NOTE TO USERS

Assessing Power to Detect Gene-Environment Interactions Using Surrogate Outcomes: A Simulation Study

Tamanna Howlader

A thesis

in

The Department

of

Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Science at
Concordia University
Montreal, Québec,Canada

August 2005

Your file Votre référence
ISBN: 0-494-10214-4
Our file Notre référence
ISBN: 0-494-10214-4

ABSTRACT

Assessing Power to Detect Gene-Environment Interactions Using Surrogate Outcomes: A Simulation Study

Tamanna Howlader

The low power of gene-environment ($G \times E$) interaction studies is of major concern in genetic-epidemiolgic research. Past research involving binary outcomes has focussed mainly on the development of efficient study designs to address this problem. This thesis explores an alternative strategy that uses quantitative "surrogates" of the "clinical" binary outcome to improve power to detect $G \times E$ interactions.

Efficiency of the quantitative "surrogate" outcome X versus the binary outcome Y is assessed for three hypothetical models of the relationship between the outcomes, and their relationships to genetic susceptibility, exposure, and other risk factors. In the first scenario, X is a risk factor of disease, and a mediator for the effect of $G \times E$ interaction. In the second scenario, X is considered a marker of disease outcome. Finally, repeated measures of the disease marker X are used to define alternative binary and quantitative outcomes.

Simulations are used to estimate the power to detect $G \times E$ interaction in models using these alternative outcomes. Some variation of such parameters as prevalence of the genetic factor and exposure, strength of the underlying $G \times E$ interaction effect on the binary outcome and surrogate, measurement errors in the outcomes, etc., is introduced to assess their impact on the power. Results indicate that under certain situations and combinations of relevant parameters, higher power can be achieved by replacing the binary outcome by a quantitative "surrogate" outcome. For example, the quantitative outcome provides higher power (36%) than the binary outcome (28%) when the effects of $G \times E$ and E on Y are transmitted *mainly* through X. The use of quantitative outcomes based on repeated measures, such as the average increase in X per year, also results in higher power (100%) for detecting strong to moderate interaction in the data relative to alternative binary outcomes ($< 82\%$).

Acknowledgements

I wish to thank my Supervisor Dr. Michal Abrahamowicz, Professor, Department of Epidemiology and Biostatistics, McGill University, for supervising this thesis work and for introducing me to an active and interesting field of research in Statistics. I am grateful to him for his constructive suggestions, his valuable time and encouragement, and for the financial support that he has given me to carry out this study.

I am indebted to my Co-supervisor, Dr. Yogendra Chaubey, Professor, Department of Mathematics and Statistics, Concordia University for reviewing the manuscript of the thesis and for his moral support. I would also like to thank the Department and Concordia University for awarding me the Graduate Entrance Scholarship and the Hydro Quebec Graduate Award, which were of immense help to me in supporting this research work.

The secretarial assistance of Ms. Georgia Panaritis at the Clinical Division of Epidemiology, Montreal General Hospital, is gratefully acknowledged. In addition, I wish to thank the Division for granting me access to some of its facilities.

Finally, I thank my husband for his patience and support, and my parents for their love and prayers for my well-being.

Dedication

To my beloved parents

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview of genetic epidemiology

Traditional epidemiology has focussed largely on the role of environmental factors (i.e. infectious, chemical, physical, nutritional, and behavioral factors) in the etiology of disease. However, studies suggest that most diseases are caused not by environmental factors or genetic factors alone, but by a complex interplay of both these factors [86],[109]. This recognition, coupled with the rapid development of molecular biology and completion of the human genome project, have given impetus to investigations of the effects of genes, and their interactions with environmental exposures. Such gene-environment ($G \times E$) interaction studies create specific analytical challenges including issues of study design, analysis and statistical power.

Genetic epidemiology studies the role of inherited causes of disease in families and populations. It is a hybrid discipline that integrates the research tools of epidemiology and human genetics [103]. Central to genetic epidemiology is the study of gene-environment interactions ($G \times E$), first considered by Haldane [44]. According to Ottman[87], gene-environment interaction is defined as "a different effect of an environmental exposure on disease risk in persons with different genotypes," or, alternatively, "a different effect of a genotype on disease risk in persons with different environmental exposures." Figure 1.1 illustrates the scope of genetic epidemiology as the interface of genetic and environmental interaction in a process leading to a disease.

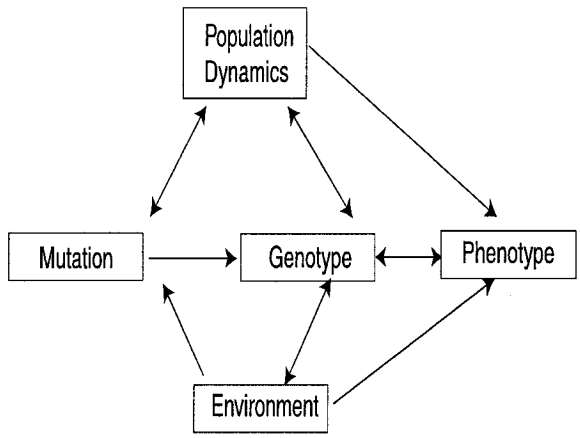

Figure 1.1: Scope of genetic epidemiology.

1.2 Overview of the basic concepts of human genetics

Genetic factors contribute to the variation of a trait. For example, individuals with specific genotypes might be predisposed towards hypertension or may have a weaker immune system which is not able to deal with a virus efficiently. To understand the role of genetic factors in the occurrence of disease in human populations, it is important to have a basic knowledge of the structure and function of genetic material, as well as of the principles underlying its transmission in families and populations.

The genetic material of higher organisms is the DNA (deoxyribonucleic acid). It is the constituent of the 23 chromosomes in the nucleus of the human cell and forms the building block of life. Each DNA molecule has a sequence of bases along it, most of which have no known function. About 3% of the bases are functional and code for polypeptide chains or molecules of ribonucleic acid (RNA) [27]. The *gene* is thus a working subunit of the DNA that contains the code for a specific product, typically a protein such as an enzyme. It occupies a fixed position on a chromosome called the *locus* and occurs in pairs at each of the loci along the pairs of chromosomes.

Different forms of a gene or DNA sequence that can exist at a single locus are called *alleles*. The two alleles at each locus of the chromosome comprise the *genotype* for that locus. If the two alleles are the same, the genotype is called *homozygous*, otherwise, it is called *heterozygous*. The set of visible or measurable (i.e., observable)

characteristics of an individual is called the *phenotype*.

1.3 Importance of $G \times E$ interaction studies

Genetic factors contribute to most human diseases, conferring susceptibility or resistance, or interacting with environmental factors [17], [87]. There is accumulating evidence that allelic variations of many gene loci play important roles in determining individual susceptibility to cancer [19], [101] and other chronic diseases [24]. In the case of cancers, although environmental agents are responsible for initiation and subsequent progression of the oncogenic process, genetic differences amongst individuals in one or more of the effectors (e.g. enzymes, receptors) involved serve as predisposing or susceptibility factors. Thus, many people exposed to a particular carcinogen (e.g. cigarette smoking) will not develop cancer even if their exposure was high while other highly susceptible individuals will develop cancer even if their exposure was low. Such differences in response among individuals exposed to the same environmental factors is observed for other diseases as well. For instance, some health conscious individuals with acceptable cholesterol levels may suffer myocardial infarction at age 40 while others might seem immune to heart disease in spite of smoking, poor diet, and obesity [45]. This suggests that manifestation of any disease, especially those chronic in nature, is very likely to be the result of an inextricable interplay of biological as well as environmental factors.

There are numerous examples of gene-environment interactions. For instance, synergy exists between variation in the lipoprotein lipase (LPL) gene and smoking on risk of coronary heart disease [109]. The relation between the recessive gene for phenylketonuria (PKU) and dietary phenylalanine in mental retardation provides another example [112]. Gene-environment interactions have also been reported for alcoholism [81], asthma [117], hypertension [66] and lung cancer [128]. Studies show that most common defects such as neural tube defects, oral clefts, and congenital cardiovascular malformations are explained not by environmental factors or genetic factors alone, but by the interaction between gene variants at multiple loci and environmental exposures [57]. When such interactions exist, the combined action of genes and environment can increase or decrease disease risk beyond that due to purely genetic and purely

environmental actions.

The goal of gene-environment studies in epidemiology is to learn how the risk of a disease changes as a joint function of genotype and exposure. They also promise the eventual capability to tailor interventions more precisely, whether at a clinical level, where the therapeutic agent prescribed or its dose may be chosen in light of an individual's genetic makeup, or at a public health level, where programs may be targeted at high-risk subpopulations [113]. From a statistical standpoint, ignoring an existing $G \times E$ interaction in an analysis can, erroneously, make the main effects of the gene and/or the environmental factor appear nonsignificant [85], so that important risk factors for the trait may be overlooked. Finally, failing to model a $G \times E$ interaction in a segregation analysis can lead to incorrect conclusions with respect to determination of the mode of inheritance [111] and estimation of the magnitude of genetic effects and allele frequencies [25].

1.4 Methodological issues in studies of $G \times E$ interactions

In recent years, there have been many studies of various $G \times E$ interactions. Some of these works are applied in nature, while others focus on methodological issues. Feasibility, is an important issue among genetic-epidemiologists since the cost of genotyping remains quite high. For studies involving rare genes or uncommon environmental exposures, and moderate strength of the interaction effect, the sample size required to achieve reasonable statistical power can be prohibitive [6], [38]. Thus, development of efficient study designs and methods of analysis that reduce required sample size has become an important area of study within research on $G \times E$ interactions.

Among the traditional epidemiologic study designs, the case-control design is widely used. Several authors have described methods for estimating power and sample size in the context of unmatched case-control studies [32], [35], [42], [51]. However, such studies may be affected by population stratification, also known as admixture, which arises when genetically diverse populations are incompletely mixed [38], [118]. Among the matched designs, the population-based case-control study [38] and family-based designs (case-sib, case-parent) [23],[121] are often used. Different matching strategies

have been proposed to increase power and feasibility for detecting $G \times E$ interaction. The flexible matching design [106], counter-matching design [6] and case-combined-control design [5] are such examples. Another study that does not use controls and under certain assumptions offers greater efficiency over the case-control design is the case-only design [59]. Several papers have discussed power and sample size calculations for these alternative designs and provide comparisons, in most cases, with the population-based case-control study [98], [105], [121].

Most studies on $G \times E$ interaction assume that the environmental factor is categorical and the genetic factor and outcome of interest are binary [32], [35], [42], [51]. However, the situation where the outcome is continuously distributed is becoming more important as researchers try to investigate the genetic basis of quantitative traits, such as blood pressure, obesity and insulin sensitivity. Methods have been described for calculation of sample size for detection of $G \times E$ interaction in the case of a continuously distributed outcome with a categorical genetic factor and continuous environmental exposure. The power in this context is found to depend on the allele frequency, the size of the main effect and magnitude of the interaction effect [72]. In planning studies to examine $G \times E$ interaction, measurement errors are often introduced by using a less precise exposure measurement or a proxy for the true outcome of interest. The effect of measurement error in exposure, outcome or genotyping on estimation and power to detect $G \times E$ interaction is an important area of research and has been investigated by some studies [122],[123].

1.5 Rationale and objectives

Large-scale databases are necessary to detect $G \times E$ interactions, and to test and confirm related hypotheses. This is particularly true when the factors under consideration are rare and the interaction effect is moderate, a common situation in studies of potential effect modification by specific, thus, relatively rare genotypes. Ensuring adequate power to detect $G \times E$ interaction is, therefore, a major concern in genetic epidemiology studies.

Most researchers concerned with this problem have focussed on optimizing the study design. However, the use of complex study designs often introduces practical

difficulties such as increased cost, complexity for subject recruitment, and computational and/or conceptual difficulty. The use of surrogate outcome measures to improve power of $G \times E$ interaction studies is a potentially interesting alternative, and an area that appears to be relatively unexplored. The strategy to use alternative outcomes could be especially attractive if an infrequent binary outcome, such as occurrence of a rare disease, is replaced by a clinically relevant quantitative "surrogate" outcome. For example, rather than using the presence of hypertension, investigators may consider a quantitative outcome such as the value of systolic blood pressure or its change over a certain time interval. Another interesting type of surrogate outcome may involve estimated risk of the clinical outcome, such as coronary heart disease, usually obtained from a multivariable model that aggregates the impact of several risk factors [55]. Indeed, statisticians working on the methodology of time-to-event analysis have demonstrated the advantages, in terms of power and efficiency, of using such surrogates to compensate for the low power of the analyses of highly censored data [67], [75], [83] [92], .

On the other hand, the disadvantages associated with the use of a dichotomized version of a continuous outcome are well known [64]. Thus, even if it is predictable that power will increase when the binary outcome is replaced by a quantitative surrogate, several practical and conceptual issues remain to be systematically addressed. First, to guide researchers in their decisions regarding the choice between a binary, more directly relevant outcome, and a quantitative surrogate, the expected gains in statistical power, offered by the latter choice needs to be quantified. Secondly, such gains will most likely depend on several aspects of the study design and the data structure, such as frequency of the binary outcome, realistic sample size, strength of the $G \times E$ impact on the outcome and on the surrogate, and errors in the measurement of both outcome and surrogate variables.

In addition, the power comparison may depend on the type of biological relationship between the surrogate and the clinical (binary) outcome. In some studies, the binary outcome may be simply defined based on the categorization of the quantitative variable. For instance, hypertension is defined as a diastolic blood pressure > 90 mmHg and abnormal renal function as serum creatinine > 1.4 mg/dl. In other situa-

tions, a quantitative variable may be an important, but not the only, determinant of the outcome, i.e. a risk factor. In yet other contexts, a quantitative variable may be a marker for either the presence of disease or for the underlying pathological process that ultimately leads to the disease occurrence. Each of these numerous alternatives has different implications for the relative power and efficiency of the analyses focussing on a quantitative surrogate rather than on a binary clinical outcome.

Moreover, there are alternative ways a quantitative variable may be measured and analyzed. For example repeated measures may be expected to increase the precision, and thus the efficiency, of the analysis [115]. One way is to simply take the mean of all available measurements, but this will be valid only if there is no systematic change over time. Alternatively, one can use repeated-measures methods such as Generalized Estimating Equations (GEE) models [70] or mixed models [14], [53] that account for inter-dependence of measurements on the same subject. However, such more complex analyses will require a carefully designed analytical strategy to account for the differences in the number of observations, and time intervals between them. Yet, another modelling strategy is to focus mostly on changes over time in the value of a quantitative "surrogate" outcome. Here, the main challenge may be to separate changes due specifically to $G \times E$ interaction from the "natural" but possibly non-linear changes due to ageing and from spurious "changes" due to regression to the mean.

In this thesis, a series of simulation experiments are used to address some of the afore-mentioned issues, related to the efficiency of using quantitative surrogate outcomes in the study of $G \times E$ interactions. Previous research, that has focussed mainly on binary outcomes [5], [6], [106], reveals that the main determinants of statistical power for detecting $G \times E$ interaction are similar regardless of study design. Although more involved designs, such as those requiring careful matching, especially within family, provide higher relative efficiency and power than unmatched or loosely matched studies [6], [17], [105], [121] many practical considerations make such matching difficult or impossible to achieve. For these reasons, all simulations in this thesis assume the simplest unmatched design, corresponding to a cross-sectional or a prospective cohort study. It is expected that the main conclusions related to relative efficiency of using

quantitative surrogate markers, will be generalizable to other study designs. Moreover, the use of a simple study design has the advantage of allowing a variety of conceptual models regarding the (presumed) underlying relationships between the "clinical" binary outcome, such as presence or incidence of a disease, and a quantitative surrogate to be studied.

Thus, the main objectives of this thesis are:

1. To design a series of simulations implementing alternative conceptual models for the relationship between a binary "clinical" outcome and a quantitative "surrogate" measure,

2. To quantify the expected gain in power from using continuous surrogate outcomes instead of clinical binary outcomes, and to assess its dependence on relevant parameters,

3. To offer researchers some guidelines for outcome selection in particular situations in the context of $G \times E$ interaction studies.

1.6 Organization of thesis

This thesis is divided into five chapters. Chapter 1 was a brief introduction to genetic epidemiology, the importance of gene-environment ($G \times E$) interactions, and the methodological issues involved in its assessment. Motivation for the current work and objectives of the thesis have also been described. Chapter 2 describes statistical concepts relating to $G \times E$ interactions and reviews current literature on power and efficiency of $G \times E$ interaction studies, and the use of surrogate outcomes. Chapter 3 deals with the set up and implementation of the simulation experiments. It describes the proposed hypothetical models for the relationship between the "clinical" outcome and the quantitative "surrogate" measure, the assumptions involved, and the method of data generation. Alternative outcomes are defined and the associated regression models are also described. Chapter 4 presents the results of the simulations, while Chapter 5 provides a discussion of the results. The S-Plus codes prepared by the author for implementing the simulations in this thesis are given in Appendix A.

Chapter 2

Review of Literature on Gene-Environment Interactions

2.1 Introduction

This chapter presents a review of scientific literature on gene-environment ($G \times E$) interactions. First, some fundamental concepts underlying the study of gene-environment interactions will be introduced. The meaning of epidemiologic interaction and the different kinds of interactions will be discussed. Secondly, it reviews some alternative definitions of $G \times E$ interaction put forth by various authors, and traces the development of $G \times E$ interaction studies over the years. Conceptual models of the relationship between gene and environment will also be described.

A brief overview of different study designs for detecting $G \times E$ interactions are presented in this chapter. In addition, issues that are particularly important in the appraisal of studies of $G \times E$ interaction, such as the feasibility, power and efficiency of the design, are briefly described. Finally, a short discussion on surrogate outcomes, and its meaning, as used in the context of this thesis, is given.

2.2 Meaning of interaction

Statistical interaction is a measure of the extent to which the effect of one factor varies with changes in the strength, level or presence/absence of other factors in an experiment. For instance, if the effect of exposure varies with age, this represents an interaction between exposure and age. Thus, interaction does not exist between

a covariate and a risk factor if the association between the covariate and outcome variable is the same across all levels of the risk factor [50]. A related concept is that of *biological* interaction which has been defined as the coparticipation of two risk factors in the same causal mechanism for disease development [95]. When assessing a biological interaction, the primary interest is to estimate the proportion of incident cases of a disease among those who are jointly exposed to both factors that may be due to the interaction of the two exposures. Many studies have suggested that for addressing public health concerns regarding disease frequency reduction, biological interactions are most relevant [62], [95], [127].

A distinction is sometimes made between *qualitative* and *quantitative* interactions. Interaction between two factors, say, X_1 and X_2 is described as qualitative if the direction of the effect of X_1 on the outcome differs depending on the value of X_2. For instance, in the context of $G \times E$ interaction studies, if the exposure is protective i.e. reduces disease risk in individuals with the low risk genotype, but the same exposure becomes a risk factor in persons with the high risk genotype, then the interaction is described as qualitative. A quantitative interaction, on the other hand, reflects changes in the magnitude of X_1 effect with X_2, which do not induce a change in the direction of the effect [30]. Epidemiologists also distinguish between *synergistic* and *antagonistic* interactions. A synergistic effect occurs when the combined effect of risk factors is greater than the sum of the effects of each factor given alone. In contrast, an antagonistic interaction results in a combined effect being weaker than the sum of the individual effects.

Epidemiologists are interested in the quantification of interactions because such effects have important public health implications [62]. They aid in predicting disease rates and provide a basis for well-informed recommendations for disease prevention. Moreover, they provide new insights into disease etiology. Detection of biologically meaningful and statistically significant interactions is therefore an important objective of epidemiologic studies .

2.3 Definitions of gene-environment interaction

The formal study of gene-environment interactions has its roots in the quantitative genetics of agriculture and animal breeding [29], [74]. In human genetics, the early work on $G \times E$ interaction can be traced to Hogben [46], [47], [48] and Haldane [44], and farther back to Galton [34]. The role of $G \times E$ interactions in human diseases has been considered by various researchers. MacMahon [76] discussed the complexity and variability of $G \times E$ interactions in human disease. Eaves [25] proposed a logistic model to evaluate disease risk in the presence of $G \times E$ interaction. Ottman [85] illustrated how epidemiologic principles could be used to investigate relationships between genetic susceptibility and other risk factors for disease. Many authors have concentrated on the methodological issues pertaining to detection of $G \times E$ interaction [6], [17], [32], [35], [38], [42], [51], [104], while others have focussed on the role of such effects in the etiology of cancer and other diseases [81], [109], [117],[128].

In the literature, $G \times E$ interaction has been defined in various ways. Caligari and Mather [15] define $G \times E$ interaction as a situation 'when, because of their genetic differences, two or more individuals, families or genotypic lines respond differently, or to different extents, to change in the environment'. Similarly, Lynch and Walsh [74] define $G \times E$ interaction as the case when 'different genotypes respond to environmental change in nonparallel ways'. Hohenboken [49] states that: 'A genotype × environment interaction exists when the differences between phenotypes due to differences in genotype differ from one environment to another. Equivalently a genotype × environment interaction exists when the magnitude or direction of effects on phenotypic differences due to specific environmental differences differ from one genotype to another.'

In statistical terms, gene-environment interaction is present when the effect of genotype on disease risk depends on the level of exposure to an environmental factor, or vice versa [20], [22], [87], (see Figure 2.1). In a $G \times E$ interaction study, the basic question is whether different genotypes have a different relationship to the phenotype in different environments. The study design therefore requires that the contrasting genotypes (e.g. AA vs aa) be studied under two or more sets of environmental conditions. If the analysis reveals a difference of genotype-phenotype relations in the different environments for persons with contrasting genotypes (as illustrated in Figure 2.1), this

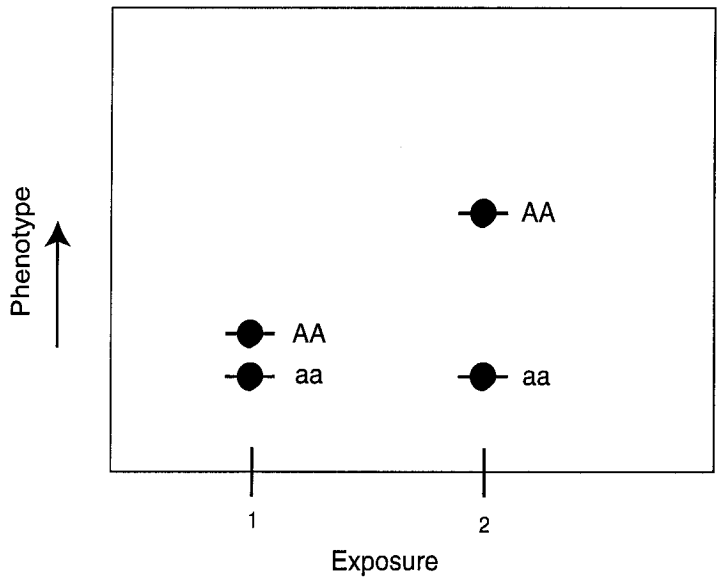

Figure 2.1: Hypothetical gene-environment interaction: shift in the relative relationship between expressed phenotypes when environment changes among individuals with different genotypes (i.e. AA vs aa).

would indicate presence of $G \times E$ interaction [22]. It is important to recognize that the presence or absence of $G \times E$ interaction on the statistical level depends critically on the scale one chooses to measure effects (i.e additive or multiplicative) [43], [120].

2.3.1 Statistical models of $G \times E$ interaction

The statistical quantification of $G \times E$ interaction is model-dependent. Therefore, conclusions regarding presence or absence of $G \times E$ interaction will depend on the statistical model chosen to represent the state of no interaction. There are two models of interaction: additive and multiplicative. If the effects of two variables meet the condition of 'no interaction on a multiplicative scale,' the data can be said to fit a 'multiplicative model,' and if their effects meet the condition of 'no interaction on an additive scale,' the data can be said to fit an 'additive model' [87].

Consider two binary factors G and E, where G represents genetic susceptibility and E represents exposure. Let G_1 and G_0 denote the presence and absence of the high risk genotype, respectively, and similarly let E_1 and E_0 denote presence and absence of the exposure. The two states of interaction may be expressed in terms of the mathematical

Table 2.1: Epidemiologic measures of effects of high-risk genotype (G) and environmental exposure (E)

I. COHORT STUDY				
Disease status	G_1		G_0	
	E_1	E_0	E_1	E_0
Affected	a	b	e	f
Unaffected	c	d	g	h
Risk (r_{ij})	$r_{11} = \frac{a}{a+c}$	$r_{01} = \frac{b}{b+d}$	$r_{10} = \frac{e}{e+g}$	$r_{00} = \frac{f}{f+h}$
Relative risk	$RR_{11} = \frac{r_{11}}{r_{00}}$	$RR_{01} = \frac{r_{01}}{r_{00}}$	$RR_{10} = \frac{r_{10}}{r_{00}}$	$RR_{00} = 1.0$ (ref.)
II. CASE-CONTROL STUDY				
Disease status	G_1		G_0	
	E_1	E_0	E_1	E_0
Case	a	b	e	f
Control	c	d	g	h
Odds ratio	$OR_{11} = \frac{ah}{cf}$	$OR_{01} = \frac{bh}{fd}$	$OR_{10} = \frac{eh}{gf}$	$OR_{00} = 1.0$ (ref.)

Source: Ottman [87], Table 1, pp.765

relationships among the risk ratios (or odd ratios). Table 2.1, Panel I, shows the data layout for a cohort study in which the effects of environmental exposure and genotype on disease risk are assessed [87]. Data from a cohort study can be used to compute four relative risks (RR) using persons with no high risk genotype and no exposure as the reference group. Thus, RR_{11} denotes the relative risk for persons with the high risk genotype and exposure, RR_{10} denotes the relative risk for persons with exposure but no high risk genotype, and so on. The corresponding table for the case-control study is shown in Panel II of Table 2.1 using the odds ratio as the measure of association.

Additive model:

On an additive scale, the effects of genotype and exposure meet the condition of 'no interaction' if

$$RR_{11} = RR_{01} + RR_{10} - 1 \tag{2.1}$$

or (for case-control studies), if

$$OR_{11} = OR_{01} + OR_{10} - 1 \tag{2.2}$$

That is, the effect of an environmental exposure differs among persons with different genotypes (interaction on an additive scale) when $r_{11} - r_{01} \neq r_{10} - r_{00}$.

Multiplicative model:

On a multiplicative scale, the effects of genotype and exposure meet the condition of 'no interaction' if

$$RR_{11} = RR_{01} \times RR_{10} \tag{2.3}$$

or (for case-control studies), if

$$OR_{11} = OR_{01} \times OR_{10} \tag{2.4}$$

If risks are measured on a multiplicative scale, the effect of an environmental exposure differs among persons with different genotypes (interaction on a multiplicative scale) when $r_{11}/r_{01} \neq r_{10}/r_{00}$.

For both additive and multiplicative models, the equations in (2.1)-(2.4) represent conditions for "synergistic" interaction if the '=' sign is replaced by '>' [87].

There has been intense debate on the question of which scale of measurement (additive or multiplicative) should be used in studies of $G \times E$ interaction [63], [65], [93], [116]. The decision regarding choice of an appropriate scale is governed by many factors, including the main objective of the investigation (discovery of etiology, public health planning, etc.) and the hypothesized pathophysiologic model [87]. According to Rothman *et al.* [95], if the primary goal is to unravel disease etiology, it may be more appropriate to use a multiplicative scale whereas if it is to predict the number of cases in the population, it may be more appropriate to use an additive scale.

2.4 Conceptual models of relationship between genotype and exposure

Recognition of the existence of different types of gene-environment relationships is important for gaining insights into disease causation so that effective disease prevention strategies may be developed. Ottman [85] proposed five biologically plausible models of the relationship between a high risk genotype and an environmental risk factor in terms of their effect on disease risk. These models are shown in Figure 2.2.

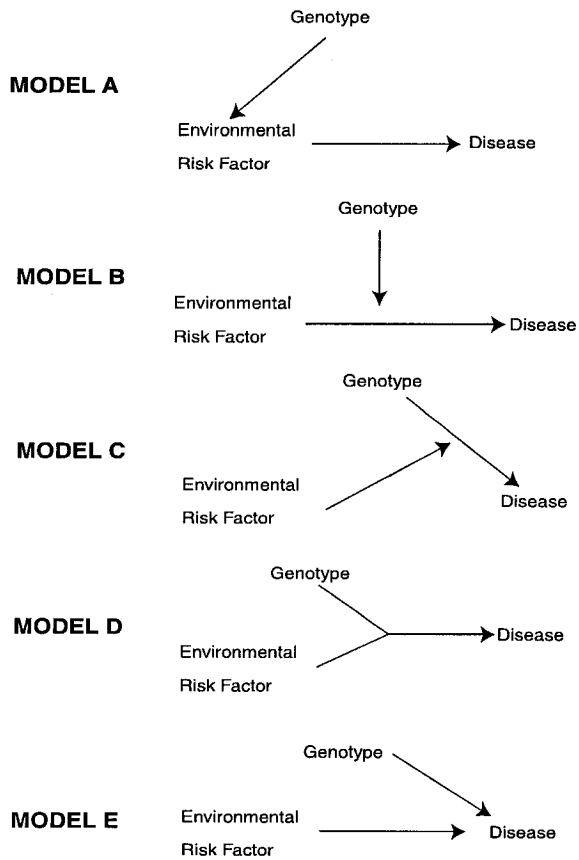

Figure 2.2: Five plausible models of relations between high risk genotype and environmental exposure, in terms of their effect on disease risk.

Model A does not represent interaction, however, it is an important mechanism through which genetic factors influence disease susceptibility. In this model, the effect of the genotype is to produce or increase expression of a "risk factor" that can also be produced nongenetically. An example is the relation of the autosomal recessive disorder, phenylketonuria (PKU), to high blood phenylalanine and mental retardation [85]. PKU results from a genetic variant that leads to deficient metabolism of the amino acid phenylalanine. In the presence of normal protein intake, phenylalanine accumulates and is neurotoxic. Individuals who are homozygous for the PKU gene (high risk genotype) have a buildup of blood phenylalanine after birth (exposure), and the high blood phenylalanine levels cause mental retardation (disease outcome). However, mental retardation can also result from exposure to high blood phenylalanine in per-

sons who do not have the high risk genotype i.e. in persons who are not homozygous for PKU. Phenylanine crosses the placenta, and high maternal blood levels produce mental retardation in the child, regardless of its genotype. It is important to note that the high blood phenylalanine level is an intervening variable [108] in Model A, and the effect of exposure is the same in persons with and without the high risk genotype. Thus, model A does *not* involve an interaction.

Model B (*E-modification* model) considers a mechanism in which the genotype exacerbates the effect of the risk factor, but there is no effect of genotype in persons without the exposure. That is, $\beta_e > 0, \beta_g = 0, \beta_{ge} > 0$. One example is the relation of xeroderma pigmentosum (high risk genotype), an autosomal recessive disorder, to ultraviolet (UV) radiation (exposure) and skin cancer (disease outcome) [85]. Excessive exposure to UV radiation increases risk for skin cancer in the general population, but individuals with xeroderma pigmentosum are deficient in an enzyme required for repair of DNA damage induced by UV radiation, and hence have even higher risk. If sun exposure could be prevented completely in these persons, they would not have increased risk for skin cancer.

In Model C (*G-modification* model), the exposure exacerbates the effect of the high risk genotype, but there is no effect of exposure in individuals with the low-risk genotype. That is, $\beta_e = 0, \beta_g > 0, \beta_{ge} > 0$. An example of this mechanism is the autosomal dominant disorder pophyria variagata [85]. Individuals with this disorder have skin problems of varying severity, including unusual sun sensitivity and a tendency to blister easily. When exposed to barbiturates, an innocuous exposure in the general population, they experience acute attacks that may involve paralysis or even death.

In Model D (*Pure interaction* model), both exposure and the high risk genotype are required to increase disease risk. That is, $\beta_e = 0, \beta_g = 0, \beta_{ge} > 0$. For example, some people with glucose- 6-phosphate dehydrogenase (G6PD) deficiency (an X-linked recessive disorder) develop severe hemolytic anemia if they eat fava beans. Dietary exposure to fava beans does not produce this reaction in individuals without G6PD deficiency [85].

Model E (*GE-modification* model) assumes that the exposure and genotype each have some effect on disease risk, and when they occur together risk is higher or lower

than when they occur alone. That is, $\beta_e > 0, \beta_g > 0, \beta_{ge} \neq 0$. The relation between a-1- antitrypsin deficiency (high risk genotype), smoking (exposure), and chronic obstructive pulmonary disease (COPD)[disease outcome] is an example [85]. Risk of COPD is increased both in nonsmokers with a-1-antitrypsin deficiency and in smokers without a-1-antitrypsin deficiency. However, risk is increased to a greater extent in smokers with a-1-antitrypsin deficiency.

2.5 Study designs for detection of $G \times E$ interaction

When the genetic factor and/or environmental exposure are rare, and the interaction effect is moderate, sample sizes required to detect $G \times E$ interactions can be prohibitive [38], [51]. Even when the required sample size is attainable, the cost involved in genotyping large samples can render such studies infeasible. Thus, search for efficient study designs that reduce required sample size has become an important area of genetic-epidemiologic research.

Most study designs for gene-environment interactions are variants of the common epidemiological cohort and case-control designs [62] (Figure 2.3). In studies of human mutations, both main designs are known to have their advantages and disadvantages [58]. In a cohort study, a group of disease-free individuals is identified, perhaps on the basis of exposure to a risk factor of concern, and then followed up over time to determine eventual disease incidence in exposed and unexposed sub-groups of the population. For a gene-environment interaction study, all of the subjects could be exposed to a particular environmental risk factor, and comparison of disease incidence between different genotypes would be of primary interest. The interaction effect $\hat{\beta}_{ge}$ is estimated by fitting a logistic regression model containing the $G \times E$ interaction term and the main effects of G and E to the full cohort data.

In studies of disease incidence, the long follow-up necessary for the cohort design can be overcome by using a retrospective cohort. However, for gene-environment interaction studies, this would necessitate archiving suitable biological samples such that genotyping of all members of the cohort could be performed [21]. In the nested case-control design (or case-control study nested in a cohort study), once sufficient cases have accrued within the cohort, appropriate controls are selected from the remainder

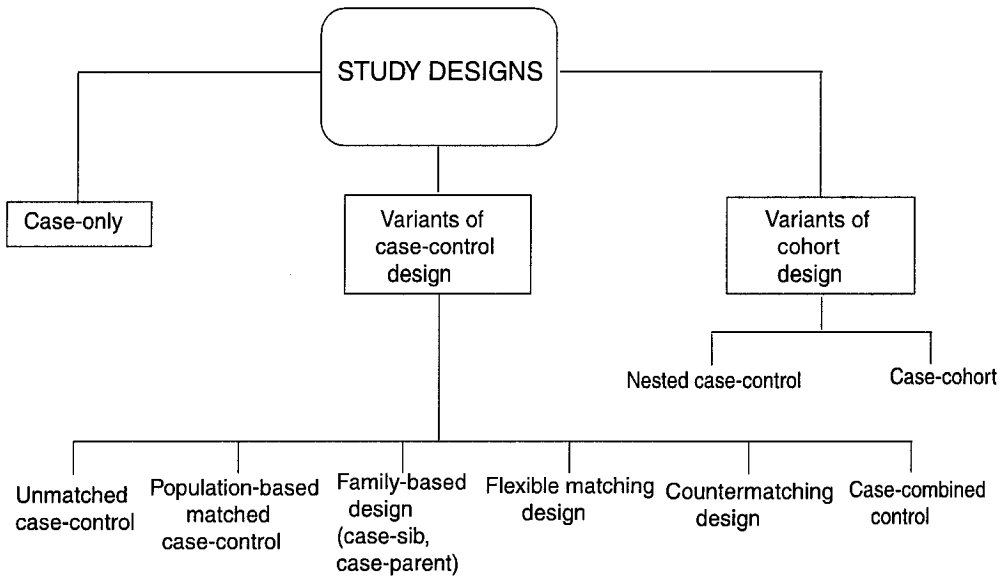

Figure 2.3: Study designs to detect gene-environment interactions.

of the cohort [21], [68]. Genotyping is required for only those cases and controls selected for the nested case-control component of the investigation. The use of nested case-control analysis of archived samples can potentially minimize the disadvantages of the cost of genotyping an entire cohort [68].

The case-control design is more commonly used in studies of gene-disease associations and gene-environment interactions for late-onset diseases [60]. In unmatched case-control designs, a major disadvantage is the low power and efficiency of these designs and bias in the estimates of genetic effects due to confounding by ethnicity [121]. If the allele frequency at a particular genetic locus varies across ethnic groups and if ethnicity (or some unobserved factor that varies by ethnicity) is a risk factor for disease independent of that locus, then failure to adequately control for ethnicity can result in false associations between the gene and the disease [40] (see Figure 2.4). This phenomenon is known as population stratification or genetic admixture [39]. To overcome these limitations, alternative designs are used that employ different matching strategies. There are three basic types of matched case-control studies:

- Population-based case-control design
- Case-sib design

- Case-parent design

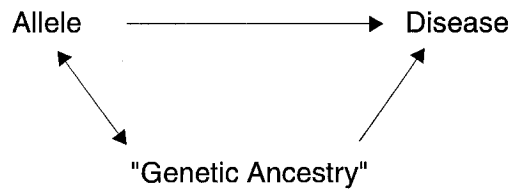

Figure 2.4: Confounding due to ethnicity or genetic ancestry.

2.5.1 Population-based case-control design

In this design, a random sample of cases (individuals with the disease of concern) is compared with a group of controls (individuals who are free of the disease at the time of the study), in terms of potentially causal genetic and exposure factors [21]. Each case is matched to one or more controls selected from the source population of the case. Adjustment for potential confounding variables is accomplished by selecting controls subject to the matching criteria (e.g. age, similar ethnic background, etc.). By using unexposed individuals with no susceptibility genotype as the referent group, odds ratios (OR's) for all other groups can be estimated under either multiplicative or additive models. The main advantage of this design is that the main effects of the environmental exposure and genetic susceptibility, as well as their interactive effect, may be estimated [42].

The power of studies with the population-based case-control design is influenced by the frequency of the exposure of interest and by the population prevalence of the susceptible genotype. It has been suggested that both need to be relatively common (i.e. $> 25\%$) for case-control studies to detect $G \times E$ interactions with a reasonable probability [42]. Thus, a disadvantage of this design is that it may not be appropriate for the study of $G \times E$ interaction involving rare genes or uncommon environmental exposures (assuming reasonable strength of the interaction effect). Moreover, population stratification adversely influences this design when there is poor ethnic matching.

Analysis

When the outcome of interest is binary, a standard method for the analysis of matched data is conditional logistic regression [12], [62]. Conditional logistic regression can be used to simultaneously model genetic and environmental main effects, as well as $G \times E$ interaction. This multivariable procedure is specifically designed for use when there are small stratum-specific sizes. Thus, it is ideally suited to matched study designs and should be used to avoid biased parameter estimates.

Consider a matched case-control study involving N cases, where the ith case is individually matched to R_i ($R_i \geq 1$) controls on one or more variables. Thus, the total number of cases and controls in the ith stratum is $m_i = 1 + R_i$. Let β_g, β_e, and β_{ge}, denote parameters for the effect of a gene (G), environmental factor (E), and $G \times E$ interaction, respectively. For a dichotomous response variable Y with $Y = 1$ indicating a case and $Y = 0$ indicating a control, consider fitting the multiple conditional logistic regression model,

$$\text{logit}[Pr(Y_{ij} = 1)] = \beta_0 + \beta_g G_{ij} + \beta_e E_{ij} + \beta_{ge} G_{ij} E_{ij} \tag{2.5}$$

where $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, m_i$. Then, the conditional likelihood for a sample of N matched sets has the form:

$$L(\beta_g, \beta_e, \beta_{ge}) = \prod_{i=1}^{N} \frac{e^{\beta_g G_{i1} + \beta_e E_{i1} + \beta_{ge} G_{i1} E_{i1}}}{\sum_{j \in M(i)} e^{\beta_g G_{ij} + \beta_e E_{ij} + \beta_{ge} G_{ij} E_{ij}}} \tag{2.6}$$

where the index '1' in the numerator of (2.6) refers to the case and the set $M(i)$ includes all subjects in matched set i [38]. Parameters β_g, β_e, and β_{ge} are estimated from (2.6) using the method of maximum likelihood (ML) [50], [62]. The ML estimates ($\hat{\boldsymbol{\beta}}$) are consistent estimates of the log-odds ratios from the logistic regression model (2.5). The baseline probability of disease in the population, corresponding to unexposed subjects with no "adverse" genotype, is given by $\frac{\exp(\beta_0)}{1+\exp(\beta_0)}$, and quantities $OR_g = \exp(\beta_g)$, $OR_e = \exp(\beta_e)$ and $OR_{ge} = \exp(\beta_{ge})$ are the genetic, environmental and interaction odds ratios, respectively.

The model without the $G \times E$ interaction term in (2.5) is the standard multiplicative odds ratio model. Departures from this model can be assessed by fitting (2.5) and calculating a Wald's test or a score test for the interaction term β_{ge}, or a likelihood

ratio test comparing (2.5) with the standard multiplicative odds ratio model. Thus, OR_{ge} is a measure of departure from the standard multiplicative odds-ratio model.

To test:

$$\begin{aligned} H_0 : \beta_{ge} &= 0 \\ H_1 : \beta_{ge} &\neq 0, \end{aligned}$$

using the score test, compute

$$\mathbf{U}'(\hat{\boldsymbol{\beta}}^{\mathbf{0}})\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}}^{\mathbf{0}})\mathbf{U}(\hat{\boldsymbol{\beta}}^{\mathbf{0}}) \sim \chi_1^2. \tag{2.7}$$

Here, $\hat{\boldsymbol{\beta}}^{\mathbf{0}}$ is the MLE of $\boldsymbol{\beta}$ under H_0, $\mathbf{U}(\boldsymbol{\beta})$ is the vector of first partial derivatives of $\ln L$ with respect to $\boldsymbol{\beta}$ where L is given in (2.6), and $\mathbf{I}(\boldsymbol{\beta}) = -\mathbf{H}(\boldsymbol{\beta})$ or $E[-\mathbf{H}(\boldsymbol{\beta})]$, where $\mathbf{H}(\boldsymbol{\beta})$ is the matrix of second partial derivatives of $\ln L$ with respect to $\boldsymbol{\beta}$. The significance of β_{ge} in (2.5) can be determined by comparing the residual chi-square in (2.7) with a chi-square with one degree of freedom. Alternatively, one may compute the Wald statistic:

$$W = \frac{\hat{\beta}_{ge}}{SE(\hat{\beta}_{ge})} \sim \mathcal{N}[0,1]. \tag{2.8}$$

$SE(\hat{\beta}_{ge})$ is calculated from $\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}})$, where $\hat{\boldsymbol{\beta}}$ is the ML estimate for the vector of parameters $\boldsymbol{\beta}$. For the likelihood ratio test, we compute:

$$\Lambda = 2(\hat{L}^1 - \hat{L}^0) \sim \chi_1^2 \text{ under } H_0. \tag{2.9}$$

Here, $\hat{L}^1 = \ln[L(\hat{\beta}_g, \hat{\beta}_e, \hat{\beta}_{ge})]$ and $\hat{L}^0 = \ln[L(\hat{\beta}_g, \hat{\beta}_e)]$ are the maximum of the log-likelihood based on (2.6) under H_1 and H_0 respectively.

2.5.2 Case-sib design

Determination of ethnicity in a large-scale epidemiology study is difficult, especially with the great diversity in cultural backgrounds that exists in urban areas where studies are most likely to be conducted. Thus, the population-based case-control design may not be optimal for the study of genes since it does not adequately match cases to controls on ethnic background [40]. In contrast, the main advantage of family-based case-control studies (case-sibling and case-parent designs) is their freedom from

population stratification. For these reasons, family-based studies are of particular interest in genetic epidemiology.

In the case-sib design, each case is matched to one or more of his/her unaffected, i.e. disease-free, siblings [39], [40]. This has the advantage that cases and controls are perfectly matched on ethnic background. For complex diseases with variable age at onset, controls are sampled from the 'risk set' consisting of those siblings who were disease-free at the age the case became affected (the index age) [40]. A sibling who is disease-free at the index age but is known to later develop the disease is also eligible as a control. Data on known environmental risk factors *at the index age* are collected for both cases and controls. Validity of the case-sibling design depends on a number of important factors. Firstly, only recent incident cases should not be considered since the age-matching requirement restricts control selection to older siblings. This could lead to confounding of the effects of environmental exposures that have secular trends or birth-order effects [40], and is a potential source of bias in estimates of $G \times E$ interaction effects [120]. Secondly, inclusion of controls who have not attained the index age can pose problems if time-dependent covariates are involved [39].

Analysis

Standard methods for the analysis of matched case-control data can be applied to the case-sibling design. These include McNemar's and Mantel Haenszel chi-squared tests and the associated estimates of the odds ratio [12]. Conditional logistic regression can be used to simultaneously model genetic and environmental main effects, as well as $G \times E$ interaction. The conditional likelihood for a sample of N case-sibling sets has the same form as in (2.6) where the index '1' refers to the case, and the set $M(i)$ includes the case and all controls from family i. If controls are matched to the case's age and selected according to the principles of risk set sampling, the quantities $R_g = \exp(\beta_g)$, $R_e = \exp(\beta_e)$, and $R_{ge} = \exp(\beta_{ge})$ can be interpreted as the corresponding hazard-rate ratios [120]. If age of onset is not a factor, then these quantities represent odds ratios.

Score, Wald and likelihood ratio tests [50] based on (2.6) can be formed as usual to test hypotheses about main or interactive effects. The conditional likelihood assumes that the disease status in sibships is independent, given their covariate information (G

and E) [39]. Thus, if there are more than two subjects per family, these tests will be valid only if disease outcomes are conditionally independent within families.

2.5.3 Case-parent design

The case-parent design has been considered by many authors for detection of $G \times E$ interaction [77], [98], [114]. In this design, 'cases' refer to the parental alleles or genotypes transmitted to an affected offspring and the 'controls' refer to the parental alleles or genotypes not transmitted. This can be cast as a 1:1 matched case-control analysis with each parent-offspring pair as a matched set and analyzed with conditional logistic regression (denoted 'allelic' transmission disequilibrium test [TDT]) [77]. Alternatively, one may consider a 1:3 matched analysis with one case genotype and three 'pseudo-siblings' through conditional logistic or log-linear regression ('genotypic' TDT) [98], [114]. 'Pseudo-siblings' refers to the three genotypes that were not transmitted to the case. For example, if the father's genotype is Aa, the mother's Aa, and the case's AA, the pseudo-sibling genotypes are Aa (paternal A, maternal a), aA and aa. Genotypes are collected from the case and his/her two parents, while environmental data are required only from the case [98].

Validity of the case-parent design depends on the assumption that parental alleles are transmitted with equal and independent probability in the population. It also assumes that the ability to recruit a case is independent of the genotype given the parental genotypes, and that the genotyped "parents" are in fact the case's biological parents [118].

Analysis

Conditional logistic regression for 1:3 matched sets provides a flexible framework for analyzing case-parent data [38], [98]. Estimation of a main environmental effect is not possible from this design since the three possible pseudo-siblings are perfectly matched to the case except for genotype. Thus, the likelihood including the genetic main effect and a $G \times E$ interaction has the form:

$$L(\beta_g, \beta_{ge}) = \prod_{i=1}^{N} \frac{e^{\beta_g G_{i1} + \beta_{ge} G_{i1} E_{i1}}}{\sum_{G_{ij}|G_{i\mathbf{P}}} e^{\beta_g G_{ij} + \beta_{ge} G_{ij} E_{i1}}} \tag{2.10}$$

where the index '1' refers to the case. The summation in the denominator of (2.10) is over the four possible genotypes that could be transmitted to an offspring given parental genotypes $G_{i\mathbf{P}}$. The quantities $R_g = \exp(\beta_g)$, and $R_{ge} = \exp(\beta_{ge})$ can be interpreted as relative risks [120].

An important assumption for valid estimation of the interaction effect β_{ge} requires that G and E are independently distributed in the population. A limitation of this design is that even if the independence assumption does hold, it is difficult to interpret $G \times E$ interaction in the absence of knowledge of the main effect of exposure. This is because without the main effect of exposure, it would not be possible to determine how the effect of the exposure on disease risk varies among individuals with and without the susceptibility genotype [118].

2.5.4 Alternative designs for rare factors or disease outcome

When either the exposure or the susceptibility genotype is rare, *multi-stage designs* may be employed to overcome the limitations of conventional case-control studies. The basic principle of these designs is to increase, in some way, the numbers of cases and/or controls with the rare factor of interest. Andrieu *et al.* [6] proposed *counter-matching* cases to controls as a technique for using available data at the time of sampling to enrich the sample for informative matched sets. Controls are selected to increase the variation in factors of interest in a case-control set relative to random sampling. The main purpose of this design is to enhance power to detect $G \times E$ interactions involving rare genes (G) or uncommon environmental exposures (E). Another multistage design is the *balanced design* [12], [119] which, rather than selecting a subset at random, selects cases and controls in order to oversample for the rare factor of interest. The oversampling is taken into account in the analysis to obtain unbiased estimates of the effects of individual factors and their interaction [6].

Stürmer and Brenner [106] introduced *flexible matching* strategies with varying proportions of a matching factor among selected controls to detect and estimate $G \times E$ interactions. Recently, Andrieu and Goldstein [5] proposed the *case-combined-control design* that uses related and unrelated controls simultaneously.

Several authors have proposed the use of *case-only designs* as an alternative to case-

control designs to study gene-environment interactions [4], [59], [118]. This design has been used in a number of studies [10], [79]. When the disease outcome of interest is rare, case-only studies offer greater precision for estimating gene-environment interactions than case-control studies of comparable size [126]. Interactions are assessed by carrying out logistic regression analysis using only cases, treating G as the outcome. Validity of the design hinges on one assumption - that the genetic and environmental factors of interest are independent of one another. A disadvantage of the case-only approach is that bias could arise because of incomplete mixing of subpopulations that differ by exposure prevalence and genotype prevalence, even if their baseline risks of disease do not differ [118]. Moreover, effects of the individual genotype and of the environment cannot be measured.

2.6 Power, efficiency and sample size considerations in studies of $G \times E$ interaction

Since investigation of interactions requires sample sizes much larger than those needed to investigate main effects [6], a major issue in the study of $G \times E$ interactions is the feasibility of the study design, in terms of ability to recruit and evaluate a number of subjects sufficient to ensure adequate statistical power to detect clinically meaningful interactions. Power and efficiency considerations are also critical for the statistical evaluation of models of interaction [35]. In the context of unmatched case-control studies, several studies have described methods for estimating sample size and power in studies of $G \times E$ interaction [32], [35], [42], [51]. The low power and efficiency of unmatched designs [105] has led to the development of various matching strategies for estimation and detection of $G \times E$ interaction.

Stürmer and Brenner [105] studied the effect of matching on an environmental risk factor on the efficiency and power to detect $G \times E$ interactions. Comparisons of sample size requirements for frequency matched versus unmatched case-control studies were made by examining differences in the number of unmatched and matched controls required to obtain a similar level of power.

Witte *et al.* [121] made some comparisons of efficiency in the case-sib, case-parent designs and matched case-control design, for a limited number of disease models. A

method for computing power and sample size to detect $G \times E$ interaction in the case-parent design was described by Schaid [98]. Feasibility of this design was compared with the unmatched case-control design. Gauderman [38] provided a general framework for the calculation of power and sample size in the context of matched case-control, case-sib and case-parent designs for a range of conceptual $G \times E$ interaction models, and studied their efficiencies relative to one another. Recently, Chatterjee *et al.* [17] proposed a novel conditional likelihood approach for the analysis of family-based studies under the assumption that genetic susceptibility and environmental exposure are distributed independently of each other within families in the source population. Based on these methods, they evaluated various family-based study designs by examining their efficiencies relative to each other, and their efficiencies compared to a population-based case-control design of unrelated subjects.

Andrieu *et al.* [6] proposed the counter-matching design to increase a study's power to detect $G \times E$ interaction when one of the factors under study is rare. Efficiency of the design was evaluated and comparisons made with a full cohort study with no matching and a standard nested case-control study. Sample sizes (number of counter-matching sets) were calculated for a design that counter- matched on surrogates of both G and E.

Power and efficiency of the flexible matching design was assessed by Stürmer and Brenner [106] under a variety of assumptions regarding the prevalence and effects of environmental exposure and the genetic susceptibility, as well as their association in the population. Results were compared with the unmatched case-control study. Saunders and Barrett [97] studied sample size requirements for this design under an optimal matching strategy. Comparisons with regard to efficiency were made with an unmatched population-based case-control study and a case-only design for a range of magnitudes of risk factor effects and frequencies.

The efficiency of the case-combined-control design relative to the classical case-control study was studied by Andrieu and Goldstein [5]. They provided examples to compare efficiency of this design with the classical case-control design.

The effect of measurement errors on the feasibility of a $G \times E$ interaction study has been investigated by Garcia-Closas and Lubin [36] for binary exposure and outcome,

and by Wong *et al.* [122] in the case of continuous exposure and outcome.

2.6.1 Methodologies for assessing feasibility and power of designs

The power of a test is the probability $1-\beta$ that the test will reject the null hypothesis H_0 when it is false. Here, β is the probability of a type II error, that is, the probability of accepting H_0 when it is false. Statistical power analysis characterizes the ability of a study to detect a meaningful effect size. It also determines the sample size required to provide a desired power for an effect of interest.

There are many factors involved in a power analysis, such as the research objective, design, data analysis method, sample size, type I error (α), effect size and variability. In $G \times E$ interaction studies, statistical power clearly depends on the prevalence of the at risk genotype and the magnitude of both absolute and relative risks. Hence, for polymorphisms present in a large proportion of the source population and with a high relative risk, only small numbers of subjects are required, whereas large samples are required when the population prevalence of the susceptible genotype is small and the relative risk is low.

In designing a $G \times E$ study, it is essential that power calculations are undertaken in advance to obtain a realistic estimate of the number of subjects necessary to ensure a statistically meaningful outcome. For some statistical models and tests, power analysis calculations are exact in the sense of utilizing a mathematical formula that expresses power directly in terms of the relevant design parameters. Such formulae typically involve either enumeration or noncentral versions of the distribution of the test statistic [16], [28]. In the absence of exact mathematical results, approximate formulae can sometimes be used. When neither exact power computations nor reasonable approximations are possible, simulation provides an increasingly viable alternative [16], [28].

Formulae for estimating minimal sample sizes, both for cohort and case-control studies, are available for standard study designs [12], [94], [99]. In $G \times E$ interaction studies, asymptotic methods such as the likelihood ratio (LR) method are commonly used for estimation of sample size and power for hypothesis testing. The steps of LR

sample size or power calculations maybe outlined as follows [13]:

1. A complete alternative hypothesis parametric model is specified which represents the true state of nature that is to be detected. This specification includes the model, the values of the parameters, and the design of the experiment.

2. Constraints are specified on the model parameters that transform the alternative hypothesis (H_1) into the null hypothesis (H_0). Usually, this includes constraining the regression parameter for $G \times E$ product to zero.

3. The LR methods use the assumption that for sufficiently large sample sizes, the distribution of twice the loglikelihood ratio is distributed approximately as a non-central χ^2. The loglikelihood $\hat{L}^1$ is obtained by fitting the data by maximum likelihood using the unconstrained model (i.e. under H_1). Similarly, $\hat{L}^0$ is obtained using a model incorporating the constraints of the null hypothesis. The degrees of freedom of the non-central χ^2 is K, the total number of constraints. If the null hypothesis is correct, the distribution of the likelihood ratio statistic $\Lambda = 2(\hat{L}^1 - \hat{L}^0)$ is approximately a central χ^2 (i.e. non-centrality parameter, $\eta = 0$).

4. The critical value, C, is computed from the two-sided significance level, α, from the cumulative central χ^2 distribution:

$$\alpha = 1 - P[\chi^2(K) \leq C]. \tag{2.11}$$

   Power is the probability that the non-central χ^2 distribution with non-centrality parameter, η, exceeds this critical value, that is

$$P[\chi^2(K;\eta) > C] \tag{2.12}$$

5. If sample size is to be determined, η is chosen to provide the desired power. The sample size is chosen to yield this value of η using the fact that the non-centrality parameter is linear in the sample size.

It is important to note that LR methods are asymptotic yielding correct values as the sample size gets large. The advantage of these methods over simulation is that

they are simple and cheap to compute. They allow rapid comparisons of different designs and provide a direct method of calculating requisite sample sizes [13].

Binary outcomes

Gauderman [38] used asymptotic methods to estimate sample size for test of $G \times E$ interaction for matched case-control, case-sib and case-parent designs. His calculations were based on the LR test statistic Λ for a conditional regression analysis of matched case-control data. For N matched sets, $N\Lambda$ is the non-centrality parameter of the χ^2 distribution under H_1. When G and E are both dichotomous, the test on interaction has one degree of freedom and N can be computed as

$$N = (z_{\alpha/2} + z_{\beta/2})^2/\Lambda \tag{2.13}$$

with α as the two-sided type I error, β as the type II error and z_u as the $(1-u)$th percentile of $\mathcal{N}(0,1)$. For a given N, power was computed as

$$1 - \beta = \Phi(\sqrt{(N\Lambda)} - z_{\alpha/2}) + \Phi(-\sqrt{(N\Lambda)} - z_{\alpha/2}) \tag{2.14}$$

Saunders and Barrett [97] used the LR method of Self *et al.* [100] to estimate sample size for the flexible matching design. LR methods have also been used by Schaid [98] for the case-parent design.

To compare efficiency of alternative designs, asymptotic relative efficiencies (ARE) are often computed. The ARE of design B relative to A is defined as the ratio of the asymptotic $G \times E$ interaction variance for design A to the corresponding variance for design B. Andrieu *et al.* [6] used this definition to compare efficiency of the counter-matched design with the classical 1:3 nested case-control and full cohort designs. Andrieu and Goldstein [5] computed ARE to compare efficiency of the case-combined control study with the classical case-control design. To evaluate the efficiency of the family-based designs relative to the case-control design, Gauderman [38] estimated the number of matched sets (N) required for both designs to achieve 80 percent power for rejecting $H_0 : \beta_{ge} = 0$ and evaluated the ARE as the ratio of N for the case-control design to N for the family-based design.

Quantitative outcome measures

The above studies considered sample size calculations for a binary environmental exposure, binary genetic factor and a binary outcome event variable. Luan *et al.* [72] described sample size calculations for the situation where the outcome variable and environmental exposure were continuously distributed. The test statistic for the $G \times E$ interaction had a non-central F-distribution under H_1. Using the non-central F-distribution and non-centrality parameter, the power to detect an interaction effect, or alternatively, the sample size necessary to detect a given interaction with fixed power and significance may be calculated. Wong *et al.* [122] took into account measurement errors in the continuous outcome and exposure in their sample size and power calculations. They presented an LR statistic for testing $H_0 : \beta_{ge} = 0$ (in the situation where there was measurement error in assessment of E and genotype could not be assessed correctly) which was approximately distributed as a non-central χ^2 under H_1 with 1 degree of freedom and non-centrality parameter ϕ_n.

Role of Simulations in Assessing Efficiency

Simulation has become an important tool in power analysis. This entails specifying hypothetical "plausible" values for model parameters and using them to randomly generate a large number of hypothetical data sets. By applying the statistical test to each data set, power is estimated as the percentage of times the null hypothesis is rejected. While the simulation approach is computationally extensive, faster-computing makes this less of an issue. A simulation-based power analysis is always a valid option, and, with a large number of data replications, it can often be more accurate than analytical approximations [16].

Simulations are increasingly being used in studies of $G \times E$ interaction. To compare the efficiency of matched and unmatched case-controls studies in $G \times E$ interaction, Stürmer and Brenner [105] simulated frequency matched and unmatched case-control studies for a wide range of scenarios regarding the prevalence of E and G in the population, their association with disease, and the strength of the interaction between these factors. Simulated samples were analyzed with multivariable logistic regression. The power of the matched and unmatched study design to detect $G \times E$ interaction

was calculated by the proportion of simulated samples in which the two-sided P value for the test of the estimated interaction parameter was smaller than 0.05. Sample size requirements were compared by observing the numbers of controls per case required to obtain certain levels of power, for a given combination of parameter values of the basic scenario.

Andrieu and Goldstein [5] used simulations to assess the efficiency of the case-combined-control design relative to a classical case-control study under a variety of assumptions. Stürmer and Brenner [106] evaluated power and efficiency of the flexible matching design under a variety of assumptions regarding the prevalence and effects of E and G as well as their association in the population. For each set of parameters, 10,000 case-control studies were simulated using varying degrees of matching and each simulated study was analyzed using unconditional logistic regression. For each degree of matching, the relative efficiency of estimation compared with the unmatched design was obtained by dividing the variance of the regression coefficient of interaction across replications in the unmatched studies by the variance observed in studies with the corresponding degree of matching. Chatterjee *et al.* [17] also used simulations to evaluate the relative efficiencies of different family-based designs and analytic methods using the population-based case-control design as the common reference point. The quantity $\tau = \hat{\beta}_{ge}/\mathrm{sd}(\hat{\beta}_{ge})$ was evaluated where $\hat{\beta}_{ge}$ and $\mathrm{sd}(\hat{\beta}_{ge})$ are the empirical mean and the empirical standard error of the estimate of β_{ge} from a given design over different simulated data sets. The asymptotic relative efficiency of the two designs in estimating the interaction was estimated by the ratio of τ^2 for the two designs. It should be noticed that this ratio is *not* equivalent to the ratio of the power obtained with the two designs.

2.6.2 Comparison of study designs based on feasibility and power

Studies comparing unmatched and matched case-control designs reveal that more than double the number of unmatched than matched controls are needed to obtain a similar level of power in detecting $G \times E$ interaction [105]. Comparisons with respect to required sample sizes have been made between the standard matched case-control stud-

ies: population-based case-control, case-sibling and case-parent designs [38]. These reveal that family designs need smaller sample sizes than the population-based case-control design. When genetic susceptibility is rare, the case-sibling design is preferable (in terms of smaller sample size) to the case-parents design [38].

Sample size requirements for the flexible matching design and case-combined-control design have also been compared to the traditional case-control design. Both are found to require smaller sample sizes than the case-control design [5], [97]. For the case-combined-control design, when the genetic factor G is rare, the required sample size is only realistic when there are strong $G \times E$ interaction and G main effects [5]. Feasibility of the counter-matching design has been evaluated [6], but its sample size requirement is yet to be compared with other designs. When the frequency of G is very small, the needed sample size is only realistic when factor E is common and the interaction effect is high.

Measurement errors have an impact on the power and feasibility of $G \times E$ interaction studies. Misclassification of a binary environmental factor biases a multiplicative interaction effect toward the null value. This result is also true for misclassification of genetic factors. As a result of misclassification, the sample size required to detect $G \times E$ interaction with a given statistical power increases [36].

The studies reported above all involve dichotomous disease and exposure variables, and define interaction on a multiplicative scale. Feasibility studies have also been made for continuous environmental exposures and disease outcomes [72], [122]. These reveal that in the absence of measurement errors, smaller sample size is required to detect a moderately strong interaction with a given level of power if the association between the exposure and outcome is strong. Power is markedly increased if the interaction is very strong. Measurement errors of the exposure and/or of the outcome, and the degree of genetic misclassification, are also determinants of sample size. Larger sample sizes are required for studies with poor assessment of the exposure and/or outcome. However, impact of misclassification in the assessment of genotype is relatively minor except when the frequency of the minor allele is low [122].

Any gain in feasibility must be balanced by power and efficiency. Studies show that matching for the environmental risk factor may often enhance power and efficiency to

detect gene-environment interactions [105]. The counter-matching design is found to be more efficient than a standard nested case-control design, the gain in efficiency being greatest for very rare risk factors. However, the study of such rare factors using the counter-matching design is unrealistic unless one is interested in very strong interaction effects [6]. Compared to traditional frequency matching, flexible matching strategies increase the power and efficiency of case-control studies, the highest gain in efficiency being obtained for a rare exposure that is a strong risk factor [106]. In spite of increased complexity for control recruitment, the case-combined-control design appears more efficient relative to the classical case-control design for detecting interactions involving rare exposures and/or genetic factors. The relative efficiency of this design decreases for common genes with moderate effects [5].

Family-based designs such as the case-sibling and case-parent designs provide greater efficiency compared to the population-based case-control design [38]. The gain in efficiency when relative controls are used usually decreases as the frequency of G increases. The case-sibling design is more efficient when studying a dominant gene, whereas a case-parent design is preferred for a recessive gene [38], [40]. However, a recent study that uses a novel conditional likelihood framework for exploiting the within-family $G-E$ independence assumption reports that the case-sibling design can be more efficient than the case-parent design even for recessive genes [17].

Estimates of $G \times E$ interactions from the case-only study have been shown to be very efficient relative to estimates obtained with a case-control study under the assumption of independence between the genetic and environmental factors [1], [59]. However, inferences about multiplicative interaction with the case-only design can be highly distorted when there is departure from the independence assumption [1], [37].

2.7 Surrogate outcomes

The terms surrogate outcome, surrogate endpoint or disease marker commonly appear in the literature of clinical studies. According to Temple [110], a surrogate endpoint is defined as a laboratory or physiologic measurement used as a substitute for a clinical endpoint that measures directly how a patient feels, functions, or survives. For example, blood pressure may be used as a surrogate outcome for stroke, degree of

atherosclerosis on coronary angiography may be considered a surrogate outcome for myocardial infarction or coronary death, etc. Surrogate outcomes or "markers" might reflect underlying disease pathophysiology, predict future events, or indicate the presence of disease or damage to an organ. A marker could also be measured to assess the progress of treatment. They are used as alternative endpoints because they are quicker to measure, thus, leading to the faster evaluation and appropriate dissemination of new treatments. An *ideal* surrogate measure is sensitive to disease evolution so that changes to the surrogate endpoint reliably predict the risk of the clinical endpoint. In clinical trials, several types of surrogate outcomes, such as composite outcomes [73] and multivariate risk scores [55], have been used to increase trial efficiency, in terms of statistical precision and power.

One reason why relatively few studies test for the impact of $G \times E$ interaction on hard clinical endpoints such as death is because sample sizes required to show interaction effects are largely prohibitive [52]. In clinical studies, where it is not feasible to have adequate statistical power for a clinical endpoint, a valid surrogate may be used as a primary outcome measure. Indeed, some studies have demonstrated the benefits of surrogate outcomes in detecting treatment effects and interactions with increased power and reduced size of trials [80], [84], [125]. However, replacing the clinical outcome by the surrogate outcome in individuals for whom the clinical outcome is observed may often lead to misleading results [31]. Thus, methods have been developed for time-to-event analysis that retain the clinical endpoint as the primary outcome, but use surrogate outcomes to provide additional information for censored subjects [67], [75], [83], [92]. These methods are known to enhance power and efficiency of the analyses of highly censored data [75].

In this thesis, the term "surrogate" outcome is used in a broader context, and is not limited to markers of disease outcome in clinical trials. We refer to "surrogate" outcomes as alternative outcome measures for the true clinical outcome. Outcome variables may be of several types: binary, continuous, categorical, counts and censored times-to-event variables. The type of outcome determines the method of analysis and the efficiency of a study. In some studies, the event of interest arises from an intrinsically binary outcome variable (e.g. death or myocardial infarction), while in

other situations, the outcome variable is measured continuously, but "dichotomized" to permit a more efficient estimation of the risk of a particular event [107]. The choice of the type of surrogate outcome may depend on several factors. For instance, the use of the dichotomized version of a continuous outcome variable may be preferred if the risk of an event is a clinically more meaningful measure than the mean of the continuous variable related to the disease [107]. However, in general, it is better to record data with the highest possible information content. For example, additional information conveyed by a continuous variable may result in a study with higher statistical power. In addition, the consequences of misclassification related to measurement error, and bias due to several values falling close to the cutoff, are potentially reduced [107]. For some diseases, the rate of change of some quantitative measurement reflects disease severity and may serve as a potential surrogate [91]. Alternatively, outcomes defined from repeated measures of a variable, such as the mean of two or more measurements, can improve precision and consequently the power of the study [115].

2.8 Summary

This chapter has reviewed some background on $G \times E$ interaction studies. Basic concepts such as definition of $G \times E$ interaction, statistical models of interaction, and types of gene-environment relationships were discussed. Study designs for detecting $G \times E$ interactions were reviewed. Special emphasis was placed on the statistical issues involved in studies of $G \times E$ interactions, and common methodologies (such as the use of simulations) for assessing power and efficiency. Lastly, a short section was presented on types of surrogate outcomes and the potential advantages of using "surrogates" instead of the true "clinical" outcome.

In the following chapter, we draw upon some of the concepts described in this chapter to formulate the main research problem of this thesis.

Chapter 3

Comparison of Alternative Surrogate Outcomes: Simulation Designs

3.1 Introduction

Statistical issues relating to $G \times E$ interaction can be complex. Ensuring adequate statistical power to detect $G \times E$ interaction is of major concern in genetic epidemiology studies, especially if the prevalence of the susceptibility genotype and/or of the environmental exposure of interest are low. The review of literature (Chapter 2) indicates that most of the recent research on enhancing power and efficiency of $G \times E$ interaction studies has focussed on optimizing the study design. Yet, the use of alternative outcome measures may be a worthwhile strategy for improving the power to detect $G \times E$ interactions.

This chapter considers various criteria for defining alternative outcome variables. The choice of the "surrogate" outcome X and its relationship with the "clinical" binary outcome Y could have important implications for power. Thus, we consider three hypothetical models of the biological relationship between an ultimate "clinical", more directly relevant binary outcome, and a quantitative "surrogate" outcome. In all investigations, it is assumed that risk of the outcome is affected by a single genetic susceptibility factor G and a single exposure E. This is an oversimplified representation of reality since most diseases, especially those complex in nature, are determined by several genes, gene variants, and exposures, as well as gene-gene and gene-environment

interactions. Nevertheless, this is a good starting point for investigations of more complex situations. The study design is an unmatched cross-sectional or prospective cohort study. In addition, we make the following assumptions for Scenarios I and II described in this chapter:

A1. Genetic susceptibility (G) and exposure (E) are binary.

A2. Prevalence of genetic susceptibility, $P(G = 1) = 0.2$, i.e the gene is relatively common.

A3. Prevalence of exposure, $P(E = 1) = 0.3$, i.e. the exposure is relatively common.

A4. G and E are independent of each other.

Simulation is an indispensable tool in research, especially, in situations where direct experimentation in the relevant real-life context is impossible. The primary objective of this simulation study is to compare efficiency of the quantitative versus binary outcomes in different plausible situations. Thus, while designing the simulation experiments, we have taken into account several aspects of the study design such as frequency of the genetic factor and exposure, sample size, strength of the $G \times E$ interaction effect on the outcome and on the surrogate, and errors in the measurement of both outcome variables. This chapter discusses the postulated models describing the relationships between G, E, X and Y, as well as the methods for data generation and analysis. All programming for data generation and analyses was done by the author in S-Plus 6.0 (R1) and Matlab 7.0 (R14) programming environments.

3.2 Scenario I: Using a risk factor as a surrogate outcome

3.2.1 Postulated model

Scenario I represents the simplest conceptual model in which the quantitative variable X is a risk factor for the binary outcome Y that indicates presence of the disease of interest. Here, probability of disease occurrence is partly affected by the value of X and partly by other risk factors. As Figure 3.1 illustrates, outcome Y depends directly

on the binary genetic factor G, and binary exposure E, and the interaction effect of G and E through some mechanism *not* involving X. Moreover, G, E and $G \times E$ affect the quantitative variable X, which is a risk factor for Y. In other words, factors G, E and $G \times E$ have their influence on the disease through two separate causal pathways: via their effects on X and via their effects on one or more unmeasured intervening variables. As an example, the relation between blood pressure (X) and coronary heart disease (Y) could be described by this model. It is conceivable that G, E and $G \times E$ may increase not only blood pressure, but also other risk factors such as blood glucose level, body weight, cholesterol level, etc., which in turn influence Y. Thus under the formulation described, X is a risk factor of the disease and a mediating factor for the effects of G, E and $G \times E$. Since X is measurable, the epidemiologic investigation of

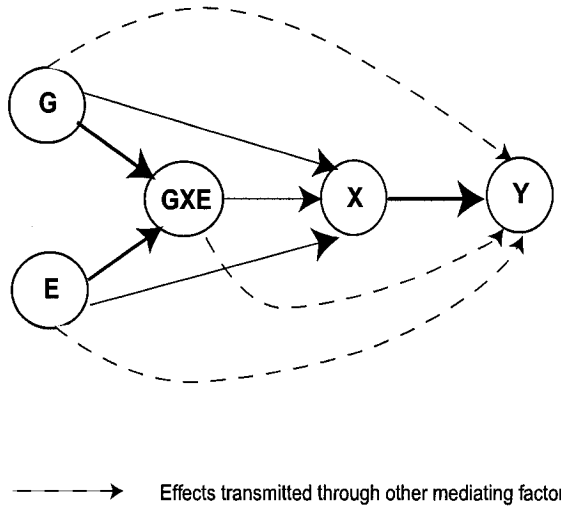

Figure 3.1: Conceptual model for relationship between quantitative surrogate outcome (X) and binary outcome (Y), under Scenario I

this model could be addressed at two levels, namely (i) the etiologic relationships of G and E to X, and (ii) the etiologic relationships of variable X, and factors G, E to the incidence of disease.

We consider four sub-scenarios of this model, obtained by altering assumptions regarding strength of the effect of E and $G \times E$ on Y via the unmeasured intervening risk factors (see Table 3.1). Case 1.1 assumes a moderate effect of E and strong effect of $G \times E$ on Y. Case 1.2 assumes a moderate effect of E and moderate effect of $G \times E$

on Y. Case 1.3 assumes that E has no influence on Y through other mediating factors, but there exists a moderate $G \times E$ interaction effect on Y. Finally, Case 1.4 assumes that the effects of both E and $G \times E$ interaction on Y are mediated entirely through X. In the latter case, it seems reasonable that the question of a possible interaction between G and E can be assessed most directly through their observable effects on variable X. The effect of measurement errors in the quantitative outcome on power to detect $G \times E$ interaction is also assessed.

3.2.2 Basic assumptions

Throughout this thesis, $G = 1$ and $E = 1$ represent, respectively, presence of the genotype and of the exposure of interest. The assumptions outlined in Section 3.1, lead to the following formal conditions used to define Scenario I.

A1. Distribution of continuous surrogate (X) depends on G with higher values occurring when $G = 1$.

A2. X increases in the presence of exposure E, with larger increases for G=1.

A3. Probability that binary outcome Y occurs [i.e. $P(Y = 1)$] depends on X, and also on G, E and $G \times E$ independently of X, i.e. by mechanisms that do *not* involve X.

A4. There are measurements errors in X.

3.2.3 Data generation

Samples of size $N = 500$ are generated. For each $i = 1, ..., 500$ subjects, G, E, X and Y are generated as follows.

B1. Generate G_i and E_i according to assumptions A2 and A3 in Section 3.1.

B2. Generate continuous X_i when $E_i = 0$ depending on value of G_i:

$$X_i \sim \mathcal{N}[0, 1], \quad \text{if } G_i = 0$$
$$X_i \sim \mathcal{N}[0.5, 1], \quad \text{if } G_i = 1$$

B3. Generate continuous X_i when $E_i = 1$ depending on value of G_i:

$$X_i \Longleftarrow X_i + 0.25\delta_i, \text{ if } G = 0$$
$$X_i \Longleftarrow X_i + 0.5\delta_i, \text{ if } G = 1$$

where $\delta_i = \exp(l_i)$ and $l_i \sim \mathcal{N}[0, 1]$.

B4. The true value of X and its observed value X^* are related by an additive error model as $X_i^* = X_i + \varepsilon_i$, where ε_i represents a random measurement error and is generated from $\mathcal{N}[0, 0.3]$.

B5. (i) Calculate

$$L_i = \text{logit}(P(Y_i = 1)) = \beta_0 + \beta_1 G_i + \beta_2 E_i + \beta_3 (G_i \times E_i) + \gamma X_i \quad (3.1)$$

for each of the sub-scenarios in Table 3.1 that determines the logarithm of odds ratio for each effect.

Table 3.1: Parameter values for four sub-scenarios of Scenario I.

Parameter	Case 1.1	Case 1.2	Case 1.3	Case 1.4
β_0	ln(0.1)	ln(0.1)	ln(0.1)	ln(0.1)
β_1	ln(3)	ln(3)	ln(3)	ln(3)
β_2	ln(1.5)	ln(1.5)	0	0
β_3	ln(3)	ln(2)	ln(2)	0
γ	ln(2)	ln(2)	ln(2)	ln(2)

(ii) Calculate $\pi_i = \frac{\exp(L_i)}{1+\exp(L_i)}$, i.e. the expected probability of the binary outcome, conditional on covariates.

(iii) Generate binary Y_i with $P(Y_i = 1) = \pi_i$

3.2.4 Data analysis

For each of the four sub-scenarios, three hundred independent samples are generated using the methodology described in Section 3.2.3. To compare the efficiency of a continuous surrogate outcome with a binary outcome, each simulated sample is analyzed using two types of models: the Multiple Linear Regression model with the continuous

dependent variable X and the Multiple Logistic Regression model with the binary outcome $Y = 1$. Power of the test for the coefficient of $G \times E$ interaction is estimated as the proportion of simulated samples where the test rejects the null hypothesis of no $G \times E$ interaction at 0.05 level of significance.

Linear regression analysis

In matrix terms, the multiple linear regression model is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3.2}$$

where, $\mathbf{X}$ is the $n \times p$ data matrix in which the first column consists of 1s, $\mathbf{Y}$ is an $n \times 1$ vector of responses, $\boldsymbol{\beta}$ is the $(p \times 1)$ vector of regression coefficients, and $\boldsymbol{\varepsilon}$ is a vector of independent normal random variables, representing errors or residuals. Estimation of parameters in the multiple linear regression model is performed by the method of least squares which consists of minimizing the quantity

$$\begin{aligned} Q &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{Y}'\mathbf{Y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \end{aligned}$$

To find the least squares estimate $\mathbf{b}$, Q is differentiated with respect $\boldsymbol{\beta}$:

$$\frac{\partial}{\partial \beta}(Q) = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

Equating to zero vector and substituting $\mathbf{b}$ for $\boldsymbol{\beta}$ gives the least squares normal equation:

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}.$$

We then find the least squares estimators as:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{3.3}$$

To test:

$$\begin{aligned} H_0 &: \beta_k = 0 \text{ against} \\ H_1 &: \beta_k \neq 0 \end{aligned}$$

we use the test statistic

$$t^* = \frac{b_k}{SE(b_k)} \sim t(n-p) \tag{3.4}$$

where p represents the number of regression coefficients in the model, including the intercept. If $\mid t^* \mid \geq t(1-\alpha/2; n-p)$, i.e. the critical value for the student t-distribution with $(n-p)$ degrees of freedom, we reject H_0 at two-tailed α significance level.

For each of the linear models presented below, b_k of primary interest corresponds to the estimated regression coefficient of $G \times E$. The standard error of the estimate and value of t^* are also obtained . We introduce a binary variable g, which takes the value 1 if the test rejects the null hypothesis of no $G \times E$ interaction at 0.05 level of significance. Thus, for a large sample size, $g = 1$ if $\mid t^* \mid \geq 1.96$, else $g = 0$. An estimate of power of the test is given by the proportion of samples that have $g = 1$.

Linear regression analysis is performed for each sample using the quantitative variable X as outcome. Two sets of X data are employed in the analysis: true X and X with measurement errors (i.e. X^*).

MODEL SI.1(i):

$$X = \alpha_0 + \alpha_1 G + \alpha_2 E + \alpha_3 (G \times E) + \varepsilon \tag{3.5}$$

MODEL SI.1(ii):

$$X^* = \beta_0 + \beta_1 G + \alpha_2 E + \beta_3 (G \times E) + \varepsilon \tag{3.6}$$

Logistic regression analysis

Logistic regression is used to model the relationship between a binary response variable and one or more predictor variables, which may be either discrete or continuous. Binary outcome data are common in medical applications. For example, the binary response variable might be whether or not a patient is alive five years after treatment for cancer or whether the patient has an adverse reaction to a new drug. The multiple logistic regression model is given by:

$$\boldsymbol{\pi} = \frac{exp(\mathbf{X}\boldsymbol{\beta})}{1 + exp(\mathbf{X}\boldsymbol{\beta})} \tag{3.7}$$

where

$$\pi_i = E(Y_i \mid \mathbf{X}_i) = P(Y_i = 1 \mid \mathbf{X}_i) \tag{3.8}$$

is the conditional probability that the outcome is present for the ith individual. The logit transform of $\pi(\mathbf{X}_i)$ leads to the linear predictor:

$$\text{logit}(\pi_i) = \mathbf{X}_i\boldsymbol{\beta} \tag{3.9}$$

The method of maximum likelihood is used to estimate the parameters of the logistic regression model. The principle of maximum likelihood states that we use as our estimate of $\boldsymbol{\beta}$ the value which maximizes the likelihood function:

$$L(\boldsymbol{\beta}) = \prod_i \pi_i^{y_i}[1-\pi_i]^{1-y_i} \tag{3.10}$$

Differentiating $\ln L(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ and setting the resulting expressions equal to zero gives the likelihood equations:

$$\mathbf{X}'\mathbf{Y} = \mathbf{X}'\hat{\boldsymbol{\pi}} \tag{3.11}$$

Numerical search procedures such as the Newton Raphson method are used to solve (3.11) to find the maximum likelihood estimates $\mathbf{b}$.

To test:

$$\begin{aligned} H_0 &: \beta_k = 0 \text{ against} \\ H_1 &: \beta_k \neq 0, \end{aligned}$$

an appropriate large-sample test statistic is:

$$W = \frac{b_k}{SE(b_k)} \sim \mathcal{N}[0,1]. \tag{3.12}$$

We reject H_0 at the 0.05 significance level if the two-sided P-value, $P(\mid Z \mid > \mid w \mid) \leq 0.05$, where Z is the standard normal variate, and w is the observed value of W.

In this thesis, the above test, known as the Wald test, is used to test for the significance of the $G \times E$ interaction parameter in the logistic regression model. The binary variable, g, takes the value 1 if the Wald test rejects the null hypothesis of no $G \times E$ interaction at 0.05 level of significance. Thus, $g = 1$ if P-value for Wald test ≤ 0.05, else $g = 0$. The proportion of samples for which g equals 1 gives an estimate of the power.

Each sample has four sets of outcome data Y corresponding to the four sub-scenarios or cases described in B5 (see Table 3.1). For each Y, the following model was fit to the sample data:

MODEL SI.2

$$\text{logit}(P(Y=1)) = \theta_0 + \theta_1 G + \theta_2 E + \theta_3 (G \times E) \tag{3.13}$$

Four estimated models are obtained for each Y and will be referred to as models SI.2(i), SI.2(ii), SI.2(iii) and SI.2(iv) respectively.

Sensitivity analysis

We investigated the effect of varying sample size (N) on the power of the tests to detect $G \times E$ interaction based on both linear and logistic regression models described in Section 3.2.4. Three hundred samples were generated and the sample size for each simulation run was varied from 500 to 1500 with increments of 200.

3.3 Scenario II: Using a marker of early disease as a surrogate outcome

A patient may have the pathology and etiology of a disease without presenting signs and symptoms. This is referred to as "silent", "latent", or "subclinical" disease [82]. People who already have the "pathological changes" that lead to a disease often experience elevated levels of some metabolic product or quantitative lab test that can serve as markers of increased disease risk. For instance, elevation of serum creatinine is a strong marker of increased vascular risk [3]. The urinary protein known as albumin is being recognized as the earliest sign of vascular damage in both the kidney and the heart [69]. Again, markers in serum are known to provide a window on the inflammatory status of an individual, and thus insight into the pathophysiology of atherosclerosis and its complications [71].

In this section, the hypothetical model assumes that the quantitative variable X is a marker of the presence of disease rather than a risk factor. It is important to consider the temporal sequence of the observed changes in X and Y. In the above discussion, it was assumed that when the patient's health status changes from healthy to subclinical latent pathology, X starts to change or starts to increase at a higher rate. Such changes in X may precede any observable change in disease status Y, thereby, indicating *future* risk of disease. However, in this study, we assume the absence of a

latent or subclinical period of the disease so that change of disease status (from $Y = 0$ to $Y = 1$) is diagnosed immediately. Thus, a change in Y occurs concurrently to certain pathological changes that lead to elevated levels of X. This means that X changes only when Y changes from 0 to 1 so that it may serve as a surrogate marker for the disease. For example, levels of the protein troponin in the blood can be used to determine if an individual has had a heart attack [56]. Similarly, the most frequent feature of acute or chronic viral hepatitis involves the elevation of serum alanine aminotransferase activity (ALT) and aspartate aminotransferase activity (AST) above the range of normal values [18].

This model might appear unrealistic for most applications since it assumes the absence of a latent period. Nevertheless, it is worth considering as a benchmark for future comparisons. Two variations of Scenario II are considered:

S1. Disease is associated with a higher *value* of X at a particular time point (as in cross-sectional studies), and

S2. Disease is associated with a higher *rate of increase* in X (over and above "natural aging") which would require measurements at two (or more) time points (cohort studies).

3.3.1 Postulated model for Case S1

Scenario I (Section 3.2) considered a simplistic situation in which X, G and E were the only risk factors for Y. However, most diseases are multifactorial and determined by a myriad of factors that have complex interrelationships. An important shortcoming of Scenario I is that it makes no provision for other risk factors, either measurable or nonmeasurable, that could have an effect on Y or that could be correlated with X or G and/or E. The assumptions in Scenario I are, therefore, restrictive, which limits applicability of its results.

Figure 3.2 presents a new causal model, consistent with Scenario II, that describes how G, E and interaction between these two factors act in the pathogenic process leading to disease. Contrary to Scenario I, here Y does affect X, i.e. X is a marker for disease, rather than a risk factor. Moreover, here G, E and $G \times E$ interaction

do *not* affect X directly, but only through their impact on the risk of Y. The model assumes that disease outcome Y is influenced by other measurable risk factors that are independent of G and E. Here, R is an aggregate "overall risk", which can be, for example, calculated from a multivariate risk score [55] representing the joint effect of these risk factors. As Figure 3.2 illustrates, level of marker X depends on the set of risk factors in R, as well as on the disease status Y.

Several issues here could have implications for power. First, is the choice of the quantitative outcome X instead of the more relevant binary outcome Y. Second, is the magnitude of misclassification errors in outcome Y versus measurement errors in marker X. We take these factors into account while evaluating the gain, in terms of power, of replacing a binary outcome by a quantitative surrogate. The efficiency of the linear versus logistic model is also evaluated for different magnitudes of the true E and $G \times E$ interaction effect on Y.

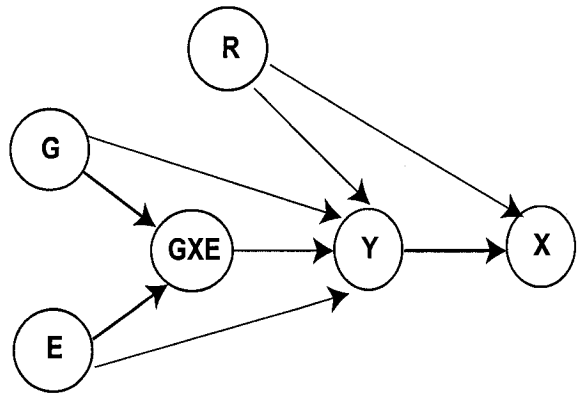

Figure 3.2: Conceptual model for relationships between G, E, $G \times E$, other factors (R) independent of G and E, quantitative surrogate outcome (X) and binary outcome (Y), under Scenario II (Case S1)

3.3.2 Postulated model for Case S2

The causal model described in Case S1 may not provide adequately for the complexity of pathogenic processes because it does account for unmeasured or *unobserved variables* that influence Y and that *may be confounders of G and/or E*. In Case S2 we

overcome this shortcoming by assuming that disease status depends additionally on some unobserved risk factors that may be grouped into three categories: risk factors correlated with neither G nor E, risk factors correlated with E only, and risk factors correlated with both G and E (see Figure 3.3). S_0, S_E and S_{GE} are multivariate risk scores representing aggregate "overall risks" for the risk factors in each group respectively.

Important goals of surrogate markers are to assess the stage or severity of disease and to determine rate of change. Cross-sectional studies may be used to measure the level of surrogate at a particular time point, which in turn reflects stage or severity of disease. For instance, cross-sectional data reveal that people with widely metastatic cancer have higher levels of angiogenesis activity compared to early stage disease or normals [11]. In cohort studies, change in surrogate marker levels reflect change in disease status or progression of disease. For instance, cohort data for the same patients show that over time the levels of angiogenesis activity rise as disease progresses.

Scenario S1 assumed a cross-sectional study in which disease was assessed at a single time point. A conceptual difference between Scenarios S1 and S2 is that in the latter, X is considered a "baseline" value at the initial assessment of disease. A second measurement, taken at a later time point, allows the *rate of change* in X to be determined. Since increase of X at a higher rate may reflect the presence of the disease of interest, it may be used as a surrogate marker for the disease. Thus, three competing outcomes emerge from scenario S2: (i) the (more relevant) binary variable Y, (ii) initial value of quantitative variable X, and (iii) the rate of change in X. Apart from the choice of outcome, misclassification errors in the diagnosis of Y and measurement errors in X, may both have important implications for power, and will be considered under this scenario. As in Scenario I and Case S1, it is of interest to compare power between the linear (with X as dependent variable) versus logistic (with $Y = 1$ as outcome) regression models while varying the true underlying effects of E and $G \times E$ on Y.

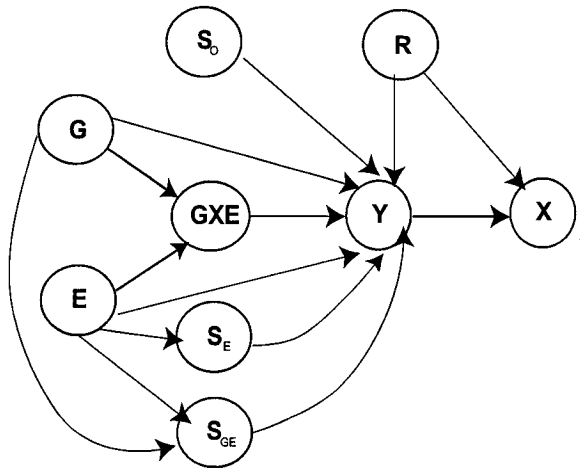

Figure 3.3: Conceptual model for relationships between G, E, $G \times E$, other observed factors (R) independent of G and E, other unobserved factors (S) that may be correlated with G and E, the quantitative surrogate outcome (X), and binary outcome (Y), under Scenario II (Case S2)

3.3.3 Basic assumptions

Important assumptions for Case S1 and Case S2, added to the "generic" assumptions listed above (Section 3.1) may be outlined as follows.

Case S1

A1. The risk of developing the disease ($Y = 1$) depends on G, E, $G \times E$ and the joint effect of other risk factors, that are independent of G and E.

A2. Marker X depends on disease status with higher values occurring when $Y = 1$.

A3. There is some misclassification of disease Y, and some errors in measurement of X.

Case S2

In addition to the assumptions for Case S1, the following assumptions are made:

A1. Disease status depends on *unobserved* risk factors (S) that are grouped into three groups:

- S_0: correlated with neither G nor E

- S_E: correlated with E only
- S_{GE} correlated with both E and G

A2. Marker X is assessed at two time points so that we have two measurements X_0 and X_1.

A3. G, R and S influence X_0 and X_1 equally so that increase in X during time period $\Delta t = t_1 - t_0$ depends on the length of time interval, exposure level at time t_0 and $G \times E$ interaction only.

A4. There are measurement errors in X at both time points, and the two errors for the same subject are independent, i.e. not correlated.

3.3.4 Data generation

For both sub-scenarios of Scenario II, samples of size $N = 500$ are generated.

Case S1

For each $i = 1, ..., 500$ subjects, G, E, Y and X are generated as follows.

B1. Generate G_i and E_i according to assumptions A2 and A3 in Section 3.1.

B2. Generate R_i from $\mathcal{N}[0, 1]$

B3. Generate true disease status Y as follows:

(i) Calculate

$$L_i = \text{logit}(P(Y_i = 1)) = \beta_0 + \beta_1 G_i + \beta_2 E_i + \beta_3 (G_i \times E_i) + \beta_4 R_i$$

for each of the scenarios in Table 3.2. Using Case 2.1 as the reference, Case 2.2 indicates a slightly weaker effect of $G \times E$ on Y. Case 2.3 represents the scenario where the $G \times E$ effect is the same as in Case 2.2, but there is no exposure effect on Y. Case 2.4 assumes no $G \times E$ or exposure effect on Y and, thus, is used to assess the empirical size of the test i.e. type-I error rate.

Table 3.2: Combination of values for parameter β under Scenario II, Case S1.

Parameter	Case 2.1	Case 2.2	Case 2.3	Case 2.4
β_0	ln(0.1)	ln(0.1)	ln(0.1)	ln(0.1)
β_1	ln(3)	ln(3)	ln(3)	ln(3)
β_2	ln(1.5)	ln(1.5)	0	0
β_3	ln(3)	ln(2)	ln(2)	0
β_4	ln(1.5)	ln(1.5)	ln(1.5)	ln(1.5)

(ii) Calculate $\pi_i = \frac{\exp(L_i)}{1+\exp(L_i)}$

(iii) Generate binary Y_i with $P(Y_i = 1) = \pi_i$

B4. Generate observed disease status Y^* from Y with sensitivity $P(Y^* = 1|Y = 1) = \eta_1$ and specificity $P(Y^* = 0|Y = 0) = \eta_0$ as follows:

(i) For $Y_i = 1$, generate b_i from $bin(1, \eta_1)$. Then

$$Y_i^* = \begin{cases} 1 & \text{if } b_i = 1 \\ 0 & \text{if } b_i = 0 \end{cases}$$

(ii) For $Y_i = 0$, generate b_i from $bin(1, \eta_0)$. Then

$$Y_i^* = \begin{cases} 0 & \text{if } b_i = 1 \\ 1 & \text{if } b_i = 0 \end{cases}$$

Five combinations of values for η_1 and η_0 are considered as shown in Table 3.3.

Table 3.3: Combination of values for sensitivity (η_1) and specificity (η_0) of the observed disease status.

	I	II	III	IV	V
η_1	1.0	0.9	0.7	0.9	0.7
η_0	1.0	0.9	0.9	0.7	0.7

B5. Generate X, conditional on Y and R, as follows:

(i) If $Y_i = 0$, $X_i \sim \mathcal{N}[\mu_{0,R}, 1]$ where $\mu_{0,R} = \theta_R \cdot R_i$

(ii) If $Y_i = 1$, $X_i \sim \mathcal{N}[\mu_{1,R}, 1]$ where $\mu_{1,R} = \theta_Y + \theta_R \cdot R_i$.

According to this formulation, X depends on R such that, on average, X increases as R increases. Larger values of X occur on average, when $Y = 1$. Six combinations of values for θ_R and θ_Y are considered in Table 3.4. Higher value of θ_Y induces a larger increment in X when $Y = 1$, i.e. a higher diagnostic value of X as a marker for $Y = 1$. Positive values of θ_R result in a positive correlation between X and R. It should be noticed that X depends on the "true" error-free status of Y, rather than on the observed Y^*.

Table 3.4: Combination of values for parameters θ_R and θ_Y.

	I	II	III	IV	V	VI
θ_R	0	0	0	0.5	0.5	0.5
θ_Y	0.5	1	1.5	0.5	1	1.5

B6. Generate observed X_i^* from the additive error model $X_i^* = X_i + \varepsilon_i$ where $\varepsilon_i \sim \mathcal{N}[0, \sigma]$. Three values for σ are considered: $\sigma = 0.05, 0.5, 1$, corresponding to the situations of almost no error, moderate error and large error, respectively. For better interpretability, one may compute the intraclass correlation (ICC) [26] for the reliability of observed X^* by:

$$ICC = \frac{\sigma_X^2}{\sigma_X^2 + \sigma^2} \tag{3.14}$$

where σ_X represents the standard deviation of the distribution of true X given in B5 of this section. Thus for $\sigma_X = 1$, ICC for $\sigma = 0.05, 0.5$ and 1 equals 0.998, 0.8 and 0.5 respectively.

Case S2

C1. G_i, E_i and R_i are generated as described in B1 and B2 of Section 3.3.4.

C2. Generate:

1. $S_{0i} \sim \mathcal{N}[0, 1]$
2. $S_{Ei} \sim \mathcal{N}[-0.3, 1]$, if $E_i = 0$; $S_{Ei} \sim \mathcal{N}[0.3, 1]$, if $E_i = 1$
3. $S_{GEi} \sim \mathcal{N}[\mu_{Si}, 1]$, where $\mu_{Si} = -0.4 + 0.3G_i + 0.5E_i$

C3. Generate true disease status Y as follows:

(i) Calculate

$$L_i = \text{logit}(P(Y_i = 1)) = \beta_0 + \beta_1 G_i + \beta_2 E_i + \beta_3 (G_i \times E_i) + \beta_4 R_i + \beta_5 S_{0i} + \beta_6 S_{Ei} + \beta_7 S_{GEi}$$

for each combination of parameter values shown in Table 3.5.

Table 3.5: Combinations of values for parameter $\boldsymbol{\beta}$ under Scenario II, Case S2.

Parameter	Case 2.1	Case 2.2	Case 2.3	Case 2.4
β_0	ln(0.1)	ln(0.1)	ln(0.1)	ln(0.1)
β_1	ln(3)	ln(3)	ln(3)	ln(3)
β_2	ln(1.5)	ln(1.5)	0	0
β_3	ln(3)	ln(2)	ln(2)	0
β_4	ln(1.5)	ln(1.5)	ln(1.5)	ln(1.5)
β_5	ln(1.3)	ln(1.3)	ln(1.3)	ln(1.3)
β_6	ln(1.5)	ln(1.5)	ln(1.5)	ln(1.5)
β_7	ln(1.75)	ln(1.75)	ln(1.75)	ln(1.75)

(ii) Calculate $\pi_i = \frac{\exp(L_i)}{1+\exp(L_i)}$

(iii) Generate binary Y_i with $P(Y_i = 1) = \pi_i$.

C4. Generate observed disease status Y^* from Y as described in B4 of Section 3.3.4.

C5. Generate true X_0, i.e. the initial (baseline) value of X, conditional on Y and R as described in B5 of this section.

C6. Generate observed X_0^*, based on X_0 generated in C5, using same method as described in B6 of this section.

C7. Generate time interval between the two measurements of X: $\Delta t_i \sim \mathcal{U}(0.5, 1.5)$.

C8. Compute increase in true value of X by:

$$\Delta X_i = [\gamma_0 + \gamma_1 E_i + \gamma_2 (E_i \times G_i)] \Delta t_i \tag{3.15}$$

where $\gamma_0 = 0.2$, $\gamma_1 = 0.1$, $\gamma_2 = 0.15$. This implies that ΔX_i depends only on exposure, the impact of which varies depending on presence or absence of the

gene: subjects with the gene have stronger reaction to exposure. The rate of increase in X for the unexposed is 0.2. The rate increases to 0.3 for those exposed but without the gene. The rate increases by an additional 50%, to 0.45 for those exposed and with the gene.

C9. Compute "true" value of X at second assessment time by:

$$X_{1i} = X_{0i} + \Delta X_i \tag{3.16}$$

C10. Observed X_1 (i.e. X_1^*) is generated using same approach as for generation of X_0^*:

$$X_{1i}^* = X_{1i} + \varepsilon_{1i}$$

where $\varepsilon_{1i} \sim \mathcal{N}[0, \sigma]$ and $\sigma = 0.05, 0.50, 1.0$. It should be noticed that ε_{1i} for X_1 is generated independently of ε_i for X_0, since it is possible that e.g. measurement error causes X_0^* to be overestimated and X_1^* to be underestimated, relative to their "true" values, for subject i.

3.3.5 Data analysis

As in Scenario I, three hundred independent random samples are generated for both Case S1 and Case S2.

Case S1:

For each combination of the parameter values given in Table 3.2, eighteen sets of values for X^* are generated according to combinations of values of θ_R, θ_Y and σ (see Section 3.3.4, Case S1). Thus, for each subject, we consider a total of 72 different values (4 different Y values times 18) for the continuous outcome X^*. The following multiple linear regression model is fit to the data for different X^*:

MODEL SII.1(i)

$$X^* = \alpha_0 + \alpha_1 G + \alpha_2 E + \alpha_3 (G \times E) + \alpha_4 R + \varepsilon \tag{3.17}$$

Twenty sets of values for Y^* are generated for combinations of values of the sensitivity (η_1) and specificity (η_0) parameters given in Table 3.3, and the parameters in Table 3.2.

The following multiple logistic regression is used to analyze each sample for different Y^*:

MODEL SII.1(ii)

$$\text{logit}(P(Y^* = 1)) = \nu_0 + \nu_1 G + \nu_2 E + \nu_3 (G \times E) + \nu_4 R \tag{3.18}$$

Case S2:

For Case S2, we consider three competing outcomes: initial assessment of quantitative marker X (i.e. X_0^*), binary outcome (observed) Y, and the time interval standardized difference in X, $Z = \frac{X_1^* - X_0^*}{\Delta t}$, which provides a crude estimate of subject-specific rate of change in X. Seventy two sets of values for X_0^* are generated for combinations of values of θ_R, θ_Y (see Table 3.4), σ and the β parameters in Table 3.5. Using the initial values of the marker as outcome, data sets are analyzed using the following multiple linear regression for different X_0^*:

MODEL SII.2(i)

$$X_0^* = \delta_0 + \delta_1 G + \delta_2 E + \delta_3 (G \times E) + \delta_4 R + \varepsilon \tag{3.19}$$

The values of X at the two assessment times generate 72 sets of standardized differences Z for the different combinations of parameter values. When rate of change in the marker is used as outcome, each sample is analyzed using:

MODEL SII.2(ii)

$$Z = \gamma_0 + \gamma_1 G + \gamma_2 E + \gamma_3 (G \times E) + \gamma_4 R + \varepsilon \tag{3.20}$$

We expect that when disease is associated with higher rate of increase in X, modelling Z as the outcome may provide greater power than using X at a single time point as the outcome. Comparison of estimated powers for models SII.2(i) and SII.2(ii) will provide some insights into this problem.

For each combination of parameter values in Table 3.5, five sets of values for Y^* are generated for the combinations of η_1 and η_0 given in Table 3.3. This results in 20 different values for Y^* in each sample. For each Y^*, the following multiple logistic model is fit to the data:

MODEL SII.2(iii)

$$\text{logit}(P(Y^* = 1)) = \lambda_0 + \lambda_1 G + \lambda_2 E + \lambda(G \times E) + \lambda_4 R \qquad (3.21)$$

For each of the models described in this section, power of the test for the coefficient of $G \times E$ interaction is estimated by the proportion of simulated samples where the appropriate test (t-test or Wald test) rejected the null hypothesis of no $G \times E$ interaction at 0.05 level of significance.

3.4 Scenario III: Repeated measures of a marker

There are alternative ways a quantitative variable may be measured and analyzed. Many studies measure a continuous covariate repeatedly over time. In some cases, this is because researchers wish to investigate the time course of a symptom or to evaluate how the effect of a treatment changes over time. Measures may also be repeated in order to obtain a more precise estimate of the characteristic of interest. Repeat assessment reduces intra-subject variability and, thus, increases study power and efficiency [115].

In Scenario III, we consider a repeated measures design, in which data on a quantitative variable of interest X, say blood pressure or value of some quantitative lab test, is collected at different points in time. For example, if an individual's blood pressure is recorded during each visit to the hospital, the resulting data is repeated measures data. Furthermore, we assume that the elevated level of the quantitative variable X may reflect the presence of disease. In other words, X is a marker of the presence of disease. For example, high levels of systolic blood pressure could indicate the presence of hypertension. This scenario extends the assumption of two assessment times of X, considered in Scenario II, to two or more assessment times but makes no explicit assumptions about the mechanism of the G, E or $G \times E$ interaction effect on X.

In some studies, the binary outcome may be defined based on categorization of the quantitative variable. For instance, hypertension is defined as blood pressure above a certain cut-off point. Whether or not an individual is categorized as being hypertensive depends on the selected threshold for blood pressure and method of dichotomizing the quantitative marker. The choice of an appropriate threshold or

method of dichotomization is typically guided by clinical relevance and other more practical considerations, often related to measurement. In this scenario, we consider two threshold values and alternative methods for classifying diseased and non-diseased individuals on the basis of repeated measures of the disease marker. We compare power for detecting $G \times E$ interaction between linear versus logistic regression models for a number of competing outcomes, all defined based on the quantitative X variable.

3.4.1 Basic assumptions

In addition to assumptions A1 and A4 in Section 3.1, we make the following assumptions pertaining to a hypothetical scenario in which X represents blood pressure and Y indicates presence or absence of hypertension.

A1. Prevalence of genetic susceptibility, $P(G = 1) = 0.3$.

A2. Lifetime prevalence of exposure, $P(E = 1) = 0.6$.

A3. Distribution of X values at birth may depend on G, with higher values occurring when $G = 1$:

$$\begin{aligned} X_{i0} &\sim \mathcal{N}[60, 10], && \text{if } G_i = 0 \\ X_{i0} &\sim \mathcal{N}[60 + \Delta_X, 10], && \text{if } G_i = 1 \end{aligned} \tag{3.22}$$

where two cases are considered: $\Delta_X = 0$ or 5.

A4. Distribution of current age of ith subject, $t_{i,0}$, is independent of any covariates.

A5. Subjects who are "exposed", become first exposed only at a certain age, S, which varies across subjects. Once exposed, subjects remain exposed for the rest of their life, which is consistent with the assumption that exposure has a permanent impact as is the case in ecological disasters [90].

A6. X increases linearly with age, and the rate (or slope) of age-dependent increase varies across subjects and may depend on G (see Figure 3.4), so that X at age $t_{i,0}$ equals:

$$X_i(t_{i,0}) = X_{i,0} + \beta_i t_{i,0} \tag{3.23}$$

where,

$$\begin{aligned} \beta_i &\sim \mathcal{N}[0.2, 0.05], && \text{if } G_i = 0 \\ \beta_i &\sim \mathcal{N}[0.2 + \Delta_G, 0.05], && \text{if } G_i = 1. \end{aligned} \tag{3.24}$$

Here two cases are considered: $\Delta_G = 0$ or 0.10. The difference in slopes for the two genetic groups as seen in Figure 3.4, prior to exposure, induces an interaction between G and age.

A7. As Figure 3.4 illustrates, the rate of change in X (i.e. β') may further increase due to exposure and/or $G \times E$ interaction. However, E and $G \times E$ will affect X only after the subject becomes exposed, that is after age S_i. Thus, if $E_i = 1$, which implies $t_{i,0} > S_i$, we get:

$$X_i(t_{i,0}) = X_i(S_i) + \beta_i'(t_{i,0} - S_i) \tag{3.25}$$

where, S_i is age at first exposure and the post-exposure slope $\beta_i' = \beta_i + \Delta_E E_i + \Delta_{GE} G_i \times E_i$ for the ith subject. Here, the following values of the relevant parameters are considered: $\Delta_E = 0.03$ and $\Delta_{GE} = 0, 0.20$ or 0.40. Thus, for a subject with $E = 1$ and $G = 0$, $\beta_i' = \beta_i + 0.03 + 0$, which implies an increase in the rate of change of X_i with exposure. However, for a subject with $E = 1$ and $G = 1$, $\beta_i' = \beta_i + 0.03 + \Delta_{GE}$, so that there is an even greater increase in the rate of change of X_i with exposure.

A8. There is error in measurement of X.

3.4.2 Data generation

This section describes the steps used to generate the data [G, E, repeated values of X and Y]. Samples of size $N = 500$ are generated. For the ith subject ($i = 1, ..., 500$), we generate several times (or ages) $t_{i,-j}$, $j = 0, ..., j_i^*$ at which X was assessed in the last six years before the "current age" $t_{i,0}$.

B1. Generate G_i and E_i according to assumptions A1 and A2 in Section 3.4.1.

B2. Generate current age $t_{i,0}$ from $\mathcal{U}[40, 60]$

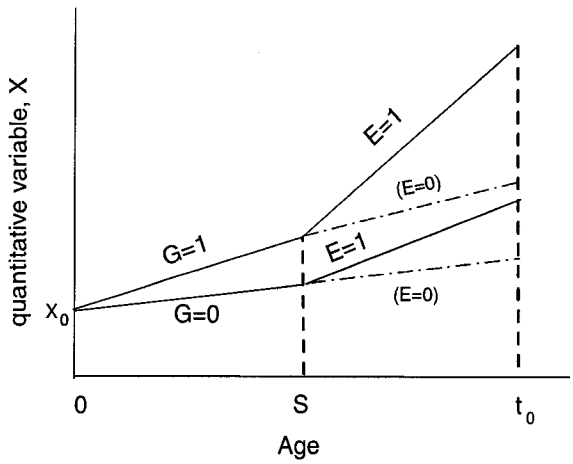

Figure 3.4: Effect of G and/or E on rate of change of X assuming no effect of G on X at birth (i.e. $\Delta_X = 0$).

B3. Generate age at first exposure S by:

$$\begin{aligned} S_i &\sim \mathcal{U}[20, 40], \text{ if } E_i = 1 \\ S_i &= 0, \qquad\quad \text{if } E_i = 0. \end{aligned} \tag{3.26}$$

B4. Subjects are considered "exposed" only at ages higher than age S. Thus, $E_{i,-j} = 1$ only if (i) $E_i = 1$ and (ii) $t_{i,-j} \geq S_i$. If $S_{i,-j} > t_{i,-j}$, where $S_{i,-j} = S_i$, change $E_{i,-j}$ and $S_{i,-j}$ to 0.

B5. Compute $X_i(t_{i,0})$ according to value of E using A3 and A6 or A7 in Section 3.4.1.

B6. For the ith subject, generate time interval between successive measurements, $dt_{i,-j}$, from $\mathcal{U}[0.5, 1.5]$

B7. Calculate age at which the previous measurement was taken by $t_{i,-1} = t_{i,0} - dt_{i,-1}$

B8. Generate random variable $U_{i,-1}$ from $bin(1, 0.3)$. If $U_{i,-1} = 1$, $X_i(t_{i,-1})$ is missing. Otherwise generate $X_i(t_{i,-1})$ as follows:

(i) Generate baseline value $X_{i,0}$ as described in A3.

(ii) If $E_{i,-1} = 0$, generate $X_i(t_{i,-1})$ as described in A6. If $E_{i,-1} = 1$, compute $X_i(S_{i,-1})$ from (3.23) and using this value, compute $X_i(t_{i,-1})$ as described in A7.

B9. Regardless of whether $t_{i,-1}$ is missing or not, carry out the following steps:

(i) Generate $dt_{i,-2}$ from $\mathcal{U}[0.5, 1.5]$ (as in B6)

(ii) Calculate $t_{i,-2} = t_{i,-1} - dt_{i,-2}$

(iii) Generate $U_{i,-2}$ from $bin(1, 0.3)$ to decide if $t_{i,-2}$ is missing or not.

(iv) If $t_{i,-2}$ is not missing, calculate $X_i(t_{i,-2})$ as described in B8.

B10. Repeat steps (i)-(iv) in B9 iteratively until $t_{i,-j} < t_{i,0} - 6$. Ignore this time point and pass to next subject $(i + 1)$. This induces "left censoring" reflecting the assumption that measurements of X are available for only past six years.

B11. For $j = 0, ..., j_i^*$, generate observed $X^*_{i,-j}$ for the ith subject from the additive error model

$$X_i^*(t_{i,-j}) = X_i(t_{i,-j}) + \varepsilon_{i,j} \tag{3.27}$$

where, $\varepsilon_{i,j} \sim \mathcal{N}[0, 0.1]$.

B12. We construct eight different outcomes based on observed X as follows:

1. Simple continuous outcome: $X_i^*(t_{i,0})$, i.e. "current value" of X.
2. Two alternative versions of simple dichotomous outcome $Y1_i^*$, based on $X_i^*(t_{i,0})$, each using a different threshold:
   (i) For each sample, find two threshold values T_1 and T_2. T_1 is chosen to be the 70th percentile of the empirical distribution of $X_i^*(t_{i,0})$ and T_2 is chosen as the 50th percentile of $X_i^*(t_{i,0})$. According to this definition, $T1 > T2$.
   (ii) Then, define two alternative binary outcomes, using either $T1$ or $T2$ as the cut-off point:

$$\begin{aligned} Y11_i^* &= 1 \text{ iff } X_i^*(t_{i,0}) > T1, \ Y11_i^* = 0 \text{ otherwise;} \\ Y12_i^* &= 1 \text{ iff } X_i^*(t_{i,0}) > T2, \ Y12_i^* = 0 \text{ otherwise.} \end{aligned}$$

3. Two alternative versions of dichotomous outcome $Y2_i$ based on at least two values above T (T_1 or T_2) in the last 6 years:
   Use all non-missing values of X^* (except at current age) available for a given subject i.
   (i) If $X_i^*(t_{i,-j}) > T1$, then $\delta_{i,j} = 1$ for $j = 1, ..., j_i^*$.
   (ii) Compute $V_i = \sum_{j=1}^{j_i^*} \delta_{i,j}$.
   (iii) $Y21_i^* = 1$, iff $V_i \geq 2$; $Y21_i^* = 0$, otherwise.
   (iv) If $T1$ is replaced by $T2$ in (i), then,
   $Y22_i^* = 1$, iff $V_i \geq 2$; $Y22_i^* = 0$, otherwise.

4. Two alternative versions of dichotomous outcomes $Y3$ and $Y4$ based on new development of "disease" (i.e. value of X above the threshold) in last 6 years:
   Here, we restrict the analysis to subjects with $X_i^*(t_{i,-j_i^*}) \leq T$ where $T = T_1$ or T_2, i.e. to subjects who were below the cut-off point at the earliest time for which X measurement is available.
   (a) (i) For subjects with $X_i^*(t_{i,-j_i^*}) \leq T1$,

   $$Y31_i^* = 1 \text{ , iff } X_i^*(t_{i,0}) > T1; \; Y31_i^* = 0, \text{ otherwise.}$$

   (ii) Similarly, for subjects with $X_i^*(t_{i,-j_i^*}) \leq T2$,

   $$Y32_i^* = 1, \text{ iff } X_i^*(t_{i,0}) > T2; \; Y32_i^* = 0, \text{ otherwise.}$$

   (b) (i) For subjects with $X_i^*(t_{i,-j_i^*}) \leq T1$, $\delta_{i,j} = 1$ if $X_i^*(t_{i,-j}) > T1$ for $j = 0, 1, ..., j_i^* - 1$
   (ii) Compute $V_i = \sum_{j=0}^{j_i^*-1} \delta_{i,j}$.
   (iii) $Y41_i^* = 1$, iff $V_i \geq 2$; $Y41_i^* = 0$, otherwise.
   (iv) If $T1$ is replaced by $T2$ in (b)(i), then $Y42_i^* = 1$, iff $V_i \geq 2$; $Y42_i^* = 0$, otherwise.

5. Continuous time-interval standardized difference (increase) in X_i^*:
   The standardized difference ΔX_i^* is calculated as,

$$\Delta X_i^* = \frac{X_i^*(t_{i,0}) - X_i^*(t_{i,-j_i^*})}{t_{i,0} - t_{i,-j_i^*}} \tag{3.28}$$

where, ΔX_i^* may be interpreted as a simple measure of average increase in X per year.

6. Rate of progression in X, $\hat{\alpha_{1i}}$:
   Regress repeated measures of X for the ith subject on age at which measurements were recorded:

$$X_i^*(t_{i,-j}) = \alpha_{0i}^* + \alpha_{1i}^* t_{i,-j}, \qquad j = 0, 1, \ldots, j_i^* \tag{3.29}$$

   The slope $\hat{\alpha_{1i}^*}$ is used as the outcome and measures the rate of progression in X for the ith individual. Note that (3.29) is similar to (3.23) in Section 3.4.1, which was used to generate X at current age in absence of exposure. Therefore, β_i is analogous to α_{1i}^*.

3.4.3 Data analysis

Three hundred independent random samples are generated for each combination of relevant parameters. Using different combinations of Δ_X, Δ_G and Δ_{GE}, 12 different scenarios are considered, each yielding different values of the continuous outcomes analyzed in models SIII(i), SIII(vi) and SIII(vii) below. In addition, two different thresholds T ($T1$ or $T2$) were used, resulting in 24 different scenarios for the outcomes analyzed in models SIII(ii)-SIII(v) below.

Thus, each simulated sample was analyzed by seven different models. All models are adjusted for the subject's age. Although the notations for X and Y in the models shown below indicate values with measurement errors (i.e. observed values), it is important to note that the same models are used to analyze true values of the generated outcomes as well.

For each generated sample, the estimated parameter of $G \times E$ interaction, its standard error, and value of the t-statistic (for linear regression) or Wald statistic (for logistic regression) are obtained as described in Section 3.2.4. The value of the binary indicator g for each sample (see Section 3.2.4) is used to estimate power of the test for $G \times E$ interaction as the proportion of simulated samples where the two-tailed test rejected the null hypothesis of no $G \times E$ interaction at 0.05 level of significance.

MODEL SIII(i):

$$X^*(t_0) = \beta_0 + \beta_1 G + \beta_2 E + \beta_3 (G \times E) + \beta_4 t_0 + \varepsilon \tag{3.30}$$

Here, $X^*(t_0)$ represents current value of the quantitative marker, such as current blood pressure.

MODEL SIII(ii):

$$\text{logit}(P(Y1^* = 1)) = \nu_0 + \nu_1 G + \nu_2 E + \nu_3 (G \times E) + \nu_4 t_0 \tag{3.31}$$

This model incorporates the same adjustments as model SIII(i). Here, $Y1$ is a binary outcome representing the state of disease which is characterized by the current value of X being above the threshold T. For instance, $Y1 = 1$ indicates presence of hypertension if current blood pressure exceeds T. Power comparisons between models SIII(i) and SIII(ii) could provide an assessment for the relative gain or loss associated with dichotomizing the continuous outcome $X^*(t_0)$.

MODEL SIII(iii):

$$\text{logit}(P(Y2^* = 1)) = \alpha_0 + \alpha_1 G + \alpha_2 E + \alpha_3 (G \times E) + \alpha_4 t_0 \tag{3.32}$$

This model is similar to the previous one except that repeated measures of the continuous outcome are taken into account while determining binary outcome $Y2$. That is, $Y2$ is based on two or more values of X above the threshold, prior to the current measurement, obtained in the last six years.

MODEL SIII(iv):

$$\begin{aligned} \text{logit}(P(Y3^* = 1)) = \lambda_0 + \lambda_1 G &+ \lambda_2 E + \lambda_3 (G \times E) + \lambda_4 t_0 \\ &+ \lambda_5 (t_0 - t_{-j^*}) + \lambda_6 X^*(t_{-j^*}) \end{aligned} \tag{3.33}$$

$Y3$ is a binary outcome that indicates disease occurrence among individuals disease-free at the earliest assessment time, if value of X at current age exceeds T. This

model adjusts the estimated effects of G, E, $G \times E$ interaction and current age t_0. Additionally, the model adjusts for (i) the follow-up duration $(t_0 - t_{-j^*})$, and (ii) the earliest observed X value $X^*(t_{-j^*})$. The former adjustment accounts for the fact that occurrence of the outcome is more likely among subjects with longer follow-up, whereas the latter accounts for subjects with higher initial X being at higher "risk" of exceeding the threshold.

MODEL SIII(v):

$$\begin{aligned}\text{logit}(P(Y4^* = 1)) = \gamma_0 + \gamma_1 G &+ \gamma_2 E + \gamma_3 (G \times E) + \gamma_4 t_0 \\ &+ \gamma_5 (t_0 - t_{-j^*}) + \gamma_6 X^*(t_{-j^*})\end{aligned} \tag{3.34}$$

Binary outcome $Y4$ is similar to $Y3$ except that disease occurrence is defined by at least two values of X exceeding the threshold for subjects initially free from disease. This model incorporates the same adjustments as model SIII(iv).

MODEL SIII(vi):

$$\Delta X^* = \mu_0 + \mu_1 G + \mu_2 E + \mu_3 (G \times E) + \mu_4 t_{-j^*} + \varepsilon \tag{3.35}$$

Here, ΔX^* represents the (standardized) average increase in X per one year increase in subject's age.

MODEL SIII(vii):

$$\hat{\alpha_1^*} = \theta_0 + \theta_1 G + \theta_2 E + \theta_3 (G \times E) + \theta_4 t_{-j^*} + \theta_5 X^*(t_{-j^*}) + \varepsilon \tag{3.36}$$

Model SIII(vii) estimates the effects of G, E and $G \times E$ on the rate of progression in X while adjusting for subject's age and initial X value, at the earliest available X measurement.

Sensitivity analysis

According to the literature review, frequency of genetic susceptibility and exposure, and magnitude of the interaction are the most important determinants of the power

Table 3.6: Parameter values under the basic setting and new setting for Scenario III (sensitivity analyses).

Parameter	Previous value	New value
$P(G=1)$	0.3	0.2
$P(E=1)$	0.6	0.3
Δ_G	0.1	0.02
Δ_{GE}	0.20	0.03
	0.40	0.06

of a study to detect $G \times E$ interaction. We explore the effects on power of changing values of some of these parameters used in the data generation. Table 3.6 gives the alternative parameter settings explored in sensitivity analysis. Parameters not shown in the table have their previous values, as in Section 3.4.1.

Table 3.6 shows that, under the sensitivity analysis setting, genetic susceptibility and exposure are less common. In the absence of exposure, subjects experience a smaller rate of increase in X with age:

$$\begin{aligned} \beta_i &\sim \mathcal{N}[0.1, 0.03], && \text{if } G_i = 0 \\ \beta_i &\sim \mathcal{N}[0.1 + \Delta_G, 0.03], && \text{if } G_i = 1. \end{aligned}$$

This rate may increase in the presence of genetic susceptibility. However, the magnitude of increase (Δ_G) is lower than what was previously assumed (see Table 3.6) so that the presence of the gene results in a smaller rate of change in X. Moreover, we consider two new values for Δ_{GE} that assume much weaker effects of $G \times E$ on rate of increase in X.

In the previous setting, subjects with current ages between 40-60 years who were exposed, became first exposed at ages between 20-40 years (Section 3.4.2). That is, first exposure occurred before the current age $t_{i,0}$ for *all subjects*. However, this assumption no longer holds in the current setting since $t_{i,0}$ is now generated from $\mathcal{U}[30, 60]$, which implies that first exposure may occur after the current age for some subjects. In addition, larger errors are assumed in the measurement of X, i.e. $\varepsilon_{i,j} \sim \mathcal{N}[0, 1]$ in (3.27). The performance of the linear regression models SIII(i) and SIII(vi) versus the logistic regression models SIII(ii) - SIII(v) is assessed under these new assumptions.

3.5 Summary

This chapter described three basic scenarios for simulations design, in the context of which the efficiency of continuous versus binary outcomes for detecting $G \times E$ interaction was studied. Scenario I assumed that the continuous outcome X was a *risk factor* for the disease ($Y = 1$) and that part of the effect of $G \times E$ on binary outcome Y was mediated by this variable. Scenario II assumed that X was a *marker* for disease and change in X followed change in disease status Y. Two variations of Scenario II were considered: (i) Y depended on the observed risk factors that were independent of G, E and $G \times E$; (ii) Y depended, additionally, on some unobserved risk factors, which could be correlated with G or E, or both G and E.

In Scenario III, we generated repeated measurements of the continuous variable X, which was a marker of disease progression. Based on these repeated measures, different types of continuous and binary outcomes were generated to investigate which outcome would yield a model with higher power for detecting $G \times E$ interaction. To assess the role of important parameters in influencing the power of the study, in sensitivity analyses two settings for these parameters were investigated.

The underlying assumptions, various sub-scenarios, corresponding to different parameter values and/or data structure, data generation and methods of analysis, were described for each scenario. In the following chapter, we shall present the results of the simulations.

Chapter 4

Results

4.1 Introduction

The choice between a continuous or binary outcome could have important implications for the power of a study to detect $G \times E$ interaction. However, the magnitude of the gain in efficiency from replacing the clinically more relevant binary outcome Y by a continuous surrogate outcome X is likely to depend on a number of factors. To explore these factors, in this thesis, a number of scenarios have been considered which vary in their assumptions about the underlying mechanisms of G, E and $G \times E$ interaction effect on X or Y, and the relationship of these factors with other risk factors for Y and/or X. Alternative models of the biological relationship between X and Y are also explored. In addition, we consider different methods of dichotomization of the continuous variable X based on repeated measures, to construct different binary outcomes. Simulations have been used to assess the efficiency of alternative outcomes to detect $G \times E$ interaction.

In Chapter 3, we discussed the scenarios, their assumptions and the data generation procedures for the simulations. Different models for analysis based on alternative binary or continuous outcome variables were also described. This chapter presents the results of the simulations for each scenario, under a variety of parameter settings. Although the simulations were performed in both S-Plus and Matlab, to avoid redundancy, we present only the results of S-Plus in this thesis.

Table 4.1: Power of test for rejecting $H_0 : \beta_{ge} = 0$ at 0.05 significance level for models SI.1(i)-SI.2(iv).

Linear Regression		Logistic Regression			
SI.1(i)	SI.1(ii)	SI.2(i)	SI.2(ii)	SI.2(iii)	SI.2(iv)
0.359	0.335	0.495	0.277	0.275	0.063

4.2 Scenario I

Section 3.2.4 described six different models for analyzing each sample in Scenario I where a quantitative variable X was considered a risk factor for the binary outcome Y. Models SI.1(i) and SI.1(ii) were linear regression models for the true values and observed values of the continuous outcome X, respectively. The remaining four models (SI.2(i), SI.2(ii), SI.2(iii) and SI.2(iv)) were logistic regression models for the four versions (or "types") of binary outcome Y, obtained from different combinations of the parameter values given in Table 3.1 of Section 3.2.3. To recapitulate, model SI.2(i) corresponds to the situation, in which there is a relatively strong *direct* effect (i.e. $\beta_3 = \ln 3$) of $G \times E$ and moderate effect of E on disease outcome Y (by "direct" we mean effects not transmitted through X). On the other hand, outcome Y in model SI.2(ii) is influenced by a slightly weaker ($\beta_3 = \ln 2$), i.e. a moderate (direct) effect of $G \times E$ interaction. The outcome in model SI.2(iii) differs from that in SI.2(ii) by assuming that there is no (direct) effect of E on Y. Lastly, model SI.2(iv) corresponds to the case where there are no direct effects of either $G \times E$ or E on Y.

The estimated power for each of the above models is reported in Table 4.1. Comparison between models SI.1(i) and SI.1(ii) reveals that there is some loss of power due to measurement errors in the quantitative risk factor X. Power comparisons between the logistic regression models indicate that when a "strong" $G \times E$ interaction effect on Y is mediated through pathways not involving X, highest power for model SI.2 is observed. Indeed, in that case using the binary outcome *improves* power compared to the quantitative "surrogate" X. The main reason is that the impact of $G \times E$ on Y is only partly mediated by X. When the direct effect for $G \times E$ is moderate, estimated power in model SI.2 declines by 44%. Comparison between models SI.2(ii) and SI.2(iii) indicates that magnitude of the *direct exposure* effect (i.e. β_2) does not

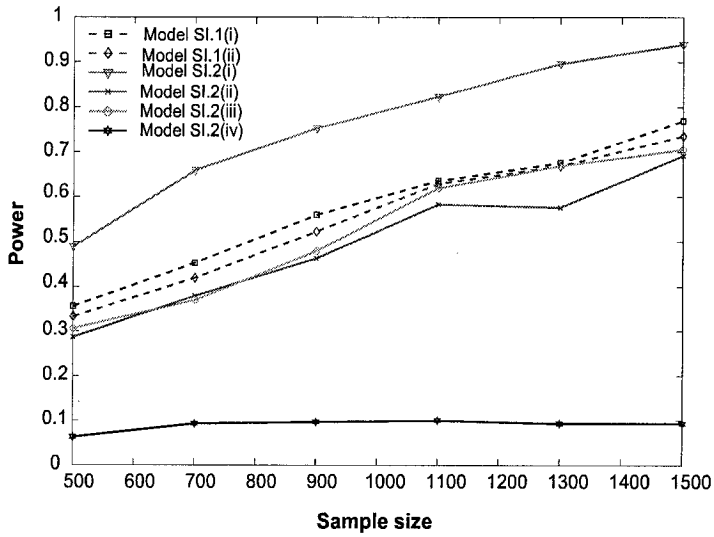

Figure 4.1: Effect of sample size variation on power for the test of $H_0 : \beta_{ge} = 0$ at 0.05 significance level for models SI.1(i)-SI.2(iv). The number of data sets generated was 300.

have a significant influence on the power to detect $G \times E$ interaction. As expected, when effects of both $G \times E$ and E are reduced to zero, probability of rejecting H_0 is close to 0.05, indicating the test has the correct size.

Based on the data generation scheme, direct power comparisons between linear and logistic regression models is of limited interest here. Nevertheless, some important results do emerge from the two types of analyses. Power for linear regression on X versus logistic regression on $P(Y = 1)$ depends mainly on:

1. Relative strength of $G \times E$ interaction effects on X (i.e. α_3 in (3.5)) versus on $P(Y = 1)$ (i.e. β_3, in (3.1)).

2. To a lesser extent, measurement errors in X.

The effect of increase in sample size on estimated power for each model is summarized in Figure 4.1. In general, power for detecting $G \times E$ interaction increases as sample size increases.

Table 4.2: Power of test for rejecting $H_0 : \beta_{ge} = 0$ at 0.05 significance level for linear regression model SII.1(i).

Combination[a]	σ^b	$(\theta_R, \theta_Y)^c$					
		(0,0.5)	(0,1)	(0, 1.5)	(0.5, 0.5)	(0.5,1)	(0.5,1.5)
Case 2.1	0.05	0.100	0.163	0.347	0.080	0.207	0.36
	0.50	0.087	0.167	0.283	0.070	0.157	0.297
	1.00	0.103	0.113	0.217	0.097	0.120	0.257
Case 2.2	0.05	0.077	0.097	0.193	0.063	0.117	0.207
	0.50	0.073	0.107	0.180	0.067	0.087	0.153
	1.00	0.083	0.080	0.113	0.060	0.080	0.143
Case 2.3	0.05	0.070	0.070	0.120	0.057	0.093	0.137
	0.50	0.070	0.087	0.107	0.047	0.063	0.113
	1.00	0.083	0.070	0.107	0.057	0.060	0.113
Case 2.4	0.05	0.050	0.037	0.037	0.057	0.057	0.063
	0.50	0.067	0.050	0.043	0.050	0.030	0.053
	1.00	0.063	0.050	0.050	0.060	0.050	0.070

[a] See Table 3.2

[b] Measurement error standard deviation for X

[c] θ_Y determines diagnostic value of X as marker for $Y = 1$; θ_R determines correlation between X and other observable risk factors R

4.3 Scenario II

Case S1

A basic difference between Scenarios I and II described in Sections 3.2 and 3.3 respectively, is that in the latter, the quantitative variable X is a marker rather than a risk factor for the disease outcome Y. This difference in the biological relationship between X and Y is likely to have some influence on power comparisons between linear and logistic regression analyses. Moreover unlike Scenario I, which made no explicit assumptions regarding the presence of other observed/unobserved risk factors of Y, the hypothetical model in Case S1 assumes that both X and Y depend on other observed risk factors (R) independent of G and E.

Both linear and logistic regression models were described in Section 3.3.5 for analyzing the continuous (surrogate) outcome X and binary outcome Y in Case S1. For each model, it is of interest to investigate the effect on power of changes in the simulation parameters θ_R, θ_Y and σ. Here, θ_R determines the correlation between X and

Table 4.3: Power of test for rejecting $H_0 : \beta_{ge} = 0$ at 0.05 significance level for logistic regression model SII.1(ii).

Combination[a]	(η_1^b, η_0^c)				
	(1,1)	(0.9,0.9)	(0.7,0.9)	(0.9,0.7)	(0.7,0.7)
Case 2.1	0.507	0.423	0.280	0.340	0.167
Case 2.2	0.223	0.203	0.157	0.123	0.100
Case 2.3	0.177	0.140	0.093	0.087	0.057
Case 2.4	0.030	0.0167	0.063	0.040	0.030

[a] See Table 3.2
[b] Sensitivity in Y
[c] Specificity in Y

R, θ_Y determines the magnitude of the change in X when $Y = 1$, and σ determines the level of measurement errors in X (Section 3.3.4). Table 4.2 summarizes the results for the linear regression model SII.1(i) under the assumptions of Case S1. Estimated power of the test for the coefficient of $G \times E$ interaction increases as θ_Y increases for a fixed θ_R. This result is apparent in Figure 4.2. The parameter θ_Y determines the impact on X of changes in Y. Thus, as expected, the greater the increase in marker X for change in disease status Y, the higher the power of the linear regression model for detecting $G \times E$ interaction. The latter finding reflects a necessary characteristic of a "good" surrogate outcome. In order to detect $G \times E$ interaction with higher power, X should undergo a large enough change when disease status changes. Estimated power does not change appreciably for a higher value of the parameter θ_R when θ_Y is fixed. In other words, correlation between X and R does not significantly affect the power to detect $G \times E$ interaction. Power declines rapidly for the linear regression model SII.1(i) as the true underlying $G \times E$ interaction effect on Y decreases. The decline is more apparent at higher values of θ_Y. Power also decreases slightly with increase in measurement error in X (Table 4.2)

The results regarding power to detect $G \times E$ interaction in the case of the logistic regression model SII.1(ii), under Case S1, are shown in Table 4.3. Similar to linear regression, estimated power for logistic regression declines as the effect of the underlying $G \times E$ interaction effect on Y (β_3, Table 3.2) becomes weaker. As expected, power also declines with increase in misclassification error of Y (Figure 4.4). Compared to

Table 4.4: Power of test for rejecting $H_0 : \beta_{ge} = 0$ at 0.05 significance level for linear regression model SII.2(i) with observed X_0 as outcome.

Combination[a]	σ^b	$(\theta_R, \theta_Y)^c$					
		(0,0.5)	(0,1)	(0, 1.5)	(0.5, 0.5)	(0.5,1)	(0.5,1.5)
Case 1	0.05	0.057	0.253	0.400	0.133	0.237	0.393
	0.50	0.063	0.223	0.323	0.133	0.207	0.343
	1.00	0.063	0.137	0.243	0.077	0.167	0.243
Case 2	0.05	0.057	0.157	0.243	0.103	0.160	0.243
	0.50	0.037	0.153	0.233	0.117	0.147	0.210
	1.00	0.037	0.083	0.170	0.067	0.120	0.163
Case 3	0.05	0.040	0.103	0.193	0.093	0.107	0.187
	0.50	0.030	0.100	0.190	0.107	0.103	0.147
	1.00	0.037	0.067	0.137	0.060	0.073	0.100
Case 4	0.05	0.030	0.037	0.100	0.067	0.043	0.077
	0.50	0.020	0.040	0.076	0.087	0.050	0.073
	1.00	0.027	0.043	0.077	0.053	0.037	0.067

[a] See Table 3.5

[b] Measurement error standard deviation for X

[c] θ_Y determines diagnostic value of X as marker for $Y = 1$; θ_R determines correlation between X and other observable risk factors R

the linear regression model for X, the logistic regression model has higher power for detecting strong $G \times E$ interaction in the data. However, this relative power advantage of the logistic regression model declines with increase in misclassification error of Y. For instance, when the sensitivity (η_1) and specificity (η_0) parameters in Table 4.3 equal 0.7 and 0.9 respectively, the linear regression model has higher power (35%) relative to the logistic regression model (28%) when $\theta_Y = 1.5$ and there exist almost no measurement errors in X (Table 4.2).

To summarize, the power for linear regression on X versus logistic regression on $P(Y = 1)$ under Case S1 depends mainly on:

1. The value of θ_Y for the association between changes in marker X and changes in Y,

2. Level of misclassification errors i.e. sensitivity and specificity in Y for logistic regression,

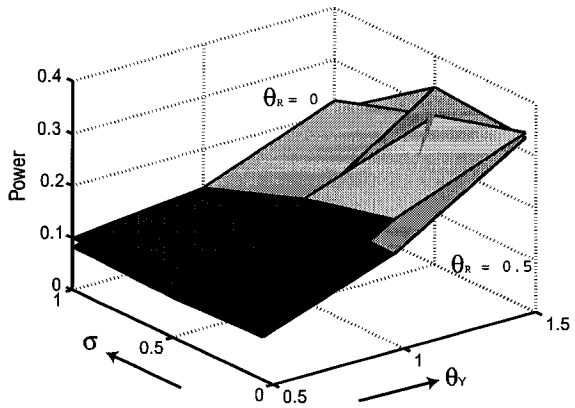

Figure 4.2: Effect of θ_Y and θ_R on estimated power for detecting $G \times E$ interaction for linear regression model SII.1(i) assuming strong $G \times E$ interaction and moderate exposure effect on Y [$\beta_2 = \ln(1.5)$, $\beta_3 = \ln(3)$; Table 3.2].

3. To a lesser extent, measurement errors in X.

Case S2

The assumptions of Case S2 differ from that of Case S1 in two major aspects: (i) outcome Y depends, additionally, on unobserved risk factors (S) which could be correlated with G and/or E, (ii) there are two repeated measurements of X. Table 4.4 shows, for different combinations of the parameter values θ_Y, θ_R and σ, the estimated power to detect $G \times E$ interaction for the linear regression model SII.2(i) that uses the *initial* value of observed X (i.e. X_0^*) as outcome. As in Case S1, estimated power for detecting $G \times E$ interaction increases as θ_Y increases for a fixed θ_R (see Figure 4.3). For a fixed θ_Y, power does not appear to change significantly when θ_R increases from 0 to 0.5.

The strength of the underlying $G \times E$ interaction effect on Y (β_3, Table 3.5) is also an important determinant of power in this model. Power decreases as β_3 decreases. The estimated power decreases only slightly for model SII.2(i) as measurement error in X_0 increases. Thus, as Figure 4.3 illustrates, there are no major differences in the factors affecting power in linear regression models SII.1(i) and SII.2(i) under Case S1

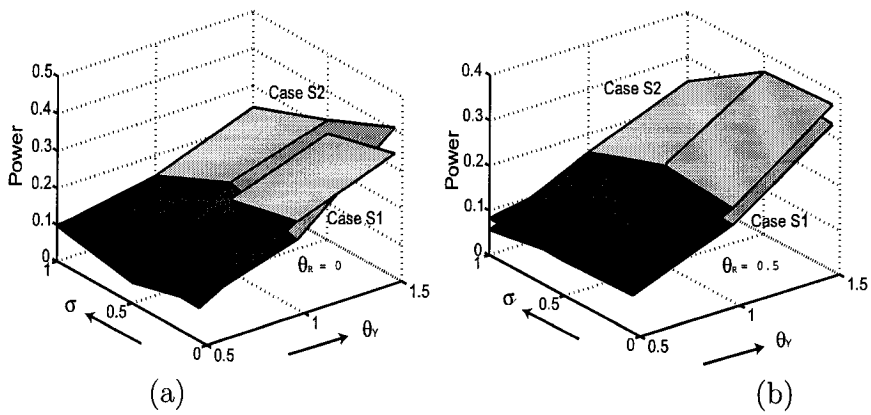

Figure 4.3: Power of test for $H_0 : \beta_{ge} = 0$ at 0.05 significance level for linear regression models SII.1(i) [Case S1] versus SII.2(i) [Case S2] assuming strong $G \times E$ interaction and moderate exposure effect on Y [$\beta_2 = \ln(1.5)$, $\beta_3 = \ln(3)$; Tables 3.2, 3.5]: (a) $\theta_R = 0$ (b) $\theta_R = 0.5$.

and Case S2, respectively.

Table 4.5: Power of test for rejecting $H_0 : \beta_{ge} = 0$ at 0.05 significance level for logistic regression model SII.2(iii).

Combination[a]	(η_1^b, η_0^c)				
	(1,1)	(0.9,0.9)	(0.7,0.9)	(0.9,0.7)	(0.7,0.7)
Case 2.1	0.400	0.390	0.243	0.327	0.193
Case 2.2	0.203	0.217	0.147	0.200	0.133
Case 2.3	0.170	0.150	0.120	0.143	0.097
Case 2.4	0.053	0.050	0.040	0.053	0.040

[a] See Table 3.5
[b] Sensitivity in Y
[c] Specificity in Y

Table 4.5 summarizes the results of power computations for the logistic regression model SII.2(iii) under Case S2. The estimated power decreases rapidly with increase in misclassification error of Y (Figure 4.4). The gradient of the decline appears to be slightly steeper for Case S1 than for Case S2, especially when the underlying $G \times E$ interaction effect on Y is strong. As in the case of model SII.2(i), power of the test to detect $G \times E$ interaction for the logistic regression model SII.2(iii) is higher when the underlying interaction effect on Y is strong (Figure 4.4). It is interesting to note that for certain combinations of the values of θ_R and θ_Y, the linear regression model

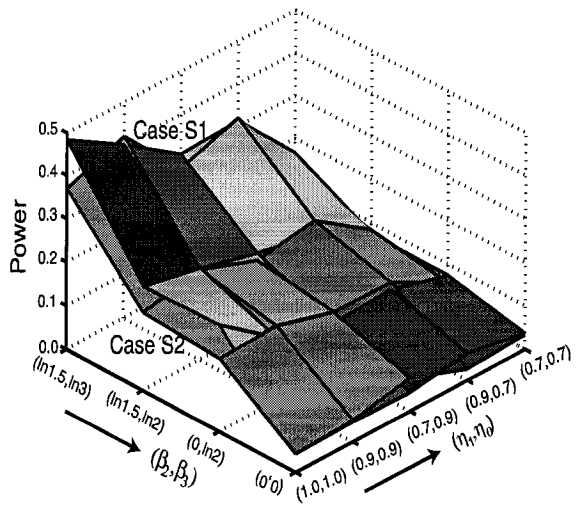

Figure 4.4: Power of test for $H_0 : \beta_{ge} = 0$ at 0.05 significance level for logistic regression models SII.1(ii) [Case S1] and SII.2(iii) [Case S2] according to combinations of sensitivity/specificity (η_1, η_0) and effects of E and $G \times E$ on Y $[\beta_2, \beta_3$; Tables 3.2, 3.5].

for X_0^*, SII.2(i), has power as high as that for the logistic regression model SII.2(iii), especially when there are almost no measurement errors in X_0. Thus, comparison of power for linear regression analysis of X_0^* versus logistic regression analysis of Y for Case S2 depends mainly on:

1. Magnitude of θ_Y, i.e. strength of the impact on the marker of the change in disease status Y,

2. Degree of misclassification error in Y,

3. To a lesser extent, measurement errors in X.

Interesting results are observed for the linear model SII.2(ii) in which the rate of change in X (referred to as Z) is the outcome. Table 4.6 shows that when measurement error in X is small ($\sigma = 0.05$, $ICC = 0.998$), the estimated power of the test to detect $G \times E$ interaction becomes 100%, irrespective of the magnitude of the underlying $G \times E$ interaction effect on Y. A possible explanation for such high power is that in the absence of moderate or large errors in X, Z is determined largely by the component $[\gamma_0 + \gamma_1 E + \gamma_2(G \times E)]$, which in turn depends on E and $G \times E$. Thus, we can expect E and $G \times E$ to be significant predictors of Z in model SII.2(ii). As measurement

error increases, the value of Z depends to a greater extent on the random component $\frac{\varepsilon_1 - \varepsilon}{\Delta t}$ and thus $G \times E$ may no longer have a significant effect on Z. This is reflected in the low values of power for $\sigma = 0.50$ ($ICC = 0.8$) and $\sigma = 1$ ($ICC = 0.5$) for all combinations of θ_R and θ_Y.

Table 4.6: Power of test for rejecting $H_0 : \beta_{ge} = 0$ at 0.05 significance level for linear regression model SII.2(ii) with time-interval standardized difference Z as outcome.

Combination[a]	σ[b]	(θ_R, θ_Y)[c]					
		(0,0.5)	(0,1)	(0, 1.5)	(0.5, 0.5)	(0.5,1)	(0.5,1.5)
Case 2.1	0.05	1.000	1.000	1.000	1.000	1.000	1.000
	0.50	0.100	0.107	0.110	0.137	0.127	0.130
	1.00	0.070	0.070	0.067	0.067	0.050	0.070
Case 2.2	0.05	1.000	1.000	1.000	1.000	1.000	1.000
	0.50	0.100	0.107	0.110	0.137	0.127	0.130
	1.00	0.070	0.070	0.067	0.067	0.050	0.070
Case 2.3	0.05	1.000	1.000	1.000	1.000	1.000	1.000
	0.50	0.100	0.107	0.110	0.137	0.127	0.130
	1.00	0.070	0.070	0.067	0.067	0.050	0.070
Case 2.4	0.05	1.000	1.000	1.000	1.000	1.000	1.000
	0.50	0.100	0.107	0.110	0.137	0.127	0.130
	1.00	0.070	0.070	0.067	0.067	0.050	0.070

[a] See Table 3.5

[b] Measurement error standard deviation for X

[c] θ_Y determines diagnostic value of X as marker for $Y = 1$; θ_R determines correlation between X and other observable risk factors R

4.4 Scenario III

4.4.1 Results of the main analysis

Scenario III involves comparison of alternative continuous and binary outcomes derived from repeated measures of the quantitative marker X. Under the assumptions of Scenario III, the rate of increase in X (from the earliest assessment time) depends on the effects of G and $G \times E$ interaction, which are determined by the simulation parameters Δ_G and Δ_{GE}, respectively.

Three types of continuous outcomes and four types of binary outcomes were considered, and analyzed using the linear and logistic regression models described in Section

Table 4.7: Power of test for rejecting $H_0 : \beta_{ge} = 0$ at 0.05 significance level for multiple linear regression models SIII(i), SIII(vi), and SIII(vii) with continuous outcomes $X(t_0)$, ΔX and $\hat{\alpha_1}$, respectively.

Outcome[a]	Δ^e_{GE}	$\Delta^f_X = 0$		$\Delta_X = 5$	
		$\Delta^g_G = 0$	$\Delta_G = 0.10$	$\Delta_G = 0$	$\Delta_G = 0.10$
$X(t_0)$[b]	0.00	0.067	0.063	0.077	0.053
	0.20	0.507	0.513	0.510	0.523
	0.40	0.977	0.973	0.963	0.963
$X^*(t_0)$	0.00	0.063	0.063	0.077	0.053
	0.20	0.503	0.513	0.510	0.523
	0.40	0.977	0.973	0.963	0.963
ΔX[c]	0.00	0.030	0.060	0.030	0.060
	0.20	1.000	1.000	1.000	1.000
	0.40	1.000	1.000	1.000	1.000
ΔX^*	0.00	0.037	0.043	0.037	0.043
	0.20	1.000	1.000	1.000	1.000
	0.40	1.000	1.000	1.000	1.000
$\hat{\alpha_1}$[d]	0.00	0.037	0.057	0.027	0.063
	0.20	1.000	1.000	1.000	1.000
	0.40	1.000	1.000	1.000	1.000
$\hat{\alpha_1^*}$	0.00	0.033	0.063	0.030	0.060
	0.20	1.000	1.000	1.000	1.000
	0.40	1.000	1.000	1.000	1.000

[a] Asterisks indicate outcomes with measurement errors
[b] Current value of marker X
[c] Time-interval standardized difference in X
[d] Rate of progression in X
[e] Impact of $G \times E$ interaction on rate of change in X
[f] Impact of G on X at birth
[g] Impact of G on rate of change in X

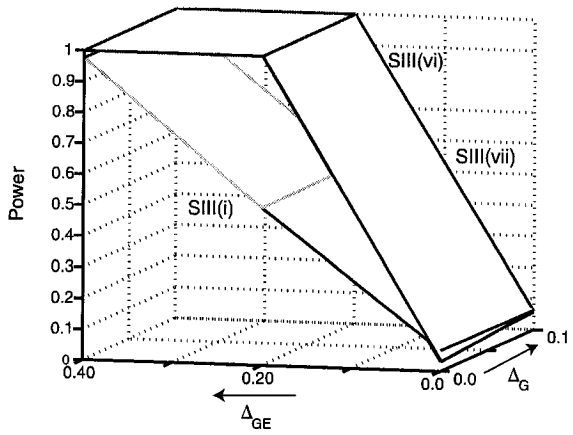

Figure 4.5: Power of test for $H_0 : \beta_{ge} = 0$ at 0.05 significance level using linear regression models SIII(i), SIII(vi) and SIII(vii) for $\Delta_X = 0$, assuming no measurement errors in X.

3.4.3. Table 4.7 gives the estimated power for the test of $G \times E$ interaction for linear regression models SIII(i), SIII(vi) and SIII(vii). To recapitulate, model SIII(i) uses the current value of the quantitative marker $X(t_0)$ as the outcome, while adjusting for the effects of G, E, $G \times E$ interaction and current age of the subject. Model SIII(vi) is similar to model SIII(i), but uses ΔX, the (standardized) average increase in X per year, as outcome and adjusts for the subject's earliest (available) measurement time. In addition to the independent variables in model SIII(vi), model SIII(vii) adjusts for the subject's initial X value and uses the rate of progression in X (i.e. $\hat{\alpha}_1$) as outcome.

For all three models, parameter Δ_{GE} affects the estimated power most (Figure 4.5). When $\Delta_{GE} = 0$, the rate of rejection of (true) H_0 is close to 0.05. Power increases from about 50% to almost 100% as Δ_{GE} increases from 0.20 to 0.40. This is expected since Δ_{GE} directly determines β', the rate of change in X due to $G \times E$ interaction (see Section 3.4.1), which in turn affects the values of X at different assessment times.

Models SIII(vi) and SIII(vii) have similar estimated powers for different parameter combinations as seen in Figure 4.5. Both have high power to detect $G \times E$ interaction even when the underlying interaction effect in the data is moderate (i.e. $\Delta_{GE} = 0.20$). In contrast, model SIII(i), which uses the current X value as outcome, requires a stronger underlying $G \times E$ interaction effect in the data to detect interactions with adequate power. Neither Δ_G nor Δ_X appear to influence estimated power for these

Table 4.8: Power of test for rejecting $H_0 : \beta_{ge} = 0$ at 0.05 significance level for multiple logistic regression models [SIII(ii)-SIII(v)] with four different binary outcomes, assuming no measurement errors in X.

Outcome *[Model]*	(Δ_X^a, T^b)	$(\Delta_G^c, \Delta_{GE}^d)$					
		(0,0)	(0,0.20)	(0, 0.40)	(0.1,0)	(0.1,0.20)	(0.1, 0.40)
Y1 *[SIII(ii)]*	(0,$T1$)	0.063	0.290	0.817	0.053	0.313	0.783
	(0,$T2$)	0.063	0.323	0.783	0.067	0.300	0.750
	(5,$T1$)	0.057	0.283	0.787	0.047	0.253	0.690
	(5,$T2$)	0.070	0.337	0.777	0.043	0.263	0.697
Y2 *[SIII(iii)]*	(0,$T1$)	0.053	0.227	0.670	0.053	0.243	0.640
	(0,$T2$)	0.063	0.260	0.680	0.063	0.237	0.623
	(5,$T1$)	0.063	0.220	0.600	0.050	0.230	0.507
	(5,$T2$)	0.070	0.260	0.597	0.057	0.187	0.530
Y3 *[SIII(iv)]*	(0,$T1$)	0.107	0.150	0.190	0.120	0.117	0.153
	(0,$T2$)	0.143	0.197	0.177	0.170	0.193	0.243
	(5,$T1$)	0.093	0.123	0.137	0.093	0.107	0.107
	(5,$T2$)	0.130	0.157	0.123	0.100	0.163	0.163
Y4 *[SIII(v)]*	(0,$T1$)	0.073	0.110	0.190	0.107	0.143	0.130
	(0,$T2$)	0.120	0.153	0.183	0.103	0.157	0.190
	(5,$T1$)	0.070	0.100	0.153	0.080	0.117	0.120
	(5,$T2$)	0.110	0.120	0.163	0.117	0.093	0.123

[a] Impact of G on X at birth
[b] Threshold value of X
[c] Impact of G on rate of change of X
[d] Impact of $G \times E$ interaction on rate of change of X

models (Table 4.7).

For Scenario III, four different types of binary outcomes were formulated to examine how the power depends on the way the outcome is operationalized. In model SIII(ii), $Y1 = 1$, if current value of X exceeds threshold T. In model SIII(iii), $Y2 = 1$ if at least two repeated measures of X before the current age exceed the threshold value. For subjects with initial X value less than the threshold, outcome $Y3 = 1$ in model SIII(iv) if the current X value exceeds the threshold. For the same restricted sample, outcome $Y4 = 1$ in model SIII(v) if two or more repeated measurements of X exceeds the threshold.

Table 4.8 shows the estimated power for detecting $G \times E$ interaction for the logistic regression models SIII(ii)-SIII(v). As expected, the type-I error rates are close to 0.05 when Δ_{GE} is 0. Obviously, power increases as Δ_{GE} increases. However, the increase

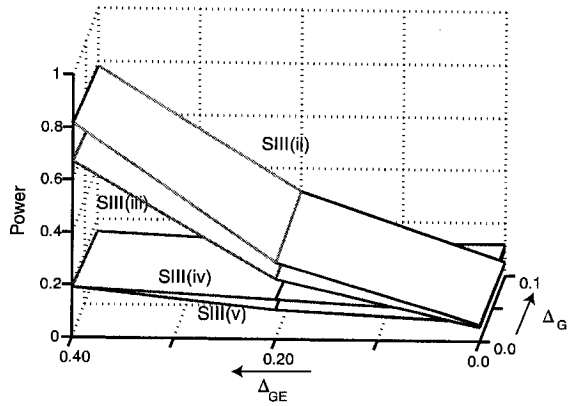

Figure 4.6: Power of test for $H_0 : \beta_{ge} = 0$ at 0.05 significance level using logistic regression models SIII(ii)- SIII(v) for $\Delta_X = 0$, $T = T1$, assuming no measurement errors.

is more pronounced for binary outcomes $Y1$ and $Y2$ corresponding to models SIII(ii) and SIII(iii), respectively. For instance, in the case of model SIII(ii), power increases from about 30% to as high as 81% as Δ_{GE} increases from 0.2 to 0.4 when $\Delta_X = 0$ and $T = T1$. For models SIII(iv) and SIII(v) with outcomes $Y3$ and $Y4$ respectively, the increase is much less, in comparison. Comparing models SIII(ii) and SIII(iii), we see that estimated power is slightly less for the latter. The uniformly low power of models SIII(iv) and SIII(v) compared to the other two logistic models (i.e. SIII(ii) and SIII(iii)) indicates that, as expected, outcomes based on the entire sample are more efficient than outcomes restricted to subjects that were disease-free at the earliest assessment time (Figure 4.6).

The threshold parameter T, as well as Δ_G and Δ_X do not have significant effects on power in either of these models. When Δ_G increases from 0 to 0.1, power decreases *slightly* in all models when Δ_{GE} is set to its highest value. For a fixed threshold T, there seems to be a *slight*, but rather systematic, decline in estimated power as Δ_X increases. Finally, measurement errors in X appear to have only a minor impact on the estimated power.

To summarize, comparison of power between the various continuous outcomes reveals that models based on outcomes ΔX, the (standardized) average increase in X, and $\hat{\alpha_1}$, the rate of progression in X (i.e. models SIII(vi) and SIII(vii)), have higher

Table 4.9: Power of test for rejecting $H_0 : \beta_{ge} = 0$ at 0.05 significance level for multiple logistic regression models [SIII(ii)-SIII(v)] with four different binary outcomes, assuming measurement errors in X.

Outcome[a] *[Model]*	(Δ_X^b, T^c)	$(\Delta_G^d, \Delta_{GE}^e)$					
		(0,0)	(0,0.20)	(0, 0.40)	(0.1,0)	(0.1,0.20)	(0.1, 0.40)
$Y1^*$ *[SIII(ii)]*	(0,$T1$)	0.040	0.290	0.807	0.077	0.297	0.783
	(0,$T2$)	0.067	0.317	0.783	0.080	0.313	0.747
	(5,$T1$)	0.057	0.330	0.803	0.047	0.267	0.707
	(5,$T2$)	0.077	0.327	0.783	0.057	0.287	0.673
$Y2^*$ *[SIII(iii)]*	(0,$T1$)	0.060	0.210	0.673	0.060	0.223	0.623
	(0,$T2$)	0.057	0.257	0.637	0.057	0.237	0.610
	(5,$T1$)	0.063	0.223	0.593	0.037	0.223	0.500
	(5,$T2$)	0.073	0.223	0.590	0.047	0.170	0.513
$Y3^*$ *[SIII(iv)]*	(0,$T1$)	0.087	0.107	0.170	0.133	0.173	0.173
	(0,$T2$)	0.080	0.097	0.130	0.110	0.163	0.170
	(5,$T1$)	0.077	0.150	0.123	0.097	0.010	0.163
	(5,$T2$)	0.113	0.097	0.143	0.103	0.100	0.127
$Y4^*$ *[SIII(v)]*	(0,$T1$)	0.153	0.193	0.240	0.187	0.267	0.253
	(0,$T2$)	0.167	0.183	0.240	0.207	0.210	0.207
	(5,$T1$)	0.150	0.177	0.207	0.153	0.193	0.220
	(5,$T2$)	0.197	0.213	0.197	0.183	0.173	0.203

[a] Binary outcomes were derived from values of X with measurement errors
[b] Impact of G on X at birth
[c] Threshold value of X
[d] Impact of G on rate of change of X
[e] Impact of $G \times E$ interaction on rate of change of X

power for detecting "moderate" $G \times E$ interaction in the data relative to the model that uses the current X value $X(t_0)$ as outcome (model SIII(i)). However, all three models are efficient in detecting strong interaction in the data. Comparison of power between the binary outcomes shows that for detecting moderate to strong $G \times E$ interaction, outcomes $Y1$ and $Y2$ based on all subjects in the data are more efficient than $Y3$ and $Y4$ which are obtained only for subjects with initial X values below the threshold. Moreover, outcomes based on the current X value ($Y1$ and $Y3$) have slightly higher power than outcomes obtained from repeated X measurements ($Y2$ and $Y4$). Overall, the simulation results for Scenario III demonstrate that for detecting moderate to strong $G \times E$ interaction effects in the data, quantitative outcomes based on variable X are more efficient than binary outcomes ($Y1$-$Y4$) derived from X.

4.4.2 Results of sensitivity analysis

In Section 3.4.3 we described modifications of some of the parameter values to assess their effects on estimated power for models SIII(i) - SIII(vi). Specifically, the prevalence of the genetic factor G and exposure E were decreased, impact of G on rate of change of X (Δ_G) and impact of $G \times E$ on rate of change of X (Δ_{GE}) were decreased. Moreover, larger errors were assumed in the measurement of X.

Table 4.10 summarizes the results for the linear regression models SIII(i) and SIII(vi) under the new parameter settings. Power is uniformly low for all parame-

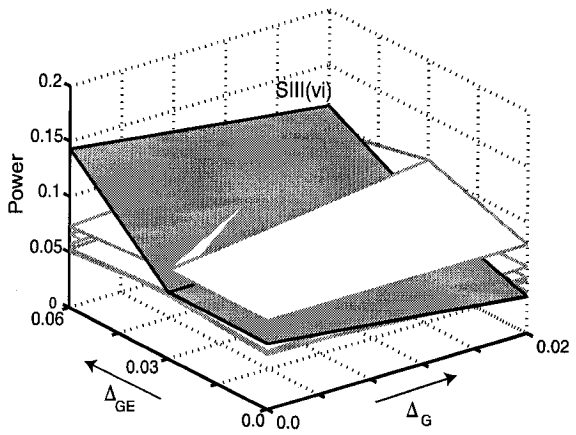

Figure 4.7: Comparison of estimated power for detecting $G \times E$ interaction in models SIII(i)-SIII(vi) under new parameter settings, with Δ_X=0 for all models and $T = T1$ for the logistic regression models.

ter combinations given in Table 4.10 when the current X value $X(t_0)$ is the outcome. The estimated power is only slightly higher for the average increase in X, ΔX, when the impact of $G \times E$ on the rate of change is weak ($\Delta_{GE} = 0.06$). Similarly, power for detecting $G \times E$ interaction is low for logistic regression analysis (Table 4.11).

For easy comparison, Figure 4.7 shows the trend in power estimates for all six models with Δ_X fixed at 0 and $T = T1$ for the logistic regression models. The observed trends are quite similar for other combinations of Δ_X and T (data not shown). Power comparisons between linear and logistic regression analysis indicate that, with reduced strength of interaction, neither of the two performs substantially better than

Table 4.10: Power of test for rejecting $H_0 : \beta_{ge} = 0$ at 0.05 significance level for multiple linear regression models SIII(i) and SIII(vi) with continuous outcomes $X^*(t_0)$ and ΔX^*, respectively, under new parameter settings.

Outcome[a]	Δ^d_{GE}	$\Delta^e_X = 0$		$\Delta_X = 5$	
		$\Delta^f_G = 0$	$\Delta_G = 0.02$	$\Delta_G = 0$	$\Delta_G = 0.02$
$X^*(t_0)$[b]	0.00	0.053	0.060	0.030	0.037
	0.03	0.057	0.057	0.043	0.050
	0.06	0.057	0.067	0.070	0.067
ΔX^*[c]	0.00	0.060	0.033	0.060	0.033
	0.03	0.060	0.063	0.060	0.063
	0.06	0.143	0.113	0.143	0.113

[a] Asterisks indicate outcomes with measurement errors
[b] Current value of marker X
[c] Time-interval standardized difference in X
[d] Impact of $G \times E$ interaction on rate of change in X
[e] Impact of G on X at birth
[f] Impact of G on rate of change in X

the other. Thus, the relative gain in power from using continuous outcomes instead of binary outcomes observed in Section 4.4.1 disappears under the new parameter settings. However, model SIII(vi) (shaded in Figure 4.7) that uses the (standardized) average increase in X as outcome seems to have relatively better power when $\Delta_{GE} = 0.06$.

Thus, the results of Scenario III simulations reveal that use of continuous outcomes provides greater power than the binary outcomes when there is moderate to strong $G \times E$ interaction in the data. The comparison of power of the two types of analyses, linear versus logistic, depends on two main factors:

1. The type of binary outcome used versus the type of continuous outcome.

2. The magnitude of the underlying $G \times E$ interaction effect in the data.

For detecting *moderate* $G \times E$ interaction, the relative power among the continuous outcomes is higher for the average increase in X and the rate of progression in X. Among the binary outcomes, the relative power for detecting *moderate* to *strong* $G \times E$ interaction is higher for outcomes defined for all subjects ($Y1$ and $Y2$) than for outcomes defined for subjects having earliest X measurement below the threshold ($Y3$

Table 4.11: Power of test for rejecting $H_0 : \beta_{ge} = 0$ at 0.05 significance level for logistic regression models SIII(ii)-SIII(v), under new parameter settings.

Outcome[a] *[Model]*	(Δ_X^b, T^c)	$(\Delta_G^d, \Delta_{GE}^e)$ (0,0)	(0,0.03)	(0, 0.06)	(0.02,0)	(0.02,0.03)	(0.02, 0.06)
$Y1^*$ *[SIII(ii)]*	(0,$T1$)	0.053	0.060	0.053	0.037	0.053	0.063
	(0,$T2$)	0.047	0.073	0.067	0.053	0.047	0.070
	(5,$T1$)	0.043	0.037	0.053	0.037	0.027	0.043
	(5,$T2$)	0.043	0.037	0.047	0.050	0.060	0.053
$Y2^*$ *[SIII(iii)]*	(0,$T1$)	0.050	0.057	0.067	0.050	0.053	0.060
	(0,$T2$)	0.057	0.063	0.070	0.063	0.047	0.053
	(5,$T1$)	0.017	0.027	0.043	0.037	0.047	0.043
	(5,$T2$)	0.040	0.050	0.043	0.040	0.053	0.047
$Y3^*$ *[SIII(iv)]*	(0,$T1$)	0.050	0.060	0.050	0.047	0.053	0.043
	(0,$T2$)	0.030	0.053	0.057	0.047	0.060	0.053
	(5,$T1$)	0.017	0.067	0.040	0.067	0.047	0.077
	(5,$T2$)	0.063	0.043	0.060	0.050	0.040	0.050
$Y4^*$ *[SIII(v)]*	(0,$T1$)	0.080	0.080	0.073	0.080	0.110	0.090
	(0,$T2$)	0.103	0.110	0.113	0.093	0.110	0.090
	(5,$T1$)	0.050	0.070	0.090	0.100	0.087	0.110
	(5,$T2$)	0.080	0.090	0.100	0.107	0.080	0.100

[a] Binary outcomes were derived from values of X with measurement errors
[b] Impact of G on X at birth
[c] Threshold value of X
[d] Impact of G on rate of change of X
[e] Impact of $G \times E$ interaction on rate of change of X

and $Y4$).

4.5 Summary

This chapter presents the results of the simulations for the three scenarios described in Chapter 3. Under Scenario I, replacing the binary outcome Y by the continuous risk factor X could result in higher power in situations where the effect of $G \times E$ interaction on Y that is mediated through other risk factors (pathways not involving X) is *weak* i.e. the interaction effect is mostly mediated through X. Under the assumptions of Scenario II, the association between changes in marker X and changes in disease status Y is the main factor that determines whether any gain in power could be achieved by using the continuous surrogate outcome. Misclassification error in Y seems to have a

more serious effect on the power in logistic regression analysis than measurement errors in X in linear regression analysis. Thus, in circumstances where misclassification error in Y is likely to be of concern, use of the continuous surrogate outcome may be a better alternative.

For Scenario III, where changes in X are assumed to reflect the change in disease status, results indicate that the most important parameter determining power is the magnitude of underlying $G \times E$ interaction in the data. For moderate and strong interaction effects, all three linear regression models were efficient. In contrast, only logistic regression models that used all cases in the analysis were successful in detecting $G \times E$ interaction with any reasonable level of power.

Sensitivity analysis showed that when genetic susceptibility and exposure are less common, interaction in the data is weak and measurement errors are large, all of the models assessed have very low power. In these circumstances, the linear regression model that uses time-interval standardized change in X as outcome may have some promise.

In conclusion, the simulation results indicate that, under most scenarios considered, use of continuous outcomes based on repeated measures of the quantitative variable X, provides better power than binary outcomes obtained from some dichotomization of this variable.

Chapter 5

Discussion and Conclusion

Power and efficiency considerations are critical for the design and feasibility of epidemiologic studies of gene-environment ($G \times E$) interactions. When the interaction effect is moderate, involving uncommon genes or environmental exposures, required sample sizes are often unattainable. Thus, much of the methodological literature on $G \times E$ interactions has focused on evaluation of different study designs to enhance statistical power. This study explored an alternative strategy for improving power without having to resort to the use of more complicated and costly study designs.

For most pathological processes, the clinical outcome of ultimate interest is binary, such as the occurrence of disease. However, in such studies available data often include also measurements of some quantitative variable that may be a risk factor or marker for this outcome. Replacing the binary outcome by this quantitative "surrogate" outcome could result in better power for disclosing $G \times E$ interactions, especially if the binary outcome is rare. However, any gain in power is likely to depend on a number of factors, such as the magnitude of the $G \times E$ interaction effect on the true binary outcome Y and on the surrogate, frequency of the binary outcome, and of the factors G and E, as well as measurement errors in both outcome variables [5], [6], [105], [122]. In addition, the biological relationship between the binary outcome and the quantitative surrogate, and the role of the latter in the pathological process will have some influence on their relative powers for detecting $G \times E$ interactions. The purpose of this thesis was to carry out a systematic quantitative study of the implications of using alternative outcomes on the power to detect $G \times E$ interactions while taking into account some of these factors.

Three basic scenarios were proposed. Each assumed that disease was determined by a single common genetic factor and a single exposure, and that both factors were binary. Scenario I considered the simplest hypothetical model of the relationship between the binary and quantitative outcomes, and their dependence on G, E and $G \times E$ interaction. Here, the surrogate X was assumed to be a risk factor for disease, and a mediating factor for the effects of G, E and $G \times E$. Alternative sub-scenarios were studied based on varying assumptions regarding the magnitudes of the $G \times E$ and E effects on binary outcome Y that were *not* transmitted through X.

Scenario II assumed that the quantitative surrogate X was a *marker* of disease incidence, and was also associated with other risk factors for the binary outcome that were independent of G and E. Elevated levels of the marker indicated the occurrence of disease. Another variation of this scenario assumed that Y depended, additionally, on some other unobserved risk factors, some of which were correlated with G and/or E. Assuming the study was prospective, two measurements were generated for X and a new quantitative outcome was defined as the rate of change in the quantitative marker.

The way a quantitative variable is measured determines, at least partly, the efficiency of analysis. Thus, the last scenario considered repeated measures of the quantitative marker, from which alternative continuous and binary outcomes were derived. Linear or logistic regression analyses were performed, depending on the type of outcome, and estimates for power were compared.

The results of the simulations reveal a number of important findings. Firstly, under the assumptions of Scenario I, the quantitative surrogate outcome provides higher power than the binary outcome when the effects of $G \times E$ and E on Y are transmitted *mainly* through X. In other words, a "good" surrogate outcome for Y to detect $G \times E$ interaction would be one that (almost) entirely mediates the effects of $G \times E$. The latter finding is consistent with the criterion for assessing surrogate endpoints of clinical outcomes in clinical trials: a perfect surrogate should completely explain the treatment effect on clinical outcome [33], [89]. Moreover, the gain in power from replacing the binary outcome by the quantitative surrogate is higher when there are no errors in measurement of X. There may be other factors that affect the power

in this scenario, although these factors have not yet been investigated. For instance, a higher prevalence of disease outcome, say close to 0.5, could result in an increased power for logistic regression. The association between X and $P(Y = 1)$ may also influence power comparisons: power for the two models may be closer when a stronger association exists.

In Scenario II, the binary outcome is a better choice for detecting $G \times E$ interaction with higher power. However, its efficiency is compromised when the degree of misclassification error in Y increases, in which case the quantitative outcome may be a better alternative. Researchers searching for biomarkers of disease outcome use sensitivity as an important criterion for evaluating markers [78], [88]. For instance, high sensitivity of cardiac troponin assays enables diagnosis of patients with unstable angina [124]. In our study, the strength of the association between changes in the quantitative marker X and changes in disease status Y is the most important factor that determines its efficiency relative to the binary outcome. The greater the increase in X when status changes from non-diseased to diseased, the higher the power of linear regression to detect $G \times E$ interaction. Thus, under the assumptions of Scenario II, replacing the binary outcome by the quantitative surrogate could result in higher power if existing methods for direct assessment of the binary outcome are crude, and accurate measurements for the quantitative outcome could be ascertained. The use of melanotransferrin, an accurate marker for mild Alzheimers disease, provides an example, since clinical diagnosis at early stages of the disease is not always reliable [61]. In the situation where Y also depends on unobserved risk factors, the relative power advantage of logistic regression declines, even in the absence of misclassification errors, when there are no measurement errors in X. Thus, even under the assumptions of Scenario II, use of alternative measures offers some promise in improving the power to detect $G \times E$ interaction.

Three types of continuous outcomes and four types of binary outcomes were considered in Scenario III. Binary outcomes were obtained through various dichotomizations of the marker. Results indicate that continuous outcomes are more efficient than binary outcomes for detecting moderate to strong $G \times E$ interaction effects. The low efficiency of binary outcomes in this scenario could perhaps be attributed to the loss

of information incurred due to discretization of the continuous variable [107]. The average increase in X per year and the rate of progression of X provide the highest power among all alternative outcomes. Thus, when repeated measures are available on a quantitative variable X, modelling the "rate" as outcome could provide better power for detecting $G \times E$ interactions than binary outcomes obtained through different dichotomizations. However, the sensitivity analysis reveals that neither linear nor logistic regression, using the outcomes considered in this study, have reasonable power to detect weak interactions in the data.

Overall, the findings of this research lead us to conclude that evaluation of alternative outcomes is a viable option for improving the power of studies to detect $G \times E$ interaction. Only a limited number of possible outcomes were assessed in this study. However, surrogate outcomes that are clinically more relevant could be developed, especially, to detect weak interactions of biological significance. In Scenario III, we considered a continuous outcome that measures the rate of progression in X. Another interesting alternative outcome could be one that dichotomizes the rate of progression to measure an "abnormally" fast progression of X. Our study involved univariate surrogate outcomes. However, it may be of interest to examine the efficiency of multivariate surrogates. Thus, the work presented in this thesis may be extended to incorporate "Multivariable Risk Scoring" [8], [55] where rather than a single X, a weighted mean of several markers/risk factors: $X_1, \ldots, X_k$, with weights reflecting their prognostic utility, may be considered.

An important scenario that has wide applicability and may be considered in future is one that assumes a latent period for the disease [2], [54], [82], [102]. Before the disease becomes observable, a pathological process may start, in which a patient's health status changes from healthy to subclinical latent pathology. With change of latent health status, the level of some factor X may start to change or increase at a higher rate. A further refinement of this scenario would be to assume a latent "severity" of the patient's latent health status (a quantitative variable) so that subjects with higher severity have both (i) faster rate of increase in X, and (ii) shorter time to diagnosis of (observable) disease Y, so that the rate of change in X is correlated with the hazard of developing the disease. Other, more complex scenarios may be considered, depending

on the disease of interest, to obtain a more comprehensive understanding of how choice of the outcome may influence the power of a $G \times E$ interaction study.

There are a myriad of factors that contribute to the development of disease. The scenarios that we have examined represent, therefore, simplified models of reality but, nevertheless, incorporate important features of the causal mechanisms of disease development and occurrence. Thus, this thesis lays the groundwork for future studies involving more complex scenarios. We hope that the findings of the present study will encourage researchers to explore this avenue in their efforts to detect $G \times E$ interactions with sufficient power.

Bibliography

[1] Albert P, Ratnasinghe D, Tangrea J, Wacholder S. Limitations of the case-only design for identifying gene-environment interactions. Am J Epidemiol 2001; 154: 687-693.

[2] André JB, Gupta S, Frank S, Tibayrenc M. Evolution and immunology of infectious diseases: Whats new? An e-debate. Infect Genet Evol 2004; 4: 69-75.

[3] Andrews P. Renal dysfunction as a marker of increased vascular risk. Br J Diabetes Vasc Dis 2004; 4: 152-155.

[4] Andrieu N, Goldstein AM. Epidemiologic and genetic approaches in the study of gene-environment interactions: An overview of available methods. Epidemiol Rev 1998; 20: 137-147.

[5] Andrieu N, Goldstein AM. The case-combined-control design was efficient in detecting gene-environment interactions. J Clinical Epidemiol 2004; 57: 662-671.

[6] Andrieu N, Goldstein AM, Thomas DC, Langholz B. Counter-matching in studies of gene-environment interaction: Efficiency and feasibility. Am J Epidemiol 2001; 153: 265-274.

[7] Armstrong BK, White E, Saracci R. Principles of Epidemiologic Measurement in Epidemiology. Oxford: Oxford University Press, 1994.

[8] Assmann G, Cullen P, Schulte H. Simple scoring scheme for calculating the risk of acute coronary events based on the 10- year follow-up of the Prospective Cardiovascular Munster (PROCAM) study. Circulation 2002; 105: 310-315.

[9] Beaty TH. Using association studies to test for gene-environment interaction asthma and other complex diseases. Clin Exper Allergy 1998; 28: 68-73.

[10] Bennett WP, Alavanja MC, Blomeke B, Vahakangas KH, Castren K, Welsh JA, Bowman ED, Khan MA, Flieder DB, Harris CC. Environmental tobacco smoke, genetic susceptibility, and risk of lung cancer in never-smoking women. J Natl Cancer Inst 1999; 91: 2009-2014.

[11] Borre M, Offersen IB, Nerstrom B, Overgaard J. Angiogenesis: Prognostic marker in prostatic cancer. Ugeskr Laeger 1999; 161: 3832-3836.

[12] Breslow NE, Day NE. Statistical Methods in Cancer Research: Volume I - The Analysis of Case-Control Studies. IARC Scientific publications: Lyon, 1980.

[13] Brown BW, Lovato J, Russell K. Asymptotic power calculations: Description, examples, computer code. Stat Med 1999; 18: 3137-3151.

[14] Bryk A, Raudenbush S. Hierarchical linear models: Applications and Data Analysis Methods. Thousand Oaks, California: Sage Publications, 2001.

[15] Caligari PDS, Mather K. Genotype-environment interaction .III. Interactions in Drosophila melanogaster. Proc R Soc Lond B 1975; 191: 387-411.

[16] Castelloe JM. Sample size computations and power analysis with the SAS system. Proc of the 25th Annual SAS Users Group Int Conf, Cary, NC: SAS Institute Inc., 2000.

[17] Chatterjee N, Kalaylioglu Z, Carroll RJ. Exploiting gene-environment independence in family-based case-control studies: increased power for detecting associations, interactions and joint effects. Genet Epidemiol 2005; 28: 138-156.

[18] Chemello L, Cavalletto L, Casarin C, Bonetti P, Bernardinello E, Pontisso P, Donada C, Belussi F, Martinelli S, Alberti A. Persistent hepatitis C viremia predicts late relapse after sustained response to interferon-alpha in chronic hepatitis C. Ann Intern Med 1996; 124: 1058-1060.

[19] Claus EB, Schildkraut JM, Thompson WD, Risch NJ. The genetic attributable risk of breast and ovarian cancer. Cancer 1996; 77: 2318-2324.

[20] Clayton D, McKeigue P. Epidemiological methods for studying genes and environmental factors in complex diseases. Lancet 2001; 358: 1356-1360.

[21] Committee on Carcinogenecity. Criteria for the design of gene-environment epidemiology studies. Joint Annual Report, 2001.

[22] Cooper RS. Gene-environment interactions and the etiology of common complex diseases. Ann Intern Med 2003; 139: 437-440.

[23] Curtis D. Use of siblings as controls in case-control association studies. Ann Hum Genet 1997; 61: 319-333.

[24] Dorman JS. Genetic epidemiology of insulin-dependent diabetes mellitus: International comparisons using molecular genetics. Ann Med 1992; 24: 393-399.

[25] Eaves LJ. The resolution of genotype $\times$ environment interaction in segregation analysis of nuclear families. Genet Epidemiol 1984; 1: 215-228.

[26] Ebel RL. Estimation of the reliability of ratings. Psychometrika 1951; 16: 407-424.

[27] Elston RC. Gene. In: Elston R, Olson J, Palmer L (eds). Biostatistical Genetics and Genetic Epidemiology. New York: John Wiley & Sons, 2002, pp. 285.

[28] Eng J. Sample Size Estimation: A glimpse beyond simple formulas. Radiology 2004; 230: 606-612.

[29] Falconer DS, Mackay TFC. Introduction to Quantitative Genetics, 3rd Ed. London: Longman, 1996.

[30] Farewell VT. Interaction. In: Elston R, Olson J, Palmer L (eds). Biostatistical Genetics and Genetic Epidemiology. New York: John Wiley & Sons, 2002, pp. 412-413.

[31] Fleming TR, DeMets DL. Surrogate endpoints in clinical trials: Are we being misled? Ann Intern Med 1996; 125: 605-613.

[32] Foppa I, Spiegelman D. Power and sample size calculations for case-control studies of gene-environment interactions with a polytomous exposure variable. Am J Epidemiol 1997; 146: 596-604.

[33] Freedman LS, Graubard BI, Schatzkin A. Statistical validation of intermediate endpoints for chronic diseases. Stat Med 1992; 11: 167-178.

[34] Galton F. Hereditary Genius, Its Laws and Consequences. London: Macmillan, 1869.

[35] Garcia-Closas M, Lubin J. Power and sample size calculations in case-control studies of gene-environment interactions: Comments on different approaches. Am J Epidemiol 1999; 149: 689-692.

[36] Garcia-Closas M, Rothman N, Lubin J. Misclassification in case-control studies of gene-environment interactions: Assessment of bias and sample size. Cancer Epidemiol Biomarkers Prev. 1999; 8: 1043-1050.

[37] Gatto NM, Campbell UB, Rundle AG, Ahsan H. Further development of the case-only design for assessing gene-environment interaction: Evaluation of and adjustment for bias. Int J of Epidemiol 2004; 33: 1014-1024.

[38] Gauderman WJ. Sample size requirements for matched case-control studies of gene-environment interaction. Stat Med 2002; 21: 35-50.

[39] Gauderman WJ, Kraft P. Family-based case-control studies. In: Elston R, Olson J, Palmer L (eds). Biostatistical Genetics and Genetic Epidemiology. New York: John Wiley & Sons, 2002, pp. 267-275.

[40] Gauderman WJ, Witte J, Thomas D. Family-based association studies. J of Nat Cancer Inst Monographs 1999; 26: 31-39.

[41] Goldstein AM, Andrieu N. Detection of interactions involving genes: Available study designs. Monogr. Natl. Cancer Inst 1999; 26: 49-54.

[42] Goldstein AM, Falk RT, Korczak JF, Lubin JH. Detecting gene-environment interactions using a case-control design. Genet Epidemiol 1997; 14: 1085-1089.

[43] Guo SW. Gene-environment interaction and the mapping of complex traits: Some statistical models and their implications. Hum Hered 2000; 50: 286-303.

[44] Haldane JBS. The interaction of nature and nurture. Ann Eugen 1946; 25: 197-205.

[45] Healthlink. The interaction of genes and environment in human disease, Article, Medical College of Wisconsin. Available at http://healthlink.mcw.edu/article/967585212.html.

[46] Hogben L. Nature and Nurture. New York: Norton, 1933.

[47] Hogben L. The limits of applicability of correlation technique in human genetics. J Genet 1933; 27: 379-406.

[48] Hogben L. The formal logic of the nature nurture issue. Acta Genetica 1951; 2: 101-140.

[49] Hohenboken WD. Genotype × environment interaction. In: Chapman AB (ed). General and Quantitative Genetics. Amsterdam: Elsevier, 1985, pp. 151-165.

[50] Hosmer DW, Lemeshow S. Applied Logistic Regression. New York: John Wiley & Sons, 1989.

[51] Hwang S, Beaty T, Liang K, Coresh J, Khoury M. Minimum sample size estimation to detect gene-environment interaction in case-control designs. Am J Epidemiol 1994; 140: 1029-1037.

[52] Ioannidis JPA, Lau J. Integrating genetics into randomized controlled trials. In: Khoury MJ, Burke W, Little J (eds). Human Genome Epidemiology: A Scientific Foundation for Using Genetic Information to Improve Health and Prevent Disease. Oxford University Press, 2004.

[53] Jennrich RI, Schluchter MD. Unbalanced repeated measures models with structured covariance matrices. Biometrics 1986; 42: 805-820.

[54] Kapoor R, Shrivastava S. Prevention of coronary artery disease from childhood. Indian Heart J 2002; 54: 726-730.

[55] Karp I, Abrahamowicz M, Bartlett G, Pilote L. Updated risk factor values and the ability of the multivariable risk score to predict coronary heart disease. Am J Epidemiol 2004; 160: 707-716.

[56] Katus HA, Remppis A, Scheffold T, Diederich KW, Kuebler W. Intracellular compartmentation of cardiac troponin T and its release kinetics in patients with reperfused and nonreperfused myocardial infarction. Am J Cardiol 1991; 67: 1360-1367.

[57] Khoury MJ. Genetic susceptibility to birth defects in humans: From gene discovery to public health action. Teratology 2000; 61: 17-20.

[58] Khoury MJ, Beaty TH, Cohen BH. Fundamentals of Genetic Epidemiology. New York: Oxford University Press, 1993.

[59] Khoury MJ, Flanders WD. Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: Case-control studies with no controls! Am J Epidemiol 1996; 144: 207-213.

[60] Khoury MJ, Little J, Burke W. Human Genome Epidemiology: A Scientific Foundation for Using Genetic Information to Improve Health and Prevent Disease. Oxford University Press, 2004.

[61] Kim DK, Seo MY, Lim SW, Kim S, Kim JW, Carroll BJ, Kwon DY, Kwon T, Kang SS. Serum melanotransferrin, p97 as a biochemical marker of alzheimers disease. Neuropsychopharmacology 2001; 25: 84-90.

[62] Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic Research: Principles and Quantitative Methods. California: Lifetime learning publications, 1982.

[63] Koopman JS. Causal models and sources of interaction. Am J Epidemiol 1977; 106: 439-444.

[64] Kramer MS. Clinical Epidemiology and Biostatistics: A Primer for Clinical Investigators and Decision-Makers. Berlin: Springer, 1988.

[65] Kupper LL, Hogan MD. Interaction in epidemiologic studies. Am J Epidemiol 1978; 108: 447-453.

[66] Kurland L. Pharmacogenetic Studies of Antihypertensive Treatment: With Special Reference to the Renin-Angiotensin-Aldosterone System. Ph.D. thesis, Department of Medical Sciences, Uppsala: Acta Universitatis Upsaliensis, 2001.

[67] Lagakos SW. Using auxiliary variables for improved estimates of survival time. Biometrics 1977; 33: 399-404.

[68] Langholz B, Rothman N, Wacholder S, Thomas DC. Cohort studies for characterizing measured genes. J Natl Cancer Inst Monogr 1999; 26: 39-42.

[69] Leoncini G, Viazzi F, Parodi D, Ratto E, Vettoretti S, Vaccaro V, Ravera M, Deferrari G, Pontremoli R. Mild renal dysfunction and cardiovascular risk in hypertensive patients. J Am Soc Nephrol 2004; 15: S88-S90.

[70] Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986; 73: 13-22.

[71] Libby P, Ridker PM. Novel inflammatory markers of coronary risk: Theory versus practice. Circulation 1999; 100: 1148-1150.

[72] Luan JA, Wong MY, Day NE and Wareham NJ. Sample size determination for studies of gene-environment interaction. Int J of Epidemiol 2001; 30: 1035-1040.

[73] Lubsen J, Kirwan B-A. Combined end points: Can we use them? Stat Med 2002; 21: 2959-2970.

[74] Lynch M, Walsh B. Genetics and Analysis of Quantitative Traits. Sunderland, MA: Sinauer Associates, Inc., 1998.

[75] Mackenzie T, Abrahamowicz M. Categorical markers at work in clinical trials as auxiliary variables: Applications to parametric and Cox's models, and the weighted log-rank family of test statistics. Can J Stat. In press.

[76] MacMahon B. Gene-environment interaction in human disease. J Psychiatr Res 1968; 6(suppl): 393-402.

[77] Maestri NE, Beaty TH, Hetmanski J, Smith EA, McIntosh I, Wyszynski DF, Liang KY, Duffy DL, VanderKolk C. Application of transmission disequilibrium tests to nonsyndromic oral clefts: Including candidate genes and environmental exposures in the models. Am J Med Genet 1997; 73: 337-344.

[78] Mahajan AP, Hogan JW, Snyder B, Kumarasamy N, Mehta K, Solomon S, Carpenter CC, Mayer KH, Flanigan TP. Changes in total lymphocyte count as a surrogate for changes in CD4 count following initiation of HAART: Implications for monitoring in resource-limited settings. J Acquir Immune Defic Syndr 2004; 36: 567-575.

[79] Marcus PM, Hayes RB, Vineis P, Garcia-Closas M, Caporaso NE, Autrup H, Branch RA, Brockmoller J, Ishizaki T, Karakaya AE, Ladero JM, Mommsen S, Okkels H, Romkes M, Roots I, Rothman N. Cigarette smoking, nacetyltransferase 2 acetylation status, and bladder cancer risk: A case-series meta-analysis of a gene-environment interaction. Cancer Epidemiol Biomarkers Prev 2000; 9: 461-467.

[80] Marian AJ, Safavi F, Ferlic L, Dunn JK, Gotto AM, Ballantyne CM. Interactions between angiotensin-I converting enzyme insertion/deletion polymorphism and response of plasma lipids and coronary arterosclerosis to treatment with fluvastatin: The lipoprotein and coronary atherosclerosis study. J Am Coll Cardiol 2000; 35: 89-95.

[81] Matsuo K, Hamajima N, Shinoda M, Hatooka S, Inoue M, Takezaki T, Tajima K. Gene-environment interaction between an aldehyde dehydrogenase-2 (ALDH2) polymorphism and alcohol consumption for the risk of esophageal cancer. Carcinogenesis 2001; 22: 913-916.

[82] Meehl PE. Comorbidity and taxometrics. Clin Psychol Sci Pract 2001; 8: 507-519.

[83] Murray S, Tsiatis AA. Using auxiliary time-dependent covariates to recover information in nonparametric testing with censored data. Lifetime Data Anal 2001; 7: 125-141.

[84] O'Brien TR, McDermott DH, Ioannidis JP, Carrington M, Murphy PM, Havlir DV, Richman DD. Effect of chemokine receptor gene polymorphisms on the response to potent antiretroviral therapy. AIDS 2000; 14: 821-826.

[85] Ottman R. An epidemiologic approach to gene-environment interaction. Genet Epidemiol 1990; 7: 177-185.

[86] Ottman R. Gene-environment interaction and public health. Am J Human Genetics 1995; 56: 821-823.

[87] Ottman R. Gene-environment interaction: Definitions and study designs. Prev Med 1996; 25: 764-770.

[88] Pares A, Deulofeu R, Gimenez A, Caballeria L, Bruguera M, Caballeria J, Ballesta AM, Rodes J. Serum hyaluronate reflects hepatic fibrogenesis in alcoholic liver disease and is useful as a marker of fibrosis. Hepatology 1996; 24: 1399-1403.

[89] Prentice RL. Surrogate endpoints in clinical trials: Definition and operational criteria. Stat Med 1989; 8: 431-440.

[90] Rachet B, Abrahamowicz M, Sasco AJ, Siemiatycki J. Estimating the distribution of lag in the effect of short-term exposures and interventions: Adaptation of a non-parametric regression spline model. Stat Med 2003; 22: 2335-2363.

[91] Rinehart BK, Terrone DA, May WL, Magann EF, Isler CM, Martin Jr JN. Change in platelet count predicts eventual maternal outcome with syndrome of hemolysis, elevated liver enzymes and low platelet count. J Matern Fetal Med. 2001; 10: 28-34.

[92] Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. Biometrics 2000; 56: 779-788.

[93] Rothman KJ. Interactions between causes. In: Modern Epidemiology. Boston, MA: Little, Brown, 1986, pp. 311-326.

[94] Rothman KJ, Boice JD. Epidemiologic Analysis with a Programmable Calculator. Chestnut Hill, MA: Epidemiology resources, 1982.

[95] Rothman KJ, Greenland S, Walker AM. Concepts of interaction. Am J Epidemiol 1980; 112: 467-470.

[96] Saracci R. Interaction and synergism. Am J Epidemiol 1980; 112: 465-466.

[97] Saunders CL, Barrett JH. Flexible matching in case-control studies of gene-environment interactions. Am J Epidemiol 2004; 159: 17-22.

[98] Schaid D. Case-parent designs for gene-environment interactions. Genet Epidemiol 1999; 16: 261-273.

[99] Schlesselman JJ. Case-control Studies: Design, Conduct, Analysis. New York: Oxford University Press, 1982.

[100] Self SG, Mauritsen RH, Ohara J. Power calculations for likelihood ratio tests in generalized linear models. Biometrics 1992; 48: 31-39.

[101] Seminara D, Obrams GI. Genetic epidemiology of cancer: A multidisciplinary approach. Genet Epidemiol 1994; 11: 235-254.

[102] Shi YL. Stochastic dynamic model of SARS spreading. Chin Sci Bull, 2003; 48: 1287-1292.

[103] Spence MA. Genetic Epidemiology. In: Elston R, Olson J, Palmer L (eds). Biostatistical Genetics and Genetic Epidemiology. New York: John Wiley & Sons, 2002, pp. 336-340.

[104] Spielman R, Ewens W. A sibship test for linkage in the presence of association: The sib transmission/disequilibrium test. Am J of Hum Genetics 1998; 62: 450-458.

[105] Stürmer T, Brenner H. Potential gain in efficiency and power to detect gene-environment interactions by matching in case-control studies. Genet Epidemiol 2000; 18: 63-80.

[106] Stürmer T, Brenner H. Flexible matching strategies to increase power and efficiency to detect and estimate gene-environment interactions in case-control studies. Am J Epidemiol 2002; 155: 593-602.

[107] Suissa S. Binary methods for continuous outcomes: A parametric alternative. J Clin Epidemiol 1991; 44: 241-248.

[108] Susser M. Causal Thinking in the Health Sciences: Concepts and Strategies in Epidemiology. New York: Oxford University Press, 1973.

[109] Talmud PJ, Hawe E, Miller GJ. Analysis of gene-environment interaction in coronary artery disease: Lipoprotein lipase and smoking as examples. Ital Heart J 2002; 3: 6-9.

[110] Temple RJ. A regulatory authority's opinion about surrogate endpoints. In: Nimmo WS, Tucker GT (eds). Clinical Measurement in Drug Evaluation. New York: Wiley, 1995.

[111] Tiret L, Abel L, Rakotovao R. Effect of ignoring genotype- environment interaction on segregation analysis of quantitative traits. Genet Epidemiol 1993; 10: 581-586.

[112] Tourian A, Sidbury JB. Phenylketonuria and hypophenylalaninemia. In: Stanbury JB, Wyngaarden JB, Fredrickson DS, Goldstein JL, Brown MS (eds). The Metabolic Basis of Inherited Diseases, 5th Ed. New York: McGraw-Hill, 1983, pp. 270-286.

[113] Umbach DM. Invited Commentary: On studying the joint effects of candidate genes and exposures. Am J Epidemiol 2000; 152: 701-703.

[114] Umbach DM, Weinberg CR. The use of case-parent triads to study joint effects of genotype and exposure. Am J Hum Genet 2000, 66: 251-261.

[115] Vickers AJ. How many repeated measures in repeated measures designs? Statistical issues for comparative trials. BMC Med Res Methodol 2003; 3: 22.

[116] Walter SD, Holford TR. Additive, multiplicative, and other models for disease risks. Am J Epidemiol 1978; 108: 341-346.

[117] Wang Z, Chen C, Niu T, Wu D, Yang J, Wang B, Fang Z, Yandava CN, Drazen JM, Weiss ST, Xu X. Association of asthma with beta(2)- adrenergic receptor gene polymorphism and cigarette smoking. Am J Respir Crit Care Med 2001; 163: 1404-1409.

[118] Weinburg C, Umbach D. Choosing a retrospective design to assess joint genetic and environmental contributions to risk. Am J Epidemiol 2000; 152: 197-203.

[119] White JE. A two stage design for the study of the relationship between a rare exposure and a rare disease. Am J Epidemiol 1982; 115: 119-128.

[120] Witte JS. Gene-environment interaction. In: Elston R, Olson J, Palmer L (eds). Biostatistical Genetics and Genetic Epidemiology. New York: John Wiley & Sons, 2002, pp. 297-299.

[121] Witte JS, Gauderman W, Thomas D. Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: Basic family designs. Am J Epidemiol 1999; 149: 693-705.

[122] Wong MY, Day NE, Luan JA, and Wareham NJ. The detection of gene-environment interaction for continuous traits: Should we deal with measurement error by bigger studies or better measurement? Int J of Epidemiol 2003; 32: 51-57.

[123] Wong MY, Day NE, Luan JA, and Wareham NJ. Estimation of magnitude in gene-environment interactions in the presence of measurement error. Stat Med 2004; 23: 987-998.

[124] Wu AHB, Apple FS, Gibler WB, Jesse RL, Warshaw MM, Valdes R. National Academy of Clinical Biochemistry Standards of Laboratory Practice: Recommendations for use of cardiac markers in coronary artery diseases. Clin Chem 1999; 45: 1104-1121.

[125] Wyse DG, Slee A, Epstein AE, Gersh BJ, Rocco Jr T, Vidaillet H, Volgman A, Weiss R, Shemanski L, Greene HL, and the AFFIRM Investigators. Alternative endpoints for mortality in studies of patients with atrial fibrillation: The AFFIRM study experience. Heart Rhythm 2004; 1: 531-537.

[126] Yang Q, Khoury MJ, Flanders WD. Sample size requirements in case-only designs to detect gene-environment interaction. Am J Epidemiol 1997; 146: 713-720.

[127] Yang Q, Khoury MJ, Sun F, Flanders WD. Case-only design to measure gene-gene interaction. Epidemiol 1999; 10: 167-170.

[128] Zhou W, Thurston SW, Liu G, Xu LL, Miller DP, Wain JC, Lynch TJ, Su L, Christiani DC. The interaction between microsomal epoxide hydrolase polymorphisms and cumulative cigarette smoking in different histological subtypes of lung cancer. Can Epi Bio Prev 2001; 10: 461-466.

APPENDIX-A

S-PLUS Codes for the Simulation Study

This Appendix contains the S-Plus program codes for simulating the hypothetical scenarios described in this thesis. The entire program for each scenario was divided into modules according to type of outcome or parameter combinations. However, only selected program codes are given here for Scenario I, Scenario II and Scenario III due to space constraints.

A-1 Simulation program for Scenario I

```
sim.SI<- function(N, S) {
# N=number of data sets generated, S= sample size
I <- 0
g1.1 <- rep(0,N)
g1.2 <- rep(0, N)
g2.1 <- rep(0, N)
⋮
g2.4 <- rep(0, N)
b <- matrix(c(log(0.1), log(3), log(1.5), log(3), log(2), log(0.1), log(3), log(1.5), log(2),
log(2), log(0.1), log(3), 0, log(2), log( 2), log(0.1), log(3), 0, 0, log(2)), 4, 5, byrow =
T)
repeat {
# Generation of data for each sample
I <- I + 1
G <- rep(0, S)
```

```
E <- rep(0, S)
X <- rep(0, S)
d <- rep(0, S)
l <- rep(0, S)
e <- rep(0, S)
G <- rbinom(S, 1, 0.2)
E <- rbinom(S, 1, 0.3)
for(i in 1:S) {
    d[i] <- rnorm(1, 0, 0.3)
    if (E[i] == 0 && G[i] == 0) {
    X[i] <- rnorm(1, 0, 1) } # true X
    else if (E[i] == 0 && G[i] == 1) {
    X[i] <- rnorm(1, 0.5, 1) } # true X
    else if(E[i] == 1 && G[i] == 0) {
    l[i] <- rnorm(1, 0, 1)
    e[i] <- exp(l[i])
    X[i] <- rnorm(1, 0, 1) + e[i] * 0.25 } # true X
    else {
    l[i] <- rnorm(1, 0, 1)
    e[i] <- exp(l[i])
    X[i] <- rnorm(1, 0.5, 1) + e[i] * 0.5 } # true X
    }
X1 <- X + d # observed X
#-----------------------
# Program for linear models
#-----------------------
lin.1<-linreg(X, G, E) # Model SI.1(i)
g1.1[I]<-lin.1$g
lin.2<-linreg(X1, G, E) # Model SI.1(ii)
g1.2[I]<-lin.2$g
# Generation of binary outcomes
```

```
XX <- matrix(0, 5, S)
XX[1, ] <- rep(1, S)
XX[2, ] <- G
XX[3, ] <- E
XX[4, ] <- G * E
XX[5, ] <- X
bXX <- bb %*% XX
L1 <- exp(bXX[1, ])
⋮
L4 <-exp(bXX[4, ])
pi.1 <- L1/(1 + L1)
⋮
pi.4 <- L4/(1 + L4)
Y1 <- rep(0, S)
⋮
Y4 <- rep(0, S)
for(j in 1:S) {
    Y1[j] <- rbinom(1, 1, pi.1[j])
    Y2[j] <- rbinom(1, 1, pi.2[j])
    Y3[j]<- rbinom(1, 1, pi.3[j])
    Y4[j] <-rbinom(1, 1, pi.4[j])
    }
#-----------------------------------
# Program for logistic regression models SI.2
#-----------------------------------
log.1<-logreg(Y1, G, E)
g2.1[I]<-log.1$g
⋮
log.4<-logreg(Y4, G, E)
g2.4[I]<-log.4$g
if(I == N)
```

```
break
}
proplin1 <- sum(g1.1)/N
proplin2 <- sum(g1.2)/N
proplog1 <- sum(g2.1)/N
⋮
proplog4 <- sum(g2.4)/N
cat("Results for Multiple Linear Regression Model SI.1(i): \n")
print(proplin1)
cat("Results for Multiple Linear Regression Model SI.1(ii): \n")
print(proplin2)
cat("Results for Multiple Logistic Regression Model SI.2(i): \n")
print(proplog1)
⋮
cat("Results for Multiple Logistic Regression Model SI.2(iv): \n")
print(proplog4)
}
```

linreg:

```
function(X, G, E){
xlreg <- glm(formula = X ~ G* E, family = gaussian)
b <- coef(xlreg)
vc <- vcov(xlreg)
vcd <- diag(vc)
std <- sqrt(vcd)
t <- b/std
if(abs(t[4]) >= 1.96) g = 1
else g = 0
list(g=g) }
```

logreg:

```
function(Y, G, E){
xlreg<- glm(formula = Y ~ G * E, family = binomial(link = logit), maxit = 50,
epsilon = 0.0001)
b <- coef(xlreg)
vc <- vcov(xlreg)
vcd <- diag(vc)
std <- sqrt(vcd)
wald <- b/std
if(wald[4] <- 0) pval <- 2 * pnorm(wald[4], 0, 1)
else pval <- 2 * (1 - pnorm(wald[4], 0, 1))
if(pval <= 0.05) g = 1
else g = 0
list(g = g) }
```

A-2 Simulation program for Scenario II, Case S1

```
sim.SIIa<-function(N, S) {
# N=number of data sets generated, S= sample size
I <- 0
b <- c(log(0.1), log(3), log(1.5), log(3), log(1.5)) # Combination 1, Table 3.2
###############BLOCK 1#################
SS <- matrix(c(1, 1, 0.9, 0.9, 0.7, 0.9, 0.9, 0.7, 0.7, 0.7), 2, 5)
theta.R <- c(0, 0.5)
theta.Y <- c(0.5, 1, 1.5)
sigma <- c(0.05, 0.5, 1)
glin.1 <- rep(0, N)
glin.2 <- glin.3 <- ... <- ... <-glin.18 <- glin.1
glog.1 <- rep(0, N)
glog.2 <- glog.3 <- glog.4 <- glog.5 <- glog.1
################ END BLOCK ################
set.seed(58670)
repeat {
```

```
# Generation of data for each sample
I <- I + 1
################BLOCK 2###################
G <- rep(0, S)
E <- rep(0, S)
R <- rep(0, S)
G <- rbinom(S, 1, 0.2)
E <- rbinom(S, 1, 0.3)
R <- rnorm(S, 0, 1)
############### END BLOCK ################
XX <- matrix(0, 5, S)
XX[1, ] <- rep(1, S)
XX[2, ] <- G
XX[3, ] <- E
XX[4, ] <- G * E
XX[5, ] <- R
################BLOCK 3###################
bXX <- b %*% XX
L <- exp(bXX)
pi <- L/(1 + L)
Y <- rep(0, S)
# Generation of true Y using combination 1
for(i in 1:S) {
   Y[i] <- rbinom(1, 1, pi[i])
   }
# Generation of observed Y
Y11e <- sim3.1(S, Y, SS[, 1])
⋮
Y15e <- sim3.1(S, Y, SS[, 5])
# Generation of true X
X11 <- sim3.2(S, Y, R, theta.R[1], theta.Y[1])
```

```
X12 <- sim3.2(S, Y, R, theta.R[1], theta.Y[2])
X13 <- sim3.2(S, Y, R, theta.R[1], theta.Y[3])
X21 <- sim3.2(S, Y, R, theta.R[2], theta.Y[1])
X22 <- sim3.2(S, Y, R, theta.R[2], theta.Y[2])
X23 <- sim3.2(S, Y, R, theta.R[2], theta.Y[3])
# Generation of observed X for three values of sigma
X111e <- sim3.3(S, X11$X, sigma[1])
X112e <- sim3.3(S, X11$X, sigma[2])
X113e <- sim3.3(S, X11$X, sigma[3])
X121e <- sim3.3(S, X12$X, sigma[1])
X122e <- sim3.3(S, X12$X, sigma[2])
X123e <- sim3.3(S, X12$X, sigma[3])
X131e <- sim3.3(S, X13$X, sigma[1])
X132e <- sim3.3(S, X13$X, sigma[2])
X133e <- sim3.3(S, X13$X, sigma[3])
X211e <- sim3.3(S, X21$X, sigma[1])
X212e <- sim3.3(S, X21$X, sigma[2])
X213e <- sim3.3(S, X21$X, sigma[3])
X221e <- sim3.3(S, X22$X, sigma[1])
X222e <- sim3.3(S, X22$X, sigma[2])
X223e <- sim3.3(S, X22$X, sigma[3])
X231e <- sim3.3(S, X23$X, sigma[1])
X232e <- sim3.3(S, X23$X, sigma[2])
X233e <- sim3.3(S, X23$X, sigma[3])
############### END BLOCK ###############
# Multiple linear regression models (SII.1(i))
################ BLOCK 4 ################
linreg.1 <- sim3.4(X111e$Xe, G, E, R)
glin.1[I] <- linreg.1$g
linreg.2 <- sim3.4(X112e$Xe, G, E, R)
glin.2[I] <- linreg.2$g
```

```
⋮
linreg.17 <- sim3.4(X232e$Xe, G, E, R)
glin.17[I] <- linreg.17$g
linreg.18 <- sim3.4(X233e$Xe, G, E, R)
glin.18[I] <- linreg.18$g
# Multiple logistic regression models (SII.1(ii))
logreg.1 <- sim3.5(Y11e$Ye, G, E, R)
glog.1[I] <- logreg.1$g
⋮
logreg.5 <- sim3.5(Y15e$Ye, G, E, R)
glog.5[I] <- logreg.5$g
############### END BLOCK ##################
if(I == N)
break
}
# Computation of estimated power for linear regression models
################ BLOCK 5 ##################
proplin.1 <- sum(glin.1)/N
proplin.2 <- sum(glin.2)/N
⋮
proplin.17 <- sum(glin.17)/N
proplin.18 <-sum(glin.18)/N
cat("Power for Multiple Linear Regression Models using combination 1: \ n")
cat("theta_R=0, theta_Y=0.5, sigma=0.05: \ n")
print(proplin.1)
cat("theta_R=0, theta_Y=0.5, sigma=0.5: \ n")
print(proplin.2)
cat("theta_R=0, theta_Y=0.5, sigma=1: \ n")
print(proplin.3)
⋮
cat("theta_R=0.5, theta_Y=1.5, sigma=1:\n")
```

```
print(proplin.18)
# Computation of estimated power for logistic regression models
proplog.1 <- sum(glog.1)/N
⋮
proplog.5 <- sum(glog.5)/N
cat("Power for Logistic Regression Models using combination 1: \n")
cat("sensitivity=1, specificity=1:\n")
print(proplog.1)
⋮
cat("sensitivity=0.7, specificity=0.7:—\n")
print(proplog.5)
################ END BLOCK ################
}
```

sim3.1:

```
function(S, Y, nu) {
# generation of observed Y from true Y
Ye <- rep(0, S)
v <- rep(0, S)
for(i in 1:S) {
    if(Y[i] == 1) {
    v[i] <- rbinom(1, 1, nu[1])
    if(v[i] == 1) Ye[i] <- 1
    else Ye[i] <- 0
    }
    else {
    v[i] <- rbinom(1, 1, nu[2])
    if(v[i] == 1) Ye[i] <- 0
    else Ye[i] <- 1
    }
}
```

```
list(Ye = Ye)
}
```

sim3.2:

```
function(S, Y, R, theta.R, theta.Y) {
# generation of true X
mu <- rep(0, S)
X <- rep(0, S)
for(i in 1:S) {
    if(Y[i] == 0) {
    mu[i] <- theta.R * R[i]
    X[i] <- rnorm(1, mu[i], 1)
    }
    else {
    mu[i] <- theta.Y + theta.R * R[i]
    X[i] <- rnorm(1, mu[i], 1)}
}
list(X = X)
}
```

sim3.3:

```
# generation of observed X
function(S, X, sigma) {
Xe <- rep(0, S)
for(i in 1:S) {
    Xe[i] <- X[i] + rnorm(1, 0, sigma)
}
list(Xe = Xe)
}
```

Note: Functions **sim3.4** and **sim3.5** are similar to the functions **linreg** and **logreg**

for fitting linear regression and logistic regression models respectively in Scenario I, except for the input variables. Three other modules to estimate power for the three remaining combinations of parameter values in Table 3.2 have not been shown. These functions are identical to the above module except for changes to the vector **b**.

A-3 Simulation program for Scenario II, Case S2

```
sim.SIIb<-function(N, S) {
# N=number of data sets generated, S= sample size
I <- 0
# Combination 1, Table 3.5
b <- c(log(0.1), log(3), log(1.5), log(3), log(1.5), log(1.3), log(1.5), log(1.75))
...
enter BLOCK 1 lines here
...
dglin.1 <- rep(0, N)
dglin.2 <- dglin.3 <-... <- dglin.18 <- dglin.1
set.seed(58670)
repeat {
I <- I + 1
...
enter BLOCK 2 lines here
...
S0 <- rep(0, S)
SE <- rep(0, S)
SGE <- rep(0, S)
S0 <- rnorm(S, 0, 1)
mu <-rep(0, S)
for(i in 1:S) {
    if(E[i] == 0) {SE[i] <- rnorm(1, -0.3, 1)}
    else {SE[i] <- rnorm(1, 0.3, 1)}
    mu[i] <- -0.4 + 0.3 * G[i] + 0.5 *E[i]
```

```
    SGE[i] <- rnorm(1, mu[i], 1)
}
XX <-matrix(0, 8, S)
XX[1, ] <- rep(1, S)
XX[2, ] <- G
XX[3, ] <- E
XX[4, ] <- G * E
XX[5, ] <- R
XX[6, ] <- S0
XX[7, ] <- SE
XX[8, ] <- SGE
...
enter BLOCK 3 lines here
...
delxt <- sim5.1(S, G, E) # generation of true delta_ X
# generation of true X at t=1
Xt11 <- X11$X + delxt$delx
Xt12 <-X12$X + delxt$delx
Xt13 <- X13$X + delxt$delx
Xt21 <- X21$X + delxt$delx
Xt22 <- X22$X + delxt$delx
Xt23 <- X23$X + delxt$delx
# generation of observed X at t=1
Xt111e <- sim5.2(S, Xt11, sigma[1])
Xt112e <- sim5.2(S, Xt11, sigma[2])
Xt113e <- sim5.2(S, Xt11, sigma[3])
⋮
Xt231e <- sim5.2(S, Xt23, sigma[1])
Xt232e <- sim5.2(S, Xt23, sigma[2])
Xt233e <- sim5.2(S, Xt23, sigma[3])
# Generation of outcome d
```

```
d1 <-(Xt111e$X1e - X111e$Xe)/delxt$t
d2 <- (Xt112e$X1e - X112e$Xe)/delxt$t
⋮
d17 <- (Xt232e$X1e - X232e$Xe)/delxt$t
d18 <-(Xt233e$X1e - X233e$Xe)/delxt$t
# models SII.2(i) and SII.2(iii)
...
enter BLOCK 4 lines here
...
# Multiple linear regression with d as dependent variable (SII.2(ii))
dlinreg.1 <- sim3.4(d1, G, E, R)
dglin.1[I] <- dlinreg.1$g
dlinreg.2 <- sim3.4(d2, G, E, R)
dglin.2[I] <-dlinreg.2$g
⋮
dlinreg.17 <- sim3.4(d17, G, E, R)
dglin.17[I] <-dlinreg.17$g
dlinreg.18 <- sim3.4(d18, G, E, R)
dglin.18[I] <- dlinreg.18$g
if(I == N)
break
}
# Computation of estimated power for linear regression models with X as dependent
variable, and for logistic regression models
...
enter BLOCK 5 lines here
...
# Computation of estimated power for linear regression models with d as dependent
variable
proplin.1 <- sum(dglin.1)/N
proplin.2 <- sum(dglin.2)/N
```

```
⋮
proplin.17 <- sum(dglin.17)/N
proplin.18 <- sum(dglin.18)/N
cat("Power for Multiple Linear Regression Models using combination 1 and d as de-
pendent variable: \n")
cat("theta_R=0, theta_Y=0.5, sigma=0.05: \n")
print(proplin.1)
cat("theta_R=0, theta_Y=0.5, sigma=0.5: \n")
print(proplin.2)
⋮
cat("theta_R=0.5, theta_Y=1.5, sigma=0.5: \n")
print(proplin.17)
cat("theta_R=0.5, theta_Y=1.5, sigma=1: \n")
print(proplin.18)
}
```

sim5.1:

```
function(S, G, E) {
g <- c(0.2, 0.1, 0.15)
x <- matrix(0, 3, S)
x[1, ] <- rep(1, S)
x[2, ] <- E
x[3, ] <- G * E
t <- rep(0, S)
for(i in 1:S) {
    t[i] <- runif(1, 0.5, 1.5)
}
delx <- (g %*% x) * t
list(delx = delx, t = t)
}
```

sim5.2:

```
function(S, X, sigma) {
# generation of observed X at time t=1
X1e <- rep(0, S)
for(i in 1:S) {
X1e[i] <- X[i] + rnorm(1, 0, sigma)
}
list(X1e = X1e)
}
```

Note: Program modules for the three remaining combinations of values of parameter vector **b** in Table 3.5 are not shown.

A-4 Simulation program for Scenario III

(a) Outcome $X^*(t_0)$ (Model SIII(i))

```
sim.SIIIa<-function(N, Sm) {
# N=number of data sets generated, Sm=sample size
I <- 0
glin1.1 <- rep(0, N)
glin1.2 <- glin1.3 <- ... <- glin1.6 <- glin1.1
set.seed(456)
repeat {
I <- I + 1
# Initialization of vectors
################ BLOCK 1 ###################
G <- rep(0, Sm)
E1 <- rep(0, Sm)
S1 <- rep(0, Sm)
```

```
X01 <- rep(0, Sm)
X02 <- rep(0, Sm)
b1 <- rep(0, Sm)
b2 <- rep(0, Sm)
X02 <- rep(0, Sm)
X02 <-rep(0, Sm)
G <- rbinom(Sm, 1, 0.3)
E1 <- rbinom(Sm, 1, 0.6)
################ END BLOCK ###################
# generation of current age
################ BLOCK 2 ###################
Z <- runif(Sm, 40, 60)
for(i in 1:Sm) {
   if(G[i] == 0) {
   # generation of baseline X values
   X01[i] <- rnorm(1, 60, 10)
   X02[i] <- rnorm(1, 60, 10)
   b1[i] <- rnorm(1, 0.2, 0.05)
   b2[i] <- rnorm(1, 0.2, 0.05)}
   else {
   X01[i] <- rnorm(1, 60, 10)
   X02[i] <- rnorm(1, 65, 10) # del_X=5
   b1[i] <- rnorm(1, 0.2, 0.05)
   b2[i] <- rnorm(1, 0.3, 0.05)} # del_G=0.1
# generation of age of first exposure
   if(E1[i] == 0) {S1[i] == 0}
   else {S1[i] < − runif(1, 20, 40)}
   }
tt <- sim6.1(Sm, Z) # repeated measurement times and missing values
u <- tt$u
u[, 1] <- 0
```

```
u[u > 0] <- NA # indicator for missing assessment times
c1<-sign(tt$t >= S1)
E <- E1 * c1
S <- S1 * c1
c2 <- sign(E == 0)
# generation of X at age before first exposure
XX0t1 <- X01 + b1 * tt$t # for del_X=0
XX0t2 <- X01 + b2 * tt$t # for del_X=0
#XX0t1 <- X02 + b1 * tt$t # for del_X=5
#XX0t2 <- X02 + b2 * tt$t # for del_X=5
XX0u1 <- XX0t1 + u
XX0u2 <- XX0t2 + u
X0u1 <- XX0u1 * c2
X0u2 <- XX0u2 * c2
c3 <-sign(E == 1)
# generation of X at age at first exposure
XXS1 <- X01 + b1 * S # for del_X=0
XXS2 <- X01 + b2 * S # for del_X=0
#XXS1 <- X02 + b1 * S # for del_X=5
#XXS2 <- X02 + b2 * S # for del_X=5
bb11 <- b1 + 0.03 * E + 0 * G * E
bb12 <- b1 + 0.03 * E + 0.2 * G * E
bb13 <- b1 + 0.03 * E + 0.4 * G * E
bb21 <- b2 + 0.03 * E + 0 * G * E
bb22 <- b2 + 0.03 * E + 0.2 * G * E
bb23 <- b2 + 0.03 * E + 0.4 * G * E
# generation of X at ages after exposure
XX1t11 <- XXS1 + bb11 * (tt$t - S)
XX1t12 <- XXS1 + bb12 * (tt$t - S)
XX1t13 <- XXS1 + bb13 * (tt$t - S)
XX1t24 <- XXS2 + bb21 * (tt$t - S)
```

```
XX1t25 <- XXS2 + bb22 * (tt$t - S)
XX1t26 <- XXS2 + bb23 * (tt$t - S)
XX1u11 <- XX1t11 + u
XX1u12 <- XX1t12 + u
XX1u13 <- XX1t13 + u
XX1u24 <- XX1t24 + u
XX1u25 <-XX1t25 + u
XX1u26 <- XX1t26 + u
X1u1 <- XX1u11 * c3
X1u2 <- XX1u12 * c3
X1u3 <- XX1u13 * c3
X1u10 <- XX1u24 * c3
X1u11 <- XX1u25 * c3
X1u12 <- XX1u26 * c3
# values of X at different assessment times for 6 combinations
#of parameters del_G and del_GE
Xt1 <- X0u1 + X1u1
Xt2 <-X0u1 + X1u2
Xt3 <- X0u1 + X1u3
Xt10 <- X0u2 + X1u10
Xt11 <- X0u2 + X1u11
Xt12 <- X0u2 + X1u12
# generation of observed values of X
e1 <- matrix(rnorm(Sm * 20, 0, 0.1), nrow = Sm )
Xte1 <- Xt1 + e1
⋮
e6 <- matrix(rnorm(Sm * 20, 0, 0.1), nrow = Sm )
Xte12 <- Xt12 + e6
############### END BLOCK ###################
# linear regression models using observed X
C1.1 <- sim3.4(Xte1[, 1], G, E[, 1], Z)
```

```
⋮
C1.6 <- sim3.4(Xte12[, 1], G, E[, 1], Z)
# linear regression models using true X
C2.1 <- sim3.4(Xt1[, 1], G, E[, 1], Z)
⋮
C2.6<- sim3.4(Xt12[, 1], G, E[, 1], Z)
glin1.1[I] <- C1.1$g
glin1.2[I] <- C1.2$g
⋮
glin2.5[I] <- C2.5$g
glin2.6[I] <- C2.6$g
if(I == N)
break
}
# Computation of estimated power
proplin1.1 <- sum(glin1.1)/N
proplin1.2 <- sum(glin1.2)/N
⋮
proplin2.5 <- sum(glin2.5)/N
proplin2.6 <- sum(glin2.6)/N
cat("Power for Multiple Linear Regression Model using true X at current age as de-
pendent variable: \n")
cat("del_G=0, del_GE=0, X: \n")
print(proplin2.1)
⋮
cat("del_G=0.02, del_GE=0.40,X: \n")
print(proplin2.6)
cat("Power for Multiple Linear Regression Model using observed X at current age as
dependent variable: \n")
cat("del_G=0, del_GE=0, Xe: \n")
print(proplin1.1)
```

```
⋮
cat("del_G=0.02, del_GE=0.40,Xe: \n")
print(proplin1.6)
}
```

sim6.1:

```
function(S, Z) {
# generation of repeated measurement times and missing values
dt <- matrix(0, S, 20)
t <- matrix(0, S, 20)
u <- matrix(0, S, 20)
for(i in 1:S) {
    I <- 0
    t[i, 1] <-Z[i]
    dt[i, 1] <- NA
    u[i, 1] <- NA
    repeat {
    I <- I + 1
    dt[i, I + 1] <- runif(1, 0.5, 1.5)
    t[i, I + 1] <- t[i, I] - dt[i, I + 1]
    u[i, I + 1] <- rbinom(1, 1, 0.3)
    if(t[i, I + 1] < Z[i] - 6) {
    for(j in (I +1):20) {
        u[i, j] <- NA
        dt[i, j] <- NA
        t[i, j] <- NA}
        break
        }
    }
}
```

```
list(t = t, u = u)
}
```

(ii) Outcome $Y1$ (Model SIII(ii))

```
sim.SIIIb<-function(N, Sm) {
I <- 0
glog1.1 <- rep(0, N)
glog2.1 <- rep(0, N)
glog1.2 <- glog1.3 <- ... <- glog1.12 <-glog1.1
glog2.2 <- glog2.3 <- ... glog2.12 <- glog2.1
set.seed(456)
repeat {
I <- I + 1
...
enter BLOCK 1 lines here
...
thrsh1 <- matrix(0, 6, 2)
thrsh2 <- matrix(0, 6, 2)
# generation of current age
...
enter BLOCK 2 lines here
...
# determining thresholds T1 and T2
thrsh1[1, ] <- quantile(Xte1[, 1], c(0.7, 0.5))
⋮
thrsh1[6, ] <- quantile(Xte12[, 1], c(0.7, 0.5))
thrsh2[1, ] <- quantile(Xt1[, 1], c(0.7, 0.5))
⋮
thrsh2[6, ] <- quantile(Xt12[, 1], c(0.7, 0.5))
# generating binary outcome Y1
```

```
outcm1 <- sim6.2(Xte1[, 1], Xte2[, 1], Xte3[ , 1], Xte10[, 1], Xte11[, 1], Xte12[ , 1],
thrsh1)
outcm2 <- sim6.2(Xt1[, 1], Xt2[, 1], Xt3[ , 1], Xt10[, 1], Xt11[, 1], Xt12[ , 1], thrsh2)
# logistic regression models based on observed X
C2a.1 <- sim3.5(outcm1$y1, G, E[, 1], Z)
⋮
C2a.12 <- sim3.5(outcm1$y12, G, E[, 1], Z)
# logistic regression models based on true X
C2b.1 <- sim3.5(outcm2$y1, G, E[, 1], Z)
⋮
C2b.12 <- sim3.5(outcm2$y12, G, E[, 1], Z)
glog1.1[I] <- C2a.1$g
⋮
glog1.12[I] <- C2a.12$g
glog2.1[I] <- C2b.1$g
⋮
glog2.12[I] <- C2b.12$g
if(I == N) break
}
# Computation of estimated power
prop1.1 <- sum(glog1.1)/N
⋮
prop1.12 <- sum(glog1.12)/N
prop2.1 <- sum(glog2.1)/N
⋮
prop2.12 <- sum(glog2.12)/N
cat("Power for Logistic Regression Model using Y1 (based on true X) as dependent
variable: \n")
cat("del_G=0, del_GE=0,T1,X:\n")
print(prop2.1)
⋮
```

```
cat("del_G=0.02, del_GE=0.06,T2,X:\n")
print(prop2.12)
cat("Power for Logistic Regression Model using Y1 (based on observed X) as dependent variable: \n")
cat("del_G=0, del_GE=0,T1,Xe:\n")
print(prop1.1)
⋮
cat("del_G=0.02, del_GE=0.06,T2,Xe:\n")
print(prop1.12)
}
```

sim6.2

```
function(XZ1, XZ2, XZ3, XZ4, XZ5, XZ6, R) {
d11 <- sign(XZ1 > R[1, 1])
⋮
d16 <-sign(XZ6 > R[6, 1])
d21 <- sign(XZ1 > R[1, 2])
⋮
d26 <- sign(XZ6 > R[6, 2])
list(y1 = d11, y2 = d12, y3 = d13, y4 = d14, y5 = d15, y6 = d16, y7 = d21, y8 = d22, y9 = d23, y10 = d24, y11 = d25, y12 = d26)
}
```