The Rare Disease Assumption: The Good, The Bad, and The Ugly

A Thesis

Presented in Partial Fulfillment of the Requirements for the Degree Master of Science in the Graduate School of The Ohio State University

By

Matthew Brems, B.A.

Graduate Program in Statistics

The Ohio State University

2015

Master's Examination Committee:

Dr. Shili Lin, Advisor

Dr. Eloise Kaizar

Abstract

In epidemiologic studies, researchers often calculate the relative risk or the odds ratio in order to quantify associations between a disease and a covariate. For decades, epidemiologists have utilized the "rare disease assumption" in order to simplify calculations by treating the odds ratio and the relative risk as approximately equal. In genetic epidemiology, the rare disease assumption is frequently made under various genetic conditions. However, in the context of genetic epidemiology, this assumption may merely be a necessary, but not a sufficient, condition. Thus relying on the rare disease assumption may lead to unintended consequences including, but not limited to, misrepresented power levels, inflated Type I error rates, and increased bias of estimators. I explore various genetic scenarios through case studies toward the goal of defining circumstances under which the rare disease assumption is sufficient and circumstances under which the rare disease assumption is merely necessary. These case studies show that the rare disease assumption can sometimes leads to good results, sometimes to bad results, and sometimes to ugly results. In particular, I explore three problems that are split into three sections based on the articles which inspired the discussion of these problems.

The first problem concerns testing for association between a binary disease and a single nucleotide polymorphism (SNP) marker using case-control data. Specifically, I consider data simulated from case/control probabilities derived based on the rare disease assumption and data simulated from case/control probabilities derived independently of the rare disease assumption. The power analysis based on the simulation study will show that association testing methods generally have higher power when data are simulated independently of the rare disease assumption, regardless of whether the underlying disease is rare or common.

The second problem deals with Hardy-Weinberg testing for a population by testing for Hardy-Weinberg equilibrium in controls only, as is customary in genetic epidemiology. If few individuals in a population have a disease, then the control group should be almost equivalent to the entire

population. Results show, however, that this method will lead to inflated Type I error rates and that the rare disease assumption is necessary, but not sufficient, for this method to be used.

The final problem involves fitting models based on paired mother/child data for detecting maternal effects. By fitting a logistic regression model, estimates of population relative risks can be found and analyzed. It has been shown that appreciable biases of these estimates still exist even when the disease considered is rare. A lengthy discussion of this problem is follows, indicating that the rare disease assumption is likely not sufficient to warrant the use of these methods.

To one of the most principled men I've ever known.

My Grandfather, William U. Brems

Semper Fidelis

Acknowledgments

This thesis and my Master's degree, plain and simple, would not have been possible without the genuine support, the tough love, and the friendship of the people below.

To my parents, Bill and Nora Brems: You deserve an enormous amount of credit for shaping me into the man I am today. The time and effort you put into raising me are second only to the values you instilled in me. I credit to you my ambition, my work ethic, my desire to be direct yet caring in all that I do, and my ability to both fight and love fiercely.

To Corey Smith: From celebrating early birthdays to commiserating over MATH 4545, you quickly became my best friend in graduate school. You made my time in school eventful and I'll cherish fondly the days we bonded over our RA experiences, the dinners we had at Panda Express, and the night we (you) showed up the Buckeye Capital Investors. (Seriously, though, the world is better with people like you in it.) I can't wait for you and Kayla to get a Corgi, though I'm glad you waited until after I left because I would have never seen you again.

To Emily Bartha, Tim Book, Katie Jones, Deborah Kunkel and Justin Strait: Thank you all so much for being amazing friends. From celebrating Scholarship weekly to making late nights and early mornings a bit more bearable, you've been through-and-through friends always ready for a laugh... or a beer. Though our respective significant others hang their heads when we talk about statistics, I appreciate

discussing our students, our favorite distributions, (Multinomial and Poisson) and how much we love/hate grad school.

To Michael Matthews, Marie Ozanne, Andrew Kamerud, and Caroline Gaker: Although you have also been incredible friends, you deserve recognition as the people who chose to live with me over the past couple of years. Thank you for tolerating my music in the shower and the pots I left soaking overnight that one time... or was it twice?

To Deron Molen, Derek Linn, Evan Tedtman, Kate McCarthy, Hilary Hauguel, Corey Mills, Alex Zimmerman, and Carley Campbell: Despite being hundreds of miles apart, I'm glad that we continued to grow as friends. I'm proud of what you've accomplished and can't wait to see what the future has in store for us!

To Dr. Elly Kaizar: I have truly enjoyed getting to work with you over the past two years. Your classes have greatly improved my understanding of statistics and your teaching style has significantly impacted how I teach. You make statistics clear to understand and make it seem so easy! I've also sincerely appreciated the job and graduate school advice you've given me when I've dropped by your office. Thank you so much!

To Dr. Shili Lin: I want to genuinely thank you for all that you've done. You pushed me hard and set unwavering expectations, which ultimately paid off - even if there were times when we thought it might not happen! I vividly remember the first time I got a "Good work!" in an email response from you, which meant all the more because you don't freely give out compliments. This has been a tough journey, but I'm a far better person and equipped for much more than I was twelve months ago!

To Nabisco, Frito-Lay, Skyline Chili, Chipotle: Thank you for providing enough sustenance in order for me to survive for the last two years on a tasty and cheap, yet unhealthy, diet.

To my future self: I'm so, so sorry that I did not take better care of my body during graduate school. I couldn't resist Oreos while writing my thesis. Please make up for this after graduation.

To Jacob Puhl: I am so very thankful I got into Ohio State, because it allowed me to meet you! Long distance hasn't been easy over the past year, but it has absolutely been worth it. I love you with every fiber of my being and every beat of my heart. I'm so proud of what you've accomplished and I can't wait to start the next chapter of my life with you and Paddington!

To Pugs, iMessage group chats, tear gas, B-Dubs, T-Dubs, Mirror Lake, schooners, Snap stories, BTC, Disco Zoo, PCB, Meiomi, 12-12-12, scaffolding, Million Dollar Cat Bracket, and various (read: all) bars on High Street: Thank you for making graduate school a well-rounded experience.

tx J

Vita

September 11, 1991	Born - Cincinnati, Ohio, USA
2009	Oak Hills High School
2013	B.A. Pure Mathematics, Applied Mathematics, Quantitative Analysis, Economics, Franklin College
2013-present	Graduate Teaching Associate, The Ohio State University.

Fields of Study

Major Field: Statistics

Table of Contents

Page

List of Tables

List of Figures

Figure **Page**

Chapter 1: Introduction

1.1 Introduction

This thesis delves into genetic epidemiology, which studies the effect of genetics on diseases. This is an important area of research as diseases are often passed down from parent to child. Before discussing the various genetic factors and their relationship with a binary disease, however, it is appropriate to step back and explain the big picture from a general perspective before diving into genetics and the specific research problems presented in this thesis.

In order to conduct any sort of statistical analysis, one must possess data on which to conduct the analysis. As such, the means by which one collects data is important. Two main methods by which an investigator can collect data are experiments and observational studies. In an experiment, the investigator manipulates a variable of interest to gauge its effect on another variable. For example, a physician interested in the impact of a new drug on a disease may split patients into two groups, administer the new drug to one group, and administer a placebo to another group. Observational studies, however, are conducted when it is unethical, difficult, or impossible to conduct an experiment. For example, if one wanted to research the impact that living under power lines had on incidence of cancer, it would be unethical to randomly select

families to live under power lines in hopes that they contracted cancer at a higher rate than families who do not live under power lines. As such, an observational study would be appropriate - rather than manipulating who lives underneath power lines, one could simply research individuals who already live under power lines and compare the incidence of cancer in that group to individuals who already do not live under power lines.

There are many types of observational studies. A survey, for example, is a very common type of observational study. In epidemiology, when an investigator is interested in finding a relationship between one variable and whether or not someone has a disease, one might conduct a "case-control" study. In case-control studies, investigators have two sets of observations: "cases" and "controls." Our variable of interest must be binary; we refer to observations that possess the characteristic of interest as cases whereas we refer to observations that do not possess the characteristic of interest as controls. In epidemiologic studies, cases are individuals affected by the disease of interest and controls are individuals not affected by the disease of interest. For example, if we are interested in studying lung cancer, we would call an individual who has lung cancer a "case" and an individual who does not have lung cancer a "control."

Epidemiologic studies lend themselves to a few quantities of interest that investigators seek to calculate in an attempt to quantify association between whether or not an individual has a disease and some other variable. Among these quantities are the "relative risk" and the "odds ratio."

The relative risk, also called the "risk ratio," is the probability of one having the disease given that they are in the "at-risk" group divided by the probability of one

having the disease given that they are not in the "at-risk" group. The "at-risk" group is the group exposed to the characteristic of interest. For example, if the independent variable is smoking, we might say that the "at-risk" group consists of smokers whereas individuals who do not smoke are in the "not at-risk" group. The odds ratio is, as the name suggests, a ratio of two odds. Specifically, the odds ratio is the ratio of the odds of having the disease given being "at-risk" to the odds of having the disease given being "not at-risk."

Examining these definitions within a table may illustrate exactly what each definition means. Consider Table 1.1.

Table 1.1: A 2-by-2 table of case-control and exposure data.

	Exposed	Not Exposed	Total
Case	p_A	p_B	$p_A + p_B$
Control	p_C	p_D	$p_C + p_D$
Total	$p_A + p_C$	$p_B + p_D$	$p_A + p_B + p_C + p_D = 1$

Assume that p_A, p_B, p_C, and p_D are parameters indicating the true proportion of the population that falls into each category. Then, the relative risk, denoted RR, is given by:

$$\begin{aligned} RR &= \frac{P(\text{case}|\text{exposed})}{P(\text{case}|\text{not exposed})} \\ &= \frac{p_A}{p_A + p_C} \Big/ \frac{p_B}{p_B + p_D} \end{aligned}$$

The odds ratio, denoted OR, is given by:

$$\begin{aligned} OR &= \frac{P(\text{case}|\text{exposed})}{P(\text{control}|\text{exposed})} \Big/ \frac{P(\text{case}|\text{not exposed})}{P(\text{control}|\text{not exposed})} \\ &= \frac{\frac{p_A}{p_A+p_C}}{\frac{p_C}{p_A+p_C}} \Big/ \frac{\frac{p_B}{p_B+p_D}}{\frac{p_D}{p_B+p_D}} \\ &= \frac{p_A}{p_C} \Big/ \frac{p_B}{p_D} \end{aligned}$$

The rare disease assumption was first discussed by Jerome Cornfield of the National Cancer Institute. Cornfield noted that one can use the odds ratio to approximate the relative risk for rare diseases. For purposes of clarity, future references to the rare disease assumption will be based on the following phrasing of the rare disease assumption:

> If the prevalence of a disease is sufficiently rare, then the odds ratio is approximately equal to the relative risk.

In genetic epidemiologic studies, the goal is to study and explain the relationship between disease status and various genetic factors. Whereas diseases like asthma, cancer, and diabetes are common, there are a significant number of rare diseases. Many of these rare diseases are caused (or believed to be caused) by genetic factors. As such, when exploring whether or not an association between disease and genetic factors exists, it is unsurprising that the rare disease assumption is applied in genetic epidemiologic studies.

In genetic studies, it can be shown that treating the odds ratio as approximately equal to the relative risk is tantamount to saying that the distribution of genotypes in the control population is equivalent to the distribution of genotypes in only the

entire population. Clearly, this simplifies many calculations. With rare diseases, a case-control study generally overrepresents the proportion of individuals who are cases within a population.

As with all assumptions, however, there are potential drawbacks. For example, in situations where one relies on a Normal distribution for data that clearly does not follow a Normal distribution, we might see decreased power of a statistical test. The rare disease assumption is not immune to these deviations. In particular, we note that relying solely on rarity of a disease in genetic contexts can lead to decreased power of a statistical test, increased type I error of a statistical test, or increased bias in estimators. The subject of this thesis is to examine the conditions under which the rare disease assumption leads to positive or negative side effects. Ultimately, it is shown that in order to treat the odds ratio as approximately equal to the relative risk in the context of genetic epidemiologic studies, knowing that a disease is rare is necessary but is not entirely sufficient. Potential other factors are explored toward the goal of establishing the conditions under which the odds ratio approximates the relative risk in a genetic context.

1.2 Terminology

There are a substantial number of technical terms that must be understood in order to comprehend many of the results in this thesis. As such, terminology that is related to genetic epidemiology and relevant to this work can be found here. Certain terms that are more appropriately covered elsewhere can be found in later chapters.

In genetic epidemiology, we seek to discover the relationship between one's genetic makeup and a disease. The genetic makeup of living beings is stored within

deoxyribonucleic acid, more commonly referred to as DNA. DNA is made up of many different molecules, the most important of which are nucleotides. There are four different types of nucleotides found in DNA: adenine, cytosine, guanine, and thymine, which are frequently abbreviated as `A`, `C`, `G`, and `T`, respectively. We often represent DNA as a sequence of these nucleotides. Consider taking a sequence of DNA from two individuals at the same location in the genome. For example, let one individual have sequence `GATTC` and the other individual have sequence `GACTC`. We see that these sequences are almost identical, however they differ at the third nucleotide in this sequence. This third nucleotide is referred to as a single nucleotide polymorphism, or a SNP. In genetic epidemiology, the goal is to look at this SNP and see if there appears to be a link between disease status and the nucleotide found at that location. For example, if 80% of the population has `C` in the third location above and 20% of the population has `T` in the third location above, and a higher proportion of the individuals who have `T` in that location have a disease than those who have `C`, this might suggest that the disease is associated with the SNP in question.

Being able to locate these SNPs is important so that researchers can establish relationships between particular SNPs and diseases. To assist with this, there are certain locations on the genome that have a known and documented DNA sequence, referred to as genetic markers. As such, researchers can more precisely define where a SNP of interest is by indicating its location relative to these genetic markers. A human's DNA is organized into 23 chromosomes, where each chromosome is identified as 1, 2, ..., 22, and a sex chromosome. Though we do not refer to it as such, this sex chromosome can be thought of as the 23rd chromosome. The sex chromosome determines the biological sex of a child. A particular location on a specific chromosome

is called a locus. We refer to these locations as genetic markers. While we can more precisely define a particular location on a given chromosome, the results in this thesis will be generalizable to any locus, thus this will suffice for this thesis.

Consider some locus, called L. L refers to a specific location on a specific chromosome. Since one's genetic makeup is inherited from his/her biological mother and father, then the genetic makeup at locus L in a child can be thought of as a combination of the genetic makeup of the father at locus L and the genetic makeup of the mother at locus L. At each locus, an individual has two alleles, where one is inherited from each parent. An allele is a variation of a gene and, for a SNP, can take on one of two forms: A and a. According to Gregor Mendel's Law of Segregation, each parent randomly passes exactly one of their two alleles to their child so that the child will inherit one allele from the father and one allele from the mother. More complicated scenarios exist where a gene takes more than two forms, but within this thesis, I assume that there are only two possible alleles at a given locus. In this case, we refer to the locus as diallelic. For simplicity, we assume that all loci being considered within this thesis are diallelic.

If an individual inherits two copies of allele A from his/her parents, then we say that individual has genotype AA. If an individual inherits one copy of allele A and one copy of allele a, we say that individual has genotype Aa. Genotype aa is defined similarly. However, we are often interested in the physical manifestation or result of having a particular genotype. A phenotype is the set of all observable traits and characteristics of an organism. In many cases, an individual's genotype will influence that individual's phenotype. For example, a human's brown eyes is part of a human's phenotype. In genetic epidemiology, researchers study disease status as part of a

human's genotype and are interested in whether or not a relationship exists between one's genotype and one's phenotype.

In a diallele SNP, one allele is usually seen more frequently than the other. We generally write A to be the more common allele and a to be the less common allele, and refer to a as the minor allele. The frequency of a in a population is called the minor allele frequency. In the context of diseases, the minor allele a is usually the allele considered to put a person "at risk" of having a disease. Let RR_1 be the relative risk of having the disease given one copy of at-risk allele a compared to having the disease given zero copies of at-risk allele a and let RR_2 be the relative risk of having the disease given two copies of at-risk allele a compared to having the disease given zero copies of at-risk allele a. RR_1 and RR_2 are known as the genotypic relative risks. The relationship between RR_1 and RR_2 gives rise to multiple genetic models.

There are an infinite number of genetic models as there are an infinite number of values that RR_1 and RR_2 can take on, but there are four models that are studied particularly frequently: dominant, recessive, additive, and multiplicative. Consider a SNP that determines whether or not one contracts breast cancer and let a be the at-risk allele. If an individual with genotype Aa or aa will necessarily have breast cancer, then having breast cancer is a dominant trait, as the presence of one or two at-risk alleles means an individual will contract breast cancer. In a dominant model, $RR_1 = RR_2 = \gamma$. If, on the other hand, an individual with genotype AA or Aa will certainly not have breast cancer, we say that having breast cancer is a recessive trait, as one can only have breast cancer if he/she has two copies of a. In a recessive genetic model, RR_1 is 1 and RR_2 is γ, where $\gamma > 1$. One can also define a relationship

between the genotypic relative risks when a model is additive (that is, the $RR_1 = \gamma$ and $RR_2 = 2\gamma - 1$) or multiplicative (that is, $RR_1 = \gamma$ and $RR_2 = \gamma^2$).

Although there are intuitive explanations for dominant, recessive, additive, and multiplicative models, genetic diseases need not behave according to these models. Since $RR_1, RR_2 \in \mathbb{R}^+$ and genetic models are defined by the values of RR_1 and RR_2, we can define many more genetic models. For example, later in this thesis we consider the situation where $RR_1 = 1.3$ and $RR_2 = 1.4$, which falls somewhere in between a purely dominant model and a purely additive model.

1.3 Structure of Thesis

Within the context of genetic epidemiology, we are interested in detecting a relationship between genotype and disease status. The rare disease assumption is relied upon in numerous circumstances where its usage makes analysis of genetic data simpler or even possible. This thesis provides an exploration of the rare disease assumption in such situations and the subsequent unintended effects toward the ultimate goal of demonstrating that rarity of a disease is necessary but not sufficient to warrant use of methods outlined in this thesis. As such, the remainder of this thesis is split into two body chapters and a lengthy discussion in the conclusion based on three applications of the rare disease assumption as it currently stands in order to assess these unintended effects:

- In Chapter 2, I discuss a number of statistical tests derived to detect an association between two categorical variables. Most of these these have been derived specifically to deal with genetic epidemiologic data and have been applied to detecting an association between binary disease status and genotype.

- In Chapter 3, I discuss Hardy-Weinberg equilibrium. In particular, I address the situation where, within a case-control study, investigators may test for Hardy-Weinberg equilibrium in controls only as a means to test for Hardy-Weinberg equilibrium in the entire population.

- In Chapter 4, I discuss the use of case-mother/control-mother pairs in an attempt to quantify the risk of particular genotypes as well as analyze the impact of maternal genetic effects on children. I examine the use of the logistic regression model to analyze these case-mother/control-mother pairs.

As the name "The Good, The Bad, and The Ugly" suggests, there will be both positive and negative aspects to the rare disease assumption: the "good" focuses on where the rare disease assumption improves the accuracy and reliability of statistical work; the "bad" entails where the rare disease assumption has a slightly negative impact; the "ugly" refers to situations where the rare disease assumption is unwarranted and its use gravely impacts the statistical analyses in question.

Chapter 2: Association Testing Using Case-Control Data

2.1 Introduction

In the context of testing the relationship between the genotypes of a genetic marker and a binary disease outcome (i.e. affected/unaffected), one might check for whether the presence of the disease is associated with some particular genotype(s). Consider a single nucleotide polymorphism (SNP) with alleles A and a, where allele a is the minor allele. Typically, the minor allele is considered to be the allele of interest as it may be the at-risk allele itself or is in linkage disequilibrium with a disease-causing locus.

2.2 Overview of Methods for Detecting SNP-Binary Disease Association

When testing for an association between two categorical variables, many turn to significance tests. Perhaps the simplest and most common method of testing for an association between two categorical variables is the chi-squared test. Though the chi-squared test is robust in detecting associations, the chi-squared merely tests for the existence of any relationship (Cornfield, 1951). More powerful tests have been developed within the context of genetic epidemiology that take into account more

information, such as the trend of relative risks. We consider the Cochran-Armitage trend test (Agresti, 2002), the maximum of the three most popular CATT settings (Freidlin et al., 2002), the genetic model selection method (Guo and Thompson, 1992), and the maximin efficiency robust test (Zheng et al., 2006).

2.2.1 Cochran-Armitage Trend Test and MAX3

As mentioned above, the chi-squared test is robust to detecting associations between two categorical variables. However, noting that there appears to be some relationship is not the same as providing inferences as to the direction of that relationship, merely that one exists. For example, in the context of a case-control study, the results of a chi-squared test may indicate that the binary outcome of disease may not be independent of the genotypes of individuals in the study. However, the relationship may be that an individual with genotype AA (or Aa or aa) is likelier to have the disease than other individuals. The chi-squared test provides no information as to details of this relationship. If a is the at-risk allele, then it is reasonable to believe that a person having two copies of the at-risk allele would have a risk of contracting the disease that is as high, or higher, than a person having only one copy of the at-risk allele - that is, the chi-squared test fails to consider the "trend" of relative risks.

The Cochran-Armitage Trend Test (CATT) serves to address this issue. This test was designed for a $2 \times I$ table where there are I different categories. For our purposes, $I = 3$ where each category is a different genotype. Cochran and Armitage used the ideas of linear regression to fit a line to describe the probability in each of the categories, implying a linear relationship among the genotypes (Agresti, 2002). That is, the relative risk of having two at-risk alleles compared to having zero at-risk alleles

is twice the relative risk of having one at-risk allele compared to having zero at-risk alleles. We might more simply write this $RR_2 = 2RR_1$. In a genetic context, Cochran and Armitage collapse the 2×3 table shown in Table 2.1 into various 2×2 tables, as shown in Table 2.2. In Table 2.1, we see that there are r_1 individuals who are cases and have genotype AA, r_2 individuals who are cases and have genotype Aa, and r_3 individuals who are cases and have genotype aa. As such, there are $r_1 + r_2 + r_3 = r$ cases. We define s_1, s_2, s_3, and s analogously for controls and let $r_1 + s_1 = n_1$, $r_2 + s_2 = n_2$, $r_3 + s_3 = n_3$ and $r + s = n$.

Table 2.1: A 2-by-3 table of SNP data.

	AA	**Aa**	**aa**	**Total**
Case	r_1	r_2	r_3	r
Control	s_1	s_2	s_3	s
Total	n_1	n_2	n_3	n

Based on the additive model, the CATT statistic is given by

$$\chi^2_{CATT} = \frac{n}{rs} \times \frac{(2r_3s - 2rs_3 + r_2s - s_2r)^2}{2n_3n + (2n_3 + n_2)(n_1 - n_3)}$$

where χ^2_{CATT} follows a χ^2 distribution with one degree of freedom (Ziegler and Konig, 2010).

A generalization was developed for genetic models for which the relative risks may not be additive. Consider the triple (x_1, x_2, x_3) where values for x_1, x_2 and x_3 characterize the underlying genetic model and $\bar{x} = \frac{\sum_{i=1}^{3} x_i}{n}$. Then, this general trend

Table 2.2: Collapsed 2-by-2 tables of genotypic frequencies.

	Dominant		**Recessive**	
	AA or Aa	**aa**	**AA**	**Aa or aa**
Case	$r_1 + r_2$	r_3	r_1	$r_2 + r_3$
Control	$s_1 + s_2$	s_3	s_1	$s_2 + s_3$
Total	$n_1 + n_2$	n_3	n_1	$n_2 + n_3$

test has test statistic

$$\chi^2_{general} = \frac{n^2}{rs} \times \frac{((\sum_{i=1}^{3} x_i r_i) - r\bar{x})^2}{\sum_{i=1}^{3} n_i (x_i - \bar{x})^2}$$

and follows the same $\chi^2_{(1)}$ as above (Ziegler and Konig, 2010). If $(x_1, x_2, x_3) = (2, 0, 0)$, then the derived test statistic is appropriate for a recessive genetic model and we denote that test statistic by χ^2_{rec}. If $(x_1, x_2, x_3) = (2, 2, 0)$, then the derived test statistic is appropriate for a dominant genetic model and we denote that test statistic by χ^2_{dom}. Thus, CATT can be generalized to include a number of popular genetic models (Ziegler and Konig, 2010).

MAX3, although discussed as a separate test in the literature, is based on the CATT statistics. Using Monte-Carlo simulations, the mean and standard deviations of χ^2_{CATT}, χ^2_{dom}, χ^2_{rec} are estimated, then the statistics are standardized. The MAX3 statistic is then given by

$$MAX3 = \max\{\chi^{2'}_{CATT}, \chi^{2'}_{dom}, \chi^{2'}_{rec}\}$$

where $\chi_a^{2'}$ is the standardized version of χ_a^2 and $a \in A = \{CATT, dom, rec\}$ (Freidlin et al., 2002). It is important to note that the asymptotic distribution is not available, but the power of MAX3 is based on simulation (Freidlin et al., 2002).

2.2.2 Genetic Model Selection

The Cochran-Armitage Trend Test is appropriate when a model is known and common - that is, if a model is additive, dominant, or recessive, a simple form for the test statistic exists. Though MAX3 provides some flexibility in misspecification of the model, often the model underlying a disease is more complex than any of the previously listed models or the true genetic model of the disease is simply unknown. In response to this issue, Zheng *et. al.* proposed a two-phase genetic model selection (GMS) process that was more robust to an unknown model than existing methods (Guo and Thompson, 1992).

The first stage was to select one of three genetic models based on the data. In order to do this, given the distribution of genotypes in cases, controls, and the population, one can divide the relative risks (RR_1 and RR_2) parameter space ($\mathbb{R}^+ \times \mathbb{R}^+$) into four regions: R_1, R_2, R_3, and R_4. In this stage, based on the value of test statistic $Z_{HWDTT} = \frac{(rs/n)^{1/2}(\hat{\Delta}_p - \hat{\Delta}_q)}{[1-n_2/n-n_1/(2n)][n_2/n+n_1/(2n)]}$, where $\hat{\Delta}_p = P(AA|D) - (P(AA|D) + P(Aa|D))^2$ and $\hat{\Delta}_q$ is defined analogously for controls, then the genetic model can be classified as recessive (R_1), dominant (R_4), or either additive or multiplicative (R_2,R_3). The second stage selects a test statistic for association Z_{MODEL} that is contingent upon the value of Z_{HWDTT} and its relation to c that is taken from a cumulative Normal for a prespecified level of significance.

Zheng and coauthors state that based on this method, for $p \geq 0.3$, the probabilities of accurately selecting the proper recessive/dominant model is above 65% and the probabilities of accurately selecting the correct additive/multiplicative model is above 85%. Zheng *et. al.* further state that these probabilities are robust to departure from Hardy-Weinberg equilibrium.

2.2.3 Maximin Robust Efficiency Test

A natural problem arising in statistics is, "Which estimator is best?" For example, in the context of regression, one might prefer to use X_1 or X_2 to predict Y, but perhaps not both. Often statisticians keep in mind that they want the mean squared error of their estimator to be low. To measure the variance, one can measure the asymptotic relative efficiency of one estimator compared to another and see which is preferable. Consider a parameter β that helps to define a particular distribution F. Then two estimators of β might be $\hat{\beta}_1$ and $\hat{\beta}_2$. We would define the asymptotic relative efficiency of $\hat{\beta}_2$ to $\hat{\beta}_1$ as:

$$ARE(\hat{\beta}_2, \hat{\beta}_1, F) = \frac{V_1(F)/n}{V_2(F)/n}$$

where $V_i(F)/n$ is the variance of $\hat{\beta}_i$. Note that $\hat{\beta}_i$ should be approximately Normally distributed with mean β and variance $\frac{V_i(F)}{n}$ for $i = 1, 2$ (Serfling, 2014).

Gastwirth developed the Maximin Efficiency Robust Test (MERT) to identify whether or not a genetic association exists when the true underlying genetic model is unknown (Gastwirth, 1985). Consider the class of optimal tests Z where an optimal test z_i exists for each plausible genetic model $i \in I$. Consider further the class of tests C for all plausible genetic models, where this consists of all asymptotically Normal, consistent tests. We seek to find an optimal test $z_i \in Z \subset C$ for our unspecified

genetic model. This optimal test can be found by satisfying

$$\inf_{i \in I} \text{ARE}(z^*, z_i, N) = \sup_{z^* \in C} \inf_{i \in I} \text{ARE}(z^*, z_i, N)$$

where N signifies a Normal distribution with mean μ and variance $\frac{\sigma^2}{n}$. The test z^* that satisfies the previous equation is the MERT (Zheng et al., 2006).

2.2.4 The W-Statistics

For incredibly rare diseases or diseases whose transmission is affected by non-genetic factors, it is often difficult to specify an exact model. As such, there exists a need to establish a method that both takes the trend of relative risks into account and is also robust under an unspecified genetic model. Chen sought to derive a set of tests that incorporate relative risk and are also robust to misspecified or unspecified genetic models (Chen, 2013). Consider Table 2.3, which represents SNP data. As above, there are r_1 individuals who are cases and have genotype AA, s_1 individuals who are controls and have genotype AA, and $r_1 + s_1 = n_1$ individuals who have genotype AA. The remaining cells are similarly defined for genotypes Aa and aa. Note that the number of cases $r = r_1 + r_2 + r_3$ and that the number of controls s and the number of observations n are analogously defined.

Suppose that (r_1, r_2, r_3) follows a multinomial distribution with r trials and probability vector $\{p_1, p_2, p_3\}$. We denote this $(r_1, r_2, r_3) \sim \text{Multinomial}(r, \{p_1, p_2, p_3\})$. Similarly suppose that $(s_1, s_2, s_3) \sim \text{Multinomial}(s, \{q_1, q_2, q_3\})$. Because the purpose of the aforementioned tests is to detect an association between disease status and genotype, one would test the hypothesis $H_0 : (p_1, p_2, p_3) = (q_1, q_2, q_3)$ versus the alternative $H_A : (p_1, p_2, p_3) \neq (q_1, q_2, q_3)$. If one concludes that $p_i \neq q_i$ for at least one $i = 1, 2, 3$, then the distribution of genotypes must be different for cases and controls.

Table 2.3: A 2-by-3 table of SNP data.

	AA	**Aa**	**aa**	**Total**
Case	r_1	r_2	r_3	r
Control	s_1	s_2	s_3	s
Total	n_1	n_2	n_3	n

This suggests that a particular allele may be responsible for increased risk of the disease.

Chen collapses the 2-by-3 table above into four distinct 2-by-2 tables. Two of the tables reflect dominant and recessive models as in Table 2.2 above. One of the remaining tables is obtained by taking Table 2.3 and removing the aa column while the final table is obtained by taking Table 2.3 and removing the AA column. From these tables, Chen defines statistics T_1, T_2, T_3, and T_4 by taking the four innermost cells and calculating the difference of the products of the diagonal elements. For example, T_2 is calculated from the table corresponding to a dominant model (Table 2.2) and is given by $T_2 = r_3 \times (s_1 + s_2) - (r_1 + r_2) \times s_3$.

Through standardization, the asymptotic properties of these statistics, and uncertainty surrounding whether a or A is the true at-risk allele, Chen derives six additional statistics that are asymptotically distributed according to a χ^2 distribution with 4 degrees of freedom. These statistics are W^{12}, W^{34}, W^{13}, W^{24}, W^{14}, and W^{23}. We refer to these henceforth as the W-statistics. Chen proposes W^{12} and W^{34} as robust

to improperly specified genetic models and powerful based on the results of his simulation study. He also proposes W^{13} and W^{24} as having these properties, but the two are asymptotically equivalent. As such, we need only consider W^{13} (Chen, 2013).

2.3 The Rare Disease Assumption and its Relationship with Odds Ratio and Relative Risk in Genetic Epidemiology

When one seeks to simulate genetic data, one must define a genotypic distribution for cases and controls. One can define case-genotypic probabilities in terms of the relative risks and the population-wide genotypic probabilities. Let D be the event that a person has the disease and ND be the event that a person does not have the disease. Then, the probability that a case has genotype AA is $P(AA|D)$ and the probability that a control has genotype AA is $P(AA|ND)$. We can define probabilities for genotypes Aa and aa analogously.

Toward the goal of seeing how the rare disease assumption relates to genotypic distributions, consider first the relationship between the odds ratio and the relative risk. Let RR_1 be the relative risk of having the disease given one copy of at-risk allele a compared to having the disease given zero copies of at-risk allele a and let RR_2 be the relative risk of having the disease given two copies of at-risk allele a compared to having the disease given zero copies of at-risk allele a. We define OR_1 and OR_2

analogously for odds ratios. Then, we have the following:

$$\begin{aligned} RR_1 &= \frac{P(D|Aa)}{P(D|AA)} \\ &= \frac{\frac{P(D|Aa)}{P(ND|Aa)P(ND|AA)}}{\frac{P(D|AA)}{P(ND|Aa)P(ND|AA)}} \\ &= \frac{\frac{P(D|Aa)}{P(ND|Aa)}}{\frac{P(D|AA)}{P(ND|AA)}} \times \frac{P(ND|Aa)}{P(ND|AA)} \\ &= OR_1 \times \frac{P(ND|Aa)}{P(ND|AA)} \\ &= OR_1 \times \frac{P(Aa|ND)}{P(Aa)} \times \frac{P(AA)}{P(AA|ND)}. \end{aligned}$$

It follows that, if $P(Aa|ND) \approx P(Aa)$ and $P(AA|ND) \approx P(AA)$, the relative risk will be approximately equal to the odds ratio. Because we focus on a diallele SNP and there are three possible genotypes, we can equivalently state that for the odds ratio and the risk ratio to be approximately equal, the distribution of genotypes in the population and the distribution of genotypes in the control group must be similar. Note that we selected the relative risk and odds ratio of Aa to AA without loss of generality, and using the relative risk and odds ratio of aa to AA will imply that the same approximate equivalency of the genotypic distributions is necessary in order for the relative risk and the odds ratio to be approximately equal.

One interpretation of the rare disease assumption is that the distribution of genotypes in the population and the distribution of genotypes in the control group are approximately equal. Thus, reliance on this fact constitutes reliance upon the rare disease assumption. In order to assess the impact of the rare disease assumption on Type I error rates and power of association tests, we seek to run two simulation studies: one where the controls are simulated relying on the rare disease assumption and one where the controls are simulated independently of the rare disease assumption.

2.4 Evaluation of Impact of Rare Disease Assumption on Type I Error Rates and the Assessment of Power of Association Tests

To evaluate the impact of the rare disease assumption, we use the publicly available software R with packages `devtools`, `Rassoc`, and `expm` to simulate 100,000 replicates of genotypes under a variety of genetic models. In particular, we evaluated the Type I error rates and power of the chi-squared test, CATT, MAX3, GMS, MERT, and W-statistics. In order to evaluate Type I error, we consider the case where $RR_1 = RR_2 = 1$. In order to evaluate power, we fix RR_2 to be 1.4 and consider RR_1 at 1, 1.1, 1.18, 1.2, 1.3 and 1.4. Note that, if $RR_2 = 1.4$, when $RR_1 = 1, 1.18, 1.2, 1.4$, the models we consider are respectively recessive, multiplicative, additive, and dominant. We consider each of these with disease population prevalence $k = 0.05$ and 0.15. We simulate three genotypic distributions under Hardy-Weinberg equilibrium (HWE) with minor allele frequency (maf) of 0.1, 0.3 and 0.5.

Most importantly, to assess the impact of relying upon the rare disease assumption, we generate probabilities under two different distributions. The genotype distribution for cases is the same regardless of reliance on the rare disease assumption. However, treating the control-genotypic distribution as approximately equivalent to the population-wide genotypic distribution, one relies upon the rare disease assumption. As such, we simulate control data using this distribution and refer to these probabilities as "RDA probabilities." We can derive the true control-genotypic distribution independently of the rare disease assumption. These probabilities are referred to as "non-RDA probabilities." The simulation study is run by simulating controls from the control-genotypic distribution relying on the rare disease assumption and

cases from the undisputed case-genotypic distribution. Then, the simulation study is run again by simulating cases in the same manner, but by simulating controls from the control-genotypic distribution that is defined by the non-RDA probabilities. Type I error rates and power are calculated for the two simulation studies and compared for the below analysis.

We briefly summarize the Type I error results here. Based on the simulation study, the differences in Type I error rates for data simulated by relying on the rare disease assumption and data simulated independently of the rare disease assumption is minimal. As seen in Table 2.4, the Type I error rates for tests using data from RDA probabilities are between 4.84% and 5.17%. In Table 2.5, the Type I error rates for tests using data from non-RDA probabilities are between 4.72% and 5.09%. As such, the Type I error rates of these tests are not significantly affected by reliance on the rare disease assumption.

When data are simulated by relying on the reare disease assumption, the power of aforementioned tests appears to be smaller than when data are simulated independently of the rare disease assumption. For example, when HWE holds, $maf = 0.1$ and $k = 0.05$, we have an average decrease of 10.77% in power across the 66 combinations of tests and relative risks. We note that this decrease in power can be as low as 0.26% and as high as 75.92%. It is clear that, if one simulated data independently of the rare disease assumption, the power of the test statistics would be represented properly and methods would be shown as more powerful than they might appear if one relied on the rare disease assumption when simulating data.

If we consider the case where HWE holds and $maf = 0.1$ as above, but where $k = 0.15$, the different representations of power become even more exaggerated. Across

Table 2.4: Type I Error Rates for Tests with Data Simulated Using the Rare Disease Assumption

	maf=0.1	**maf=0.1**	**maf=0.3**	**maf=0.3**	**maf=0.5**	**maf=0.5**
	k=0.05	**k=0.15**	**k=0.05**	**k=0.15**	**k=0.05**	**k=0.15**
W12	0.04948	0.05048	0.05157	0.04863	0.04954	0.04914
W34	0.04984	0.05068	0.0516	0.04895	0.04947	0.04881
W13	0.0493	0.05054	0.05171	0.04883	0.04932	0.04896
Chi-Squared	0.04953	0.04838	0.05027	0.04982	0.05005	0.05006
MAX3	0.04907	0.04909	0.05072	0.05089	0.05025	0.0499
GMS	0.0486	0.04854	0.04878	0.05009	0.05039	0.04903
CATT	0.05005	0.0495	0.04996	0.05076	0.05044	0.04957
MERT	0.04936	0.04985	0.04892	0.05029	0.05011	0.04925

the 66 combinations of tests and relative risks, the average difference in power is 17.49%, with a minimum of 0.51% and a maximum of 83.05%. We can also see that, for the W-statistics and the Cochran-Armitage Trend Test, the loss of power is monotonically increasing as RR_1 increases from 1 to 1.4. This exaggeration of power loss stands to reason, as the rare disease assumption requires that the prevalence of a disease is sufficiently rare. Since $k = 0.15$ would be considered a common disease, it would be unwise to rely on the rare disease assumption in such a situation.

A set of figures are presented to illustrate the different representations of power of methods based on RDA probabilities versus non-RDA probabilities for different values of maf and k. For clarity, only four tests were included in each figure. In each

Table 2.5: Type I Error Rates for Tests with Data Simulated Independently of the Rare Disease Assumption

	maf=0.1	**maf=0.1**	**maf=0.3**	**maf=0.3**	**maf=0.5**	**maf=0.5**
	k=0.05	**k=0.15**	**k=0.05**	**k=0.15**	**k=0.05**	**k=0.15**
W12	0.04894	0.04957	0.05034	0.05011	0.04991	0.05057
W34	0.04886	0.0493	0.04994	0.04919	0.05088	0.05052
W13	0.04869	0.04928	0.05035	0.04984	0.05012	0.05062
Chi-Squared	0.0499	0.04797	0.05066	0.04954	0.05033	0.04989
MAX3	0.04855	0.04765	0.0494	0.05024	0.04869	0.04937
GMS	0.04722	0.04725	0.05022	0.04943	0.05057	0.05011
CATT	0.0502	0.0488	0.0503	0.05054	0.05038	0.049
MERT	0.04985	0.04968	0.0508	0.04999	0.04902	0.04999

figure, the solid lines indicate tests were conducted on data simulated based on the rare disease assumption and the dashed lines indicate that tests were conducted on data simulated independently of the rare disease assumption.

First consider Figure 2.1, where we examine the three robust W-statistics and CATT. When comparing a solid line to the dashed line of the same color, we note each dashed line is entirely above the solid line of the same color. This indicates that for a multiplicative model (that is, where $RR_1 = 1.18$ and $RR_2 = 1.4$), the power is greater when the data are simulated independently of the rare disease assumption, irrespective of the other factors, including maf and prevalence.

Next consider Figure 2.2, where we examine the chi-squared test, MAX3, GMS, and MERT under the same multiplicative genetic model considered in Figure 2.1. Again the dashed lines are entirely above the corresponding solid lines, indicating the power is uniformly higher when working independently of the rare disease assumption. As a result, if the underlying genetic model of a disease is believed to be multiplicative, the power of these methods will not be correctly reflected when the data are simulated under the rare disease assumption.

Figures 2.3 and 2.4 both concern a dominant genetic model, where $RR_1 = RR_2 = 1.4$. In Figure 2.3 we once again note that the power of these significance tests are represented to be higher when data are simulated independently of the rare disease assumption. In Figure 2.4, we see that the methods do not appear uniformly more powerful when data are simulated independently of the rare disease assumption. Specifically, for $maf = 0.5$ and $k = 0.05$, all four tests actually appear to have higher power better when data are simulated while relying on the rare disease assumption. However, the tests appear to have, at most, 5% more power when data are simulated based on the rare disease assumption than when data are simulated independently of the rare disease assumption. As such, an overall recommendation would be to simulate these data without relying upon the rare disease assumption in order for the representation of power to be accurate.

Based on the direction of the lines on the figures, there appears to be a general trend that, as maf and k increase, the power of these tests increase as well. Although Figure 2.3 appears to defy this trend, note that the scale of the y-axis is significantly different from the scale of the other figures as the power of these tests is high for all considered values of maf and k. Based on these graphs, it appears that maf and

k may have some impact on the power of a statistical test and that this should be explored further.

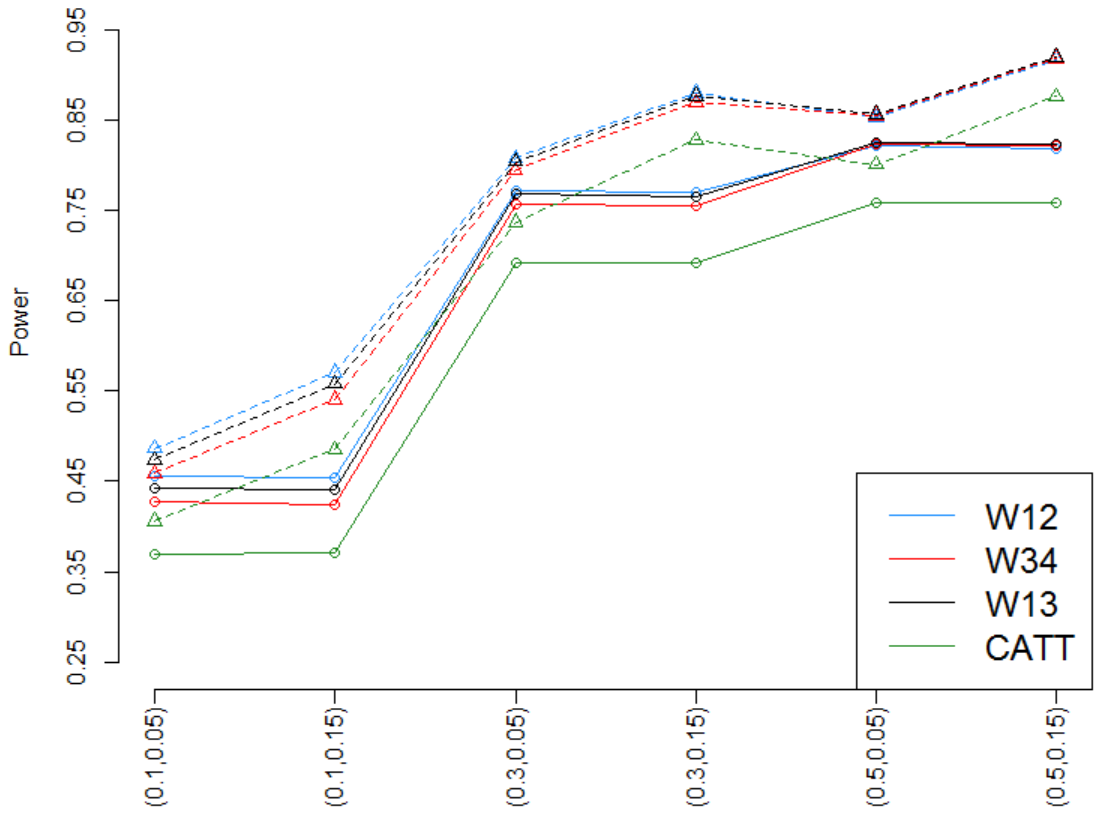

Figure 2.1: Power by (maf, prevalence) for Test Statistics under $RR_1 = 1.18$ and $RR_2 = 1.4$ (solid line indicates RDA, dashed line indicates non-RDA)

Based on the results of the simulation study, if a population is in HWE, it is inadvisable to simulate data based on the rare disease assumption when testing for equivalent genotypic distributions between the case and control populations as this

will almost always cause an underrepresentation of the power of the methods being examined.

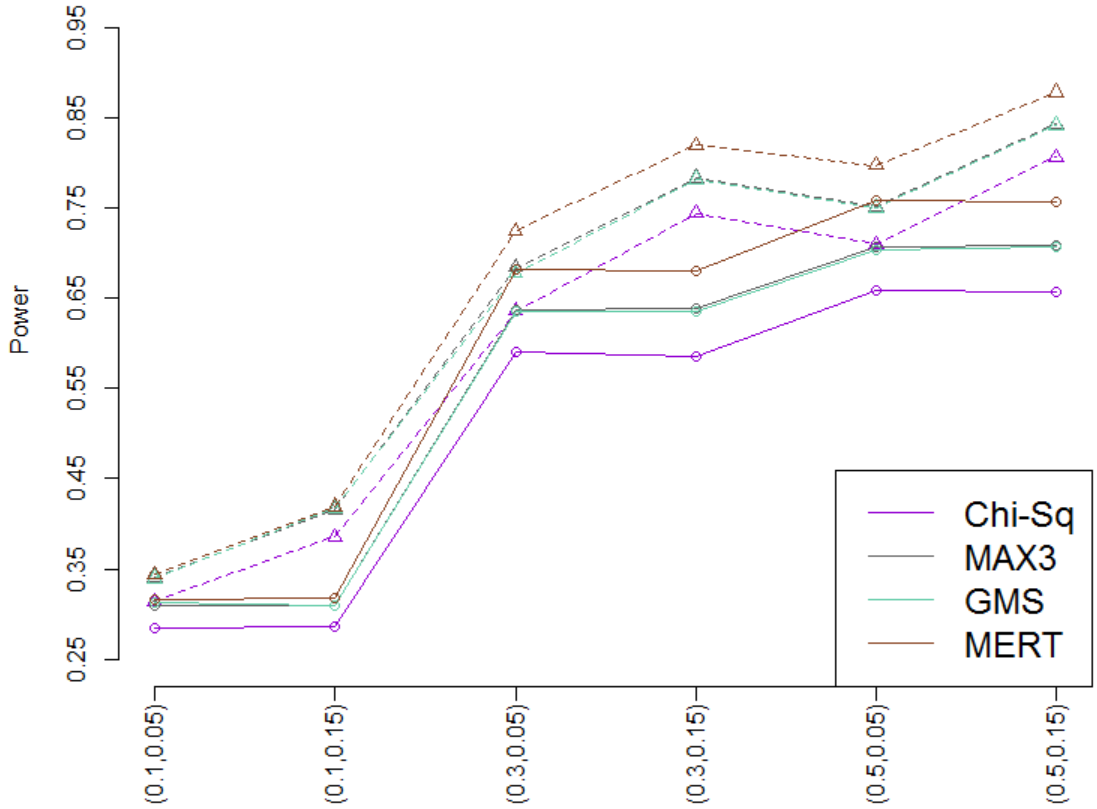

Figure 2.2: Power by (maf, prevalence) for Test Statistics under $RR_1 = 1.18$ and $RR_2 = 1.4$

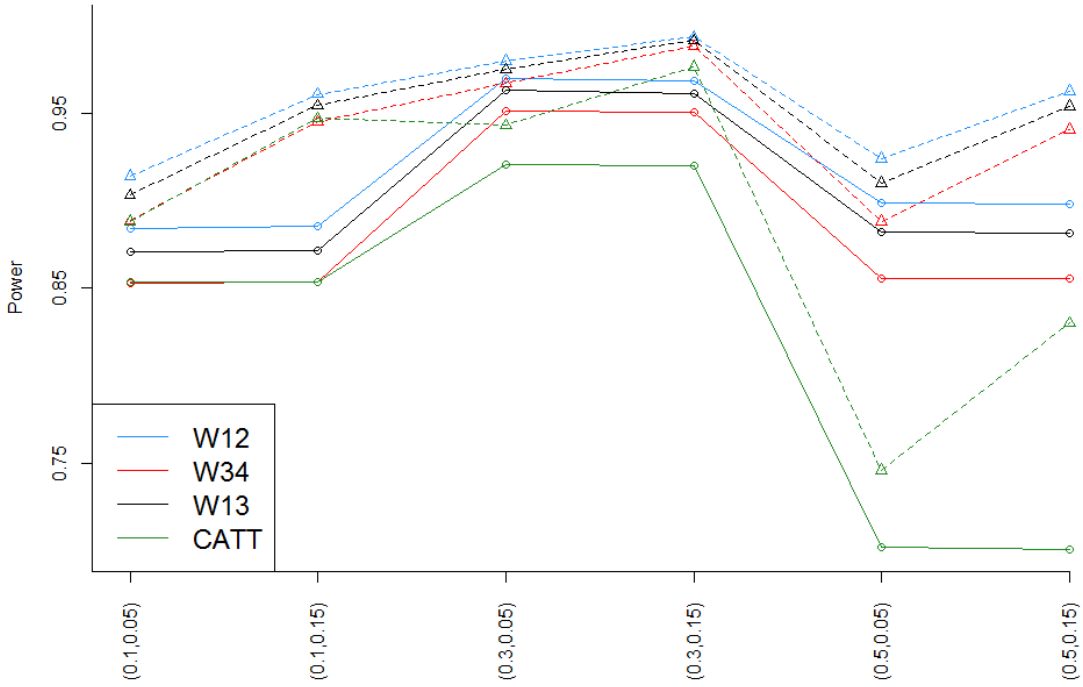

Figure 2.3: Power by (maf, prevalence) for Test Statistics under $RR_1 = RR_2 = 1.4$

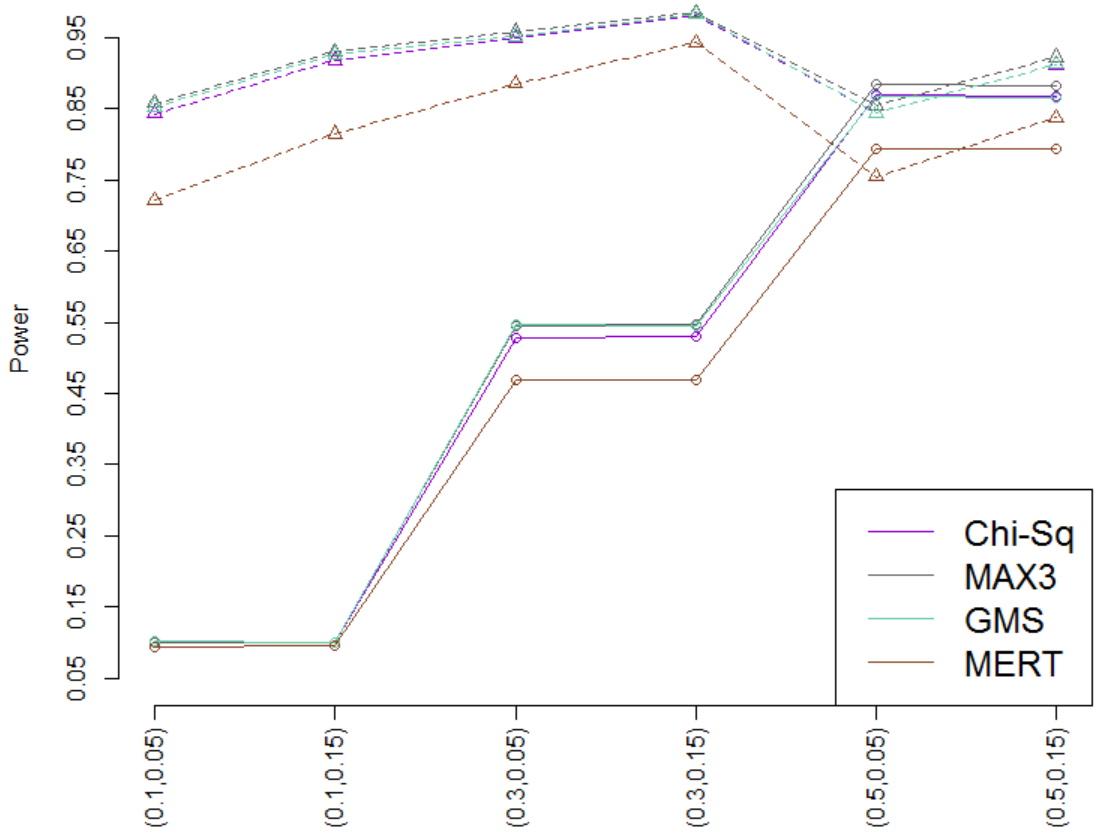

Figure 2.4: Power by (maf, prevalence) for Test Statistics under $RR_1 = RR_2 = 1.4$

Chapter 3: Testing for Hardy-Weinberg Equilibrium in Controls: Is it justifiable?

3.1 Introduction

When working with genetic data, often one of the first tests run is a test for Hardy-Weinberg equilibrium (HWE). HWE is a state of a population's genetics that occurs when there is random mating and when there are no evolutionary forces acting upon the population. Although the required conditions are never satisfied in reality, we say that HWE holds if the assumptions below are not too severely violated. If a population is in HWE, then allele frequencies and genotype frequencies are constant from one generation to the next and are related through a simple function, allowing for us to significantly simplify our calculations and increase the power of some tests.

In order for HWE to hold, one must make five assumptions about the genetics underlying population in question:

1. All mating occurs randomly.

2. There is no genetic drift. (This is equivalent to stating that the population size is infinitely large.)

3. There is no natural selection.

4. There are no mutations.

5. The allele and genotype frequencies are constant after one generation of mating.

Some textbooks or lists may have slightly different assumptions or may have more, however of these five assumptions, the most important is that mating occurs randomly. If these five assumptions hold, we will state that a population is in Hardy-Weinberg equilibrium. This allows us to simplify many calculations. For example, consider a diallele single nucleotide polymorphism with alleles A and a. If a population is in HWE, then we can use the genotype frequencies to calculate the allele frequencies and vice versa (Reece et al., 2014).

3.2 Methods for Testing for Hardy-Weinberg Equilibrium

Because many individuals rely on Hardy-Weinberg equilibrium in analyses, it is important to make sure the HWE assumptions are valid. As with many other assumptions, tests exist by which we can measure the likelihood that the assumption holds. The most commonly used test for HWE is Pearson's chi-squared test where the distribution of genotypes in the population in question is compared to the expected distribution of genotypes - that is, the distribution of genotypes under HWE. This is unsurprising because it provides means by which one can compare a categorical variable to its expected distribution. More sophisticated tests, however, have developed over time to meet the needs of geneticists. Emigh published a comparison of various methods including the standard chi-squared test, a conditional chi-squared tests, and the Freeman-Tukey and Mantel-Li tests (Emigh, 1980). These tests are asymptotic parametric tests, which may not be satisfied if the sample size is small. Issues also arise when certain genotypic frequencies are small, as is often the case with rare

diseases. As such, robust alternatives were desired. Guo and Thompson proposed and compared two algorithms for calculating the exact significance level of Hardy-Weinberg tests that did not rely on asymptotic arguments and were computationally superior to existing methods (Zheng and Ng, 2008). Wigginton *et. al.* discuss exact tests of Hardy-Weinberg equilibrium in populations, drawing from Fishers exact test for cross-tabular data (Wigginton et al., 2005).

These focus on testing for Hardy-Weinberg equilibrium in the entire population, as classical Hardy-Weinberg tests are designed for random samples. However, in data collection, rarely does one attain a truly random sample. Thus, a question arises when a population is clearly not homogeneous. For example, population stratification may occur due to a mixture racial or ethnic groups where one may expect differences among the groups in terms of their genetic makeup. Schaid *et. al.* suggest an exact test for HWE across strata and demonstrate its superiority to previous tests, including minimum exact p-value tests (Schaid et al., 2006). A similar, but not quite analogous, occurrence in sampling is clustering. Whereas stratification will divide a sample into groups based on a variable of interest, clustering will divide a sample into groups typically based on ease or accessibility. Strata are generally regarded as entities with different characteristics, but clusters are entities that, optimally, are similar. As such, a population that is sampled in clusters must be addressed differently than a random sample or a stratified sample. In surveys, complex designs arise when one combines both stratification and clustering. Li and Graubard incorporated appropriate weights and extensions of Pearson's X^2 statistic based on the Rao-Scott correction and quadratic test statistics to develop a Wald statistic for testing HWE in complex surveys (Li and Graubard, 2009).

One may also want to test for Hardy-Weinberg equilibrium in the context of various studies. For example, Ziegler and coauthors argue that one could test for HWE in a study that combines case-control data and cohort data in a meta-analysis context or data from exactly one case-control or cohort study (Ziegler et al., 2011). We focus our attention on case-control studies where the outcome of interest is a binary variable indicating simply whether or not one has the disease in question.

3.3 Testing for Hardy-Weinberg Equilibrium in Case-Control Studies

When testing for Hardy-Weinberg equilibrium, one assumes the that the population is in HWE as the null hypothesis and tests for significant deviation from HWE as the alternative hypothesis. Consider Table 3.1.

Table 3.1: Expected Distribution of Genotypes under HWE and Observed Distribution of Genotypes

Genotype	**Observed**	**Expected**
AA	n_{AA}	$nq_{AA} = nq_A^2$
Aa	n_{Aa}	$nq_{Aa} = 2nq_Aq_a$
aa	n_{aa}	$nq_{aa} = nq_a^2$

One can calculate the deviation from HWE by computing

$$X^2 = \sum_{g \in G} \frac{(n_g - nq_g)^2}{nq_g}$$

where $G = \{AA, Aa, aa\}$ is the set of genotypes, n_g indicates the observed count of individuals with genotype $g \in G$, q_g indicates the probability that an individual has genotype $g \in G$, q_a is the minor allele frequency, and $1 - q_a = q_A$ is the probability that a randomly chosen allele is A. X^2 follows a χ^2 distribution with one degree of freedom. A small p-value indicates that there is significant deviation from the theorized distribution of genotypes and provides evidence that HWE does not hold.

In the context of case-control studies, one can test HWE separately within cases and controls. However, it is unwise to test within the case population as cases are expected to deviate from HWE proportions if the test locus is in linkage disequilibrium except under a model of multiplicative relative risks (Clayton, 1999). In case-control studies, one may believe that if a disease is rare enough, then we can test for HWE in the entire population by testing for HWE in the controls (Wang and Shete, 2010). This intuitively makes sense - if very few people in the population have a disease, then the population and the control population are nearly identical. This relies upon the rare disease assumption as it implies that the distribution of genotypes in the control population is approximately equal to the distribution of genotypes in the entire population. What we will show, however, is that rarity of the disease is necessary but not sufficient to test for HWE in controls only.

In order to assess whether its warranted to test HWE in a population by testing the controls only, especially if the disease is rare, we conduct a simulation study. We assume that Hardy-Weinberg holds and examine the Type I error rates of the HWE test under a variety of scenarios. For each scenario considered in terms of prevalence of the disease (k), minor allele frequency (maf) and a set of relative risks, we simulated 1,000 cases and 1,000 controls under HWE. In particular, the distribution of genotypes

in the control population was constructed by using the true control probabilities. The observed frequencies were compared to expected frequencies by way of the chi-squared test and a p-value was recorded. 10,000 replicates of this study were run to empirically estimate the Type I error rate of using this test. As expected, HWE is violated in the cases, as discussed in the literature (Bourgain et al., 2004), and therefore we only present our results for the controls.

In Figure 3.1, we consider where the minor allele frequency maf ranges from 0.1 to 0.5 in increments of 0.1. We hold RR_2 constant at 1.4 and consider RR_1 at 1, 1.1, 1.18, 1.2, 1.3, and 1.4. Note again that when $RR_1 = 1, 1.18, 1.2$, and 1.4, the model is exactly recessive, multiplicative, additive, and dominant.

From Figure 3.1, it is immediately apparent that, for $k = 0.15$, the Type I error rates are vastly inflated - in four circumstances the error rate is more than double the nominal significance level and in one case it is as high as 11.87%! Note that the nominal significance level, $\alpha = 0.05$, is surpassed by the Type I error rates for nearly every combination of maf and RR_1. For $k = 0.15$, Type I error seems to be an increasing function of maf and RR_2, particularly when maf is lower than 0.4. Based on this information alone, we are inclined to conclude that testing for Hardy-Weinberg equilibrium in controls is extremely inadvisable when the disease is not rare. Now consider only the case where the disease is rare - that is, $k = 0.05$. It is immediately apparent that the Type I error rate is much better controlled. We attribute these Type I error rates to random variability in the simulated data and consider any "inflation" negligible.

In Figure 3.2, we similarly consider where the minor allele frequency maf ranges from 0.1 to 0.5 in increments of 0.1. However, in this case we hold RR_2 constant at

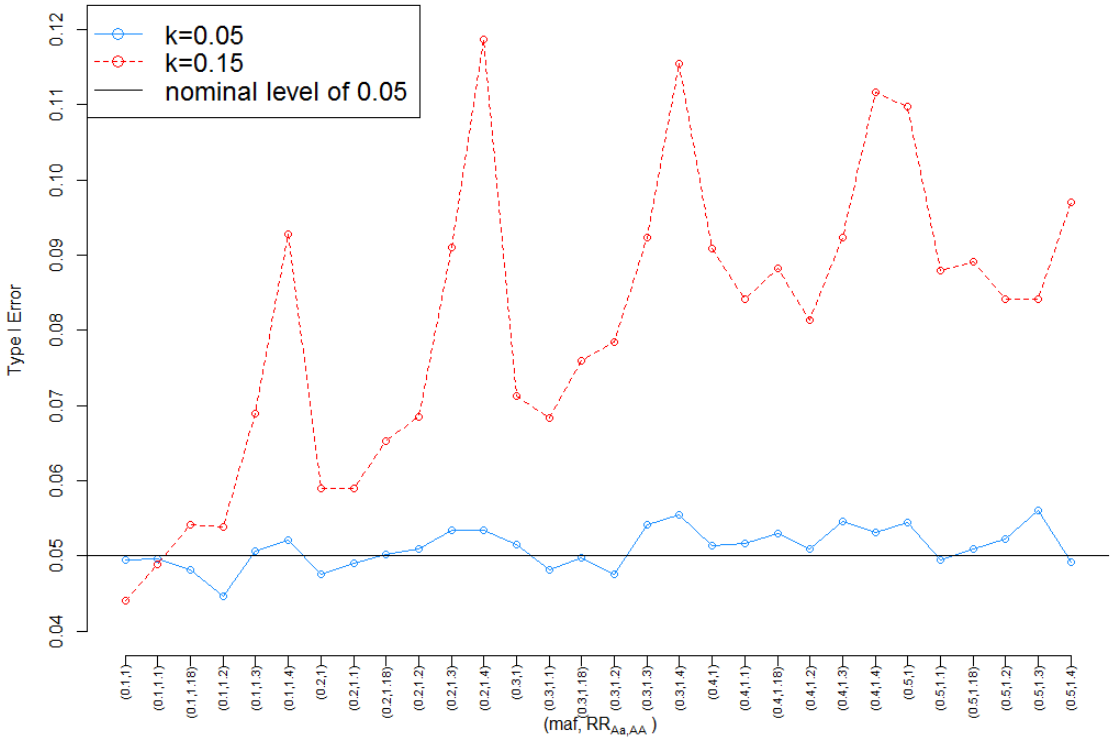

Figure 3.1: Type I Error Rates for Control Populations with $RR_2 = 1.4$

2 and consider RR_1 from 1 to 2 in increments of 0.1. There is increased granularity in this graph as there are nearly twice as many observations at each level of maf in Figure 3.2 compared to Figure 3.1.

First and foremost, though the Type I error rates visually appear to lie closer to the nominal level of 0.05, we note that the y-axis in Figure 3.2 is different from Figure 3.1 in that it extends from 0.04 to 0.40 in Figure 3.2 rather than from 0.04 to 0.12. There appears to be a similar trend in this graph as in Figure 3.1: as maf and RR_1 increase, the Type I error will in general increase.

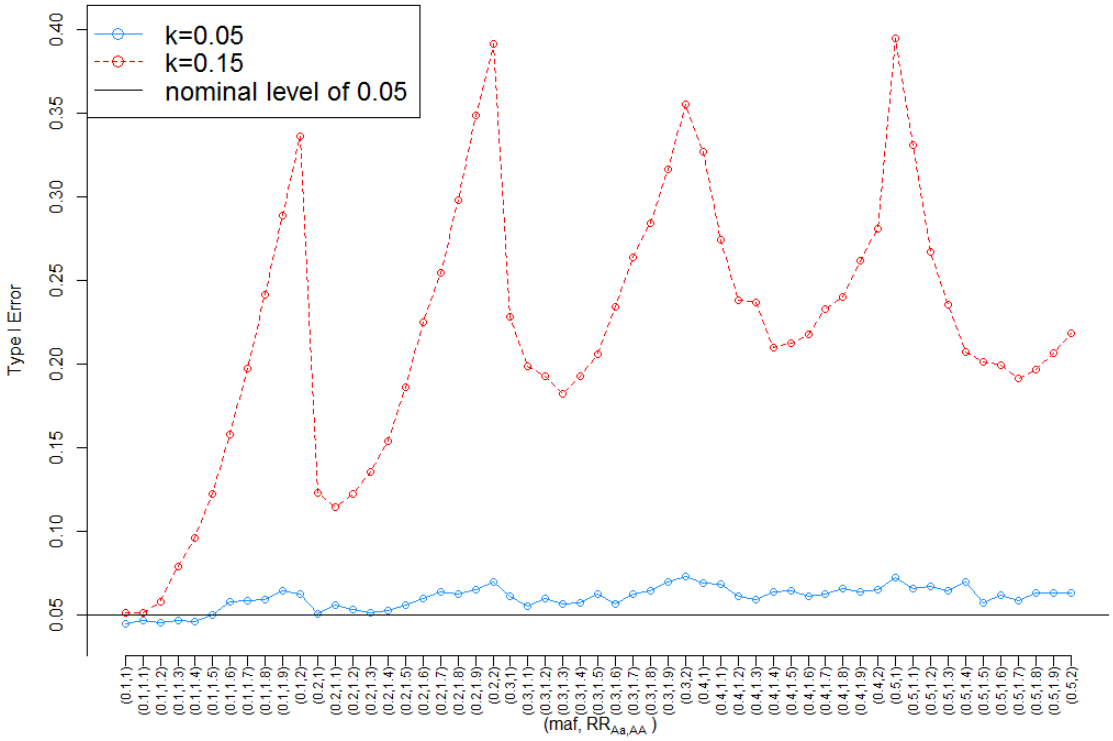

Figure 3.2: Type I Error Rates for Control Populations with $RR_2 = 2$

For $k = 0.15$, we see that the Type I error rate reaches as high as 39.5% and as low as 5.1%. The disparity between the Type I error curves for $k = 0.15$ and $k = 0.05$ are further evidence that the rare disease assumption is required for this method of HWE testing. Focusing only on $k = 0.05$, the Type I error rate peaks at 7.3%. Of the 55 combinations of maf and RR_1, there are only six where the Type I error rate is below our nominal level $\alpha = 0.05$ and occur when $maf = 0.1$ and RR_1 is 1.5 or below. As such, the Type I error rates are significantly inflated above 0.05.

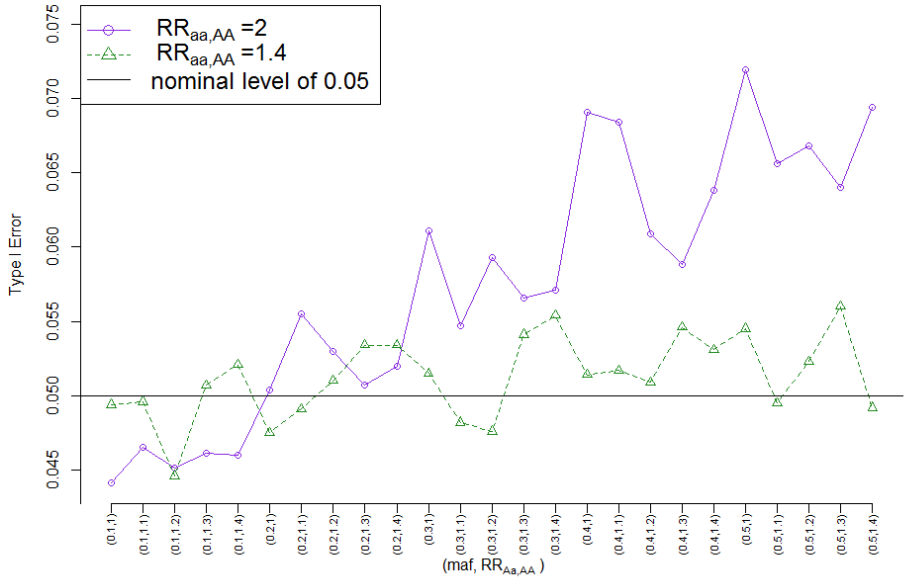

Figure 3.3: Type I Error Rates for $k = 0.05$

Examining Figure 3.3, one can better see the impact of maf and RR_1 on Type I error: as maf and RR_1 increase, Type I error will, on average, increase. Thus, as the underlying model of the disease shifts farther from a recessive model and closer to a dominant model, the Type I error rate increases. We also note that simulations conducted where $RR_2 = 2$ appears to generally have greater Type I error rates than simulations conducted where $RR_1 = 1.4$. This suggests that, as the relative risks increase, the effective Type I error rates will become more inflated.

As evidenced by the simulation study above, testing population-wide HWE by examining only controls requires the rare disease assumption. However, the fact that Type I error rates are still inflated stands to show that the rare disease assumption is not sufficient.

Chapter 4: Conclusion

4.1 Results

This thesis considered the rare disease assumption and showed the direct impact that the rare disease assumption has on genetics analysis methods. In particular, two methods have been explored:

- Significance Testing for Association between Two Categorical Variables
- Testing for Hardy-Weinberg Equilibrium in a Population by Testing Controls

In each of these chapters, the rare disease assumption is applied in order to simplify calculations. For example, in a case-control study, one can estimate the odds ratio but cannot meaningfully estimate the relative risk. The rare disease assumption is relied upon in order to use the odds ratio to approximate the relative risk. In the context of Hardy-Weinberg equilibrium, using the rare disease assumption to test for HWE in the controls as a proxy for testing for HWE in the population at large may prove beneficial. In many cases, however, there are unanticipated effects of relying upon this assumption.

When conducting hypothesis testing to detect an association between two categorical variables, simulating the data while relying upon the rare disease assumption

may cause the methods to appear less powerful than if the data were simulated independently of the rare disease assumption. This misrepresentation of power was exacerbated when prevalence of the disease k was 0.15 compared to when $k = 0.05$, suggesting that the rare disease assumption is necessary. However, there are still significant power misrepresentations when $k = 0.05$. In this case, it would be prudent to simulate the data independently of the rare disease assumption rather than relying upon the rare disease assumption when simulating data.

When testing for HWE in controls as a means to test for HWE in the population as a whole, the type I error rates were significantly inflated for common diseases. For a disease with prevalence $k = 0.15$, $RR_2 = 2$, and a nominal type I error rate of $\alpha = 0.05$, type I error rates reached up to nearly 40%. In the case where $RR_2 = 1.4$, the type I error rates for $k = 0.15$ were still inflated (up to approximately 12%) but were significantly better than the case where $RR_2 = 2$. For both $RR_2 = 1.4$ and $RR_2 = 2$, when $k = 0.05$, the Type I error rates were much closer to $\alpha = 0.05$ than for a common disease. This implies that the rare disease assumption is necessary for this method of HWE testing. However, in comparing the Type I error rates for rare diseases under both of the aforementioned settings of relative risk, the Type I error rates are larger when $RR_2 = 2$, increasing to above 7%. This indicates that relative risk plays a role in Type I error rates. This is further supported by the fact that, as RR_1 increases, Type I error rates tend to increase as well. It is also important to note that, as maf increases, Type I error rates will in general increase.

4.2 Discussion

In Chapter 2, reliance upon the rare disease assumption when simulating data was found to be ill-advised. Using the rare disease assumption may lead to misrepresented power of association testing methods. Though the misrepresentation of power was more significant for common diseases, simulating data independently of the rare disease assumption showed to be the appropriate course.

In Chapter 3, we saw that the Type I error rates in Chapter 3 were significantly more inflated for common diseases. It is clear that the methods used in this chapter require the rare disease assumption.

Perhaps the most strongly recurring theme from the previously mentioned problems is the role of the minor allele frequency. In Chapter 2, when maf increased, the power of statistical tests appeared to increase. In Chapter 3, when the rare disease assumption was used in testing for Hardy-Weinberg equilibrium, as maf increased, the Type I errors increased as well. With respect to the Hardy-Weinberg problem in Chapter 3, Type I error rates were far better controlled for small maf. Ultimately, the simulation studies and discussion of the results in this thesis indicate that the minor allele frequency plays a significant role in the statistical analysis of methods relying upon the rare disease assumption.

Based on the simulation studies, other factors seemed to affect power and Type I error rate. For example, the misrepresentation of power in Chapter 2 became more significant as the relative risks of having one and two copies of at-risk allele a increased. In Chapter 3, the periodic shape of the Type I error graphs suggest that RR_1 is positively correlated with Type I error rate. In addition, as Type I error rates are more inflated when $RR_2 = 2$ than when $RR_2 = 1.4$.

4.3 Impact of the Rare Disease Assumption on Parameter Estimation in Studying Maternal Effects

A problem discussed at length by Yang and Lin (2009) is the use of the rare disease assumption when using case-mother/control-mother pairs to study the impact of maternal genetic effects on presence or absence of a disease in the child. It was shown that relying upon the rare disease assumption in this case would cause there to be significant biases in parameter estimates when fitting log-linear models.

4.3.1 Introduction

When considering the relationship between binary disease status and genotype, one natural question to ask is whether or not that relationship is affected by which alleles are passed from which parents. For example, Prader-Willi syndrome is caused by a mutation inherited from one's father while Angelman syndrome is caused by a mutation inherited from one's mother. These two diseases are very different; Prader-Willi is marked by significant obesity arising from a slow metabolism and excessive hunger and Angelman syndrome is similar to autism and cerebral palsy. The fascinating part is that both syndromes are caused by the exact same mutation at the exact same locus on chromosome 15 (Ziegler and Konig, 2010). As such, the only genetic factor affecting whether one is afflicted with Prader-Willi syndrome or with Angelman syndrome is, "From which parent was the mutation inherited?" As such, the parent from whom an allele is inherited may impact the phenotype of a disease. This effect is known as imprinting, where paternal imprinting indicates that the effect of the paternal allele is suppressed and maternal imprinting is defined analogously

(Ziegler and Konig, 2010). As such, we would say that Prader-Willi syndrome arises from maternal imprinting and Angelman syndrome arises from paternal imprinting.

In order to account for this issue and to gain a greater understanding of imprinting and similar genetic effects, a researcher may design a study by using a case-parent triad design, where information is collected from a mother, father, and child. However, recruiting fathers is often far more difficult than recruiting mothers and maternity is easier to establish than paternity (Shi et al., 2008). As such, robust alternatives are desired. A number of imputation techniques to treat missing fathers as missing data have been developed, but models have also been developed that rely solely on genetic data gathered from child-mother pairs. Called a case-mother/control-mother study, both children who have the disease and children who do not have the disease are recruited to the study, then the mother's genetic information is gathered along with the child's. This paired data allows one to explore maternal genetic effects (Ainsworth et al., 2011).

4.3.2 Establishing the Genotype Distribution of Child-Mother Pairs

In order to examine the relationships among the mother's genotype, the child's genotype, and the presence or absence of a disease in the child, we examine the number of at-risk alleles a possessed by the mother and child. The goal is to estimate four relative risks that quantify the risk that the child will have the disease given certain maternal or child genotypes. In particular, we seek to estimate the relative risk of the child possessing one copy of a, the relative risk of the child possessing two copies of a, the relative risk of the child with mother possessing one copy of a and the relative risk of the child with mother possessing two copies of a. In order to

distinguish among these relative risks, we add some notation. Let the relative risk of the child possessing one copy of at-risk allele a be R_1 and the relative risk of the mother possessing one copy of at-risk allele a be S_1. We use R_2 and S_2 in the case where the child and/or mother have two copies of allele a.

As we have a binary response variable - that is, whether or not a child is affected by the disease - we can fit a logistic regression model using the aforementioned relative risks as explanatory variables. Let Y be a random variable such that $Y = 1$ if the child has the disease and $Y = 0$ if the child does not have the disease. Then consider the logistic regression model

$$\ln\left(\frac{P(Y=1|C=c,M=m)}{P(Y=0|C=c,M=m)}\right) = \mu + \beta_1 I_{(C=1)} + \beta_2 I_{(C=2)} + \gamma_1 I_{(M=1)} + \gamma_2 I_{(M=2)},$$

where $C = c$ indicates that the child has c copies of at-risk allele a and $M = m$ indicates that the mother has m copies of at-risk allele a. This model is discussed extensively by Shi *et. al.* (2008).

In the above model, there are parameters β_1, β_2, γ_1, and γ_2. Shi *et. al.* define the following:

$$\begin{aligned}
\beta_1 &= \ln(R_1), \\
\beta_2 &= \ln(R_2), \\
\gamma_1 &= \ln(S_1), \\
\gamma_2 &= \ln(S_2).
\end{aligned}$$

Note that β_1, β_2, γ_1 and γ_2 are parameters of a logistic regression model and are thus the log-odds ratios. $\ln(R_1)$, $\ln(R_2)$, $\ln(S_1)$, and $\ln(S_2)$ are the natural logarithms of the relative risks. Thus Shi equates the relative risks and odds ratios.

In defining the distribution of the case-mother and control-mother pairs, we must consider the distribution of the mother's genotypes as well as the distribution of the child's genotypes. Assume, for simplicity, that the population in question is in Hardy-Weinberg equilibrium and thus we can express genotypic frequencies in terms of allele frequencies. Further assume that the minor allele frequency is p. Then the joint probability that the mother and father have $(M = m, F = f)$ alleles is given in Table 4.1.

Table 4.1: Probability that mother and father have (m, f) copies of at-risk allele a.

$\mu_{\mathbf{mf}}$	$\mathbf{F = 0}$	$\mathbf{F = 1}$	$\mathbf{F = 2}$
$\mathbf{M = 0}$	$(1-p)^4$	$2p(1-p)^3$	$p^2(1-p)^2$
$\mathbf{M = 1}$	$2p(1-p)^3$	$4p^2(1-p)^2$	$2p^3(1-p)$
$\mathbf{M = 2}$	$p^2(1-p)^2$	$2p^3(1-p)$	p^4

We denote these probabilities as μ_{mf}, where m is the number of at-risk alleles the mother has and f is the number of at-risk alleles the father has. From this, we can derive the probability that a child will have $c = 0, 1, 2$ copies of the at-risk allele, assuming that at-risk alleles are inherited from the mother and father with equal probability. (More complex situations arise when inbreeding parameters are nonzero, but we set the inbreeding parameters equal to zero, simplifying our calculations.) For example, if a mother has one copy of a and a father has two copies of a, then the child has a 50% chance of inheriting a from the mother and a 100% chance of receiving a copy of a from the father.

Working within the context of case-mother/control-mother pairs, we can define the joint distribution of the child/mother genotypes based on the possible genotypes of the father. If a child has one copy of the disease allele and the mother has two copies of the disease allele, then the father must have had either zero copies or one copy of the disease allele. Logically, if the father had two copies, he must have passed one to the child. However, the child has one copy of a and must have inherited this from the mother, as the mother has two copies of a. As such, the father having two disease alleles is impossible. We can define the probability of the child/mother genotypes as a linear combination of the μ_{mf} parameters described in Table 4.1. In this case, if the father had zero copies and the mother had two copies, then the child would have exactly one copy with probability 1. If the father had one copy and the mother had two copies, then the child would have exactly one copy with probability 50%. Thus, we would say that the probability of $(M = 2, C = 1)$ is $\mu_{20} + \frac{1}{2}\mu_{21}$. In order to avoid confusion, we will denote the probability of event $(M = m, C = c)$ as $P(M = m, C = c) = \kappa_{mc}$. We use similar logic to define all κ_{mc} probabilities in terms of the μ_{mf} parameters from Table 4.1 and list the κ_{mc} probabilities in Table 4.2.

Table 4.2: Probability that mother and child have (m, c) copies of at-risk allele a as defined by μ_{mf} parameters.

$\kappa_{\mathbf{mc}}$	$\mathbf{C = 0}$	$\mathbf{C = 1}$	$\mathbf{C = 2}$
$\mathbf{M = 0}$	$\mu_{00} + \frac{1}{2}\mu_{01}$	$\frac{1}{2}\mu_{01} + \mu_{02}$	0
$\mathbf{M = 1}$	$\frac{1}{2}\mu_{10} + \frac{1}{4}\mu_{11}$	$\frac{1}{2}(\mu_{10} + \mu_{11} + \mu_{12})$	$\frac{1}{4}\mu_{11} + \frac{1}{2}\mu_{12}$
$\mathbf{M = 2}$	0	$\mu_{20} + \frac{1}{2}\mu_{21}$	$\frac{1}{2}\mu_{21} + \mu_{22}$

Now that we have the joint child/mother genotypic distribution established, we seek to define the genotypic distribution of case-mother pairs and the genotypic distribution of control-mother pairs. Because case-control studies are retrospective (that is, events like disease presence are examined after the onset of disease), it is appropriate to include additional information about the disease by conditioning on affection status (Yang and Lin, 2009).

Table 4.3: Expected frequencies of case-mother pairs as established in (Yang and Lin, 2009).

	$\mathbf{C = 0}$	$\mathbf{C = 1}$	$\mathbf{C = 2}$
$\mathbf{M = 0}$	$B[\mu_{00} + \frac{1}{2}\mu_{01}]$	$BR_1[\frac{1}{2}\mu_{01} + \mu_{02}]$	0
$\mathbf{M = 1}$	$BS_1[\frac{1}{2}\mu_{10} + \frac{1}{4}\mu_{11}]$	$BR_1S_1[\frac{1}{2}(\mu_{10} + \mu_{11} + \mu_{12})]$	$BR_2S_1[\frac{1}{4}\mu_{11} + \frac{1}{2}\mu_{12}]$
$\mathbf{M = 2}$	0	$BR_1S_2[\mu_{20} + \frac{1}{2}\mu_{21}]$	$BR_2S_2[\frac{1}{2}\mu_{21} + \mu_{22}]$

The distribution of control-mother pairs is similar to the distribution of case-mother pairs, but we incorporate the prevalence of the disease.

Table 4.4: Expected frequencies of control-mother pairs.

	$\mathbf{C = 0}$	$\mathbf{C = 1}$	$\mathbf{C = 2}$
$\mathbf{M = 0}$	$B(1-\delta)[\mu_{00} + \frac{1}{2}\mu_{01}]$	$B(1-\delta R_1)[\frac{1}{2}\mu_{01} + \mu_{02}]$	0
$\mathbf{M = 1}$	$B(1-\delta S_1)[\frac{1}{2}\mu_{10} + \frac{1}{4}\mu_{11}]$	$B(1-\delta R_1 S_1)[\frac{1}{2}(\mu_{10} + \mu_{11} + \mu_{12})]$	$B(1-\delta R_2 S_1)[\frac{1}{4}\mu_{11} + \frac{1}{2}\mu_{12}]$
$\mathbf{M = 2}$	0	$B(1-\delta R_1 S_2)[\mu_{20} + \frac{1}{2}\mu_{21}]$	$B(1-\delta R_2 S_2)[\frac{1}{2}\mu_{21} + \mu_{22}]$

4.3.3 The Rare Disease Assumption and Bias in Parameter Estimates

When constructing a model, one must keep the idea of bias in mind. Bias arises when an estimate of a parameter is different from the true parameter value. As such, minimizing bias is, in general, a good idea, and a model that has large amounts of bias in its parameter estimates might be considered a poor model. Given a parameter θ, one can calculate relative bias as $\frac{\hat{\theta}-\theta}{\theta} \times 100\%$, which measures how far the estimate is from the true parameter value relative to the true parameter value and is expressed as a percentage.

In Tables 4.3 and 4.4, B is a normalizing constant that will ensure the sum of all expected frequencies will be the total number of control-mother pairs. In this case, B is equal to 1 minus the prevalence of the disease. Under the rare disease assumption, B will be approximately 1 as the prevalence of the disease will be low. If $B = 1$, one could use the population-wide distribution (probabilities found in Table 4.2) as the distribution of the control-mother pairs. This method was used by Shi *et. al.* (2008).

However, this is problematic as some of the cells are incredibly difficult to estimate even when the disease is rare and the sample size is large. For example, consider the cell pertaining to $(M = 2, C = 2)$. This is given by

$$\begin{aligned} P(M = 2, C = 2) &= \kappa_{22} \\ &= \frac{1}{2}\mu_{21} + \mu_{22} \\ &= \frac{1}{2}p^3(1-p) + p^4 \end{aligned}$$

If the frequency of the at-risk allele a is small, then κ_{22} will also be very close to 0 (Yang and Lin, 2009). For example, if $p = 0.1$, then $\kappa_{22} = 0.00055$. In order for the

expected number of child/mother pairs with $M = 2$ and $C = 2$ to be 1, one must have a sample of more than 1800 individuals. Therefore, estimation can be very difficult when p is low and this can give rise to significant biases, even when the disease itself is rare. If one considers the case where $p = 0.9$, the same potential for bias exists for the cell corresponding to $(M = 0, C = 0)$. It has been shown that biases of parameter estimates relying on the rare disease assumption are significant - sometimes higher than 50% (Yang and Lin, 2009). This suggests that, in addition to needing the rare disease assumption, certain conditions on the minor allele frequency must also be imposed in order to apply the rare disease assumption in this setting.

In summary, when assessing maternal genetic effects through case-mother/control-mother studies by constructing models, there is a significant risk of bias in parameter estimation under certain genetic conditions. The probability of a mother and father having m and f at-risk alleles respectively can be defined entirely by the minor allele frequency p. The probability that the mother and father each have two copies of the disease allele is given by p^4 and the probability that the mother and father each have zero copies of the disease allele is given by $(1 - p)^4$. It is clear that for large or small p, either μ_{00} or μ_{22} will be very small. Because these μ_{mf} parameters give rise to the distributions of case child-mother and control child-mother pairs, attempting to estimate counts can quickly become problematic for extreme values of p, even if a disease is rare and with a large sample size. This will almost certainly cause the parameter estimates in our model to be significantly biased, which indicates that the model poorly describes the data and should not be used.

4.4 Future Work

A natural direction in which to head is toward a better understanding of the relationship between the minor allele frequency and the Type I error of tests or bias of estimators. Though there is empirical evidence to suggest that certain restrictions on the minor allele frequency may be necessary in order to use HWE testing methods outlined in Chapter 3 and case-mother/control-mother modeling as discussed previously in this chapter, conducting more thorough simulation studies or gathering real data may help to establish exactly what values of p are appropriate when using these methods.

Perhaps better than simulations or real data would be to head toward a rigorous treatment of the minor allele frequency. A proof could demonstrate exactly how the minor allele frequency affects these methods would be beneficial in order to understand when it is (and, perhaps more importantly, when it is not) proper to apply methods that are negatively affected when the maf takes on certain values.

Though relative risks are generally unknown, one may also want to explore more rigorously the relationship between the relative risks and these statistical measures, as the data suggest that relative risks may have an effect on type I error and power. For example, one important question is "For what settings of relative risk is Method X appropriate?," where Method X refers to any of the methods in this thesis.

4.5 Conclusion

Ultimately, this thesis served to examine statistical analyses of genetic data in a case-control study and the impact of the rare disease assumption on these analyses. Although rarity of a disease is sufficient to treat the odds ratio and relative risk as

approximately equal in general epidemiology, the goal of this thesis is to show that rarity of a disease may not be sufficient within the context of genetic epidemiology. An understanding of this allows us to capitalize on the good as well as understand when the rare disease assumption should not be relied upon by itself in order to mitigate or avoid entirely bad and ugly effects of the rare disease assumption.

Through the simulation studies and discussions of the problems, it is clear that the rare disease assumption is a necessary, but not sufficient, condition for using some of the methods outlined in this thesis, and further exploration of the impact of the rare disease assumption on these genetic analyses may be warranted.

Bibliography

A. Agresti. *Categorical Data Analysis*. Wiley, Second edition, 2002.

H.F. Ainsworth, J. Unwin, D.L. Jamison, and H.J. Cordell. "Investigation of maternal effects, maternal-fetal interactions and parent-of-origin effects (imprinting), using mothers and their offspring". *Genetic Epidemiology*, 35(1):19–45, January 2011.

C. Bourgain, M. Abney, D. Schneider, C. Ober, and M.S. McPeek. "Testing for Hardy-Weinberg Equilibrium in Samples With Related Individuals". *Genetics*, 168 (4):2349–2361, December 2004.

Z. Chen. "Association tests through combining p-values for case control genome-wide association studies". *Statistics & Probability Letters*, 83(8):1854–1862, August 2013.

D. Clayton. "A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission". *American Journal of Human Genetics*, 65(1):1170–1177, July 1999.

J. Cornfield. "A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast, and cervix". *Journal of the National Cancer Institute*, 11(6):1269–75, June 1951.

T.H. Emigh. "A comparison of tests for Hardy-Weinberg Equilibrium". *Biometrics*, 36(4):627–642, December 1980.

B. Freidlin, G. Zheng, Z. Li, and J.L. Gastwirth. "Trend tests for case-control studies of genetic markers: power, sample size and robustness". *Human Heredity*, 54(3): 146–152, July 2002.

J.L. Gastwirth. "The Use of Maximin Efficiency Robust Tests in Combining Contingency Tables and Survival Analysis". *Journal of the American Statistical Association*, 80(390):380–384, June 1985.

S.W. Guo and E. Thompson. "Performing the Exact Test of Hardy-Weinberg Proportion for Multiple Alleles". *Biometrics*, 48(2):361–372, June 1992.

Y. Li and B. Graubard. "Testing Hardy-Weinberg Equilibrium and Homogeneity of Hardy-Weinberg Equilibrium using Complex Survey Data". *Biometrics*, 65(4): 1096–1104, December 2009.

J. Reece, L. Urry, M. Cain, S. Wasserman, P. Minorsky, and R. Jackson. *Biology*. Pearson, Tenth edition, 2014.

D. Schaid, A. Batzler, G. Jenkins, and M. Hildebrandt. "Exact Tests of Hardy-Weinberg Equilibrium and Homogeneity of Disequilbrium across Strata". *American Journal of Human Genetics*, 79(6):1071–1080, December 2006.

R. Serfling. Asymptotic Relative Efficiency in Estimation. In *International Encyclopedia of Statistical Sciences*, pages 68–72. Springer, 2014.

M. Shi, D.M. Umbach, S.H. Vermeulen, and C.R. Weinberg. "Making the most of case-mother/control-mother studies". *American Journal of Epidemiology*, 168(5): 541–547, September 2008.

J. Wang and S. Shete. "Using Both Cases and Controls for Testing Hardy-Weinberg Proportions in a Genetic Association Study". *Human Heredity*, 69(3):212–218, March 2010.

J.E. Wigginton, D.J. Cutler, and G.R. Abecasis. "A Note on Exact Tests of Hardy-Weinberg Equilibrium". *American Journal of Human Genetics*, 76(5):887–893, May 2005.

J. Yang and S. Lin. "Technical Report". *Department of Statistics, The Ohio State University*, 2009.

G. Zheng and H.K.T. Ng. "Genetic model selection in two-phase analysis for case-control association studies". *Biostatistics*, 9(3):391–399, July 2008.

G. Zheng, B. Freidlin, and J.L. Gastwirth. "Comparison of robust tests for genetic association using case-control studies". In *International Encyclopedia of Statistical Sciences*, pages 253–265. Springer, 2006.

A. Ziegler and I. Konig. *A Statistical Approach to Genetic Epidemiology: Concepts and Applications*. Wiley, Second edition, 2010.

A. Ziegler, K. Van Steen, and S. Wellek. "Investigating Hardy-Weinberg equilibrium in case-control or cohort studies or meta-analysis". *Breast Cancer Research and Treatment*, 128(1):197–201, July 2011.