Stony Brook University

The impact of genotype misclassification errors on the power to detect a genetic association and gene-environment interaction with Cox proportional hazards modeling

A Dissertation Presented by

Lin Tung

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Statistics)

Stony Brook University

August 2007

Stony Brook University

The Graduate School

Lin Tung

We, the dissertation committee for the above candidate for the Doctor of Philosophy degree, hereby recommend acceptance of this dissertation.

Stephen J. Finch, Dissertation Advisor
Professor, Applied Mathematics & Statistics

John Reinitz, Chairperson of Defense
Associate Professor, Applied Mathematics & Statistics

Hongshik Ahn
Associate Professor, Applied Mathematics & Statistics

Derek Gordon
Associate Professor, Department of Genetics
Rutgers, The State University of New Jersey

This dissertation is accepted by the Graduate School.

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

The impact of genotype misclassification errors on the power to detect a genetic association and gene-environment interaction with Cox proportional hazards modeling

by

Lin Tung

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Statistics)

Stony Brook University

2007

Genetic model parameters determine the power and sample size required to detect a genetic association and gene-environment interaction with Cox proportional hazards modeling in cohort genetic studies. Detecting a genetic indirect association is largely dependent on the difference in the allele frequencies between the underlying functional variant and the marker locus. Similar to case-control genetic association studies, the increase in sample size to detect a genetic indirect association is approximately $1/r^2$, where the linkage disequilibrium parameter r^2 is the square of the correlation between alleles in coupling at the disease and marker locus. Detecting a gene-environment interaction for dominant and recessive modes of inheritance for a direct association study is feasible for common disease allele frequencies ($p_d > 0.10$) and moderate effect sizes.

The impact of each genotyping misclassification error upon the increase in sample size required to maintain constant Type I and II error rates can be calculated mathematically through a first-order linear Taylor series expansion. We find that, consistent with previous genotyping errors research in case-control genetic association studies, any misclassification of the more common homozygote is the most deleterious in terms of increase in sample size (or equivalently loss of power) to detect a genetic indirect association. The total required sample size to detect a genetic indirect association in the presence of genotyping errors may be partitioned into the sample size required for a direct association study, the increase in sample size due to an indirect association study, and the increase in sample size due to genotyping errors. For high r^2 (typically $r^2 > 0.95$) and approximately equal allele frequencies, the increase in sample size due to genotyping error rates as low as 2% is larger than the increase in sample size due to an indirect association study. In the detection of a gene-environment interaction, any misclassification of a subject without the at-risk genotype as having the at-risk genotype is the most deleterious. Such errors require an indefinitely large increase in required sample size as the SNP minor allele frequency approaches 0.

I dedicate this work to my husband, Omar Guey, and my mother, who have both provided me with unconditional tremendous love and support.

Table of Contents

List of Figures vi

List of Tables vii

1. Background 1

2. Methods 5

3. Genetic association results 16

4. G × E interaction only model results 29

5. Full G × E interaction model results 39

6. Discussion and Conclusion 49

References 51

Appendix 1 58

Appendix 2 60

Appendix 3 62

List of Figures

3.1a	Required sample size for 80% power	25
3.1b	Required sample size for 95% power	25
3.2a	% increase due to an indirect association study, $p_d > p_A$	26
3.2b	% increase due to an indirect association study, $p_d < p_A$	26
3.3	Distribution of relative error $R(\tilde{\varepsilon})$ for genotyping error rates 0.5% to 2%	27
3.4	Simulation results (genetic association model)	28
4.1a	Taylor series approximation for dominant MOI	36
4.1b	Taylor series approximation for recessive MOI	36
4.2a	Sum of %MSSN coefficients for dominant MOI	37
4.2b	Sum of %MSSN coefficients for recessive MOI	37
4.3	Simulation results (G × E interaction only model)	38
5.1a	Relative error for dominant MOI	44
5.1b	Relative error for recessive MOI	44
5.2a	Taylor series approximation for dominant MOI	45
5.2b	Taylor series approximation for recessive MOI	45
5.3a	Sum of %MSSN coefficients for dominant MOI	46
5.3b	Sum of %MSSN coefficients for recessive MOI	46
5.4	Simulated powers of the full interaction model versus the interaction only model	47
5.5	Simulation results (Full G × E interaction model)	48

List of Tables

3.1	Tabulation of sample size for an indirect genetic association study	21
3.2	Relationship between N_I / N_D and r^2 for a genetic association study	22
3.3	Bounds of %MSSN coefficients (genetic association model)	23
3.4	Example design settings (genetic association model)	24
4.1	Example design settings (G × E interaction only model)	33
4.2a	Relative error for dominant MOI with $p_d = 0.7$ and $\Delta = 1.6$	34
4.2b	Relative error for recessive MOI with $p_d = 0.2$ and $\Delta = 1.6$	34
4.3	Bounds of %MSSN coefficients (G × E interaction only model)	35
5.1	Example design settings (Full G × E interaction model)	42
5.2	Bounds of %MSSN coefficients (Full G × E interaction model)	43

Chapter 1 – Background

Survival analysis is a common statistical technique used to detect genetic associations and gene-environment ($G \times E$) interactions in prospective genetic cohort studies. The Cox proportional hazards (PH) model is the most widely-used survival analysis regression model in a proportional hazards setting. This model is of particular importance when studying the relationship between genetic and environmental factors on disease risk and progression. Among the most influential genetic association findings using Cox PH modeling are an association between P-glycoprotein expression and survival in a sample of adolescent cancer patients with neuroblastoma (Chan, et al. 1991), an association between E-cadherin expression and survival in a sample of prostate cancer patients (Umbas, et al. 1994), and an association between the chemokine reception 5 (CKR5) genotype and HIV-1 disease progression (Dean, et al. 1996). Example $G \times E$ interaction findings using this model are an association between the cytochrome-p450ia1 (CY1A1) genotype and smoking status on lung cancer survival length (Goto, et al. 1996), an association between apolipoprotein E gene (APOE) epsilon 4 carriers and smoking status on cardiovascular disease risk (Talmud, et al. 2005), and a moderation of the alcohol dehydrogenase 1C (ADH1C) genotype on the relationship between alcohol consumption and coronary heart disease (Younis, et al. 2005).

$G \times E$ interaction studies characterize the interplay between genetic and environmental effects. Understanding the relationship between genes and environment with respect to disease risk and progression enables improvement in the estimation of genetic and environmental main effects by consideration of their joint interaction, insight into the genetic and environmental effects on biological pathways pertinent to disease, and facilitates the design of new treatments customized for individuals based on their genotypes (Hunter 2005). Among the most prominent $G \times E$ interaction findings are a moderation of the influence of stressful live events on depression by a functional polymorphism in the promoter region of the serotonin transporter gene (5-HTT) (Caspi, et al. 2003), an association between polymorphisms in the N-acetyltransferase (NAT) genes with red meat consumption on colorectal cancer risk (Chen, et al. 1998), and an association between APOE epsilon 4 carriers and estrogen usage on cognitive decline among a cohort Medicare-eligible women (Yaffe, et al. 2000).

Ottman (1996) summarizes the current definitions and study designs of $G \times E$ interaction for genetic epidemiology. The genetic factor (*G*) is a broadly defined high-risk genotype, such as an autosomal gene or a polygenic model. Similarly, the environmental factor (*E*) is broadly defined to include an expansive range of environmental risks, such as an exposure (physically, chemically, or biologically), a behavior pattern, or a life event. Assuming dichotomous independent and dependent variables, the $G \times E$ interaction may be estimated as an odds ratio (OR) or relative risk (RR) of disease risk and may be additive or multiplicative, depending on the scale of measurement. The null hypothesis for an additive $G \times E$ interaction is $OR_{G\times E} = OR_G + OR_E - 1$ or $RR_{G\times E} = RR_G + RR_E - 1$ depending on the study design. The null hypothesis for a multiplicative $G \times E$ interaction is $OR_{G\times E} = OR_G \times OR_E$ or $RR_{G\times E} = RR_G \times RR_E$ depending on the study design. Ottman (1996) further describes differing models of $G \times E$ interaction upon disease risk. The models include a direct

effect of genotype on disease risk exacerbated by the environment (Model I), a direct effect of the environment on disease risk exacerbated by the genotype (Model II), a direct effect of $G \times E$ interaction on disease risk but no individual effects of genotype or environment (Model III), and a direct effect of both genotype and environment in addition to their interaction upon disease risk (Model IV).

Power and sample size calculations are critical in the design of $G \times E$ interaction studies. Luan et al. (2001) provide sample size calculations corresponding to a Model II $G \times E$ interaction for a continuous dependent variable and environmental covariate, in the context of simple linear regression where the $G \times E$ interaction estimate is the ratio of the slopes dependent on the genotypes. Hwang et al. (1994) provide sample size calculations corresponding to a Model II $G \times E$ interaction for binary genetic and environmental factors in a case-control setting. Foppa and Spiegelman (1997) provide sample size calculations corresponding to a Model IV $G \times E$ interaction for a binary genetic factor and polytomous environmental factor in a case-control setting, using multivariate logistic regression. García-Closas and Lubin (1999) show that the formulas provided by both Hwang et al. (1994) and Foppa and Spiegelman (1997) underestimate the minimum sample size necessary for a given power for large $G \times E$ interaction effect sizes. García-Closas and Lubin (1999) suggest a power and sample size approach described by Lubin and Gail (1990). Goldstein et al. (1997) perform power and sample size calculations using the Lubin and Gail (1990) approach for detecting $G \times E$ interactions of complex diseases using case-control designs. Goldstein et al. (1997) suggest that in light of the unreasonably large sample size necessary to detect a $G \times E$ interaction for uncommon genes or environmental factors in a standard case-control design, researchers should consider more efficient alternative study designs.

Although power and sample size calculations have been extensive in genetic association and $G \times E$ interaction studies with respect to cross-sectional statistical methods (e.g., chi-square tests of independence, simple linear regression, and logistic regression), there has not been much design research in the context of survival analysis. It is natural to extend the current study designs to survival analysis as the time to survival event is a critical biostatistical quantity that may be genetically associated. This research examines design considerations to detect genetic associations and $G \times E$ interactions (Models III and IV) explicitly in the survival analysis context. The log-rank test statistic will be used to detect a genetic association, as it is equivalent to the score test statistic for a Cox proportional hazards model with binary covariates (Hosmer and Lemeshow 1999). Cox PH modeling will be used to detect a Model III and IV $G \times E$ interaction. These study designs require the sample size to consist of a representative sample of the population, unlike the popular case-control or partial case-control collection designs.

Schoenfeld (1983) was the first to derive a sample size formula for a Cox proportional hazards model with a binary covariate. The sample size formula was based on the asymptotic distribution of the score test statistic, which as noted previously, is equivalent to the log-rank test statistic in a proportional hazards setting. Lakatos (1988) derived a sample size formula under pragmatic complex trial conditions to allow for non-proportional hazards, loss to follow up, non-compliance and drop-in. Hsieh and Lavori (2000) extend upon the Schoenfeld sample size formula to derive a sample size formula for a Cox proportional hazards model with a non-binary covariate. Their sample size formula allows for a multivariate Cox model where the correlation between additional

covariates and the non-binary covariate of interest is accounted for in a variance inflation factor. This variance inflation factor is then multiplied by the required sample size for the Cox model with a single non-binary covariate.

Although there has been a lot of research devoted to sample size formulas to compare the survival distributions between two groups, there has been relatively little research devoted to the sample size formula derivation to compare more than two groups. Ahnn and Anderson (1995) extend the Schoenfeld sample size formula and derive a sample size formula that compares three or more groups of equal proportion using the log-rank test. This formula assumes proportional hazards and a uniform censoring distribution across groups. Ahnn and Anderson (1998) extend the Lakatos sample size formula to compare three or more groups of equal allocations. Halabi and Singh (2004) extend upon the Ahnn and Anderson (1995) sample size formula to allow for unequal allocation of groups. This formula also assumes proportional hazards and identical censoring distributions across groups.

Misclassification errors are present in the majority of data sets and can affect the validity of a study. It is well documented that environmental misclassification errors are a primary source of bias in epidemiological studies (Rothman 1998). A genotype misclassification error occurs when an observed genotype does not correspond to the true genotype. Genotyping errors are result from interactions between DNA molecules, low quality or quantity of DNA, biochemical anomalies, and human error (Pompanon, et al. 2005). Genotyping error rates can be estimated with genotyping replication using an implicit reference genotype (Rice and Holmans 2003), testing for deviation from Hardy-Weinberg Equilibrium (HWE) (Hosking, et al. 2004), or through non-Mendelian errors in a case-parent design (Douglas, et al. 2002; Geller and Ziegler 2002; Gordon, et al. 1999). Studies have reported genotyping error rates ranging from 0.2% to 15%. It is common for laboratories to report genotyping error rates between 0.5% and 1% (Pompanon, et al. 2005). Lincoln and Lander (1992), Douglas et al. (2002) and Sobel et al. (2002) summarize the integration of genotyping misclassification errors in statistical genetics and describe several realistic empirical error models.

There has been considerable research on the impact of genotyping errors in case-control genetic association studies (Ahn, et al. 2007; Gordon and Finch 2005; Gordon, et al. 2002; Kang, et al. 2004a; Kang, et al. 2004b). It has been shown that genotyping misclassification errors lead to biased estimates and a decrease in power in genetic association studies (Gordon and Finch 2005). The impact of each individual SNP genotyping error type for the chi-square test of independence and the linear trend test was examined in Kang et al. (2004b) and Ahn et al. (2007), respectively. They determined which SNP genotyping misclassification error was most deleterious in terms of increase in required sample size.

Many advocate a double-sampling procedure to provide information regarding genotyping error rates and incorporate this information in the calculation of asymptotic power (Gordon, et al. 2004; Ji, et al. 2005; Tintle, et al. 2007). Double-sampling procedures resample a random subset of the data with a gold standard measurement (e.g., a perfect classification method). Test statistics incorporating information from double sampling procedures have been derived for association studies (e.g. Barral, et al. 2005; Gordon, et al. 2004; Rice and Holmans 2003).

Misclassification errors in $G \times E$ interaction studies may lead to an increase in the required sample size and an attenuation of the $G \times E$ interaction estimate. García-Closas et al. (1999) show that differential misclassification of a dichotomous environmental factor with respect to the disease biases the multiplicative interaction effect toward the null value in case-control studies when the genotype and environmental factors are independent for the controls and the environmental misclassification is non-differential with respect to the genotype. Non-differential misclassification occurs when the rate of misclassification is the same for all subjects. Differential misclassification occurs when the rate of misclassification differs between subjects conditional on some other information (e.g., case or control). García-Closas et al. (1999) also show that misclassification of genotype or environmental risk factors may substantially increase the sample size necessary for case-control studies, particular for small effect sizes. Deitz et al. (2004) illustrate the impact of genotyping misclassification errors upon sample size in an $G \times E$ interaction study of the NAT2 gene, smoking, and bladder cancer. They compare the genotyping errors of NAT2 of a 3-single nucleotide polymorphism (SNP) genotyping assay with a more precise 11-SNP assay, concluding that the 11-SNP assay substantially decreases the sample size necessary with consequent decrease in the cost of the study. Wong et al. (2003) extend the $G \times E$ interaction sample size calculations of Luan et al. (2001) for continuous traits to calculate the increase in sample size necessary in the presence of genetic and environmental non-differential measurement errors. They consider the same simple linear regression model for a continuous dependent variable and environmental covariate, in which the $G \times E$ interaction estimate is the ratio of slopes dependent on the genotype. They show that the increase in sample size necessary due to both genetic and environmental errors can be substantial, suggesting that studies should invest in better measurement, if possible, rather than increasing the sample size to handle the errors. Wong et al. (2004) provide a method to estimate the $G \times E$ interaction in the presence of genetic and environmental non-differential measurement errors for a continuous dependent and environmental variable. They adjust the crude $G \times E$ interaction estimate in their simple linear regression model using a correction factor obtained through a validation study.

This research assesses the practicality in terms of sample size to detect a genetic association and $G \times E$ interaction with survival analysis techniques. The impact of each genotyping error upon power and sample size is quantified mathematically to determine those genetic error rates that are most deleterious. Results are confirmed through simulation studies. Chapter 2 details the method in which these research questions will be answered for the various models. Chapter 3 considers direct and indirect genetic association studies using the log-rank test statistic. Chapters 4 and 5 consider the $G \times E$ interaction only model (Model III) and full $G \times E$ interaction model (Model IV), respectively, for direct association studies. Chapters 4 and 5 consider dominant and recessive patterns of inheritance.

Chapter 2 – Methods

This chapter begins with the notation that will be used in the subsequent chapters. Assume that there exists survival analysis genetic data for a di-allelic disease locus with a high-risk allele *d* and a low-risk allele +. Furthermore, assume that there is survival genetic data for a di-allelic SNP marker locus with a less common allele *A* that is in coupling with the high-risk allele *d* and a more common allele *B*.

Survival model parameters

$G_i =$ genotype indicator function. Given a direct association study, G_i represents the disease genotype *i* covariate, where $i = 1$ for the *d*+ genotype and $i = 2$ for the *dd* genotype (genotype ++ is the reference genotype). Given an indirect association study, G_i represents the SNP marker genotype *i* covariate, where $i = 1$ for the *AB* genotype and $i = 2$ for the *AA* genotype (genotype *BB* is the reference genotype).

For the $G \times E$ interaction only model and full model (Chapters 4 and 5), *G* is an indicator function such that $G = 1$ if the subject has the at-risk disease genotype (*dd* or *d*+ for dominant models, *dd* for recessive models where *d* is the disease allele and + is the wild-type allele); $G = 0$ otherwise.

$E =$ environment covariate. Let *E* be transformed into a *Z*-scale metric such that *E* is assumed to be normally distributed with mean 0 and variance 1.

Furthermore, G and E are assumed to be independent.

$G_i \times E =$ gene-environment interaction covariate. The $G \times E$ interaction only model and the full model (Chapters 4 and 5, respectively) consider direct association studies. That is, the functional variant is directly observed.

$t =$ a continuous random variable indicating the subject's survival time. It is assumed to be exponentially distributed.

$\lambda(t) =$ hazard function of survival time *t*. The hazard function is defined to be $\frac{f(t)}{1-F(t)}$, where $f(t)$ is the probability density function of *t* and $F(t)$ is the cumulative density function of *t*. Conceptually, the hazard function is the instantaneous probability of the survival event, given that the subject has survived to time *t*. Examples of survival events are disease onset, disease relapse, disease progression, disease remission and death. For the purposes of this work, the survival event is defined to be disease onset. The hazard function of an exponentially distributed random variable with mean ξ is $1/\xi$. That is, the hazard function does not depend on *t* for exponential distributions.

$\lambda_0(t)$ = baseline hazard function. This is the hazard function for those subjects with all covariates equal to 0.

$\Lambda(t) =$ cumulative hazard function of survival time *t*. The cumulative hazard function is defined to be $\int_0^t \lambda(t)dt$.

$\eta =$ censoring proportion, the proportion of subjects who did not experience the survival event. It is assumed that the censoring proportions across differing genotype groups are identical.

Genetic model parameters in the context of survival analysis

Frequency parameters:

p_A = SNP marker *A* allele frequency; $p_A \le \frac{1}{2}$

p_d = disease locus *d* allele frequency

π_0 = frequency of ++ genotype = p_d^2

π_1 = frequency of *d*+ genotype = $2p_d(1-p_d)$

π_2 = frequency of *dd* genotype = $(1-p_d)^2$

Hardy-Weinberg equilibrium (HWE) is assumed at the disease locus.

Disequilibrium parameters:

ρ = coefficient of maximal disequilibrium $[0 \le \rho \le 1]$

$R^2_{\max} = (\min[(1-p_A)p_d, p_A(1-p_d)])^2 / [p_d(1-p_d)p_A(1-p_A)]$

$r^2 = \rho \times R^2_{max}$

$D = \sqrt{r^2 p_A(1-p_A)p_d(1-p_d)}$, disequilibrium (non-standardized)

Survival penetrances and genotypic relative risks:

$f_0(t) = \Lambda_0(t \mid G = ++) =$ Pr(survival event occurring in the time interval $[0,t]$ | genotype ++ at disease locus)

$f_1(t) = \Lambda_1(t \mid G = d+) =$ Pr(survival event occurring in the time interval $[0,t]$ | genotype *d*+ at disease locus)

$f_2(t) = \Lambda_2(t \mid G = dd) =$ Pr(survival event occurring in the time interval $[0,t]$ | genotype *dd* at disease locus). These functions will be referred to as survival penetrances as they are analogous to the conventional penetrance functions.

$r_1 = f_1(t) / f_0(t) = f_1 / f_0$

$r_2 = f_2(t) / f_0(t) = f_2 / f_0$. These are analogous to the genotypic relative risks (GRR) (Schaid and Sommer 1993). They are independent with respect to time, given an exponentially distributed failure time.

Survival prevalence:

$\varphi(t) =$ Pr(survival event occurring in the time interval $[0,t]$) and is analogous to prevalence. Hardy-Weinberg equilibrium (HWE) is assumed at the disease locus so that $\varphi(t) = (1-p_d)^2 f_0(t) + 2p_d(1-p_d) f_1(t) + p_d^2 f_2(t)$. The censoring proportion may be calculated from the survival prevalence as $\eta = 1 - \varphi(t)$.

Haplotype frequencies:

The haplotype frequencies h_{ij} of disease allele *i* and marker allele *j* are functions of p_d, p_A and *D*. Specifically,

$h_{+A} = (1-p_d)p_A - D$

$h_{+B} = (1-p_d)(1-p_A) + D$

$h_{dA} = p_d p_A + D$

$h_{dB} = p_d(1-p_A) - D$

Conditional survival probabilities for marker genotypes:

$P_1(t) = \Pr(\text{marker genotype } AA \mid \text{survival event occurring in the time interval } [0,t]) =$

$$\left[1 \Big/ \varphi(t) \right] \left\{ (h_{+A})^2 f_0 + 2(h_{+A})(h_{dA}) f_1 + (h_{dA})^2 f_2 \right\}$$

$P_2(t) = \Pr(\text{marker genotype } AB \mid \text{survival event occurring in the time interval } [0,t]) =$

$$\left[2 \Big/ \varphi(t) \right] \left\{ (h_{+A})(h_{+B}) f_0 + (h_{+A} h_{dB} + h_{dA} h_{+B}) f_1 + (h_{dA})(h_{dB}) f_2 \right\}$$

$P_3(t) = \Pr(\text{marker genotype } BB \mid \text{survival event occurring in the time interval } [0,t]) =$

$$\left[1 \Big/ \varphi(t) \right] \left\{ (h_{+B})^2 f_0 + 2(h_{+B})(h_{dB}) f_1 + (h_{dB})^2 f_2 \right\}$$

Genotyping error parameters

Disease genotype error parameters (Chapters 4 and 5):

The genotyping misclassification errors are assumed to be independent and identically distributed (i.i.d) with respect to the survival event. The six parameter disease genotyping error model is displayed in Table 2.1. The error parameters are defined as $e_{ij} = \Pr(\text{observed genotype } j \mid \text{true genotype } i)$ where $i, j \in \{0,1,2\}$, 0 = ++ genotype, 1 = *d*+ genotype , and 2 = *dd* genotype.

Table 2.1: Disease locus error parameters e_{ij}

True genotype	Observed genotype		
	*dd**	*d*+**	++*
dd	$1-e_{21}-e_{20}$	e_{21}	e_{20}
d+	e_{12}	$1-e_{12}-e_{10}$	e_{10}
++	e_{02}	e_{01}	$1-e_{02}-e_{01}$

The *ij*th cell denotes the conditional probability of observing the *j*th genotype conditional on the *i*th true genotype, where 0 = genotype ++, 1 = genotype *d*+, and 2 = genotype *dd*.

SNP marker genotyping error parameters (Chapter 3):

The SNP genotyping misclassification errors are assumed to be assumed to be non-differential. The SNP error parameters are defined as $\varepsilon_{ij} = \Pr(\text{observed marker genotype } j \mid \text{true marker genotype } i)$ where $i, j \in \{1,2,3\}$ with 1 = *AA* genotype, 2 = *AB* genotype and 3 = *BB* genotype.

Table 2.2: SNP marker locus error parameters ε_{ij}

True genotype	Observed genotype		
	*AA**	*AB**	*BB**
AA	$1-\varepsilon_{21}-\varepsilon_{31}$	ε_{12}	ε_{13}
AB	ε_{21}	$1-\varepsilon_{12}-\varepsilon_{32}$	ε_{23}
BB	ε_{31}	ε_{32}	$1-\varepsilon_{13}-\varepsilon_{23}$

The *ij*th cell denotes the conditional probability of observing the *j*th genotype conditional on the *i*th true genotype, where 1 = genotype *AA*, 2 = genotype *AB*, and 3 = genotype *BB*.

Percent increase in minimum sample size requirements necessary (%MSSN)
To differentiate between the coefficients of the percent increase in the minimum sample size necessary (%MSSN) for the three Cox models presented in the subsequent chapters, the following notation will be used:

Genetic association model (Chapter 3) – Let C_{ij} denote %MSSN associated with a 1% increase in the SNP error parameter ε_{ij} as defined by Kang et al. (2004a). The sum of marker %MSSN coefficients is defined as $C_{12}+C_{13}+C_{21}+C_{23}+C_{31}+C_{32}$. The misclassification of one homozygote to another homozygote may be rare in practice (Miller, et al. 2002); however, these misclassifications are included for theoretical completeness.

Model III $G\times E$ interaction (Chapter 4) – Let D_{ij} denote the %MSSN associated with a 1% increase in the disease locus error parameter e_{ij} for the interaction only model. The sum of the disease %MSSN coefficients for this model is defined as $D_{01}+D_{02}+D_{10}+D_{12}+D_{20}+D_{21}$.

Model IV $G\times E$ interaction (Chapter 5) – Let F_{ij} denote the %MSSN associated with a 1% increase in the disease locus error parameter e_{ij} for the full model. The sum of disease %MSSN coefficients for this model is defined as $F_{01}+F_{02}+F_{10}+F_{12}+F_{20}+F_{21}$.

<u>*Gene only model (Chapter 3)*</u>

Log-rank test statistic and sample size formula

The log-rank test statistic is used to compare the survival curves between three or more groups. Let *K* denote the total number of groups and $t_1 < t_2 < \ldots < t_m$ be the *m* rank ordered distinct survival times. Let d_{ki} denote the number of survival events belonging to group *k* (*k* = 1,2,…*K*) at survival time t_i and n_{ki} denote the number of at-risk subjects belonging to group *k* (*k* = 1,2,…*K*) at survival time t_i. Furthermore, let d_i and n_i denote the total number of survival events across all *k* groups and the number of at-risk subjects across all *k* groups at time t_i. The log-rank test statistic formula is $\Psi(0)' I(0)^{-1}\Psi(0)$ where $\Psi(0)$ is a score vector of length $k-1$ whose elements are defined as $\Psi_k(0)=\sum_{i=1}^{m}\left(d_{ki}-n_{ki}/n_i\right)$ for group *k* = 2,3…*K*, and $I(0)$ is a $(k-1)\times(k-1)$ information matrix whose elements are defined as $I_{kl}(0)=\sum_{i=1}^{m}\frac{n_{kl}}{n_i}\left(\delta_{kl}-n_{li}/n_i\right)$ for group *k*,*l* = 2,3…*K* where $\delta_{kl}=1$ if $k=l$ and $\delta_{kl}=0$ if $k\neq l$ (Lawless 1982). This test statistic is asymptotically distributed as χ^2_{k-1} under the null hypothesis that the survival curves between groups are equivalent.

We consider the following log-rank model to detect a genetic association:

$$\lambda_j(t \mid G_{1j}, G_{2j}) = \lambda_{0j}(t) exp\left(\beta_{1j} G_{1j} + \beta_{2j} G_{2j}\right)$$

where $j = D$ or I for direct association and indirect association study respectively. A functional polymorphism in lieu of the disease locus is observed in indirect association studies. The direct association model is a special case of the indirect association model when there is perfect disequilibrium ($r^2 = 1$). The direct association model compares the survival distributions between differing disease locus genotypes, and the indirect association model compares the survival distributions between differing SNP marker genotypes. The baseline hazard, $\lambda_{0D}(t)$, is the risk of those subjects with disease genotype ++ for the direct association model, and $\lambda_{0I}(t)$ is the risk of those subjects with marker genotype *BB* for the indirect association model. The model coefficients β_{ij} represent the log hazard ratio for genotype *i.* The hazard ratio is analogous to relative risk and is the ratio of hazard functions between subjects with differing genotypes (Hosmer and Lemeshow 1999). For example, the hazard ratio between subjects with marker genotype *AB* and genotype *BB* is $\frac{\lambda(t \mid G_{1I} = 1, G_{2I} = 0)}{\lambda(t \mid G_{1I} = 0, G_{2I} = 0)} = \frac{\lambda_0(t) exp(\beta_{1I})}{\lambda_0(t)} = exp(\beta_{1I})$.

We use the Halabi and Singh (2004) sample size formula for the log-rank test statistic with unequal allocations. The null and alternative hypotheses are $H_{0j} : \beta_{1j} = \beta_{2j} = 0$, $H_{1j} : \beta_{1j} \neq 0$ or $\beta_2 \neq 0$. This sample size formula is applicable to genetic association studies where it is common for there to be an unequal proportion between groups with differing genotypes. The sample size formula is given by:

$$N_j = \frac{\Theta^2_{(\upsilon=2,\alpha,1-\beta)}}{(1-\eta)\left[\pi_{1j}(1-\pi_{1j})\left(log\left(\lambda_{1j}/\lambda_{0j}\right)\right)^2 + \pi_{2j}(1-\pi_{2j})\left(log\left(\lambda_{2j}/\lambda_{0j}\right)\right)^2 - 2\pi_{1j}\pi_{2j}\left(log\left(\lambda_{1j}/\lambda_{0j}\right)\right)\left(log\left(\lambda_{2j}/\lambda_{0j}\right)\right)\right]}$$

where $j = D$ or I for direct association and indirect association models respectively, Θ^2 is the noncentrality parameter for a noncentral chi-square distribution with a given level of significance α, power $1-\beta$, and degrees of freedom $\upsilon = 2$. The SAS routine CNONCT can be used to calculate Θ^2; for example $\Theta^2_{\upsilon=2,\alpha=0.01,\beta=0.2} = 13.88$ and $\Theta^2_{\upsilon=2,\alpha=0.01,\beta=0.05} = 20.65$. The overall censoring proportion is η. The genotype frequencies π_{1j} and π_{2j} denote the expected proportion of subjects allocated to genotypes *dd* and *d+* for a direct association model and genotypes *AB* and *AA* for an indirect association model. The corresponding hazard ratios of these genotype groups with respect to their baseline genotype group are denoted as $\lambda_{1j} / \lambda_{0j}$ and $\lambda_{2j} / \lambda_{0j}$. The total required sample size necessary with these given design parameters is N_j.

<u>*Gene-environment interaction only model (Chapter 4)*</u>

Cox proportional hazards (PH) model and sample size formula

The gene-environment ($G \times E$) interaction is modeled for a direct association study using the following Cox PH model: $\lambda(t \mid G \times E) = \lambda_0(t) exp\left[\gamma \times (G \times E)\right]$. The model coefficient γ represents the log hazard ratio for a one unit change in $G \times E$. The

hazard ratio is interpreted in the same manner as in log-rank model. For example, given a dominant MOI, the hazard ratio between subjects with the at-risk genotype, *dd* or *d+*, and subjects without the at-risk genotype, ++, is $\frac{\lambda(t\,|\,(G=dd\ or\ d+)\times E)}{\lambda(t\,|\,(G=++)\times E)}=\frac{\lambda_0(t)exp(\gamma\times(G\times E))}{\lambda_0(t)}=exp(\gamma\times(G\times E))$. The log hazard ratio between subjects with the disease genotype and without the disease genotype is $\gamma\times(G\times E)$ for both dominant and recessive mode of inheritance (MOI).

To investigate the required sample size to detect a $G\times E$ interaction using Cox PH modeling, the Hsieh and Lavori (2000) sample size formula for a Cox PH model with nonbinary covariates is used. The sample size formula states that the total required sample size, *N*, for a given level of significance α and power $1-\beta$ is:

$$N=\frac{(z_{1-\alpha}+z_{1-\beta})^2}{(1-\eta)\sigma^2_{G\times E}\,log^2\Delta}$$

where $z_u=\Phi^{-1}(u)$ with Φ^{-1} being the inverse of the standard normal cumulative distribution (for example, $z_{0.80}=0.842$, $z_{0.95}=1.645$ and $z_{0.99}=2.326$), $\sigma^2_{G\times E}=var(G\times E)$, η = censoring rate, and $log\,\Delta=\gamma$ is the log hazard ratio associated with a one-unit change in $G\times E$.

Let τ denote the at-risk genotype frequency so that $\tau=p_d^2+2p_d(1-p_d)$ for a dominant MOI and $\tau=p_d^2$ for a recessive MOI. Then $\sigma^2_{G\times E}=var_D(G\times E)=\tau$. Let Δ denote the hazard ratio for the $G\times E$ interaction and be defined as the product of the genotypic relative risks (GRR) and a one unit change in *E*. That is, for a dominant MOI,

$$\Delta_D=HR(G\times E)=GRR(dd\ or\ d+,++)\times((E+1)-1)=\frac{f_2(t)\pi_2+f_1(t)\pi_1}{(\pi_2+\pi_1)f_0(t)}.$$

For a recessive MOI,

$$\Delta_R=HR(G\times E)=GRR(dd,d+\ or\ ++)\times((E+1)-1)=\frac{f_2(t)(\pi_0+\pi_1)}{f_0(t)\pi_0+f_1(t)\pi_1}.$$

Then $\Delta_D=r_1(t)=r_2(t)$ for a dominant MOI and $\Delta_R=r_2(t)$ for a recessive MOI.

<u>*Full model (Chapter 5)*</u>

Cox proportional hazards (PH) model and sample size formula

The full model contains genetic and environmental main effects and their interaction: $\lambda(t\,|\,G,E)=\lambda_0(t)exp[\gamma_1G+\gamma_2E+\gamma_3GE]$ for a direct association study. The model coefficients are interpreted in a manner similar to previous sections.

To investigate the required sample size to detect a $G\times E$ interaction for a model that includes genetic and environmental main effects, the Hsieh and Lavori (2000) sample size formula for a Cox PH model with nonbinary covariates and variance inflation factor (VIF) can be used. The VIF approximates the ratio between the variance of the regression coefficient of interest in a model without other covariates and a model with correlated covariates. The sample size formula states that the total required sample size, *N*, for a given level of significance α and power $1-\beta$ is:

$$N = \frac{\left(z_{1-\alpha} + z_{1-\beta}\right)^2}{(1-\eta)\sigma_{G\times E}^2 \, log^2 \Delta}\left(\frac{1}{1-R^2_{G\times E\bullet G,E}}\right)$$

where $VIF = 1/\left(1-R^2_{G\times E\bullet G,E}\right)$ and $R^2_{G\times E\bullet G,E}$ is the amount of variance explained by regressing $G\times E$ on covariates *G* and *E*. We may calculate $R^2_{G\times E\bullet G,E}$ using the known bivariate correlations between the three variables (*G*, *E*, and $G\times E$) as $R^2_{G\times E\bullet G,E} = \frac{r^2_{G\times E,G} + r^2_{G\times E,E} - 2r_{G\times E,G}r_{G\times E,E}r_{G,E}}{1-r^2_{G,E}}$ (Healey 1993). All other parameters, including the $G\times E$ hazard ratio were defined in the previous section.

Both sample size formulas (Hsieh and Lavori 2000; Halabi and Singh 2004) assume the censoring distributions among all subjects and groups to be identical and the hazard ratios to be independent with respect to time (proportional hazards assumption). Futhermore, the sample size formula assumes the alternative hypothesis to be in close proximity to the null hypothesis.

<u>*Increase in required sample size*</u>

Indirect association study

We assess the increase in required sample size to maintain a constant specified power for a fixed level of significance due to an indirect association study. We define π_{1j} and π_{2j} to be the expected genotype frequency such that $\pi_{1D} = \Pr(d+) = 2p_d(1-p_d)$ and $\pi_{2D} = \Pr(dd) = p_d^2$ for a direct association study and $\pi_{1I} = \Pr(AA)$ and $\pi_{2I} = \Pr(AB)$ for an indirect association study. Furthermore, we define the hazard ratios to be $\frac{\lambda_{1D}}{\lambda_{0D}} = \frac{f_1}{f_0} = r_1$ and $\frac{\lambda_{2D}}{\lambda_{0D}} = \frac{f_2}{f_0} = r_2$ for a direct association study and

$$\frac{\lambda_{1I}}{\lambda_{0I}} = \frac{P_{01}(t)\varphi(t)\big/Pr(AA)}{P_{03}(t)\varphi(t)\big/Pr(BB)} = \frac{P_{01}(t)}{P_{03}(t)}\frac{Pr(BB)}{Pr(AA)} \text{ and } \frac{\lambda_{2I}}{\lambda_{0I}} = \frac{P_{02}(t)\varphi(t)\big/Pr(AB)}{P_{03}(t)\varphi(t)\big/Pr(BB)} = \frac{P_{02}(t)}{P_{03}(t)}\frac{Pr(BB)}{Pr(AB)}$$

for an indirect association study. Note that the hazard ratios for the indirect association study are functions of the conditional probabilities ($P_{0i}(t), P_{1i}(t)$), which are in turn functions of the survival penetrances and haplotype frequencies. Thus, we are able to compute the indirect association study design parameters from p_d, p_A, r_1, r_2, $\varphi(t)$ and ρ. There is an attenuation in the hazard ratios for an indirect association study unless there is perfect disequilibrium (e.g., $r^2 = 1$). With these parameters, we calculate the required sample size for a direct association study (N_D), the required sample size for its affiliated indirect association study (N_I), and the increase in required size due to an indirect association study (N_I/N_D). We fix the Type I and Type II error rates so that the noncentrality parameter Θ^2 remains constant. We also assume the censoring distributions to be identical in the direct association and indirect association models. Thus

the increase in required sample size is independent of the specified level of significance, power and censoring proportion.

Increase in required sample size due to genotyping misclassification errors

The increase in sample size to maintain a target level of significance and power due to genotyping misclassification error is examined similarly to Kang et al. (2004b). The marker genetic model parameters are calculated in the presence of error using the law of total probability (see Kang, et al. 2004b for details). The parameters in the presence of error are denoted with asterisks. For example, $\pi_1^* = \Pr(AA^*)$ denotes the probability of observing genotype *AA* in the presence of errors. The total required sample size in the presence of error is thus:

$$N_G^* = \frac{\Theta^2_{(\upsilon=2,\alpha,1-\beta)}}{(1-\eta)\left[\pi_1^*(1-\pi_1^*)\left(log\left(\lambda_1^*/\lambda_0^*\right)\right)^2 + \pi_2^*(1-\pi_2^*)\left(log\left(\lambda_2^*/\lambda_0^*\right)\right)^2 - 2\pi_1^*\pi_2^*\left(log\left(\lambda_1^*/\lambda_0^*\right)\right)\left(log\left(\lambda_2^*/\lambda_0^*\right)\right)\right]}.$$

It is assumed that the noncentrality parameter and censoring proportion remain constant. Thus the increase in required sample size due to SNP genotyping misclassification error for log-rank model is:

$$HS(\tilde{\varepsilon}) = \frac{N_G^*}{N_G} = \frac{\left[\pi_1(1-\pi_1)\left(log\left(\lambda_1/\lambda_0\right)\right)^2 + \pi_2(1-\pi_2)\left(log\left(\lambda_2/\lambda_0\right)\right)^2 - 2\pi_1\pi_2\left(log\left(\lambda_1/\lambda_0\right)\right)\left(log\left(\lambda_2/\lambda_0\right)\right)\right]}{\left[\pi_1^*(1-\pi_1^*)\left(log\left(\lambda_1^*/\lambda_0^*\right)\right)^2 + \pi_2^*(1-\pi_2^*)\left(log\left(\lambda_2^*/\lambda_0^*\right)\right)^2 - 2\pi_1^*\pi_2^*\left(log\left(\lambda_1^*/\lambda_0^*\right)\right)\left(log\left(\lambda_2^*/\lambda_0^*\right)\right)\right]}$$

This formula is denoted as $HS(\tilde{\varepsilon})$ to represent the increase in sample size due to genotyping errors using the Halabi and Singh (2004) sample size formula. Note that $HS(\tilde{\varepsilon})$ is not a function of the target level of significance, power, and censoring proportion.

Similarly, the required sample size in the presence of errors for a given level of significance α and power $1-\beta$ for the $G\times E$ only model and full model is:

$$N_{G\times E}^* = \frac{\left(z_{1-\alpha} + z_{1-\beta}\right)^2}{(1-\eta)\sigma_{G\times E}^{2*} log^2\Delta^*}\left(\frac{1}{1-R^{2*}_{G\times E\bullet G,E}}\right)$$

where the VIF (i.e., $1/\left(1-R^2_{G\times E\bullet G,E}\right)$) is equal to 1 for the $G\times E$ only model (Chapter 4) and is greater than 1 for the full model (Chapter 5). It is also assumed that the censoring pattern remain unchanged in the presence of errors. The minimum increase in required sample size to maintain a constant level of significance and power is thus the ratio of the required sample size with errors over the required sample size without errors:

$$HL(\tilde{e}) = \frac{N_{G\times E}^*}{N_{G\times E}} = \frac{\left(z_{1-\alpha} + z_{1-\beta}\right)^2/(1-\eta)\left(1-R^{2*}_{G\times E\bullet G,E}\right)\sigma_{G\times E}^{2*} log^2\Delta^*}{\left(z_{1-\alpha} + z_{1-\beta}\right)^2/(1-\eta)\left(1-R^2_{G\times E\bullet G,E}\right)\sigma_{G\times E}^2 log^2\Delta} = \frac{\left(1-R^2_{G\times E\bullet G,E}\right)\sigma_{G\times E}^2 log^2\Delta}{\left(1-R^{2*}_{G\times E\bullet G,E}\right)\sigma_{G\times E}^{2*} log^2\Delta^*}.$$

This formula is denoted as $HL(\tilde{e})$ to represent the increase in sample size due to genotyping errors using the Hsieh and Lavori (2000) sample size formula.

Linear Taylor series approximation of the increase in required sample size

To quantify mathematically which genotyping misclassification errors are most deleterious with respect to power and sample size, the %MSSN coefficients are derived

from a first-order Taylor series expansion about 0 with respect to the error parameters (ε_{ij} for the log-rank model and e_{ij} for the Cox model). For example, for the log-rank model, $\frac{N_G^*}{N_G} \approx \left(\frac{N_G^*}{N_G}\right)\Big|_{\varepsilon_{ij}=0\forall i,j} + \sum_{i\neq j}^{3}\sum_{j=}^{2} \frac{d}{d\varepsilon_{ij}}\left(\frac{N_G^*}{N_G}\right)\Big|_{\varepsilon_{ij}=0\forall i,j} \varepsilon_{ij}$, where $\frac{d}{d\varepsilon_{ij}}\left(\frac{N_G^*}{N_G}\right)\Big|_{\varepsilon_{ij}=0\forall i,j}$ is the partial first derivative of N_G^*/N_G with respect to the error parameter ε_{ij}, evaluated at $\varepsilon_{ij}=0$ for all $i\neq j\in\{1,2,3\}$. To calculate the first-order Taylor series about the error parameters e_{ij} for the Cox model, simply replace the ε_{ij}'s with e_{ij}'s. Note that $\left(\frac{N_G^*}{N_G}\right)\Big|_{\varepsilon_{ij}=0\forall i,j} = 1$, which is intuitive but may be easily shown mathematically. The marker %MSSN coefficients are defined as $C_{ij} = \frac{d}{d\varepsilon_{ij}}\left(\frac{N_G^*}{N_G}\right)\Big|_{\varepsilon_{ij}=0\forall i,j}$. Thus the increase in required sample size for the log-rank model can be approximated as $\frac{N_G^*}{N_G} \approx T(\tilde{\varepsilon}) = 1 + \sum_{i\neq j}^{3}\sum_{j=1}^{3} C_{ij}\varepsilon_{ij}$ for design evaluation purposes, where $T(\tilde{\varepsilon})$ denotes the Taylor series approximation to $HS(\tilde{\varepsilon})$.

Similarly, the disease %MSSN coefficients are defined as $D_{ij} = \frac{d}{de_{ij}}\left(\frac{N_{G\times E}^*}{N_{G\times E}}\right)\Big|_{e_{ij}=0\forall i,j}$ and $F_{ij} = \frac{d}{de_{ij}}\left(\frac{N_{G\times E}^*}{N_{G\times E}}\right)\Big|_{e_{ij}=0\forall i,j}$ for the $G\times E$ interaction only model (Model III) and the full $G\times E$ interaction model (Model IV) respectively. Thus the increase in required sample size for the Cox PH model can be approximated as $\frac{N_{G\times E}^*}{N_{G\times E}} \approx S(\tilde{e}) = 1 + \sum_{i\neq j}^{2}\sum_{j=0}^{2} D_{ij}e_{ij}$ and $\frac{N_{G\times E}^*}{N_{G\times E}} \approx S(\tilde{e}) = 1 + \sum_{i\neq j}^{2}\sum_{j=0}^{2} F_{ij}e_{ij}$ for $G\times E$ interaction Models III and IV respectively, where $S(\tilde{e})$ denotes the Taylor series approximation to $HL(\tilde{e})$. All %MSSN coefficients are independent with respect to the specified level of significance and power and censoring rate as it is assumed that these parameters remain constant in the presence of error.

Adequacy of Taylor series

The adequacy of the Taylor approximation is evaluated through its relative error with respect to $HS(\tilde{\varepsilon})$ for the log-rank model and $HL(\tilde{e})$ for the Cox models. The relative error is defined as $R(\tilde{\varepsilon}) = |T(\tilde{\varepsilon}) - HS(\tilde{\varepsilon})| / HS(\tilde{\varepsilon})$ for the log-rank model and $R(\tilde{e}) = |S(\tilde{e}) - HL(\tilde{e})| / HL(\tilde{e})$ for the Cox models. The relative error is computed for the design parameters specified in the next section. A uniform error model is considered in which the six genotyping misclassification errors are equal (e.g., $\varepsilon = \varepsilon_{12} = \varepsilon_{13} = \varepsilon_{21} = \varepsilon_{23} = \varepsilon_{31} = \varepsilon_{32}$ for marker genotyping errors and

$e = e_{01} = e_{02} = e_{10} = e_{13} = e_{20} = e_{21}$ for disease genotyping errors). Uniform error rates ranging from 0.5% to 5% are considered.

Design parameters considered

The following range of design parameters is considered in evaluating the %MSSN coefficients and increase in sample size:

Indirect association study (Chapter 3):

$0 < p_d \leq 0.5$

$0 < p_A \leq 0.5,\ p_A = p_d \pm \delta$ where $0 \leq \delta \leq 0.1$

$0.8 \leq \rho \leq 1$

$1.25 \leq \lambda_1 / \lambda_0 \leq 2$

$1.25 \leq \lambda_2 / \lambda_0 \leq 2$ (Note that $\lambda_1 / \lambda_0 \leq \lambda_2 / \lambda_0$ by design and for dominant models, $\lambda_1 / \lambda_0 = \lambda_2 / \lambda_0$ and for recessive models, $\lambda_1 / \lambda_0 = 1, \lambda_2 / \lambda_0 > 1$)

$0 < \varepsilon_{ij} \leq 0.05$

Direct association study (Chapters 4 and 5):

$0 < p_d \leq 0.7$

$1 < \Delta \leq 2$

$0 < e_{ij} \leq 0.05$

The upper and lower bounds are calculated for all of the %MSSN coefficients for these design settings and situations that lead to an indefinitely large %MSSN coefficient are identified.

Simulation study

A simulation study is performed to confirm that the sample sizes (N) for the log-rank model and the Cox models yield the specified power and that the increases in required sample size (N^*) due to genotyping misclassification error maintains a constant power. Exponential failure times with a uniform censoring distribution are simulated in R-2.4.0, with each power simulation consisting of 10,000 replications. The following high and low design parameter settings are considered for the simulation study in which the specified level of significance is 1% and the specified power is 80%:

Indirect association study (Chapter 3):

$p_d = 0.2,\ 0.3$

$p_A = 0.2,\ 0.3$

$\rho = 0.8,\ 1$

$r_1 = 1.25,\ 1.75$

$r_2 = 1.5,\ 2$

$\eta = 0.10,\ 0.30$

$\varepsilon = 0.005,\ 0.01$ where $\varepsilon = \varepsilon_{12} = \varepsilon_{13} = \varepsilon_{21} = \varepsilon_{23} = \varepsilon_{31} = \varepsilon_{32}$

Direct association study (Chapters 4 and 5), dominant MOI:

$p_d = 0.2,\ 0.3,\ 0.4$

$\Delta = 1.25,\ 1.50,\ 1.75,\ 2.00$

$\eta = 0.10,\ 0.30$

$e = 0.005,\ 0.01$ where $e = e_{01} = e_{02} = e_{10} = e_{13} = e_{20} = e_{21}$

We determine whether the simulated powers were significantly different from the expected power using the test statistic $z = \dfrac{\left| power_{simulated} - power_{expected} \right| - 1/(20000)}{\sqrt{power_{expected}(1 - power_{expected})/10000}}$, where z approximates a standard normal distribution for large samples (Fleiss, et al. 2003). A Bonferroni adjusted significance level is used to account for the multiple design settings. Regression analyses are performed to determine whether the increase in sample size maintains a constant simulated power.

Chapter 3 – Genetic Association Results

Impact on required sample size due to an indirect association study

It is possible to detect a genetic association with a survival event for an indirect association study with common allele frequencies ($0.1 \le p_d$, $0.1 \le p_A$), a high coefficient of maximal disequilibrium ρ and moderate hazard ratios. Consistent with the design of any experiment, required sample size is dependent upon the specified level of significance, power, and effect size(s). Figures 3.1a and 3.1b display the required sample size for an indirect association study as a function of the disease allele frequency p_d and coefficient of maximal disequilibrium ρ for powers of 80% and 95% respectively given a 1% level of significance. Table 3.1 tabulates the required sample size for Figures 3.1a and 3.1b. The design settings in both figures and tabulation are $p_A = p_d + 0.05$, censoring rate of 30%, and genotypic relative risks of $r_1 = 1.5$ and $r_2 = 2$. These design parameters require a total sample size no greater than 3000 for a specified power of 80% and no greater than 4200 for a specified power of 95%. Furthermore, when $0.2 \le p_d \le 0.5$, the total required sample size is less than 700 for a 80% target power and less than 1000 for a 95% target power. The tabulations show that for common allele frequencies ($0.1 \le p_d$, $0.1 \le p_A$), designing a study with 95% power does not require unrealistic sample sizes. However, as the disease allele frequency p_d approaches its lower bound of 0, the expected genotype group proportion of the less common SNP marker genotype (genotype *AA*) becomes extremely small, thus causing the sample size to increase drastically in order to detect any effect in this genotype group. Also, there is a slight increase in sample size as the coefficient of maximal disequilibrium ρ moves away from 1. The required sample size may be calculated for any design setting from the Halabi and Singh (2004) sample size formula (see Methods).

The increase in required sample size to detect a genetic association with an indirect association study is dependent on the difference in allele frequencies $\delta = |p_d - p_A|$ and the coefficient of maximal disequilibrium ρ. An indirect association study leads to an attenuation of effect size (in the hazard ratios) leading to an increase in sample size, unless there is perfect disequilibrium. The increase in required sample size due to an indirect association study is not dependent on the specified level of significance, power and censoring rate. Figures 3.2a and 3.2b show the percent increase in required sample size due to an indirect association study as a function of the absolute difference in allele frequencies (δ) with design parameters $\rho = 0.9$ and genotypic relative risks of $r_1 = 1.25$ and $r_2 = 1.75$. Figure 3.2a considers $p_d > p_A$ and Figure 3.2b considers $p_d < p_A$. Figure 3.2a shows that for $p_d = 0.15$, there is a nonlinear percent increase in required sample size due to δ when $p_d > p_A$. The percent increases in required sample size as a function of δ is linear in all other scenarios. There is less of an impact of δ upon the increase in required sample size as the disease allele frequency increases. For example, if $p_d = 0.15$ and $p_A = 0.10$, the percent increase in required sample size is 75% whereas if $p_d = 0.35$ and $p_A = 0.30$, the percent increase in required

sample size is 40%. There is an infinite percent increase in required sample size as the $p_d \downarrow 0$ due to an indirect association study. Thus, when designing an indirect association study in the context of survival analysis, it is critical that the marker allele frequency be close to the disease allele frequency, especially for small disease allele frequencies. These results are consistent with Zondervan and Cardon (2004), who found that the required sample size for an indirect association study is highly dependent on the interplay between the disease and minor SNP marker allele frequencies.

The increase in sample size due to an indirect association study for a case-control genetic association study is approximately equal to $1/r^2$, where r^2 is the squared correlation coefficient between the SNP marker locus and disease locus (Pritchard and Przeworski 2001). This is also valid for genetic association studies within the framework of survival analysis. The increase in sample size due to an indirect association study (N_I / N_D) using the log-rank test statistic is approximately equal to $1/r^2$. This finding is robust to effect sizes, as shown in Table 3.2. The estimated correlation between N_I / N_D and $1/r^2$ across all design parameters specified in Methods is 0.978 with a 95% confidence interval (0.975, 980) indicating that there is strong agreement between N_I / N_D and $1/r^2$.

Impact on required sample size due to genotyping misclassification errors

We approximate the increase on required sample size due to genotyping misclassification errors through a first-order linear Taylor series approximation, denoted as $T(\tilde{\varepsilon})$. The relative error of the approximation $T(\tilde{\varepsilon})$ with respect to the actual increase in required sample size $HS(\tilde{\varepsilon})$ for a uniform error model in which $\varepsilon_{ij} = \varepsilon$ $\forall i \neq j \in \{1,2,3\}$ is on average 3.7% (median 2.5%) for common allele frequencies ($0.1 \leq p_d$, $0.1 \leq p_A$) for error rates up to 5%. The linear Taylor series approximation $T(\tilde{\varepsilon})$ may only serve as a lower bound to the increase in required sample $HS(\tilde{\varepsilon})$ as the allele frequencies approach their lower bounds of 0 as there is a nonlinear and boundless increase in required sample size for these scenarios. When the error rates are less than 2%, the relative error is on average 0.9% (median 0.5%) for common allele frequencies ($0.1 \leq p_d$, $0.1 \leq p_A$). Figure 3.3 displays the distribution of the relative error for genotyping misclassification rates of 0.5% - 2% in increments of 0.5% for $p_d = 0.25$ and 0.35 and $p_A = p_d + \delta$ such that $|\delta| \leq 0.10$ for all design settings specified in Methods. Since the relative error is sufficiently small, $T(\tilde{\varepsilon})$ may be used for design evaluation purposes.

Table 3.3 summarizes the minimal and maximal impact on the increase in sample size to maintain a constant power of each SNP genotyping misclassification error for the design parameters considered. The %MSSN coefficients C_{ij} are functions of the SNP marker genotype frequencies and conditional survival probabilities for marker genotypes (see Appendix 1 for explicit formulas). They are not dependent on the specified significance level, power, and censoring rate. All genotyping misclassification errors have finite bounds with respect to their %MSSN coefficients except for any

misclassification of the more common homozygote (C_{31} and C_{32}). When $r_1 = 1.25$ and $r_2 = 2$ and the allele frequencies are 0.5 ($p_d = p_A = 0.5$) with perfect disequilibrium ($r^2 = 1$), a misclassification of the less common homozygote to a heterozygote (C_{12}) reaches its upper coefficient bound of 1. With the same settings, a misclassification of the heterozygote to the more common homozygote (C_{23}) and a misclassification of the more common homozygote to the heterozygote (C_{32}) reach their lower coefficient bounds of 0.5 and 0.2 respectively. When $r_1 = r_2 = 2$ and there is perfect disequilibrium ($r^2 = 1$) for allele frequencies of 0.5 ($p_d = p_A = 0.5$), the misclassification of the heterozygote to the more common homozygote (C_{23}) reaches its maximal coefficient bound of 4.5 while the misclassification of the more common homozygote to the less common homozygote (C_{31}) reaches its minimum coefficient bound of 1.1. A misclassification of the less common homozygote to the more common homozygote (C_{13}) obtains its upper bound of 3 when the allele frequencies are 0.5 with perfect disequilibrium and $r_1 = 1.5$ and $r_2 = 2$. As $p_d \downarrow 0$ and $p_A \downarrow 0$ with $\rho = 0.8$ and $r_2 = 2$, a misclassification of the heterozygote as the less common homozygote (C_{21}) reaches its lower coefficient bound of 0 when $r_1 = 2$ and upper coefficient bound of 2 when $r_1 = 1.25$. As $p_A \downarrow 0$, any misclassification of the less common homozygote (C_{12} and C_{13}) reaches its lower coefficient bound of 0. However, this causes an indefinite increase in the coefficients associated with any misclassification of the more common homozygote (C_{31} and C_{32}).

Consistent with previous genotyping misclassification research (Ahn, et al. 2007; Kang, et al. 2004a; Kang, et al. 2004b), there is noTable limiting behavior in the %MSSN coefficients as the minor SNP allele approaches 0. As noted previously, the %MSSN coefficients corresponding to a misclassification of the less common homozygote (C_{12} and C_{13}) and a misclassification of the heterozygote to the less common homozygote (C_{21}) have a lower bound of 0 as $p_A \downarrow 0$. The %MSSN coefficients associated with any misclassification of the more common homozygote (C_{31} and C_{32}) increase without bound as $p_A \downarrow 0$. As a result, the sum %MSSN increases indefinitely as the minor SNP allele approaches 0. Furthermore, the %MSSN coefficient associated with a misclassification of the more common homozygote to the less common homozygote (C_{31}) is the largest coefficient in every possible design and increases at a faster rate for recessive modes of inheritance (e.g., when there is a large difference between the hazard ratios). This result is robust across differing study designs and is consistent with previous genotyping misclassification research in a case-control setting (Kang, et al. 2004a; Kang, et al. 2004b). This consistent finding emphasizes the need to correctly genotype the more common homozygote. Failure to do so may lead to deleterious loss in power or increase in required sample size.

Simulation study results

Figure 3.4 shows the results of the indirect genetic association simulation study. The horizontal axis represents the simulated power with a sample size of *N*, the required

sample size for an indirect association study with perfect genotyping classification. The vertical axis represents the simulated power with a sample size of N^*, the required sample size for an indirect association study in the presence of SNP genotyping misclassification error rates of 0.5% and 1%. The solid line represents a perfect correspondence between the simulated power with sample sizes of N and N^*. The scatter plot of the simulated powers with sample sizes N and N^* falls nicely around the solid line.

The Halabi and Singh (2004) sample size formula yielded slightly higher simulated powers than specified power. Our simulation study considered minor SNP allele frequencies of 0.2 and 0.3 causing the proportion of the reference genotype group (genotype *BB*) to be at least half of the sample. This yielded higher simulated powers than specified power, consistent with the simulation study performed by Halabi and Singh (2004). Furthermore, designs with a high censoring rate of 30% yielded a simulated power that was 5% higher on average than expected.

A linear regression model was fit regressing the simulated power using sample size N^* on the simulated power using sample size N. This model yields an intercept estimate of 0.048 with a standard error of 0.029, which differs nonsignificantly from zero, and a slope estimate of 0.94 with a standard error of 0.035. The estimated standardized beta-coefficient of the slope (e.g., the Pearson correlation estimate between the simulated powers with sample sizes N and N^*) is 0.95 with a 95% confidence interval (CI) of (0.928, 0.971). The intercept estimate differed nonsignificantly from 0 and the correlation estimate was near 1 indicating that the increase in required sample size due to genotyping errors maintains a constant simulated power. We conclude that the simulation study confirms the accuracy of the sample size formulas and the increase in sample size due to genotyping errors maintains a constant power for a fixed level of significance.

One of the critical assumptions of the Halabi and Singh (2004) sample size formula is that the alternative hypothesis be close to the null hypothesis. Thus their sample size formula is only appropriate for small effect sizes ($r_1, r_2 \leq 2$). For larger effect sizes, it is highly recommended that one perform simulations to determine the required sample size for a given design setting. The Halabi and Singh (2004) sample size formula yields sample sizes which tend to underestimate the specified power for larger effect sizes. It is difficult to run simulations with small sample sizes, as there may not be enough events per covariate for the test statistic to converge.

Comparison of N_I / N_D and N_I^* / N_I

We may partition the total required sample size for an indirect association study in the presence of genotyping misclassification errors as $N_I^* = N_D \left(\frac{N_I}{N_D} \right) \left(\frac{N_I^*}{N_I} \right)$ where N_D is the required sample size for a direct association study, N_I / N_D is the impact upon required sample size due to an indirect association study and N_I^* / N_I is the impact upon required sample size due to genotyping misclassification errors. Table 3.4 displays example design settings and the impact upon sample size due to an indirect association study and genotyping misclassification errors. For genotyping error rates less than 2%, N_I / N_D is always greater than or equal to N_I^* / N_I unless the allele frequencies are equal

($p_d = p_A$) and ρ is high (generally $\rho > 0.90$). For genotyping error rates between 2% and 5%, N_I^* / N_I is greater than N_I / N_D when $p_d = p_A$, otherwise N_I / N_D is consistently greater than N_I^* / N_I. As the allele frequencies approach 0, N_I / N_D and N_I^* / N_I increase indefinitely with N_I / N_D greater than N_I^* / N_I except when $p_d = p_A$. As the difference between allele frequencies increases, the difference between N_I / N_D and N_I^* / N_I increases substantially as well. Table 3.4 shows that in situations in which $p_d \neq p_A$, there is at least a 20% difference between increase in sample size due to an indirect association study and the increase in sample size due to genotyping errors. Furthermore, when $p_d = p_A$, the increase in sample size due to an indirect association study and genotyping errors are similar, with the increase in sample size due to genotyping errors slightly larger than the increase in sample size due to an indirect association study.

We find that in most situations, unless the marker allele frequency is approximately equal to the disease gene frequency and the coefficient of maximal disequilibrium is high, the percent increase in sample size due to an indirect association study outweighs the percent increase in sample size due to genotyping misclassification error. There are also scenarios in which N_I is quite large (for example for small allele frequencies and/or small effect sizes) and any additional increase in sample size due to genotyping errors may make the design unfeasible. When the SNP marker locus and disease locus are in perfect disequilibrium, then $N_I / N_D = 1$ and the total required sample size in the presence of genotyping errors is solely contingent upon the required sample size for a direct association study and the increase in required sample size due to genotyping misclassification errors. Gordon et al. (2003) found that there is an interaction between genotyping misclassification errors and disequilibrium upon the required sample size for a case-control genetic association study. Specifically, they found that the impact upon required sample size due to genotyping misclassification errors is higher for low levels of LD and lower for high levels of LD. We find the same result here – the total required sample size for an indirect association study in the presence of genotyping misclassification errors is contingent on the size of the LD parameter and the genotyping misclassification error rate.

Table 3.1: Tabulation of sample size for an indirect genetic association study

N_I		Model parameters			N_I		Model parameters		
80% power	95% power	p_d	p_A	ρ	80% power	95% power	p_d	p_A	ρ
2821	4197	0.05	0.10	0.80	517	769	0.30	0.35	0.80
2696	4010	0.05	0.10	0.84	492	732	0.30	0.35	0.84
2581	3840	0.05	0.10	0.88	469	698	0.30	0.35	0.88
2477	3685	0.05	0.10	0.92	449	667	0.30	0.35	0.92
2381	3542	0.05	0.10	0.96	430	639	0.30	0.35	0.96
1235	1838	0.10	0.15	0.80	491	730	0.35	0.40	0.80
1180	1756	0.10	0.15	0.84	466	694	0.35	0.40	0.84
1130	1681	0.10	0.15	0.88	445	661	0.35	0.40	0.88
1084	1613	0.10	0.15	0.92	425	632	0.35	0.40	0.92
1042	1550	0.10	0.15	0.96	406	604	0.35	0.40	0.96
830	1235	0.15	0.20	0.80	480	715	0.40	0.45	0.80
793	1179	0.15	0.20	0.84	456	679	0.40	0.45	0.84
758	1128	0.15	0.20	0.88	435	647	0.40	0.45	0.88
727	1081	0.15	0.20	0.92	415	617	0.40	0.45	0.92
698	1039	0.15	0.20	0.96	396	590	0.40	0.45	0.96
657	977	0.20	0.25	0.80	483	719	0.45	0.50	0.80
626	932	0.20	0.25	0.84	459	683	0.45	0.50	0.84
599	891	0.20	0.25	0.88	437	649	0.45	0.50	0.88
573	853	0.20	0.25	0.92	416	619	0.45	0.50	0.92
550	818	0.20	0.25	0.96	398	591	0.45	0.50	0.96
567	843	0.25	0.30	0.80	415	617	0.50	0.50	0.80
540	803	0.25	0.30	0.84	394	585	0.50	0.50	0.84
516	767	0.25	0.30	0.88	374	557	0.50	0.50	0.88
493	734	0.25	0.30	0.92	357	530	0.50	0.50	0.92
473	703	0.25	0.30	0.96	340	506	0.50	0.50	0.96

This tabulates the required sample size for an indirect association study for specified powers of 80% and 95% with the following design parameters: 1% significance level, 30% censoring rate, and $r_1 = 1.5$ and $r_2 = 2$.

Table 3.2: Relationship between N_I / N_D and r^2 for a genetic association study

p_d	p_A	r_1	r_2	ρ	r^2	$1/r^2$	N_I / N_D
0.20	0.25	1.25	1.50	0.85	0.64	1.569	1.505
0.20	0.25	1.25	1.50	0.90	0.68	1.482	1.423
0.20	0.25	1.25	1.50	0.95	0.71	1.404	1.350
0.30	0.35	1.25	1.50	0.85	0.68	1.478	1.441
0.30	0.35	1.25	1.50	0.90	0.72	1.396	1.361
0.30	0.35	1.25	1.50	0.95	0.76	1.323	1.290
0.40	0.45	1.25	1.50	0.85	0.69	1.444	1.422
0.40	0.45	1.25	1.50	0.90	0.73	1.364	1.342
0.40	0.45	1.25	1.50	0.95	0.77	1.292	1.270
0.20	0.25	1.50	1.75	0.85	0.64	1.569	1.499
0.20	0.25	1.50	1.75	0.90	0.68	1.482	1.414
0.20	0.25	1.50	1.75	0.95	0.71	1.404	1.338
0.30	0.35	1.50	1.75	0.85	0.68	1.478	1.464
0.30	0.35	1.50	1.75	0.90	0.72	1.396	1.376
0.30	0.35	1.50	1.75	0.95	0.76	1.323	1.297
0.40	0.45	1.50	1.75	0.85	0.69	1.444	1.468
0.40	0.45	1.50	1.75	0.90	0.73	1.364	1.375
0.40	0.45	1.50	1.75	0.95	0.77	1.292	1.292
0.20	0.20	1.25	2.00	0.85	0.85	1.177	1.154
0.20	0.20	1.25	2.00	0.90	0.90	1.111	1.097
0.20	0.20	1.25	2.00	0.95	0.95	1.053	1.046
0.30	0.30	1.25	2.00	0.85	0.85	1.177	1.158
0.30	0.30	1.25	2.00	0.90	0.90	1.111	1.100
0.30	0.30	1.25	2.00	0.95	0.95	1.053	1.047
0.40	0.40	1.25	2.00	0.85	0.85	1.177	1.164
0.40	0.40	1.25	2.00	0.90	0.90	1.111	1.103
0.40	0.40	1.25	2.00	0.95	0.95	1.053	1.049

Table 3.3: Bounds of %MSSN coefficients (genetic association model)

%MSSN coefficients		Bounds		Design Setting				
				p_d	p_A	ρ	r_1	r_2
C_{12}	Misclassification of *AA*	Lower	0		$\downarrow 0$			
	genotype as *AB*	Upper	1	0.5	0.5	1	1.25	2
C_{13} [a]	Misclassification of *AA*	Lower	0		$\downarrow 0$			
	genotype as *BB*	Upper	3	0.5	0.5	1	1.5	2
C_{21}	Misclassification of *AB*	Lower	0	$\downarrow 0$	$\downarrow 0$	0.8	2	2
	genotype as *AA*	Upper	2	$\downarrow 0$	$\downarrow 0$	0.8	1.25	2
C_{23}	Misclassification of *AB*	Lower	0.5	0.5	0.5	1	1.25	2
	genotype as *BB*	Upper	4.5	0.5	0.5	1	2	2
C_{31} [a]	Misclassification of *BB*	Lower	1.1	0.5	0.5	1	2	2
	genotype as *AA*	Upper	∞		$\downarrow 0$			
C_{32}	Misclassification of *BB*	Lower	0.2	0.5	0.5	1	1.25	2
	genotype as *AB*	Upper	∞		$\downarrow 0$			

[a] We note that a misclassification of one homozygote to another homozygote has a small probability of occurring in practice as reported by Miller et al. (2002). However, the limits of these %MSSN coefficients are included for completeness.

This table displays the upper and lower bounds for each %MSSN coefficient, and the respective design parameters for where these bounds occur. When a design setting is missing, it may take any value in its range. For example, C_{32} has an infinite upper bound for any setting in which $p_d \downarrow 0$.

Table 3.4: Example design settings (genetic association model)

$p_d \neq p_A$								
p_d	p_A	ρ	r_1	r_2	ε	N_D	N_I / N_D	N_I^* / N_I
0.15	0.20	0.85	1.50	2.00	0.01	515	1.5246	1.0863
0.15	0.20	0.85	1.50	2.00	0.02	515	1.5246	1.1755
0.15	0.20	0.90	1.50	2.00	0.01	515	1.4439	1.0856
0.15	0.20	0.90	1.50	2.00	0.02	515	1.4439	1.1742
0.20	0.15	0.85	1.50	2.00	0.01	422	1.7130	1.0957
0.20	0.15	0.85	1.50	2.00	0.02	422	1.7130	1.1901
0.20	0.15	0.90	1.50	2.00	0.01	422	1.6248	1.0947
0.20	0.15	0.90	1.50	2.00	0.02	422	1.6248	1.1884
0.15	0.10	0.85	1.75	1.75	0.01	316	2.0465	1.0820
0.15	0.10	0.85	1.75	1.75	0.02	316	2.0465	1.1617
0.15	0.10	0.90	1.75	1.75	0.01	316	1.9400	1.0807
0.15	0.10	0.90	1.75	1.75	0.02	316	1.9400	1.1595
0.20	0.25	0.85	1.75	1.75	0.01	275	1.5121	1.0637
0.20	0.25	0.85	1.75	1.75	0.02	275	1.5121	1.1320
0.20	0.25	0.90	1.75	1.75	0.01	275	1.4203	1.0632
0.20	0.25	0.90	1.75	1.75	0.02	275	1.4203	1.1310
$p_d = p_A$								
p_d	p_A	ρ	r_1	r_2	ε	N_D	N_I / N_D	N_I^* / N_I
0.15	0.15	0.90	1.50	2.00	0.01	515	1.1017	1.0911
0.15	0.15	0.90	1.50	2.00	0.02	515	1.1017	1.1821
0.15	0.15	0.95	1.50	2.00	0.01	515	1.0483	1.0902
0.15	0.15	0.95	1.50	2.00	0.02	515	1.0483	1.1804
0.20	0.20	0.90	1.50	2.00	0.01	422	1.1051	1.0794
0.20	0.20	0.90	1.50	2.00	0.02	422	1.1051	1.1626
0.20	0.20	0.95	1.50	2.00	0.01	422	1.0498	1.0787
0.20	0.20	0.95	1.50	2.00	0.02	422	1.0498	1.1613
0.15	0.15	0.90	1.75	1.75	0.01	316	1.1109	1.0661
0.15	0.15	0.90	1.75	1.75	0.02	316	1.1109	1.1342
0.15	0.15	0.95	1.75	1.75	0.01	316	1.0526	1.0652
0.15	0.15	0.95	1.75	1.75	0.02	316	1.0526	1.1326
0.20	0.20	0.90	1.75	1.75	0.01	275	1.1201	1.0611
0.20	0.20	0.90	1.75	1.75	0.02	275	1.1201	1.1260
0.20	0.20	0.95	1.75	1.75	0.01	275	1.0569	1.0605
0.20	0.20	0.95	1.75	1.75	0.02	275	1.0569	1.1248

This table shows example design settings and their impact upon sample size due to an indirect association study (N_I / N_D) and genotyping misclassification errors (N_I^* / N_I) given a 1% significance level, 80% power, and 30% censoring rate.

Figure 3.1a: Required sample size for 80% power

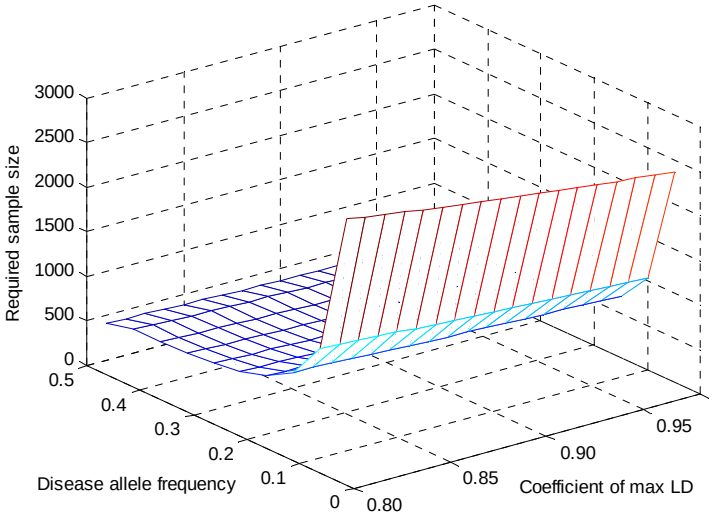
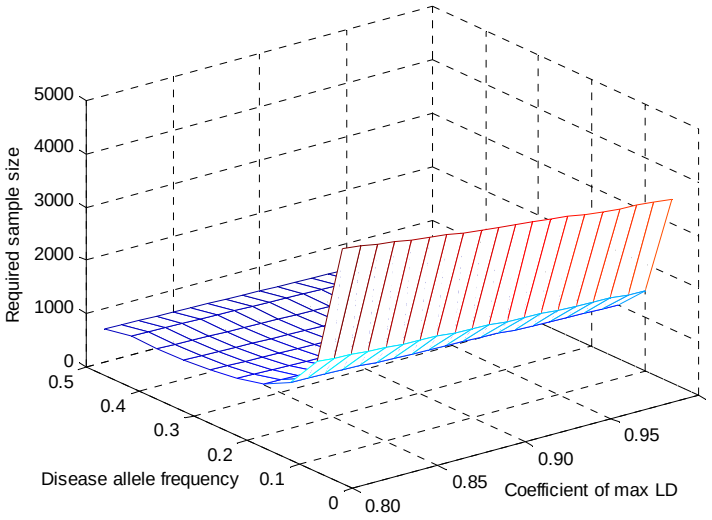


Figure 3.1b: Required sample size for 95% power



These graphs displays the required sample size for an indirect association study for specified powers of 80% and 95% with the following design parameters: 1% significance level, $p_A = p_d + 0.05$, 30% censoring rate, and $r_1 = 1.5$ and $r_2 = 2$.

Figure 3.2a: % increase due to an indirect association study, $p_d > p_A$

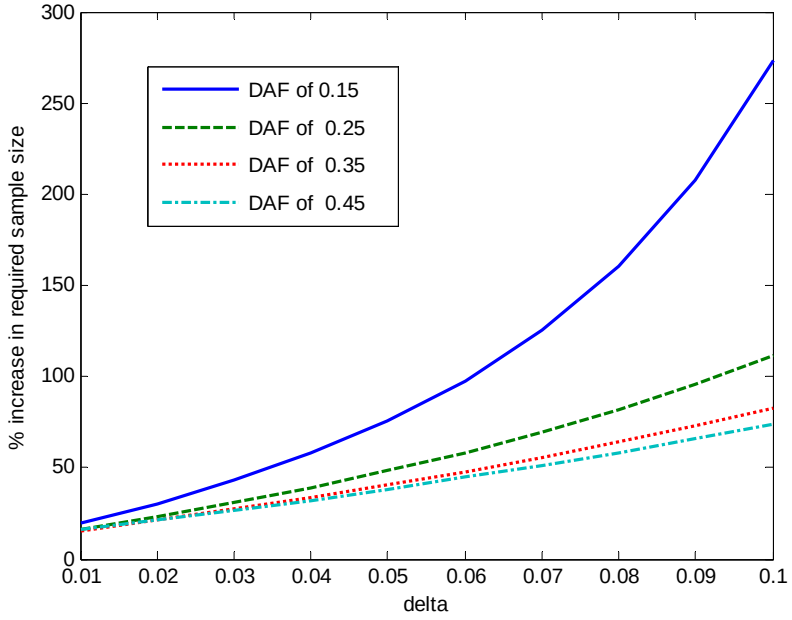
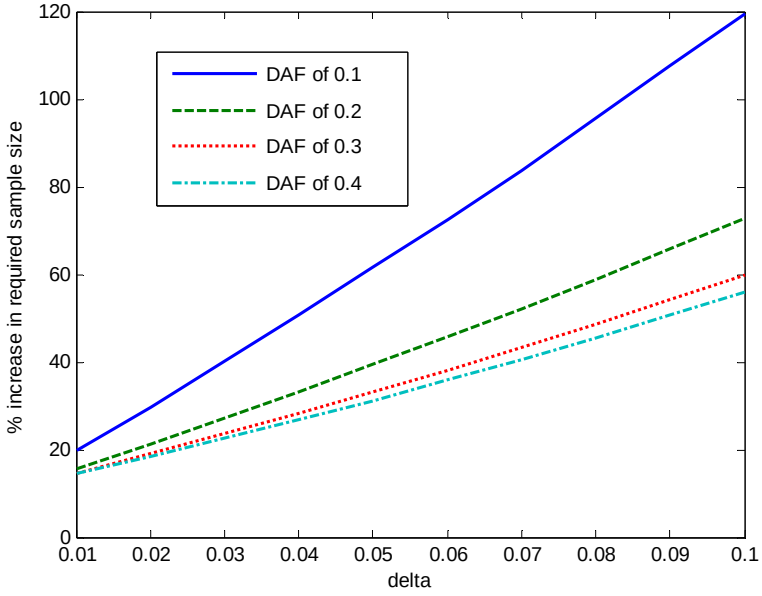
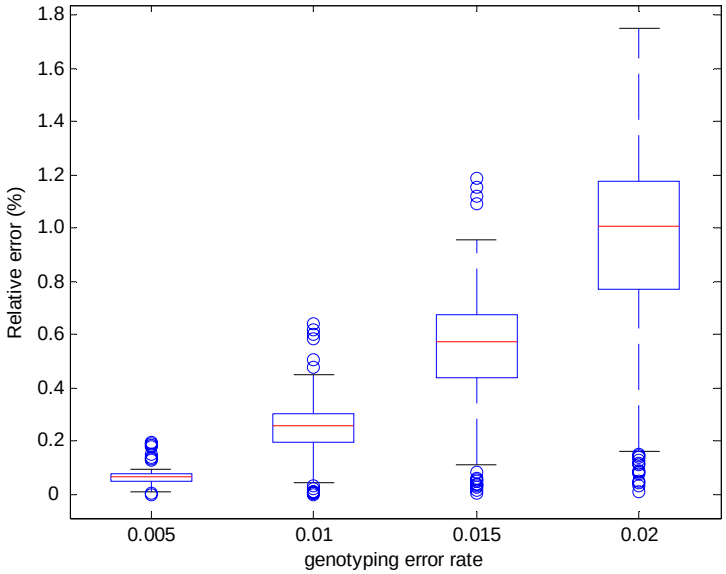


Figure 3.2b: % increase due to an indirect association study, $p_d < p_A$



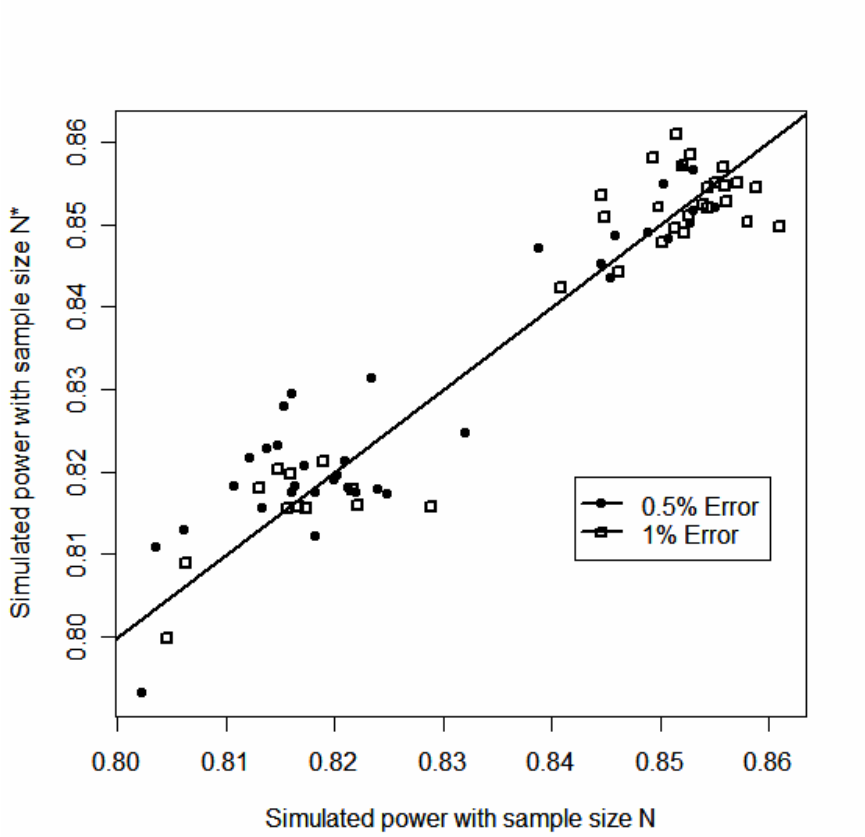
The design parameters are $\rho = 0.9$ and genotypic relative risks of $r_1 = 1.25$ and $r_2 = 1.75$.

Figure 3.3: Distribution of relative error $R(\tilde{\varepsilon})$ for genotyping error rates 0.5% to 2%



These boxplots display the distributions of the relative error $R(\tilde{\varepsilon}) = |T(\tilde{\varepsilon}) - HS(\tilde{\varepsilon})| / HS(\tilde{\varepsilon})$ for genotyping error rates 0.5%, 1.0%, 1.5% and 2.0% with p_d = 0.25 and 0.35 and $p_A = p_d + \delta$ such that $|\delta| \leq 0.10$ for all design settings specified in Methods (Chapter 2).

Figure 3.4: Simulation results (genetic association model)



The horizontal axis represents the simulated power with a sample size of *N*, the required sample size for an indirect association study without genotyping misclassification errors. The vertical axis represents the simulated power with a sample size of *N**, the required sample size for an indirect association study in the presence of genotyping error rates of 0.5% and 1%. The solid line represents a perfect correlation between the simulated powers of sample sizes *N* and *N**.

Chapter 4 – G × E Interaction Only Model Results

It is feasible to detect a Model III $G \times E$ interaction with Cox PH modeling for direct association studies. As in the design of any experiment, the total required sample size is contingent upon the target level of significance, power, and effect size. When designing studies that utilize survival analysis techniques such as the Cox PH model, there is an added layer of complexity created by the specification of the censoring rate. Example design settings and their respective required sample sizes to detect a Model III $G \times E$ interaction are displayed in Table 4.1. The table shows required sample sizes for target powers of 80% and 95%, given a 1% level of significance. Small disease allele frequencies and small effect sizes (e.g., hazard ratios) require large sample sizes. There is an interactive effect between p_d and the hazard ratio (Δ) such that substantially larger sample sizes are required when p_d and Δ are both small. The required sample size for a dominant MOI is considerably less than the required sample size for a recessive MOI, as to be expected. This is because the expected at-risk genotype frequency for a dominant MOI is $2p_d(1-p_d)$ greater than the expected at-risk genotype frequency for a recessive MOI. These genotype frequencies explain the drastic differences in required sample size for the two patterns of inheritance. Recessive models with small disease allele frequencies require large sample sizes regardless of effect size, due to the expected proportion of subjects with the at-risk genotype.

The relative error $R(\tilde{e})$ of the Taylor approximation $S(\tilde{e})$ is small enough so that $S(\tilde{e})$ can be used for design evaluation. The increase in required sample size due to genotyping misclassification errors is independent with respect to the specified level of significance, power and censoring proportion. Moreover, $S(\tilde{e})$ is a first-order expansion and therefore approximates a *linear* increase in sample size due to genotyping errors. However, there is a nonlinear increase when $p_d > 0.5$ for a dominant MOI and when $p_d \downarrow 0$ for a recessive MOI. Under these circumstances, $S(\tilde{e})$ provides a lower bound for $HL(\tilde{e})$.

Tables 4.2a and 4.2b display the $R(\tilde{e})$ and the design requirements for a dominant and recessive MOI, respectively. Table 4.2a considers $p_d = 0.7$ and $\Delta = 1.6$ assuming a uniform error model in which $e_{ij} = e \ \forall i \neq j \in \{0,1,2\}$ and shows that for genotyping error rates less than 2%, the $R(\tilde{e})$ for a dominant MOI is less than 7% for these design settings. The $R(\tilde{e})$ for a dominant MOI increases as the disease allele frequency approaches 0.7 so that Table 4.2a displays the worst case scenario. Table 4.2b considers $p_d = 0.2$ and $\Delta = 1.6$ assuming a uniform error model and shows that for genotyping error rates less than 2%, the $R(\tilde{e})$ for a recessive MOI is less than 2.5% for these design settings. The $R(\tilde{e})$ for a recessive MOI obtains its upper bound of 22.3% as $p_d \downarrow 0$.

Figures 4.1a and 4.1b graph $HL(\tilde{e})$ and $S(\tilde{e})$ as functions of $e_{ij} = e \ \forall i \neq j \in \{0,1,2\}$ for a dominant and recessive MOI, respectively. The parameter specifications are $p_d = 0.2$ and $\Delta = 1.6$ for both models. As noted previously, $HL(\tilde{e})$ and

$S(\tilde{e})$ are not functions of the specified level of significance, power and censoring proportion. Figures 4.1a and 4.1b document visually that the relative error of the Taylor approximation is small enough that it can be used for design evaluation. For a dominant MOI (Figure 4.1a), $R(\tilde{e}) < 0.17\%$ when $e \le 1\%$ and $R(\tilde{e}) < 0.68\%$ when $e \le 2\%$. For a recessive MOI (Figure 4.1b), $R(\tilde{e}) < 0.74\%$ when $e \le 1\%$ and $R(\tilde{e}) < 2.42\%$ when $e \le 2\%$.

The %MSSN coefficients D_{ij} are strictly functions of the disease genotype frequencies π_i where $i = 0,1,2$ for genotypes ++, *d*+ and *dd* respectively and genotypic relative risks $r_1(t)$ and $r_2(t)$. They are not functions of the specified level of significance, power and censoring proportion. The %MSSN coefficients D_{ij} are explicitly stated in Appendix 2.

A %MSSN coefficient will equal zero when the corresponding genotyping misclassification error does not affect the genotype covariate *G*. For example, given a dominant MOI, $D_{12} = D_{21} = 0$ since a misclassification of the *d*+ genotype to the *dd* genotype or vice versa yields the same result of the subject having the at-risk genotype (*G* = 1). Similarly, given a recessive MOI, $D_{01} = D_{10} = 0$ since a misclassification of the *d*+ genotype to the ++ genotype or vice versa yields the same result of the subject not having the at-risk disease genotype (*G* = 0).

It can be seen from the explicit formulas for the %MSSN coefficients stated in Appendix 2 that for a dominant MOI, the coefficient associated with the misclassification of the ++ genotype to the *dd* genotype and the coefficient associated with the misclassification of the ++ genotype to the *d*+ genotype are equivalent ($D_{02} = D_{01}$). Thus, the %MSSN coefficient of any misclassification of a subject without the at-risk genotype has the same impact on the increase in required sample size to maintain a constant power and level of significance given a dominant MOI. The reverse is true for recessive models. For a recessive MOI, the coefficient associated with the misclassification of the *dd* genotype to the ++ genotype and the coefficient associated with the misclassification of the *dd* genotype to the *d*+ genotype are equivalent ($D_{20} = D_{21}$). Thus, the %MSSN coefficient of any misclassification of a subject with the at-risk genotype has the same impact on the increase in required sample size to maintain a constant power and level of significance given a recessive MOI.

Figures 4.2a and 4.2b show the interaction between p_d and Δ upon the sum of %MSSN coefficients for a dominant and recessive MOI, respectively. These Figures show that small disease allele frequencies and hazard ratios yield a large sum of %MSSN coefficients for both MOI. There are several distinct patterns of %MSSN coefficients between the differing patterns of inheritance. Given a dominant MOI, an interaction between larger disease allele frequencies and hazard ratios lead to a high sum of %MSSN coefficients. Furthermore, there is a nonlinear increase in the sum of the %MSSN coefficients for $p_d > 0.5$ (see Figure 4.2a). This does not hold for a recessive MOI. Small disease allele frequencies yield a large sum of %MSSN coefficients for a recessive MOI. There is a slight interaction between p_d and Δ for a recessive MOI so that smaller disease allele frequencies and hazard ratios lead to a high sum of %MSSN coefficients.

As $p_d \downarrow 0$, the sum of %MSSN coefficients for a recessive MOI is substantially larger than the sum of %MSSN coefficients for a dominant MOI. As p_d increases ($p_d > 0.5$), the sum of %MSSN coefficients for a recessive MOI is substantially less than the sum of %MSSN coefficients for a dominant MOI. However, the sum of the %MSSN coefficients as $p_d \downarrow 0$ given a recessive MOI is greater than the sum of the %MSSN coefficients as p_d increases above 0.5 given a dominant MOI.

Table 4.3 specifies the upper and lower bounds of each %MSSN coefficient D_{ij} and their respective design parameters. As noted earlier, the %MSSSN coefficients are independent with respect to the specified level of significance, power and censoring proportions so that the coefficients may be calculated from p_d, $r_1(t)$ and $r_2(t)$. Given a dominant MOI, the coefficients associated with the misclassification of the *d*+ genotype and the *dd* genotype to the ++ genotype (D_{20} and D_{10}) obtain their respective upper bounds of 14 and 16.3 when $p_d = 0.7$ and $r_1(t) = r_2(t) = 2$. Also for a dominant MOI, the coefficients associated with the misclassification of the ++ genotype to the *d*+ genotype and *dd* genotype (D_{01} and D_{02}) obtain their lower bounds of 0 when $p_d = 0.7$. Given a recessive MOI, the coefficients associated with the misclassification of the *dd* genotype to the ++ genotype and *d*+ genotype (D_{20} and D_{21}) obtain their respective upper bounds of 3 when $p_d = 0.7$ and $r_1(t) = 1,\ r_2(t) = 2$. With the same design settings for recessive models, the coefficients associated with the misclassification of the ++ genotype or *d*+ genotype to the *dd* genotype (D_{02} and D_{12}) reach their respective lower bounds of 3.5 and 2.5. Also for a recessive MOI, the *lower* bound of the %MSSN coefficients associated with any misclassification of the ++ genotype (D_{02} and D_{12}) is approximately equal to *upper* lower of the %MSSN coefficients associated with any misclassification of the *dd* genotype (D_{20} and D_{21}).

Consistent with Chapter 3 results, there is noTable behavior as $p_d \downarrow 0$. Given a dominant MOI, the coefficients associated with the misclassification of the *d*+ genotype and the *dd* genotype to the ++ genotype (D_{10} and D_{20}) reach their lower bounds of 1 and 0, respectively, as $p_d \downarrow 0$. Similarly, given a recessive MOI, the coefficients of the misclassification of the *dd* genotype to the ++ genotype and the *d*+ genotype (D_{20} and D_{21}) attain their lower bounds of 1 as $p_d \downarrow 0$. The %MSSN coefficients of any misclassification of a subject without an at-risk genotype to a subject with an at-risk genotype (D_{01} and D_{02} for a dominant MOI; D_{02} and D_{12} for a recessive MOI) increase without bound as $p_d \downarrow 0$. Although this is true in both patterns of inheritance, the coefficients increase at a significantly faster rate for a recessive MOI. Although D_{01} and D_{02} increase indefinitely as $p_d \downarrow 0$ for a dominant MOI, the impact upon increase in required sample size due to genotyping misclassification errors is also deleterious for $p_d > 0.5$ as noted previously. The coefficient for a recessive model corresponding to the

misclassification of the ++ genotype to the *dd* genotype (D_{02}) is the largest coefficient under all design settings and increases at the fastest rate as $p_d \downarrow 0$.

Simulation study results

Figure 4.3 displays the results from the Model III $G \times E$ interaction simulation study. As in Chapter 3, the horizontal axis represents the simulated power with a sample size of *N*, the required sample size with perfect genotyping classification and the vertical axis represents the simulated power with a sample size of *N**, the required sample size in the presence of genotyping misclassification errors. The solid line represents a perfect concordance between the simulated powers of sample sizes *N* and *N**. The scatter plot of simulated powers with sample sizes *N* and *N** are close to the solid line. A regression analysis of the simulated power of sample size *N** on the simulated power of sample size *N* yields an intercept estimate of 0.07 with a standard error of 0.038, which differs nonsignificantly from zero, and a slope estimate of 0.912 with a standard error of 0.050. The estimated standardized beta-coefficient of the slope (e.g. the estimated Pearson correlation coefficient) is 0.936 with a 95% confidence interval of (0.890, 0.964). The intercept estimate did not differ significantly from zero and the estimated correlation between the simulated powers is near 1, indicating that the increase in sample size due to genotyping errors maintains a constant power.

Similar to the Halabi and Singh (2004) sample size formula, one of the critical assumptions of the Hsieh and Lavori (2000) sample size formula is that the alternative hypothesis be near the null hypothesis (e.g. small effect size). Designs with hazard ratios greater than 1.5 yield simulated powers significantly less than the specified power. These correspond to powers less than 75% in Figure 4.3. As the hazard ratio increases, the simulated power decreases thereby substantially underestimating the specified power. When using this design formula, simulations are highly recommended to ensure that the sample size yields adequate power.

Table 4.1: Example design settings (G × E interaction only model)

Model parameters		N given a dominant MOI		N given a recessive MOI	
p_d	Δ	80% power	95% power	80% power	95% power
0.10	1.20	2642	4029	50192	76558
0.20	1.20	1394	2127	12548	19140
0.30	1.20	984	1501	5577	8507
0.40	1.20	784	1196	3137	4785
0.50	1.20	669	1021	2008	3062
0.10	1.40	776	1183	14737	22479
0.20	1.40	409	624	3684	5620
0.30	1.40	289	441	1637	2498
0.40	1.40	230	351	921	1405
0.50	1.40	197	300	590	899
0.10	1.60	398	606	7553	11520
0.20	1.60	210	320	1888	2880
0.30	1.60	148	226	839	1280
0.40	1.60	118	180	472	720
0.50	1.60	101	154	302	461
0.10	1.80	254	388	4829	7366
0.20	1.80	134	205	1207	1842
0.30	1.80	95	144	537	818
0.40	1.80	76	115	302	460
0.50	1.80	64	98	193	295
0.10	2.00	183	279	3473	5297
0.20	2.00	97	147	868	1324
0.30	2.00	68	104	386	589
0.40	2.00	54	83	217	331
0.50	2.00	46	71	139	212

This table shows example design settings and their respective required sample sizes for a specified significance level of 1% and a censoring rate of 30%.

Table 4.2a: Relative error for dominant MOI with $p_d = 0.7$ and $\Delta = 1.6$

	Taylor series approximation		Hsieh and Lavori Formula		
ε_{ij}	MSSN	N^*	MSSN	D^*	$R(\tilde{e})$
0.005	1.1347	89	1.1410	89	0.0055
0.01	1.2693	99	1.2951	100	0.0199
0.015	1.4040	109	1.4635	113	0.0407
0.02	1.5387	119	1.6471	127	0.0658
0.025	1.6733	130	1.8470	143	0.0940
0.03	1.8080	140	2.0646	160	0.1243
0.035	1.9427	150	2.3011	177	0.1558
0.04	2.0773	160	2.5581	197	0.1879
0.045	2.2120	171	2.8371	219	0.2203
0.05	2.3467	181	3.1400	243	0.2527

Table 4.2b: Relative error for recessive MOI with $p_d = 0.2$ and $\Delta = 1.6$

	Taylor series approximation		Hsieh and Lavori formula		
ε_{ij}	MSSN	D^*	MSSN	D^*	$R(\tilde{e})$
0.005	1.0826	1083	1.0845	1084	0.0018
0.01	1.1651	1166	1.1732	1173	0.0069
0.015	1.2477	1247	1.2660	1266	0.0145
0.02	1.3302	1330	1.3632	1363	0.0242
0.025	1.4128	1413	1.4652	1466	0.0358
0.03	1.4953	1496	1.5720	1570	0.0488
0.035	1.5779	1577	1.6841	1684	0.0631
0.04	1.6604	1660	1.8018	1801	0.0785
0.045	1.7430	1743	1.9254	1926	0.0948
0.05	1.8255	1826	2.0554	2056	0.1119

These graphs display the relative error $R(\tilde{e}) = |S(\tilde{e}) - HL(\tilde{e})| / HL(\tilde{e})$ for differing modes of inheritance assuming a uniform error model $e_{ij} = e \ \forall i \neq j \in \{0,1,2\}$.

Table 4.3: Bounds of %MSSN coefficients (G × E interaction only model)

Dominant MOI		Bounds		p_d	$r_1(t)=r_2(t)$
D_{10}	Misclassification of the *d*+ genotype → ++ genotype	Lower	1	↓ 0	
		Upper	14	0.7	2
D_{20}	Misclassification of the *dd* genotype → ++ genotype	Lower	0	↓ 0	
		Upper	16.3	0.7	2
D_{01}	Misclassification of the ++ genotype → *d*+ genotype	Lower	0	0.7	
		Upper	∞	↓ 0	
D_{02}	Misclassification of the ++ genotype → *dd* genotype	Lower	0	0.7	
		Upper	∞	↓ 0	

Recessive MOI		Bounds		p_d	$r_1(t)=1, r_2(t)$
D_{20}	Misclassification of the *dd* genotype → ++ genotype	Lower	1	↓ 0	
		Upper	3	0.7	2
D_{21}	Misclassification of the *dd* genotype → *d*+ genotype	Lower	1	↓ 0	
		Upper	3	0.7	2
D_{02}	Misclassification of the ++ genotype → *dd* genotype	Lower	3.5	0.7	2
		Upper	∞	↓ 0	
D_{12}	Misclassification of the *d*+ genotype → *dd* genotype	Lower	2.5	0.7	2
		Upper	∞	↓ 0	

Note that $D_{12} = D_{21} = 0$ for dominant models and $D_{01} = D_{10} = 0$ for recessive models. Recall that the %MSSN coefficients are not functions of the target level of significance, power and censoring proportion.

Figure 4.1a: Taylor series approximation for dominant MOI

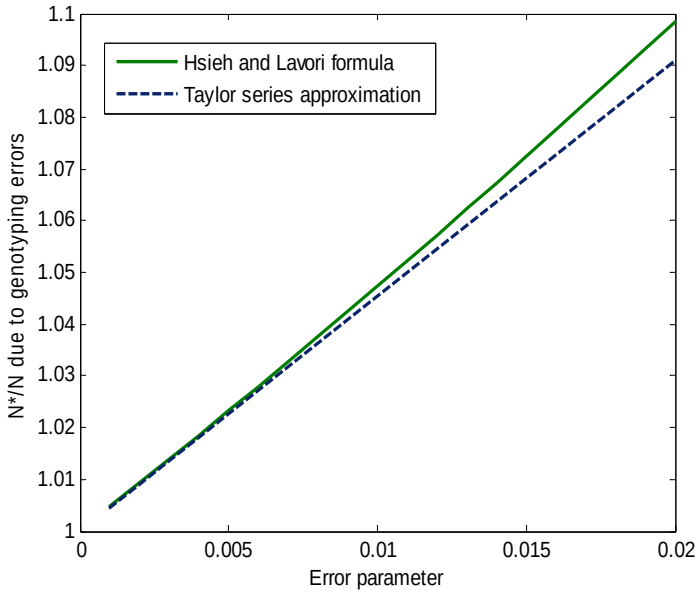
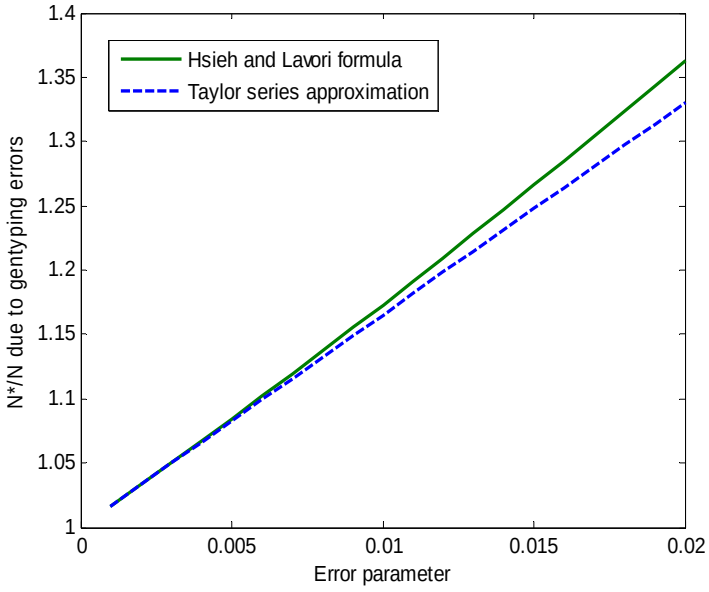


Figure 4.1b: Taylor series approximation for recessive MOI



These graphs display $HL(\tilde{e})$ and $S(\tilde{e})$ as functions of $e_{ij} = e \ \forall i \neq j \in \{0,1,2\}$ given $p_d = 0.2$ and $\Delta = 1.6$.

Figure 4.2a: Sum of %MSSN coefficients for dominant MOI

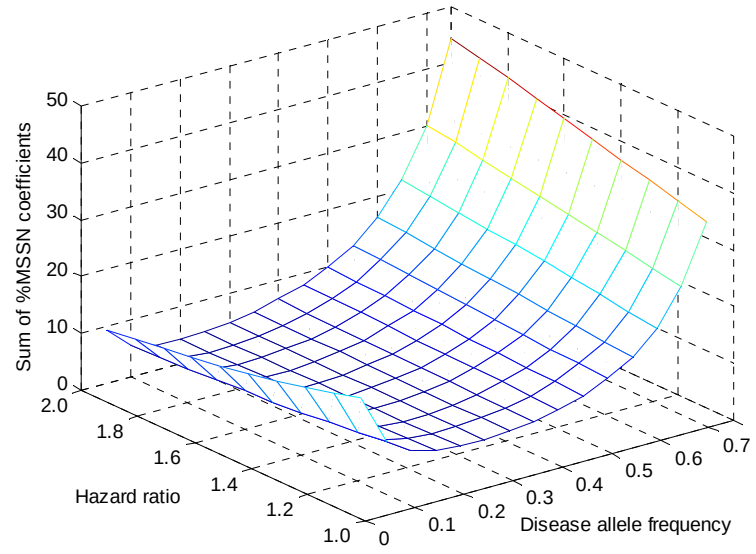


Figure 4.2b: Sum of %MSSN coefficients for recessive MOI

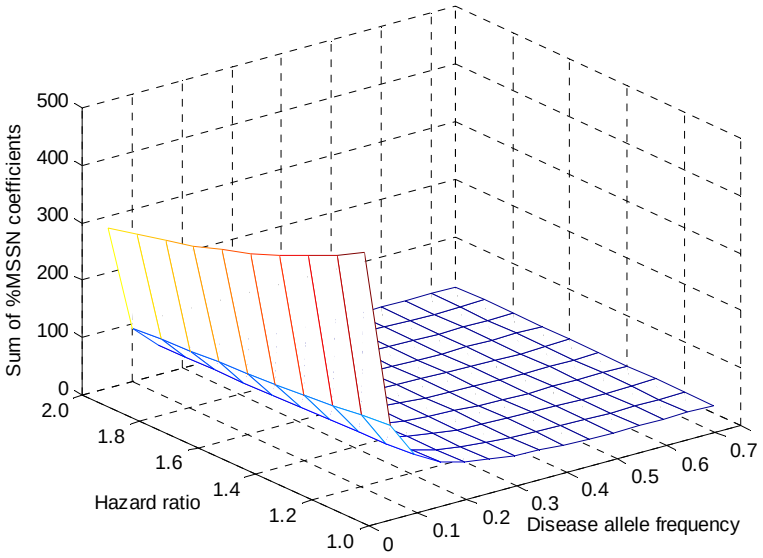
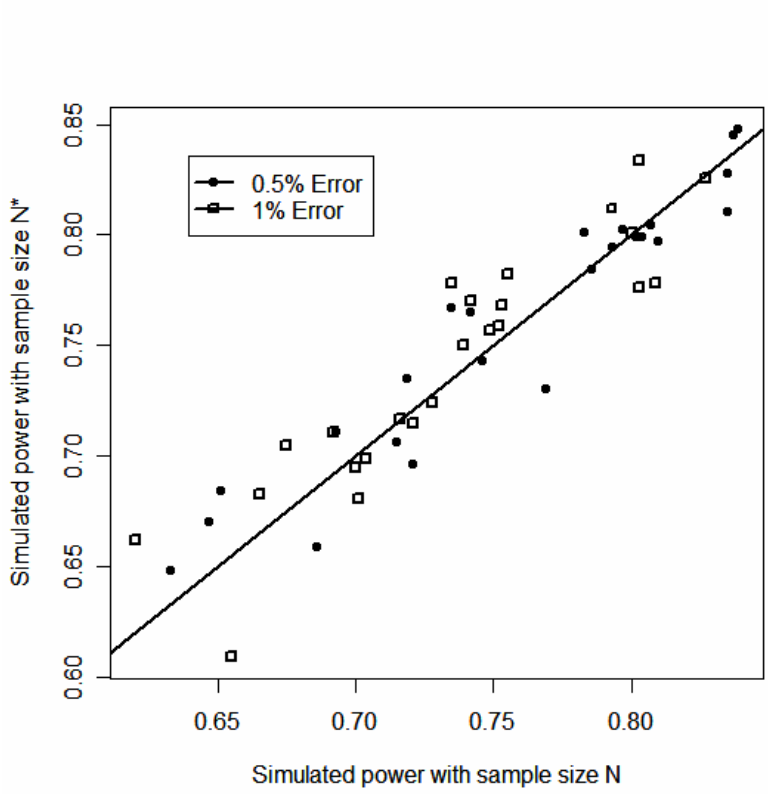


Figure 4.3: Simulation results (G × E interaction only model)



The horizontal axis represents the simulated power with a sample size of *N*, the required sample size with perfect genotyping classification and the vertical axis represents the simulated power with a sample size of *N**, the required sample size in the presence of genotyping misclassification errors. The solid line represents a perfect concordance between the simulated powers of sample size *N* and *N**.

Chapter 5 – Full G × E Interaction Model Results

This chapter addresses the design considerations in detecting a $G \times E$ interaction when genetic and environmental marginal effects are included in the model (Model IV or full model) given a direct association study. The required sample size to detect a Model IV $G \times E$ interaction is equal to the required sample size to detect a Model III $G \times E$ interaction multiplied by a variance inflation factor (VIF) (Hsieh and Lavori 2000, for details refer to Chapter 2). Thus, for a fixed sample size, one has more power to detect a $G \times E$ interaction through the interaction only model (Model III) than the full model (Model IV). However, the full model is more commonly used as researchers often are interested in estimating genetic and environmental marginal effects, in addition to their interaction. It is still feasible to detect a $G \times E$ interaction in the full model but this design requires a larger sample size as indicated by the VIF in Table 5.1. The VIF is dependent upon genotype frequency and is thus larger for dominant models, since the expected at-risk genotype frequency for a dominant MOI is $2p_d(1-p_d)$ greater than the expected at-risk genotype frequency for a recessive MOI (as noted in Chapter 4). This design exhibits the same characteristics as the interaction only model; namely, recessive models usually require a substantially greater sample size, and small disease allele frequencies and hazard ratios yield unreasonably large sample sizes.

The distribution of the relative error $R(\tilde{e})$ for a dominant and recessive MOI is displayed in Figure 5.1. The $R(\tilde{e})$ is less than 1.6% for dominant models and less than 36.9% for recessive models. The distribution of $R(\tilde{e})$ given a recessive MOI is highly right skewed and has many outliers corresponding to designs with small disease allele frequencies. The mean and median $R(\tilde{e})$ for a recessive MOI are 2.2% and 0.36% respectively. The distribution of the relative error for a dominant MOI exhibits a normal distribution, with a mean of 0.32% and standard deviation of 0.27% (median 0.23%).

The increase in required sample size due to genotyping errors $HL(\tilde{e})$ is approximated with the Taylor series approximation $S(\tilde{e})$. The discrepancy between $HL(\tilde{e})$ and $S(\tilde{e})$ for a design setting of $p_d = 0.2$ and $\Delta = 1.6$ is portrayed in Figure 5.2 given a dominant and recessive MOI. The $R(\tilde{e})$ is less than 1.0% for a dominant MOI (Figure 5.2a) and less than 2.9% for a recessive MOI (Figure 5.2b). Thus, the $R(\tilde{e})$ indicates that $S(\tilde{e})$ is an adequate approximation of $HL(\tilde{e})$ and can be used for design evaluation purposes.

The %MSSN coefficients (F_{ij}) for the full model are strictly functions of p_d, $r_1(t)$ and $r_2(t)$ similar to the interaction only model (see Appendix 3 for explicit formulas). Note that $F_{12} = F_{21} = 0$ for a dominant MOI and $F_{01} = F_{10} = 0$ for a recessive MOI (for details, see Chapter 4). Also similar to the %MSSN coefficients for the interaction only model, $F_{02} = F_{01}$ for a dominant MOI and $F_{20} = F_{21}$ for a recessive MOI.

Figures 5.3a and 5.3b display the effects of p_d and Δ upon the sum of %MSSN coefficients for a dominant and recessive MOI respectively. These graphs demonstrate similar behavior as the sum of the %MSSN coefficients in the interaction only model. There remains a pronounced interaction between larger hazard ratios and disease allele

frequencies for a dominant MOI and a significant increase in the sum of the %MSSN coefficients as $p_d \downarrow 0$ for a recessive MOI. As exhibited by the clear quadratic form of the %MSSN coefficients given a dominant MOI (Figure 5.3a), the sum of the %MSSN coefficients increases more dramatically as $p_d \downarrow 0$ for the full model than the sum of the %MSSN coefficients for the interaction only model (see Figure 4.2a). Similar to the interaction only model, the sum of the %MSSN coefficients are substantially greater for a recessive MOI as $p_d \downarrow 0$, indicating that genotyping misclassification errors in these instances cause deleterious impacts on power and sample size.

Table 5.2 specifies the upper and lower bounds of each %MSSN coefficient F_{ij} and their respective design parameters. As expected, the %MSSN coefficients F_{ij} obtain their upper and lower bounds at the same design settings as in Model III. Given a dominant MOI, the coefficients associated with the misclassifications of the *d*+ and *dd* genotype to the ++ genotype (F_{10} and F_{20}) attain their upper bounds of 10 and 11 respectively when $p_d = 0.7$ and $\Delta = 2$. With the same design settings and given a recessive MOI, the coefficients associated with any misclassification of the *dd* genotype to the ++ or *d*+ genotype (F_{20} and F_{21}) attain their upper bounds of 3 and the coefficients associated with misclassifications of the ++ and *d*+ genotype to the *dd* genotype (F_{02} and F_{12}) achieve their lower bounds of 0.25 and 1 respectively. For a dominant MOI when $p_d = 0.7$, the %MSSN coefficient associated with any misclassification of the ++ genotype (F_{02} and F_{01}) obtains its lower bound of 0. As $p_d \downarrow 0$ for a dominant MOI, the coefficients associated with misclassifications of the *d*+ genotype and *dd* genotype to the ++ genotype (F_{10} and F_{20}) achieve their respective lower bounds of 1 and 0. Similarly, as $p_d \downarrow 0$ for a recessive MOI, the coefficients associated with any misclassification of the *dd* genotype (F_{20} and F_{21}) attain their lower bounds of 0. Any misclassification of a subject without the at-risk genotype results in an indefinite increase in their associated %MSSN coefficient (F_{02} and F_{01} for a dominant MOI; F_{02} and F_{12} for a recessive MOI) as $p_d \downarrow 0$. Furthermore, these misclassifications are more deleterious for recessive MOI than dominant MOI. The misclassification of the ++ genotype to the *dd* genotype (F_{02}) for a recessive MOI is the largest coefficient in every design setting.

Simulation study results

Figure 5.4 compares the simulated powers between the sample sizes calculated for the $G \times E$ interaction only model (Model III) and the full $G \times E$ interaction model (Model IV) from the Hsieh and Lavori (2000) sample size formula for a specified power of 80%. The solid line again represents a perfect correspondence between these simulated powers with differing $G \times E$ interaction models. The average Model III simulated power was 0.746 (standard deviation (SD) = 0.059) and the average Model IV simulated power was 0.663 (SD = 0.054). For $G \times E$ hazard ratios of 1.25, the mean ± SD simulated powers for Model III and Model IV were 0.815 ± 0.0193 and 0.735 ± 0.0161,

respectively. For $G \times E$ hazard ratios of 1.50, the mean ± SD simulated powers for Model III and Model IV were 0.773 ± 0.0285 and 0.666 ± 0.0200, respectively. For $G \times E$ hazard ratios of 1.75, the mean ± SD simulated powers for Model III and Model IV were 0.727 ± 0.0200 and 0.662 ± 0.0216, respectively. For $G \times E$ hazard ratios of 2.00, the mean ± SD simulated powers for Model III and Model IV were 0.671 ± 0.0309 and 0.591 ± 0.0159, respectively. The Model IV simulated power was on average 0.083 (SD = 0.036) less than the Model III simulated power, differing significantly from the Model III simulated power (paired $t = 15.933$, df = 47, $p < 0.0001$).

The VIF significantly underestimated the increase in sample size necessary to maintain a constant power given the inclusion of correlated covariates in the model. The simulated power is highly dependent upon the size of the hazard ratio, with larger hazard ratios yielding less power. The sample size for Model III significantly underestimates the specified power for hazard ratios ≥ 1.75 and additionally, the VIF significantly underestimates the increase in sample size for Model IV. Thus, the sample size for Model IV significantly underestimates the specified power for all hazard ratios considered. It is imperative that researchers perform simulations to confirm that their sample size yields sufficient power.

Figure 5.5 displays results of the full $G \times E$ interaction model genotyping misclassification errors simulation study. As noted previously, the Hsieh and Lavori (2000) sample size formula significantly underestimates the target power of 80% in all design settings. However, this Figure shows that the increase in sample size due to genotyping misclassification errors maintains a constant power even in situations where the simulated power notably underestimates the specified power. The solid line represents a line with a slope of 1 and an intercept of 0. Regressing the simulated power with sample size N^* on the simulated power with sample size N yields a significant ($p < 0.001$) intercept estimate of 0.132 with a standard error of 0.0365, and a slope estimate of 0.801 with a standard error of 0.0548. The estimated standardized beta-coefficient of the slope (e.g. the Pearson correlation coefficient between these simulated powers) is 0.907 with a 95% CI of (0.839-0.947). This strong correlation can be seen by the proximity of the points to the solid line for both genotyping error rates of 0.5% and 1%. The intercept estimate of the linear regression model suggests that the simulated powers with sample size N^* are slightly higher than the simulated powers with sample size N, suggesting that the increase in sample size due to genotyping errors is somewhat conservative.

Table 5.1: Example design settings (Full G × E interaction model)

Design		Dominant MOI			Recessive MOI		
p_d	Δ	VIF	N 80% power	N 95% power	VIF	N 80% power	N 95% power
0.1	1.2	1.2346	3261	4975	1.0101	50699	77332
0.2	1.2	1.5625	2179	3323	1.0417	13071	19937
0.3	1.2	2.0408	2009	3064	1.0989	6128	9348
0.4	1.2	2.7778	2179	3323	1.1905	3735	5696
0.5	1.2	4.0000	2677	4083	1.3333	2677	4083
0.1	1.4	1.2346	958	1461	1.0101	14886	22706
0.2	1.4	1.5625	640	976	1.0417	3838	5854
0.3	1.4	2.0408	590	900	1.0989	1799	2745
0.4	1.4	2.7778	640	976	1.1905	1097	1673
0.5	1.4	4.0000	786	1199	1.3333	786	1199
0.1	1.6	1.2346	491	749	1.0101	7629	11637
0.2	1.6	1.5625	328	500	1.0417	1967	3000
0.3	1.6	2.0408	302	461	1.0989	922	1407
0.4	1.6	2.7778	328	500	1.1905	562	857
0.5	1.6	4.0000	403	614	1.3333	403	614
0.1	1.8	1.2346	314	479	1.0101	4878	7440
0.2	1.8	1.5625	210	320	1.0417	1258	1918
0.3	1.8	2.0408	193	295	1.0989	590	899
0.4	1.8	2.7778	210	320	1.1905	359	548
0.5	1.8	4.0000	258	393	1.3333	258	393
0.1	2.0	1.2346	226	344	1.0101	3508	5350
0.2	2.0	1.5625	151	230	1.0417	904	1379
0.3	2.0	2.0408	139	212	1.0989	424	647
0.4	2.0	2.7778	151	230	1.1905	258	394
0.5	2.0	4.0000	185	283	1.3333	185	283

This table displays example design settings and their respective required sample sizes for a specified significance level of 1% and a censoring rate of 30%.

Table 5.2: Bounds of %MSSN coefficients (Full G × E interaction model)

Dominant MOI		Bounds		p_d	$r_1(t)=r_2(t)$
F_{10}	Misclassification of the *d*+ genotype → ++ genotype	Lower	1	↓0	
		Upper	10	0.7	2
F_{20}	Misclassification of the *dd* genotype → ++ genotype	Lower	0	↓0	
		Upper	11	0.7	2
F_{01}	Misclassification of the ++ genotype → *d*+ genotype	Lower	0	0.7	
		Upper	∞	↓0	
F_{02}	Misclassification of the ++ genotype → *dd* genotype	Lower	0	0.7	
		Upper	∞	↓0	

Recessive MOI		Bounds		p_d	$r_1(t)=1, r_2(t)$
F_{20}	Misclassification of the *dd* genotype → ++ genotype	Lower	1	↓0	
		Upper	3	0.7	2
F_{21}	Misclassification of the *dd* genotype → *d*+ genotype	Lower	1	↓0	
		Upper	3	0.7	2
F_{02}	Misclassification of the ++ genotype → *dd* genotype	Lower	0.25	0.7	2
		Upper	∞	↓0	
F_{12}	Misclassification of the *d*+ genotype → *dd* genotype	Lower	1	0.7	2
		Upper	∞	↓0	

The coefficients $F_{12} = F_{21} = 0$ for a dominant MOI and $F_{01} = F_{10} = 0$ for a recessive MOI.

Figure 5.1a: Relative error for dominant MOI

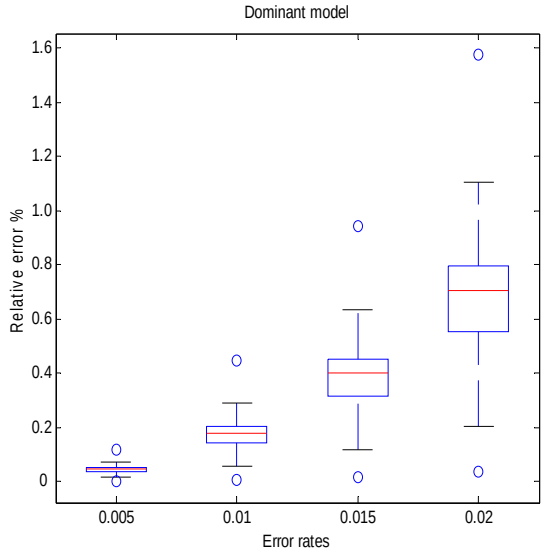


Figure 5.1b: Relative error for recessive MOI

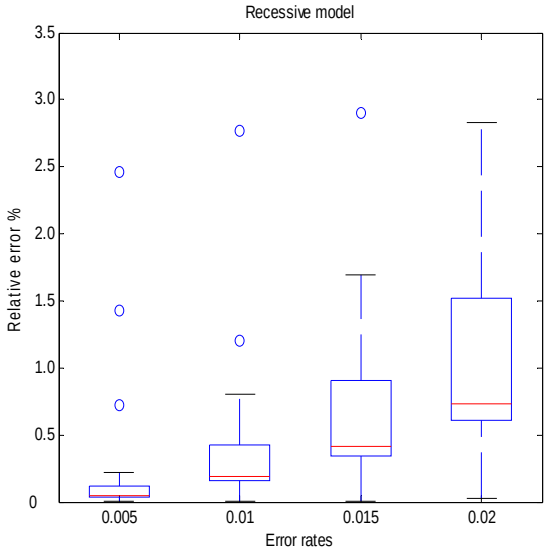


Figure 5.2a: Taylor series approximation for dominant MOI

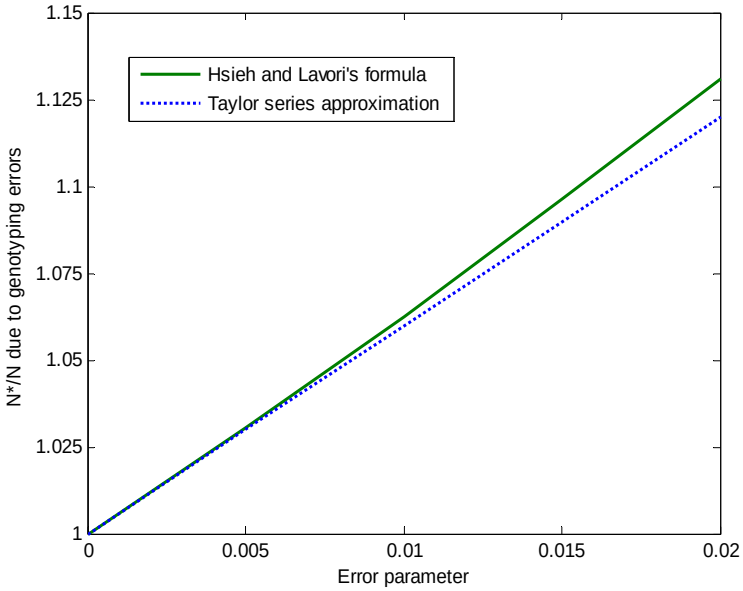
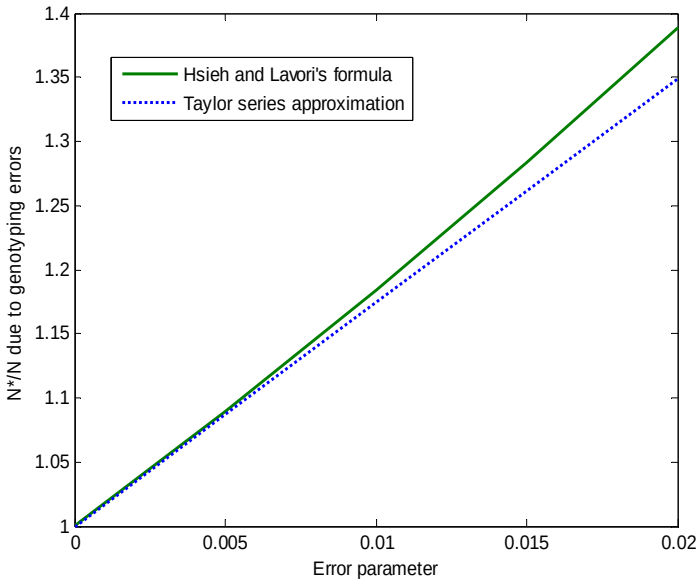


Figure 5.2b: Taylor series approximation for recessive MOI



These graphs display $HL(\tilde{e})$ and $S(\tilde{e})$ as functions of $e_{ij} = e \ \forall i \neq j \in \{0,1,2\}$ given $p_d = 0.2$ and $\Delta = 1.6$.

Figure 5.3a: Sum of %MSSN coefficients for dominant MOI

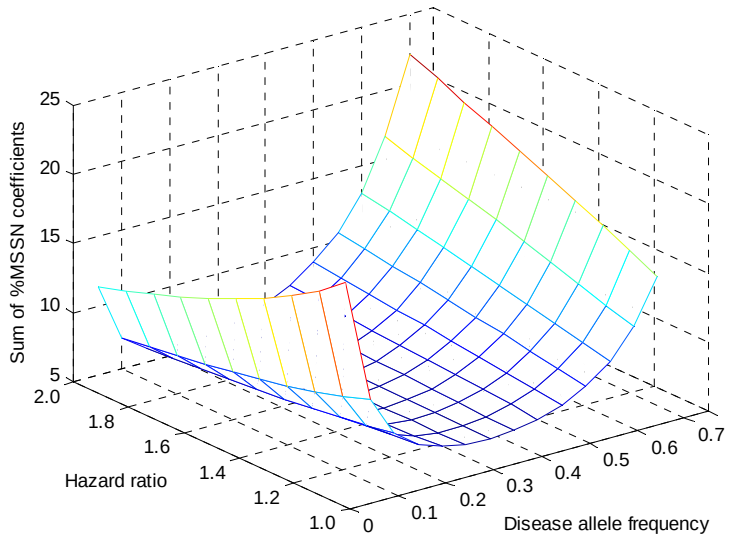


Figure 5.3b: Sum of %MSSN coefficients for recessive MOI

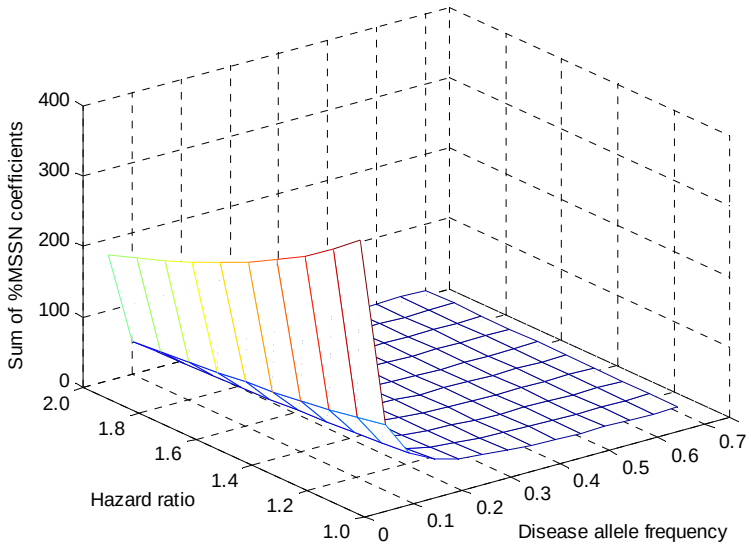
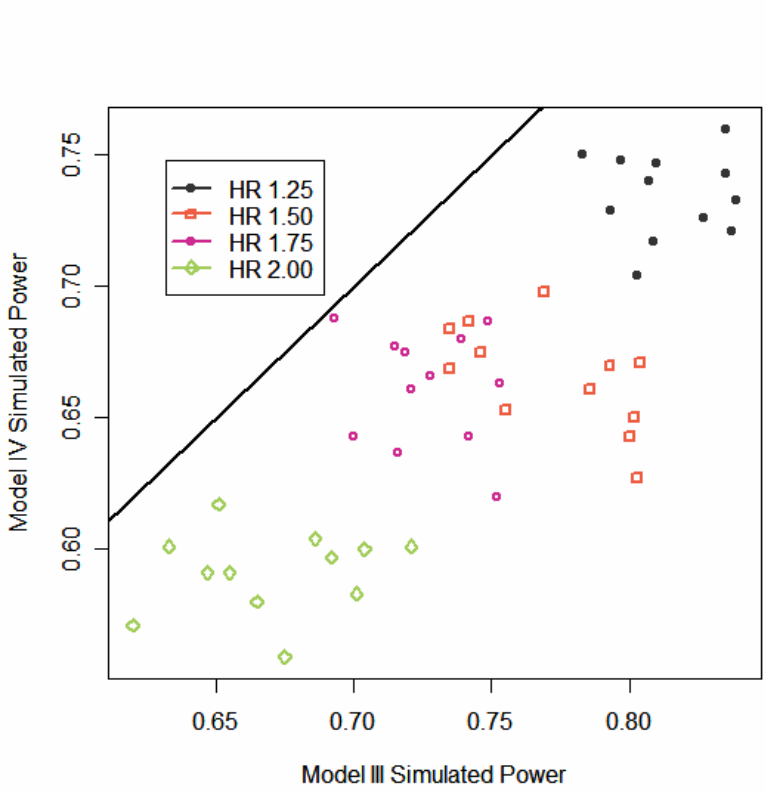
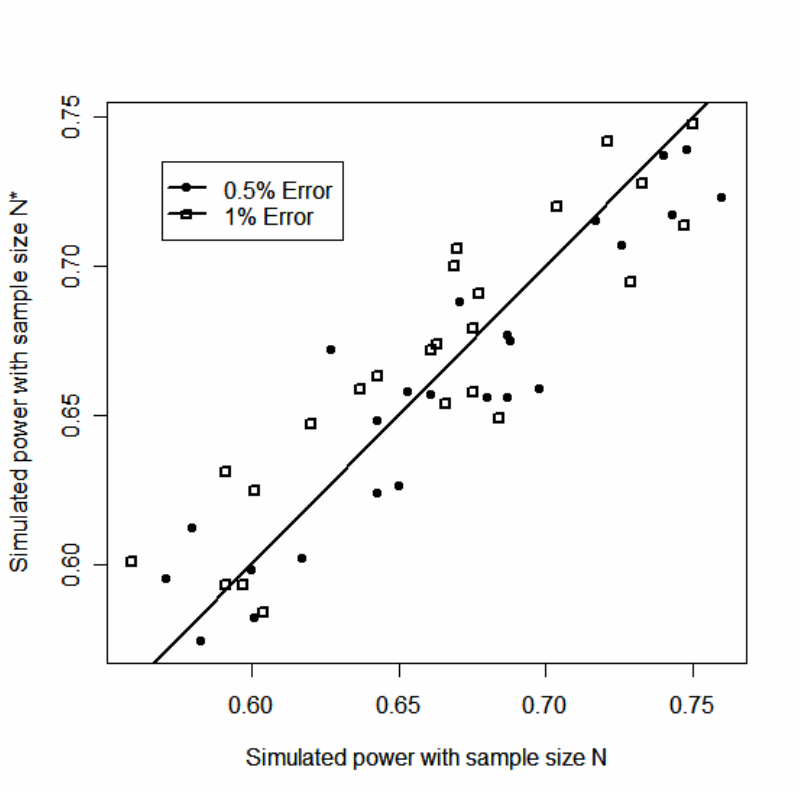


Figure 5.4: Simulated powers of the full interaction model versus the interaction only model



The horizontal axis represents the simulated power with the required sample size for the $G \times E$ interaction only model (Model III) and the vertical axis represents the simulated power with the sample size required for the full $G \times E$ interaction model (Model IV) with the solid line representing a perfect concordance between the two simulated powers.

Figure 5.5: Simulation results (Full G × E interaction model)



Similar to Figure 4.3, the horizontal axis represents the simulated power with a sample size of *N*, the required sample size with perfect genotyping classification and the vertical axis represents the simulated power with a sample size of *N**, the required sample size in the presence of genotyping misclassification errors. The solid line represents a perfect concordance between the simulated powers of sample size *N* and *N**.

Chapter 6 – Discussion and Conclusion

Gene-environment interactions will be of considerable importance in future scientific research. Significant findings will greatly impact disease treatment, disease prevention techniques and methods, policy making, dietary guidelines, to name a few as we begin to understand the integration of nature and nurture. It is important that statistical methods consider genotyping misclassification errors, as these errors can have detrimental effects on power and sample size. This research shows that non-differential genotyping errors bias the estimates towards the null and thus increase the required sample size to maintain a constant power similar to previous misclassification research (García-Closas, et al. 1999).

One limitation of Chapters 4 and 5 is that it did not consider indirect association studies which allow researchers to examine the polymorphism acting as a surrogate for the causal locus with respect to disease status (Cordell and Clayton 2005). The $G \times E$ interaction was modeled in these chapters with a single model coefficient for a dominant or recessive MOI. There lacks an appropriate sample size formula for a model in which the $G \times E$ interaction models > 2 genotypes with more than one coefficient. A simulation study can be performed to investigate the impact of genotyping errors for a two parameter indirect association $G \times E$ interaction model.

One of the limitations of designing a genetic cohort study to investigate the possibilities of $G \times E$ interactions is the large representative sample size. Partial-collection study designs have been suggested to increase the efficiency in detecting a $G \times E$ interaction for cross-sectional studies. Partial-collection $G \times E$ interaction studies include case-parent trio designs, matched case-control designs, case-sibling(s) designs, and case-only designs (Andrieu, et al. 2005; Gauderman 2002b; Schaid 1999; Umbach and Weinberg 1997; Yang, et al. 1997). Although these partial-collection designs can be more efficient, they are limited by their assumption of genotype and environment independence and are only applicable to multiplicative scales. Despite the limitations of the partial-collection designs, their efficiency makes them an attractive alternative for researchers interested in $G \times E$ interactions (Lui, et al. 2004). However, longitudinal studies are imperative to understanding the influence of genetic and environmental effects upon disease progression and therefore it would be extremely useful to extend the partial-collection designs to survival analysis or other longitudinal methods.

The sample size formulas considered for this research (Halabi and Singh 2004; Hsieh and Lavori 2000) assume the alternative hypothesis be near the null hypothesis (e.g. small effect size). The sample size formulas underestimate the specified power for larger hazard ratios (> 2 for the Halabi and Singh formula and > 1.75 for the Hsieh and Lavori formula). Moreover, the required sample size for the full $G \times E$ interaction model (Model IV) according to the Hsieh and Lavori formula (2000) is significantly less than the specified power, even for small effect sizes. This is due to the sample size increase underestimate of the VIF, which yields a (simulated) power on average 0.083 less than the Model III simulated power. Researchers should perform simulations especially for the full $G \times E$ interaction model to ensure that their sample size yields sufficient power to detect the hypothesized effect.

Chapter 3 illustrated the robustness of the LD parameter r^2 in the survival analysis framework. The increase in sample size due to an indirect association study

(N_I / N_D) using the log-rank test statistic is approximately equal to $1/r^2$, similar to case-control genetic association studies. It can be conjectured that the increase in sample size due to an indirect association study is approximately equal to $1/r^2$ for more complex Cox PH models as well.

This work found that for an indirect association, the misclassification of the more common homozygote to the less common homozygote has the most deleterious impact upon the power of the study than all other SNP genotyping errors. Furthermore, the %MSSN coefficient associated with any misclassification of the more common homozygote increases without bound as the minor SNP allele frequency goes to 0. This finding is consistent with genetic association cross-sectional studies utilizing the chi-square test statistic and linear trend test (Kang, et al. 2004a; Ahn, et al. 2007). This work also found that that for direct association studies, the %MSSN coefficient associated with any misclassification of a subject without the at-risk genotype to an at-risk genotype increase indefinitely as the disease allele frequency approaches 0. These results find that a misclassification of the most common genotype has the worst impact on power and sample size as it contaminates the other genotypes whose frequencies are smaller.

Future work should consider the impact of differential genotyping misclassification errors and the impact of environmental measurement (or misclassification) error in the detection of a $G \times E$ interaction using survival analysis techniques. The environmental covariate if often derived from the respondent's self-repot which is subject to further bias (e.g. recall bias, telescoping bias). It would be informative to model gene expression and environment as time-varying covariates and examine the impact of errors in these scenarios. There is no doubt that understanding $G \times E$ interactions will be an important tool therefore it is imperative that well-designed studies have sufficient power that adjusts for potential biases created by misclassification errors.

References

Abel L, Muller-Myhsok B. 1998. Maximum-likelihood expression of the transmission/disequilibrium test and power considerations. American Journal of Human Genetics 63(2):664-667.

Ahn K, Haynes C, Kim W, Fleur R, Gordon D, Finch S. 2007. The effects of SNP genotyping errors on the power of the Cochran-Armitage linear trend test for case/control association studies. Annals of Human Genetics 71(Pt 2):249-61.

Ahnn S, Anderson SJ. 1995. Sample-size determination for comparing more than 2 survival distributions. Statistics in Medicine 14(20):2273-2282.

Ahnn S, Anderson SJ. 1998. Sample size determination in complex clinical trials comparing more than two groups for survival endpoints. Statistics in Medicine 17(21):2525-2534.

Akazawa K, Nakamura T, Moriguchi S, Shimada M, Nose Y. 1991. Simulation program for estimating statistical power of Cox's proportional hazards model assuming no specific distribution for the survival time. Computer Methods and Programs in Biomedicine 35(3):203-12.

Akey JM, Zhang K, Xiong MM, Doris P, Jin L. 2001. The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures. American Journal of Human Genetics 68(6):1447-1456.

Andrieu N, Dondon M, Goldstein A. 2005. Increased power to detect gene-environment interaction using siblings controls. Annals of Epidemiology 15(9):705-11.

Barral S, Haynes C, Levenstien M, Gordon D. 2005. Precision and type I error rate in the presence of genotype errors and missing parental data: a comparison between the original transmission disequilibrium test (TDT) and TDTae statistics. BMC Genetics 6 Supplement 1:S150.

Beer DG, Kardia SLR, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen GA, Gharib TG, Thomas DG and others. 2002. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nature Medicine 8(8):816-824.

Bender R, Augustin T, Blettner M. 2005. Generating survival times to simulate Cox proportional hazards models. Statistics in Medicine 24(11):1713-23.

Bernardo M, Harrington D. 2001. Sample size calculations for the two-sample problem using the multiplicative intensity model. Statistics in Medicine 20(4):557-79.

Bernstein D, Lagakos SW. 1978. Sample size and power determination for stratified clinical trails. Journal of Statistical Computation and Simulation. p 65-73.

Bross I. 1954. Misclassification in 2 x 2 tables. Biometrics 10(4):478-486.

Casella G, Berger RL. 2002. Statistical inference. Pacific Grove, Calif.: Thomson Learning.

Caspi A, McClay J, Moffitt T, Mill J, Martin J, Craig I, Taylor A, Poulton R. 2002. Role of genotype in the cycle of violence in maltreated children. Science 297(5582):851-4.

Caspi A, Sugden K, Moffitt TE, Taylor A, Craig IW, Harrington H, McClay J, Mill J, Martin J, Braithwaite A and others. 2003. Influence of life stress on depression: Moderation by a polymorphism in the 5-HTT gene. Science 301(5631):386-389.

Chang YPC, Kim JDO, Schwander K, Rao DC, Miller MB, Weder AB, Cooper RS, Schork NJ, Province MA, Morrison AC and others. 2006. The impact of data

quality on the identification of complex disease genes: experience from the Family Blood Pressure Program. European Journal of Human Genetics 14(4):469-477.
Chan HSL, Haddad G, Thorner PS, Deboer G, Lin YP, Ondrusek N, Yeger H, Ling V. 1991. P-glycoprotein expression as a predictor of the outcome of therapy for neuroblastoma. New England Journal of Medicine 325(23):1608-1614.
Chen J, Stampfer MJ, Hough HL, Garcia-Closas M, Willett WC, Hennekens CH, Kelsey KT, Hunter DJ. 1998. A prospective study of N-acetyltransferase genotype, red meat intake, and risk of colorectal cancer. Cancer Research 58(15):3307-3311.
Cordell HJ, Clayton DG. 2005. Genetic epidemiology 3 - Genetic association studies. Lancet 366(9491):1121-1131.
Dean M, Carrington M, Winkler C, Huttley GA, Smith MW, Allikmets R, Goedert JJ, Buchbinder SP, Vittinghoff E, Gomperts E and others. 1996. Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Science 273(5283):1856-1862.
Deitz A, Rothman N, Rebbeck T, Hayes R, Chow W, Zheng W, Hein D, García-Closas M. 2004. Impact of misclassification in genotype-exposure interaction studies: example of N-acetyltransferase 2 (NAT2), smoking, and bladder cancer. Cancer Epidemiology Biomarkers and Prevention 13(9):1543-6.
Douglas J, Skol A, Boehnke M. 2002. Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. American Journal of Human Genetcis 70(2):487-95.
Draper NR, Smith H. 1998. Applied regression analysis. New York ; Chichester: John Wiley.
Edwards B, Haynes C, Levenstien M, Finch S, Gordon D. 2005. Power and sample size calculations in the presence of phenotype errors for case/control genetic association studies. BMC Genetics 6(1):18.
Fleiss JL, Levin BA, Paik MC. 2003. Statistical methods for rates and proportions. Hoboken: Wiley.
Foppa I, Spiegelman D. 1997. Power and sample size calculations for case-control studies of gene-environment interactions with a polytomous exposure variable. American Journal of Epidemiology 146(7):596-604.
García-Closas M, Lubin J. 1999. Power and sample size calculations in case-control studies of gene-environment interactions: comments on different approaches. American Journal of Epidemiology 149(8):689-92.
Garcia-Closas M, Rothman N, Lubin J. 1999. Misclassification in case-control studies of gene-environment interactions: assessment of bias and sample size. Cancer Epidemiology Biomarkers and Prevention 8(12):1043-50.
García-Closas M, Thompson W, Robins J. 1998. Differential misclassification and the assessment of gene-environment interactions in case-control studies. American Journal of Epidemiology 147(5):426-33.
Gauderman W. 2002a. Sample size requirements for association studies of gene-gene interaction. American Journal of Epidemiology 155(5):478-84.
Gauderman W. 2002b. Sample size requirements for matched case-control studies of gene-environment interaction. Statistics in Medicine 21(1):35-50.
Geller F, Ziegler A. 2002. Detection rates for genotyping errors in SNPs using the trio

design. Human Heredity 54(3):111-7.
George S, Desu M. 1974. Planning the size and duration of a clinical trial studying the time to some critical event. Journal of Chronic Disease 27(1):15-24.
Goldstein A, Falk R, Korczak J, Lubin J. 1997. Detecting gene-environment interactions using a case-control design. Genetic Epidemiology 14(6):1085-9.
Gordon D, Finch SJ. 2005. Factors affecting statistical power in the detection of genetic association. Journal of Clinical Investigation 115(6):1408-1418.
Gordon D, Finch SJ, Nothnagel M, Ott J. 2002. Power and sample size calculations for case-control genetic association tests when errors are present: Application to single nucleotide polymorphisms. Human Heredity 54(1):22-33.
Gordon D, Heath S, Ott J. 1999. True pedigree errors more frequent than apparent errors for single nucleotide polymorphisms. Human Heredity 49(2):65-70.
Gordon D, Levenstien M, Finch S, Ott J. 2003. Errors and linkage disequilibrium interact multiplicatively when computing sample sizes for genetic case-control association studies. Pacific Symposium on Biocomputing:490-501.
Gordon D, Yang Y, Haynes C, Finch S, Mendell N, Brown A, Haroutunian V. 2004. Increasing power for tests of genetic association in the presence of phenotype and/or genotype error by use of double-sampling. Statistical Applications in Genetic and Molecular Biology 3:Article26.
Goto I, Yoneda S, Yamamoto M, Kawajiri K. 1996. Prognostic significance of germ line polymorphisms of the CYP1A1 and glutathione S-transferase genes in patients with non-small cell lung cancer. Cancer Research 56(16):3725-3730.
Govindarajulu US, Spiegelman D, Miller KL, Kraft P. 2006. Quantifying bias due to allele misclassification in case-control studies of haplotypes. Genetic Epidemiology 30(7):590-601.
Halabi S, Singh B. 2004. Sample size determination for comparing several survival curves with unequal allocations. Statistics in Medicine 23(11):1793-1815.
Haupts S, Ledergerber B, Boni J, Schupbach J, Kronenberg A, Opravil M, Flepp M, Speck RF, Grube C, Rentsch K and others. 2003. Impact of genotypic resistance testing on selection of salvage regimen in clinical practice. Antiviral Therapy 8(5):443-454.
Healey JF. 1993. Statistics, a tool for social research. Belmont, Calif.: Wadsworth Pub. Co.
Hosking L, Lumsden S, Lewis K, Yeo A, McCarthy L, Bansal A, Riley J, Purvis I, Xu C. 2004. Detection of genotyping errors by Hardy-Weinberg equilibrium testing. European Journal of Human Genetics 12(5):395-9.
Hosmer DW, Lemeshow S, Kim S. 1999. Applied survival analysis : regression modeling of time to event data. New York ; Chichester: Wiley.
Hsieh FY, Lavori PW. 2000. Sample-size calculations for the Cox proportional hazards regression model with nonbinary covariates. Controlled Clinical Trials 21(6):552-560.
Hunter D. 2005. Gene-environment interactions in human diseases. Nature Review Genetics 6(4):287-98.
Hu P, Tsiatis A, Davidian M. 1998. Estimating the parameters in the Cox model when covariate variables are measured with error. Biometrics 54(4):1407-19.
Hwang S, Beaty T, Liang K, Coresh J, Khoury M. 1994. Minimum sample size

estimation to detect gene-environment interaction in case-control designs. American Journal of Epidemiology 140(11):1029-37.
Jaffee S, Caspi A, Moffitt T, Dodge K, Rutter M, Taylor A, Tully L. 2005. Nature X nurture: genetic vulnerabilities interact with physical maltreatment to promote conduct problems. Development and Psychopathology 17(1):67-84.
Ji F, Yang YN, Haynes C, Finch SJ, Gordon D. 2005. Computing asymptotic power and sample size for case-control genetic association studies in the presence of phenotype and/or genotype misclassification errors. Statistical Applications in Genetics and Molecular Biology 4:26.
Kang SJ, Finch SJ, Haynes C, Gordon D. 2004a. Quantifying the percent increase in minimum sample size for SNP genotyping errors in genetic model-based association studies. Human Heredity 58(3-4):139-144.
Kang SJ, Gordon D, Finch SJ. 2004b. What SNP genotyping errors are most costly for genetic association studies? Genetic Epidemiology 26(2):132-141.
Khoury M, Flanders W. 1996. Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls! American Journal of Epidemiology 144(3):207-13.
Lachin J, Foulkes M. 1986. Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. Biometrics 42(3):507-19.
Lai RZ, Zhang H, Yang YN. 2007. Repeated measurement sampling in genetic association analysis with genotyping errors. Genetic Epidemiology 31(2):143-153.
Lakatos E. 1988. Sample sizes based on the log-rank statistic in complex clinical-trials. Biometrics 44(1):229-241.
Lakatos E, Lan K. 1992. A comparison of sample size methods for the logrank statistic. Statistics in Medicine 11(2):179-91.
Latouche A, Porcher R, Chevret S. 2004. Sample size formula for proportional hazards modelling of competing risks. Statistics in Medicine 23(21):3263-74.
Lawless JF. 1982. Statistical models and methods for lifetime data. New York ; Chichester: Wiley.
Lincoln S, Lander E. 1992. Systematic detection of errors in genetic linkage data. Genomics 14(3):604-10.
Liu X, Fallin M, Kao W. 2004. Genetic dissection methods: designs used for tests of gene-environment interaction. Current Opinion in Genetics and Development 14(3):241-5.
Li Y, Scott W, Hedges D, Zhang F, Gaskell P, Nance M, Watts R, Hubble J, Koller W, Pahwa R and others. 2002. Age at onset in two common neurodegenerative diseases is genetically controlled. American Journal of Human Genetcis 70(4):985-93.
Luan J, Wong M, Day N, Wareham N. 2001. Sample size determination for studies of gene-environment interaction. International Journal of Epidemiology 30(5):1035-40.
Lubin J, Gail M. 1990. On power and sample size for studying features of the relative odds of disease. American Journal of Epidemiology 131(3):552-66.
Miller CR, Joyce P, Waits LP. 2002. Assessing allelic dropout and genotype reliability

using maximum likelihood. Genetics 160(1):357-366.
Morris RW, Kaplan NL. 2004. Testing for association with a case-parents design in the presence of genotyping errors. Genetic Epidemiology 26(2):142-154.
Ottman R. 1996. Gene-environment interaction: definitions and study designs. Preventive Medicine 25(6):764-70.
Pompanon F, Bonin A, Bellemain E, Taberlet P. 2005. Genotyping errors: Causes, consequences and solutions. Nature Reviews Genetics 6(11):847-859.
Pritchard JK, Przeworski M. 2001. Linkage disequilibrium in humans: Models and data. American Journal of Human Genetics 69(1):1-14.
Rice K, Holmans P. 2003. Allowing for genotyping error in analysis of unmatched case-control studies. Annals of Human Genetics 67(Pt 2):165-74.
Richardson D. 2003. Power calculations for survival analyses via Monte Carlo estimation. American Journal of Industrial Medicine 44(5):532-9.
Ross SM. 1998. A first course in probability. Upper Saddle River, N.J.: Prentice Hall.
Rothman KJ, Greenland S. 1998. Modern epidemiology. Philadelphia, PA: Lippincott-Raven.
Rubinstein L, Gail M, Santner T. 1981. Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. Journal of Chronic Disease 34(9-10):469-79.
Rutter M, Silberg J. 2002. Gene-environment interplay in relation to emotional and behavioral disturbance. Annual Review of Psychology 53:463-490.
Schaid D. 1999. Case-parents design for gene-environment interaction. Genetic Epidemiology 16(3):261-73.
Schaid DJ, Sommer SS. 1993. Genotype relative risks - methods for design and analysis of candidate-gene association studies. American Journal of Human Genetics 53(5):1114-1126.
Schmoor C, Sauerbrei W, Schumacher M. 2000. Sample size considerations for the evaluation of prognostic factors in survival analysis. Statistics in Medicine 19(4):441-52.
Schoenfeld DA. 1983. Sample-size formula for the proportional-hazards regression-model. Biometrics 39(2):499-503.
Schoenfeld DA, Borenstein M. 2005. Calculating the power or sample size for the logistic and proportional hazards models. Journal of Statistical Computation and Simulation 75(10):771-785.
Sham P. 1998. Statistics in human genetics. London New York: Arnold ;Wiley.
Simon R, Radmacher MD, Dobbin K. 2002. Design of studies using DNA microarrays. Genetic Epidemiology 23(1):21-36.
Sobel E, Papp J, Lange K. 2002. Detection and integration of genotyping errors in statistical genetics. American Journal of Human Genetics 70(2):496-508.
Talmud PJ, Stephens JW, Hawe E, Demissie S, Cupples LA, Hurel SJ, Humphries SE, Ordovas JM. 2005. The significant increase in cardiovascular disease risk in APOE epsilon 4 carriers is evident only in men who smoke: Potential relationship between reduced antioxidant status and ApoE4. Annals of Human Genetics 69:613-622.
Thomas D. 2000. Case-parents design for gene-environment interaction by Schaid. Genetic Epidemiology 19(4):461-3.

Tintle N, Gordon D, McMahon F, Finch S. 2007. Using duplicate genotyped data in genetic analyses: testing association and estimating error rates. Statistical Applications in Genetic and Molecular Biology 6:Article4.
Tu IP, Whittemore AS. 1999. Power of association and linkage tests when the disease alleles are unobserved. American Journal of Human Genetics 64(2):641-649.
Tung L, Gordon D, Finch S. 2007. The impact of genotype misclassification errors on the power to detect a gene-environment interaction using cox proportional hazards modeling. Human Heredity 63(2):101-10.
Umbach D, Weinberg C. 1997. Designing and analysing case-control studies to exploit independence of genotype and exposure. Statistics in Medicine 16(15):1731-43.
Umbas R, Isaacs WB, Bringuier PP, Schaafsma HE, Karthaus HFM, Oosterhof GON, Debruyne FMJ, Schalken JA. 1994. Decreased e-cadherin expression is associated with poor-prognosis in patients with prostate-cancer. Cancer Research 54(14):3929-3933.
Vaeth M, Skovlund E. 2004. A simple approach to power and sample size calculations in logistic regression and Cox regression models. Statistics in Medicine 23(11):1781-92.
van de Vijver MJ, He YD, van 't Veer LJ, Dai H, Hart AAM, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ and others. 2002. A gene-expression signature as a predictor of survival in breast cancer. New England Journal of Medicine 347(25):1999-2009.
Wong M, Day N, Luan J, Chan K, Wareham N. 2003. The detection of gene-environment interaction for continuous traits: should we deal with measurement error by bigger studies or better measurement? International Journal of Epidemiology 32(1):51-7.
Wong M, Day N, Luan J, Wareham N. 2004. Estimation of magnitude in gene-environment interactions in the presence of measurement error. Statistics in Medicine 23(6):987-98.
Wray NR. 2005. Allele frequencies and the 12 measure of linkage disequilibrium: Impact on design and interpretation of association studies. Twin Research and Human Genetics 8(2):87-94.
Yaffe K, Haan M, Byers A, Tangen C, Kuller L. 2000. Estrogen use, APOE, and cognitive decline - Evidence of gene-environment interaction. Neurology 54(10):1949-1953.
Yang Q, Khoury M, Flanders W. 1997. Sample size requirements in case-only designs to detect gene-environment interaction. American Journal of Epidemiology 146(9):713-20.
Younis J, Cooper JA, Miller GJ, Humphries SE, Talmud PJ. 2005. Genetic variation in alcohol dehydrogenase 1C and the beneficial effect of alcohol intake on coronary heart disease risk in the Second Northwick Park Heart Study. Atherosclerosis 180(2):225-232.
Zhen B, Murphy J. 1994. Sample size determination for an exponential survival model with an unrestricted covariate. Statistics in Medicine 13(4):391-7.
Zheng G, Tian X. 2005. The impact of diagnostic error on testing genetic association in case-control studies. Statistics in Medicine 24(6):869-82.
Zondervan KT, Cardon LR. 2004. The complex interplay among factors that influence allelic association (vol 5, pg 89, 2004). Nature Reviews Genetics 5(3):238-238.

Zou GH, Zhao HY. 2004. The impacts of errors in individual genotyping and DNA pooling on association studies. Genetic Epidemiology 26(1):1-10.

Appendix 1

Formulas for %MSSN coefficients for the log-rank model

The %MSSN coefficients, C_{ij}, are functions of the conditional survival probabilities for marker genotypes in the context of survival analysis ($P_{01}(t), P_{02}(t), P_{03}(t)$ – see Methods for details) and SNP marker genotype frequencies. Let π_{AA}, π_{AB} and π_{BB} denote the SNP marker genotype frequencies for genotypes *AA*, *AB* and *BB* respectively. Also, let Δ_1 denote the hazard ratio between subjects with genotypes *AA* and *BB* and let Δ_2 denote the hazard ratio between subjects with genotypes *AB* and *BB*. These hazard ratios are functions of the conditional survival probabilities for marker genotypes and marker genotype frequencies. Specifically, $\Delta_1 = \frac{P_{01}(t)\pi_{BB}}{P_{03}(t)\pi_{AA}}$ and $\Delta_2 = \frac{P_{02}(t)\pi_{BB}}{P_{03}(t)\pi_{AB}}$. The %MSSN coefficients are independent with respect to the target level of significance, power and censoring proportion.

The explicit %MSSN coefficients are:

$$C_{12} = \Omega \times \left[log\,\Delta_1^2\left(\pi_{AA} - 2\pi_{AA}^2\right) + log\,\Delta_2^2\left(2\pi_{AA}\pi_{AB} - \pi_{AA}\right) - 2\left(\pi_{AA}\pi_{AB} - \pi_{AA}^2\right) log\,\Delta_1\, log\,\Delta_2 \right.$$
$$\left. + 2\left(\frac{\pi_{AA}}{\pi_{AB}^2} - \frac{P_{01}(t)}{P_{02}(t)\pi_{AB}}\right)\left(\pi_{AA}\pi_{AB}^2\, log\,\Delta_1 - \pi_{AB}^2(1-\pi_{AB})\, log\,\Delta_2\right)\right]$$

$$C_{13} = \Omega \times \left[log\Delta_1^2\left(\pi_{AA} - 2\pi_{AA}^2\right) - 2\pi_{AA}\pi_{AB}\, log\Delta_1\, log\Delta_2 + 2\left(\frac{P_{01}(t)\pi_{BB} - P_{03}(t)\pi_{AA}}{P_{03}(t)\pi_{AA}\pi_{BB}}\right)\left(\pi_{AA}^2(1-\pi_{AA})\, log\Delta_1 - \pi_{AA}^2\pi_{AB}\, log\Delta_2\right)\right.$$
$$\left. + 2\left(\frac{\pi_{AA}}{\pi_{BB}^2} - \frac{P_{01}(t)}{P_{03}(t)\pi_{AB}}\right)\left(\pi_{AA}\pi_{AB}^2\, log\Delta_1 - \pi_{AB}^2(1-\pi_{AB})\, log\Delta_2\right)\right]$$

$$C_{21} = \Omega \times \left[log\,\Delta_1^2\left(2\pi_{AA}\pi_{AB} - \pi_{AB}\right) + log\,\Delta_2^2\left(\pi_{AB} - 2\pi_{AB}^2\right) - 2\left(\pi_{AA}\pi_{AB} - \pi_{AB}^2\right) log\,\Delta_1\, log\,\Delta_2 + \right.$$
$$\left. 2\left(\frac{\pi_{AB}}{\pi_{AA}^2} - \frac{P_{02}(t)}{P_{01}(t)\pi_{AA}}\right)\left(\pi_{AA}^2(1-\pi_{AA})\, log\,\Delta_1 - \pi_{AA}^2\pi_{AB}\, log\,\Delta_2\right)\right]$$

$$C_{23} = \Omega \times \left[log\Delta_2^2\left(\pi_{AB} - 2\pi_{AB}^2\right) - 2\pi_{AA}\pi_{AB}\, log\Delta_1\, log\Delta_2 + 2\left(\frac{P_{03}(t)\pi_{AB} - P_{02}(t)\pi_{BB}}{P_{03}(t)\pi_{AB}\pi_{BB}}\right)\left(\pi_{AA}\pi_{AB}^2\, log\Delta_1 - \pi_{AB}^2(1-\pi_{AB})\, log\Delta_2\right)\right.$$
$$\left. - 2\left(\frac{\pi_{AB}}{\pi_{AA}\pi_{BB}} - \frac{P_{02}(t)}{P_{03}(t)\pi_{AA}}\right)\left(\pi_{AA}^2(1-\pi_{AA})\, log\Delta_1 - \pi_{AA}^2\pi_{AB}\, log\Delta_2\right)\right]$$

$$C_{31} = \Omega \times$$
$$\left[log\,\Delta_1^2\left(2\pi_{AA}\pi_{BB} - \pi_{BB}\right) + 2\pi_{AB}\pi_{BB}\, log\,\Delta_1\, log\,\Delta_2 + 2\left(\frac{\pi_{BB}}{\pi_{AA}^2} - \frac{P_{03}(t)}{P_{01}(t)\pi_{AA}}\right)\left(\pi_{AA}^2(1-\pi_{AA})\, log\,\Delta_1 - \pi_{AA}^2\pi_{AB}\, log\,\Delta_2\right)\right]$$

$$C_{32} = \Omega \times$$
$$\left[log\,\Delta_2^2\left(2\pi_{AB}\pi_{BB} - \pi_{BB}\right) + 2\pi_{AA}\pi_{BB}\, log\,\Delta_1\, log\,\Delta_2 - 2\left(\frac{\pi_{BB}}{\pi_{AB}^2} - \frac{P_{03}(t)}{P_{02}(t)\pi_{AB}}\right)\left(\pi_{AA}\pi_{AB}^2\, log\,\Delta_1 - \pi_{AB}^2(1-\pi_{AB})\, log\,\Delta_2\right)\right]$$

where $\Omega = \left[\frac{1}{\pi_{AA}(1-\pi_{AA}) log\, \Delta_1^2 + \pi_{AB}(1-\pi_{AB}) log\, \Delta_2^2 - 2\pi_{AA}\pi_{AB}\, log\, \Delta_1\, log\, \Delta_2} \right]$.

Appendix 2

Formulas for %MSSN Coefficients for the $G \times E$ interaction only model (Model III)

The %MSSN coefficients D_{ij} are strictly functions of the genotype frequencies π_i where $i = 0,1,2$ for genotypes ++, *d*+ and *dd* respectively and genotyping relative risks $r_1(t)$ and $r_2(t)$. They are not functions of the specified level of significance, power and censoring rate.

The explicit %MSSN coefficients for a dominant MOI are:

$D_{12} = D_{21} = 0$

$$D_{10} = \frac{\pi_1}{\pi_0^2(\pi_1+\pi_2)(\pi_1 r_1(t)+\pi_2 r_2(t))} \times \left[\frac{2}{log\left(\frac{\pi_1 r_1(t)+\pi_2 r_2(t)}{(\pi_1+\pi_2)}\right)} \{\pi_0\pi_1\pi_2(r_1(t)+r_2(t))(r_1(t)-1) \right.$$

$$\left. +(\pi_2^2 r_2(t)+\pi_1^2 r_1(t))\pi_0(r_1(t)-1)+\pi_0^2\pi_2(r_1(t)-r_2(t))\} + \{\pi_0^2(\pi_1 r_1(t)+\pi_2 r_2(t))\} \right]$$

$$D_{20} = \frac{\pi_2}{\pi_0^2(\pi_1+\pi_2)(\pi_1 r_2(t)+\pi_2 r_2(t))} \times \left[\frac{2}{log\left(\frac{\pi_1 r_2(t)+\pi_2 r_2(t)}{(\pi_1+\pi_2)}\right)} \{\pi_0\pi_1\pi_2(r_1(t)+r_2(t))(r_2(t)-1) \right.$$

$$\left. +(\pi_2^2 r_2(t)+\pi_1^2 r_1(t))\pi_0(r_2(t)-1)+\pi_0^2\pi_1(r_2(t)-r_1(t))\} + \{\pi_0^2(\pi_1 r_1(t)+\pi_2 r_2(t))\} \right]$$

$$D_{01} = D_{02} =$$

$$\frac{1}{\pi_0(\pi_1+\pi_2)(\pi_1 r_1(t)+\pi_2 r_2(t))} \times$$

$$\left[\left(\frac{2}{log\left(\frac{\pi_1 r_1(t)+\pi_2 r_2(t)}{(\pi_1+\pi_2)}\right)} \{\pi_0^2(\pi_2(r_2(t)-1)+\pi_1(r_1(t)-1))\} \right) - \{\pi_0^2(\pi_1 r_1(t)+\pi_2 r_2(t))\} \right]$$

The explicit %MSSN coefficients for a recessive MOI are:

$$D_{01} = D_{10} = 0$$

$$D_{02} = \frac{\pi_0}{(\pi_0+\pi_1)(\pi_0+\pi_1 r_1(t))\pi_2^2 r_2(t)} \times \left[\left(\frac{2}{log\left(\frac{r_2(t)(\pi_0+\pi_1)}{\pi_0+\pi_1 r_1(t)} \right)} \{\pi_0\pi_1\pi_2(r_2(t)-1)(r_1(t)+1) \right.\right.$$

$$\left.\left. + \left(\pi_0^2\pi_2 + \pi_1^2\pi_2 r_1(t)\right)(r_2(t)-1) + \pi_1\pi_2^2 r_2(t)(r_1(t)-1)\} \right)\right.$$

$$\left. - \{\pi_2 r_2(t)(\pi_0^2 + \pi_1^2 r_1(t)) + \pi_0\pi_1\pi_2 r_2(t)(r_1(t)+1)\} \right]$$

$$D_{12} = \frac{\pi_1}{(\pi_0+\pi_1)(\pi_0+\pi_1 r_1(t))\pi_2^2 r_2(t)} \times \left[\left(\frac{2}{log\left(\frac{r_2(t)(\pi_0+\pi_1)}{\pi_0+\pi_1 r_1(t)} \right)} \{\pi_0\pi_1\pi_2(r_2(t)-r_1(t))(r_1(t)+1) \right.\right.$$

$$\left.\left. + \left(\pi_0^2\pi_2 + \pi_1^2\pi_2 r_1(t)\right)(r_2(t)-r_1(t)) + \pi_0\pi_2^2 r_2(t)(r_1(t)-1)\} \right)\right.$$

$$\left. - \{\pi_2 r_2(t)(\pi_0^2 + \pi_1^2 r_1(t)) + \pi_0\pi_1\pi_2 r_2(t)(r_1(t)+1)\} \right]$$

$$D_{20} = D_{21} =$$

$$\frac{1}{(\pi_0+\pi_1)(\pi_0+\pi_1 r_1(t))\pi_2 r_2(t)} \times \left[\left(\frac{2}{log\left(\frac{r_2(t)(\pi_0+\pi_1)}{\pi_0+\pi_1 r_1(t)} \right)} \{\pi_2^2 r_2(t)(\pi_0(r_2(t)-1) + \pi_1(r_2(t)-r_1(t)))\} \right)\right.$$

$$\left. + \{\pi_2 r_2(t)(\pi_0^2 + \pi_1^2 r_1(t)) + \pi_0\pi_1\pi_2 r_2(t)(r_1(t)+1)\} \right]$$

Appendix 3

Formulas for %MSSN Coefficients for the full $G \times E$ interaction model (Model IV)

The %MSSN coefficients F_{ij} are strictly functions of the genotype frequencies π_i where $i = 0,1,2$ for genotypes ++, *d*+ and *dd* respectively and genotyping relative risks $r_1(t)$ and $r_2(t)$. They are not functions of the specified level of significance, power and censoring rate.

The explicit %MSSN coefficients for a dominant MOI are:

$$F_{12} = F_{21} = 0$$

$$F_{10} = \frac{1-\left(1-\pi_1^2\right)\left(\pi_1+\pi_2\right)}{\pi_0\pi_1\left(\pi_1+\pi_2\right)} + \frac{2\pi_1}{log\left(\frac{\pi_2 r_2(t)+\pi_1 r_1(t)}{\left(\pi_1+\pi_2\right)}\right)}\left\{\frac{r_1(t)}{\pi_2 r_2(t)+\pi_1 r_1(t)}+\frac{r_1(t)}{\pi_0}-\frac{1}{\pi_0\left(\pi_1+\pi_2\right)}\right\}$$

$$F_{20} = \frac{1-\left(1-\pi_2^2\right)\left(\pi_1+\pi_2\right)}{\pi_0\pi_2\left(\pi_1+\pi_2\right)} + \frac{2\pi_2}{log\left(\frac{\pi_2 r_2(t)+\pi_1 r_1(t)}{\left(\pi_1+\pi_2\right)}\right)}\left\{\frac{r_2(t)}{\pi_2 r_2(t)+\pi_1 r_1(t)}+\frac{r_2(t)}{\pi_0}-\frac{1}{\pi_0\left(\pi_1+\pi_2\right)}\right\}$$

$$F_{01} = F_{02} = \frac{\pi_0^2-1}{\pi_0\left(\pi_1+\pi_2\right)} - \frac{2}{log\left(\frac{\pi_2 r_2(t)+\pi_1 r_1(t)}{\left(\pi_1+\pi_2\right)}\right)}\left\{\frac{\pi_0}{\pi_2 r_2(t)+\pi_1 r_1(t)}-\frac{\pi_0}{\left(\pi_1+\pi_2\right)}\right\}$$

The explicit %MSSN coefficients for a recessive MOI are:

$$F_{01} = F_{10} = 0$$

$$F_{02} = \frac{\pi_0\left(2\pi_2-1\right)}{\pi_2\left(1-\pi_2\right)} - \frac{2\pi_0}{log\left(\frac{\left(\pi_0+\pi_1\right)r_2(t)}{\pi_0+\pi_1 r_1(t)}\right)}\left\{\frac{1}{\pi_0+\pi_1 r_1(t)}+\frac{1}{\pi_2 r_2(t)}-\frac{1}{\pi_2\left(\pi_0+\pi_1\right)}\right\}$$

$$F_{12} = \frac{\pi_1\left(2\pi_2-1\right)}{\pi_2\left(1-\pi_2\right)} - \frac{2\pi_1}{log\left(\frac{\left(\pi_0+\pi_1\right)r_2(t)}{\pi_0 r_0(t)+\pi_1 r_1(t)}\right)}\left\{\frac{r_1(t)}{\pi_0+\pi_1 r_1(t)}+\frac{r_1(t)}{\pi_2 r_2(t)}-\frac{1}{\pi_2\left(\pi_0+\pi_1\right)}\right\}$$

$$F_{20} = F_{21} = \frac{1-2\pi_2}{1-\pi_2} - \frac{2}{log\left(\frac{\left(\pi_0+\pi_1\right)r_2(t)}{\pi_0+\pi_1 r_1(t)}\right)}\left\{\frac{\pi_2}{\left(\pi_0+\pi_1\right)}-\pi_2 r_2(t)\right\}$$