'=g, E=e, X=x)}{P(G'=g, E=e, X=x)} \\ &= \frac{P(E=e \mid G=g^*, G'=g, X=x) P(G=g^*, G'=g, X=x)}{P(E=e \mid G'=g, X=x) P(G'=g, X=x)} \\ &= \frac{P(E=e \mid X=x) P(G=g^*, G'=g, X=x)}{P(E=e \mid X=x) P(G'=g, X=x)} \\ &= P(G=g^* \mid G'=g, X=x). \end{aligned}$$

Thus, equation (C1) reduces to

$$\begin{aligned} P(D=1 \mid G'=g, E=e, X=x) \\ &= \sum_{g^*} P(D=1 \mid G=g^*, E=e) P(G=g^* \mid G'=g, X=x). \end{aligned}$$

And, under the no-interaction risk model $P(D=1 \mid G=g, E=e) \propto \psi(g) \exp(\eta(e))$, we obtain X-adjusted GRRs for G' of

$$\begin{aligned} GRR_i(x) &= \frac{P(D=1 \mid G'=i, E=e, X=x)}{P(D=1 \mid G'=i-1, E=e, X=x)} \\ &= \frac{\sum_{g^*} \psi(g^*) P(G=g^* \mid G'=i, X=x)}{\sum_{g^*} \psi(g^*) P(G=g^* \mid G'=i-1, X=x)};\ i=1,2. \end{aligned}$$

Each of the two GRRs for G' take on as many different values as there are values of X. In other words, in analyses of G', there are as many genetic main effects as there are values of X. For the case of a binary X, this is achieved by the GRRs in equation (4.2).