Unbiased Penetrance Estimates with Unknown Ascertainment Strategies

Kristen L. Gore

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2014

ABSTRACT

Unbiased Penetrance Estimates with Unknown Ascertainment Strategies

Kristen L. Gore

Allelic variation in the genome leads to variation in individuals' production of proteins. This, in turn, leads to variation in traits and development, and, in some cases, to diseases. Understanding the genetic basis for disease can aid in the search for therapies and in guiding genetic counseling. Thus, it is of interest to discover the genes with mutations responsible for diseases and to understand the impact of allelic variation at those genes.

A subject's genetic composition is commonly referred to as the subject's *genotype*. Subjects who carry the gene mutation of interests are referred to as *carriers*. Subjects who are afflicted with a disease under study (that is, subjects who exhibit the *phenotype*) are termed *affected* carriers. The age-specific probability that a given subject will exhibit a phenotype of interest, given mutation status at a gene is known as *penetrance*.

Understanding penetrance is an important facet of genetic epidemiology. Penetrance estimates are typically calculated via maximum likelihood from family data. However, penetrance estimates can be biased if the nature of the sampling strategy is not correctly reflected in the likelihood. Unfortunately, sampling of family data may be conducted in a haphazard fashion or, even if conducted systematically, might be reported in an incomplete fashion. Bias is possible in applying likelihood methods to reported data if (as is commonly the case) some unaffected family members are not represented in the reports.

The purpose here is to present an approach to find efficient and unbiased penetrance estimates in cases where there is incomplete knowledge of the sampling strategy and incomplete information on the full pedigree structure of families included in the data. The method may be applied with different conjectural assumptions about the ascertainment strategy to balance the possibly biasing effects of wishful assumptions about the sampling strategy with the efficiency gains that could be obtained through valid assumptions.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to thank God for guiding my steps throughout this journey.

I would also like to thank Professor Daniel Rabinowitz, my advisor, for always being supportive, kind, patient, and a true blessing in my life. You were the perfect advisor, and because of your steadfast involvement with me throughout this process, I have grown professionally.

Additionally, I would also like to thank the members of my thesis committee: Professors David Madigan, Richard Davis, Daniel Rabinowitz, Yuanjia Wang, and José Blanchet. I value your insight and expertise. The feedback you provided was beneficial and augmented my research.

I would also like to thank Dr. Ruth Ottman, Director of the Sergievsky Center of Columbia University, for providing data for my study and answering my many questions to better understand common practices and the complexity surrounding genetic epidemiological studies.

I am grateful to Dood, Regina, and Tian for your guidance and mentoring. Shawn, from the beginning to the end, my experience at Columbia has been enhanced by your patience, encouragement, and friendship.

I cannot imagine how I could have completed my journey through graduate school without the unwavering support of my family. Through all of my trials, you were there to

provide endless laughter, a shoulder to cry on, and the encouragement I needed to continue striving toward my goal. Words cannot express how thankful I am to have such an amazing family.

I would also like to thank my church family at Almond Branch Baptist Church (ABBC) for keeping me in their prayers throughout this process and being so warm and welcoming on my trips home.

My love for statistics was fostered by multiple math and science teachers who provided the academic foundation I needed to persevere, especially Mrs. Lyanne Haislip, Mrs. Teresa Carter, Mr. Clarence Raiford, Mrs. Karen Cue, Dr. Evon Hixon, Dr. Prabha Ramakrishnan, Dr. Jamila Simpson, Dr. Jackie Hughes-Oliver, Dr. Kimberly Weems, Dr. Roger Woodard, Dr. Anantha Aiyyer, Professor William Hunt, and Dr. Marcia Gumpertz. Thank you for igniting my passion for science and perpetuating my quest for knowledge.

I extend sincere gratitude to Commander Larry Laughlin (US Navy, Retired) for his countless recommendations, his encouragement, and for being an invaluable support to me during my time in the Navy Junior Reserve Officers Training Corps program and many years beyond. Your enthusiasm always brightens my day, and I am forever grateful for your impact on my life.

Finally, as I transition toward new endeavors I will always remember the contributions from the myriad of individuals who motivated me through encouraging words and acts of kindness. My graduate school experience will forever be a memorable chapter in my life, and I am honored that each of you had a role in this experience. Thank you.

I would like to dedicate this to my wonderful family, especially Grandaddy. Although God called him home before he was able to see my mom and I earn our PhD's, we know he's watching from Heaven. I hope I've made him proud.

Chapter 1

Introduction

A subject's genetic composition is commonly referred to as the subject's *genotype*. Subjects who carry the gene mutation of interests are referred to as "carriers." Subjects who are afflicted with a disease under study (that is, subjects who exhibit the *phenotype*) are termed *affected* carriers. The age-specific probability that a given subject will exhibit a phenotype of interest, given mutation status at a gene, is known as *penetrance*.

In a population-based sample, the age-specific proportion affected, stratified by genotype, is unbiased for penetrance. However, samples in which allelic variation and disease status are available are generally not population-based samples. Rather, they are the result of complex sequential ascertainment strategies that may only be incompletely reported. Such samples provide a challenge for the estimation of penetrance. Methods for addressing these challenges are the topic of this dissertation.

1.1 Probands

Family data used in penetrance analyses may have been collected for other purposes such as linkage analyses. In such studies typically there is an affected subject, known as a *proband*, who brings the family to the attention of the researchers [12]. Some cases may exist in which multiple family members simultaneously come to attention of the researchers. These subjects comprise the *proband class*. If a family contributing study

data contains multiple affected members, absent a report on the ascertainment of the family, any subset of the affected members of that family could conceivably have been as the proband class.

Family data used in penetrance analyses are obtained through a variety of means. These include self-referral, physician referral, the use of data collected for linkage or association analyses, etc. [4]. Thus, members of the pedigree who do not live near the clinic or within a reasonable distance of a facility which the researchers will use to ascertain families would not make likely members of the proband class [13]. Sobel and Elston suggest, in such cases, defining a proband sampling frame (PSF), a list of plausible probands from a given pedigree [7].

1.2 Ascertainment

When formulating a likelihood for the observed data, conditioning on all aspects of the data pertinent to the families' ascertainment in the likelihood is sufficient for unbiased estimates of the penetrance parameter [6] [15]. In this dissertation, the penetrance parameter will be denoted as π, and the probability of a subject being a carrier will be denoted by p. The assumptions that 1) a non-carrier cannot be affected, 2) affected subjects must be carriers, 3) if neither parent is a carrier, then none of their progeny can be carriers, and 4) subjects' phenotypes are conditionally independent given their genotypes will be made throughout this dissertation.

Sampling schemes commonly used for inclusion of families in a study include complete ascertainment, single ascertainment, and multiple ascertainment. The following paragraphs will briefly summarize these common schemes.

Complete Ascertainment

Under complete ascertainment, every affected individual becomes a proband [16] [15]. Thus, the probability that a family of size N is ascertained is

$$1 - (1 - \pi)^N$$

For reasons stated in previous chapters, this is often not a valid assumption and necessitates the defining of a proband sampling frame.

Single Ascertainment

Under single ascertainment, every affected individual has an equal probability of being brought to the attention of the researcher. As such, the probability of a family being ascertainment is related to the number of affected members [15]. The probability that a family of size N with k affected members is ascertained is

$$1 - (1 - \rho)^k,$$

where ρ is the probability that an affected individual would be brought to the attention of the researcher.

Multiple Ascertainment

Multiple ascertainment is a sort of hybrid of complete and single ascertainment. Unlike single ascertainment, multiple ascertainment allows for more than one proband per pedigree. In an effort to reduce the standard error associated with multiple ascertainment, some researchers have generated penetrance estimates by treating single and complete ascertainment as limiting cases of multiple ascertainment. Under this framework, the probability of being a proband, ρ, would be approximately 0 under single ascertainment and 1 under complete ascertainment. However, it has been shown that these are not

necessarily the limiting cases of multiple ascertainment [8] [12].

The assumptions for the multiple ascertainment model are often violated because they include that family members achieve proband status independently of each other and that the probability of being a proband is the same for all members of each family [15]. However, if one were to allow for less stringent proband status assumptions, the model would become underdetermined because there would be too many additional parameters to estimate in the model [15].

1.3 Sequential Sampling

If the proband(s)' family meets the preset inclusion criteria set by the researchers (for example more than two affected siblings or more than three affected family members), the family is typically ascertained via some form of sequential sampling. Under sequential sampling, subjects in the selected pedigree are brought into the study in stages according to preset inclusion criteria designated by the researcher(s). For instance, starting with the proband class, researchers may choose to sample all first degree relatives (i.e. parents, offspring, and siblings) of all affected subjects in an iterative fashion.

Most models implicitly assume that the probability of one family member being a proband is independent of another family member being the proband for that family [12]. However, if sequential sampling is implemented, the proband class can play a significant role in terms of which subjects are sampled. Vieland and Hodge (1995) refer to the class of sampling schemes which depend on the proband as "proband-dependent" (PD) sampling schemes [11]. Sequential sampling schemes are often PD sampling schemes. In some cases, different proband classes for the same family can yield different portions of the complete pedigree and subsequently yield vastly different penetrance estimates under different sequential sampling schemes. This problem becomes further complicated when 1) probands are distant relatives from the same family but are ascertained separately [2] and/or 2) the proband status is not accurately recorded or recorded at all.

Both PD and sequential sampling pose potential for biased sampling if the nature of the sampling is not reflected in the likelihood. Even if the researchers define the PSF for a given study, properly specifying the "true" probability of the subject(s) in the proband class comprising the proband class can be difficult. Additionally, other forms of often uncontrollable missingness emerge in sequential sampling. At every stage of the iterative sampling process, some subjects of the complete pedigree who would be eligible to enter the study under the sampling scheme fail to enter the study for a variety of reasons. These reasons can range from unwillingness to enter the study to geographical limitations [14]. The nature of the missingness may be deterministic, random, or both. In some cases, only incomplete information is available on subjects who do agree to participate in the study.

Depending on the proband class and the nature of the sampling scheme, the missingness which occurs at one step of the sampling process, particularly if the missingness occurs in one of the early steps of the iterative sampling process, can significantly affect the portion of the complete pedigree that is ultimately sampled. By extension, this can affect the amount of phenotypic and genotypic information available to the researchers. Thus, the missingness which occurs during various steps of an iterative sampling process can result in omission of information in the likelihood. If the missingness mechanism is not accounted for in the likelihood,then biased penetrance estimates may result.

1.4 Conditioning Strategies

The concept of conditioning on the data pertinent to ascertainment has been widely discussed in the genetic epidemiology literature and is a key aspect to the genotype-restricted likelihood (GRL) method proposed by Bonaiti et al. [1] [3]. Under this method, the conditional distribution of the genotypes of the sampled individuals are modeled, given the phenotypes of the sampled subjects and the event that the proband is a carrier using the Elston-Stewart algorithm and the Weibull model for penetrance [3]. However, in general, the penetrance estimates generated under the full data likelihood

differ from the estimates generated under the model which feature conditioning. Hodge (1988) discusses how likelihoods which feature conditioning on some (often unobserved) portion of the data use less information than is available and consequently yield larger standard errors than do likelihoods that use the complete data. If the information being conditioned on would be unobserved anyway, then no information will be lost. However, when aspects of the ascertainment are known conditioned on in the likelihood, there may be an inevitable reduction in efficiency [5].

In practice, properly conditioning on the data pertinent to ascertainment may be difficult, especially when the proband status of the family members is unrecorded, the ascertainment scheme is unknown, and/or the incomplete pedigree is reported. There may also be other unknown factors which influence which pedigrees are selected for the study and which subjects within each eligible family are likely to be sampled. These issues preclude the conditioning strategies mentioned in Section 1.4. Incorrect assumptions about the ascertainment process can lead to highly biased penetrance estimates [12] [15]. The incorrect specification of the likelihood will likely result in biased penetrance estimates, biased scores/estimating equations under the null hypothesis, and incorrect standard errors.

To remedy this, Shute and Ewens (1988) proposed the ascertainment-assumption-free (AAF) likelihood [9] [12]. The AAF likelihood conditions on the portion of the data that is "relevant to ascertainment" [9][10]. It should be noted that precisely specifying the data pertaining to the ascertainment process can be difficult. Let m be the number of children in a given family, x_o be the observed pedigree structure, d_1 be the portion of the observed data that relates to the ascertainment process, d_2 be the portion of the observed data that does not relate to the ascertainment process, and $n(x_o, d_1, d_2)$ be the number of families in the families with m children and data d_1 and d_2. Lastly, let $q_{x_o}(D)$ be the probability of observing phenotypic/genotypic data D for a family with pedigree structure x_o. Then the likelihood may be expressed as

$$
\begin{aligned}
L &= \prod_{x_o}\prod_{d_1}\prod_{d_2}\left[\frac{P_{x_o}(A,d_1,d_2)}{P_{x_o}(A)}\right]^{n(x_o,d_1,d_2)} \\
&= \prod_{x_o}\prod_{d_1}\prod_{d_2}\left[\frac{q_{x_o}(d_1,d_2)P_{x_o}(A|d_1)}{\sum_{j\in\{d_1\}} q_{x_o}(j)P_{x_o}(A|j)}\right]^{n(x_o,d_1,d_2)}
\end{aligned}
$$

The AAF likelihood is invariant of ascertainment process because rather than assuming that the probability of familial ascertainment is the probability that at least one member of the pedigree is affected, the method models the familial ascertainment by an arbitrary function of the number of family members and affected family members [9][10]. The AAF method also assumes a known distribution for the probability of ascertainment. However, while the AAF does provide a way of temporarily surpassing the issue of uncertainty on the ascertainment scheme, this method also yields higher standard errors than would be incurred under full knowledge of the ascertainment [9]. It is also possible that the AAF method could yield higher penetrance estimates than methods which incorporate knowledge of a class of plausible ascertainment schemes.

1.5 The Likelihood

Complete information for a family included in a study minimally consists of the complete pedigree structure, genotypic, phenotypic, and possibly covariate information for every subject in the pedigree.

For the i^{th} subject in a pedigree, let Y_j denote the indicator that subject j is affected, and let G_j indicate that subject j is a carrier. The assumption that affected subjects must be carriers will be made throughout this paper. It should be noted that the representation of genotype need not be dichotomous. For a given trait, subjects can be heterozygotous or homozygotous, and subjects can have several different possible alleles. The dichotomous coding of genotype is the common case of SNP data.

Full Ascertainment

For family i with N_i members ascertained under full ascertainment (i.e. if the entire pedigree is brought into the study and phenotypic and genotypic information is available for all subjects in the pedigree), the likelihood of observing phenotypes $Y_1, \ldots, Y_{N_i}$ given the family members' genotypes is

$$\prod_{j:G_j=1} \pi^{y_j}(1-\pi)^{1-y_j}$$

The first subscript will denote the family to which the subject belongs. For ease of notation, define $\boldsymbol{Y}_i$ to be $(Y_{i,1}, \ldots, Y_{i,N_i})$ and $\boldsymbol{G}_i$ to be $(G_{i,1}, \ldots, G_{i,N_i})$. For independent families $1, \ldots, n$, the probability of observing phenotypes $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ given genotypes $\boldsymbol{G}_1, \ldots, \boldsymbol{G}_n$ may be expressed as follows:

$$\begin{aligned} & P\left\{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n | \boldsymbol{G}_1, \ldots, \boldsymbol{G}_n\right\} \\ = & \prod_{i=1}^{n} \left\{ \prod_{j:g_{i,j}=1}^{N_i} \pi^{y_{i,j}}(1-\pi)^{1-y_{i,j}} \right\} \end{aligned}$$

The full likelihood of observing phenotypes $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ and genotypes $\boldsymbol{G}_1, \ldots, \boldsymbol{G}_n$ is given by

$$\begin{aligned} & P\left(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n, \boldsymbol{G}_1, \ldots, \boldsymbol{G}_n\right) \\ = & \prod_{i=1}^{n} P(\boldsymbol{G}_i) \left\{ \prod_{j:g_j=1}^{N_i} \pi^{y_{i,j}}(1-\pi)^{1-y_{i,j}} \right\} \end{aligned}$$

The probability of a subject and subsequent offspring being a carrier is dependent upon the genetic makeup of the subject's parents, also known as *founders*. Under Mendelian inheritance and the case in which two alleles are present at a specific locus, for a given trait, subjects receive one allele from the mother and one allele from the father. If both parents carry one dominant allele and one recessive allele, then under Mendelian inheritance, each of the parents' offspring has a 25% probability of being a *homozygous dominant* carrier (that is, being a carrier of two dominant alleles), 50% probability of being *heterozygous* carrier (that is, being a carrier of one dominant allele and one recessive allele), and a 25% probability of being a *homozygous recessive* carrier (that is, being a

carrier of two recessive alleles). If one parent carries two recessive alleles, and the other carries two dominant alleles, then all all offspring will be heterozygous carriers. For a given trait, if both parents are homozygous dominant carriers , then the offspring will be homozygous dominant carriers with probability 1, and if both parents are homozygous recessive carriers, then the offspring will, too, be homozygous recessive carriers with probability 1. If one parent is a homozygous dominant carrier and the other is a heterozygous carrier, then each offspring has a 50% probability of being a homozygous dominant carrier and a 50% probability of being a heterozygous carrier. If one parent is a homozygous recessive carrier and the other is a heterozygous carrier, then each offspring has a 50% probability of being a homozygous recessive carrier and a 50% probability of being a heterozygous carrier. Thus, the probability of a given genotypic configuration may be expressed as a conditional probability for each subject, where each subject's respective probability of being a carrier is dependent upon the subject's founders' genetic makeup. Accordingly, the joint likelihood of the observed phenotype and genotypes for each family in the study may be expressed as the product of the founder genotypes and the respective Mendelian transmission probabilities.

For commonly used ascertainment conditioning strategies, the likelihood may be succinctly expressed. The remainder of this section is devoted to likelihood expressions under complete, single, multiple, and sequential ascertainment.

Complete Ascertainment

If families $1, \ldots, n$ enter a study through complete ascertainment, the likelihood of observing phenotypic and genotypic information $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n, \boldsymbol{G}_1, \ldots, \boldsymbol{G}_n$ may be expressed

as follows:

$$
\begin{aligned}
& P\left(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n, \boldsymbol{G}_1, \ldots, \boldsymbol{G}_n | \text{Families } 1, \ldots, n \text{ Ascertained via Complete Ascertainment}\right) \\
= & \prod_{i=1}^{n} P(\boldsymbol{G}_i) \left\{ \frac{\prod_{j:g_{i,j}=1}^{N_i} \pi^{y_{i,j}} (1-\pi)^{1-y_{i,j}}}{\prod_{i=1}^{n} 1 - (1-\pi)^{N_i}} \right\}
\end{aligned}
$$

Single Ascertainment

If families $1, \ldots, n$ enter a study through single ascertainment, the likelihood of observing phenotypic and genotypic information $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n, \boldsymbol{G}_1, \ldots, \boldsymbol{G}_n$ may be expressed as follows:

$$
\begin{aligned}
& P\left(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n, \boldsymbol{G}_1, \ldots, \boldsymbol{G}_n | \text{Families } 1, \ldots, n \text{ Ascertained via Single Ascertainment}\right) \\
= & \prod_{i=1}^{n} P(\boldsymbol{G}_i) \left\{ \prod_{j:g_j=1}^{N_i} \pi^{y_{i,j}} (1-\pi)^{1-y_{i,j}} \left[1 - (1-\rho)^{\sum_{j=1}^{N_i} y_{ij}} \right]^{-1} \right\}
\end{aligned}
$$

Multiple Ascertainment

The joint likelihood expression under multiple ascertainment is substantially more complicated due to its allowance of a subset of affected subjects to serve as the proband class. Let $\boldsymbol{G}_f^i$ denote the genotypes of the members of the founders for the i^{th} family, and let $\boldsymbol{G}_{-f}^i$ denote the genotypes of the rest of the subjects in family i. If families $1, \ldots, n$ enter a study through multiple ascertainment, the likelihood of observing phenotypic and genotypic information $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n, \boldsymbol{G}_1, \ldots, \boldsymbol{G}_n$ may be expressed as follows:

$$
\begin{aligned}
& P\left(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n, \boldsymbol{G}_1, \ldots, \boldsymbol{G}_n | \text{Families } 1, \ldots, n \text{ Ascertained via Multiple Ascertainment}\right) \\
= & \prod_{i=1}^{n} \frac{\sum_{\boldsymbol{Y}_i \text{ compatible}} P(\boldsymbol{G}_f^i) P(\boldsymbol{G}_{-f}^i | \boldsymbol{G}_f^i) P(\boldsymbol{Y}_i | \boldsymbol{G}_f^i, \boldsymbol{G}_{-f}^i) P(\text{ascertainment event})}{P(\text{ascertainment event})},
\end{aligned}
$$

where the summation in the numerator is taken over all phenotypic configurations compatible with the genotypic configurations $(\boldsymbol{G}_f^i, \boldsymbol{G}_{-f}^i)$.

Sequential Sampling

Under sequential sampling, family members are brought into the study on an iterative basis based on a preset sampling rule. The sampling stops when there are no more family members eligible to enter the study. For family i, let C_0^i denote the members of the proband class, and let C_j^i be the individuals of the pedigree who are eligible to enter the study, based on the phenotypes (and other information relevant to ascertainment) of the family members already sampled by the $(j-1)^{th}$ iteration. Let R_i be the number of iterations of the sampling process under the preset sampling rule, i.e. the minimum integer such that $C_{R_i+1}^i = \varnothing$. Assume that the following conditions outlined by Cannings and Thompson (1977) are met:

1. "the pedigree structure is independent of the mode of inheritance"
2. "the choice of individuals to be examined depends only on phenotypes already observed"
3. "all individuals whose types are examined are included [that is, reported] in the analysis" [2].

Let $\boldsymbol{Y}_{i,C_r}$ and $\boldsymbol{G}_{i,C_r}$ denote the phenotypes and genotypes of the family members comprising C_r^i, respectively. Then under single ascertainment and iterative sampling featuring phenotype-based inclusion rules, the likelihood for observed data for the i^{th} family may be expressed as

$$\begin{aligned} & P\left(\boldsymbol{Y}_{i,C_0^i}, \ldots, \boldsymbol{Y}_{i,C_{R_i}^i}, \boldsymbol{G}_{i,C_0^i}, \ldots, \boldsymbol{G}_{i,C_{R_i}^i}\right) \\ = \; & P(\boldsymbol{Y}_{i,C_0^i}, \boldsymbol{G}_{i,C_0^i}) \prod_{j=1}^{R_i} P(\boldsymbol{Y}_{i,C_j^i}, \boldsymbol{G}_{i,C_j^i} | \boldsymbol{Y}_{i,C_{j-1}^i}) \\ = \; & P(\boldsymbol{Y}_{i,C_0^i}, \boldsymbol{G}_{i,C_0^i}) \prod_{j=1}^{R_i} P(\boldsymbol{G}_{i,C_j^i}) P(\boldsymbol{Y}_{i,C_j^i} | \boldsymbol{G}_{i,C_j^i}), \end{aligned}$$

For a given sampling class C_j^i, the conditional probability $P(Y_{i,C_j^i}|G_{i,C_j^i})$ follows the form of $P(\boldsymbol{Y}_i|\boldsymbol{G}_i)$ outlined in the full ascertainment section. Unless $\bigcup_{j\in\{0,\ldots,R_i\}} C_j^i$ is the complete underlying pedigree for family i, the likelihood for family i generated under sequential sampling will be biased. The joint likelihood across families $1,\ldots,n$ is given by

$$\prod_{i=1}^{n} \boldsymbol{P}(Y_{i,C_0^i}, G_{i,C_0^i}) \prod_{j=1}^{R_i} P(\boldsymbol{G}_{i,C_j^i}) P(\boldsymbol{Y}_{i,C_j^i}|\boldsymbol{G}_{i,C_j^i}).$$

Knowledge of the underlying pedigree is still necessary for all of the methods mentioned so far. Ewens and Elston (2012) argue, "Without knowledge of the true underlying pedigree structure (including who are the unobserved members of the pedigree) it is not possible to write down a correct likelihood and the ascertainment correction problem becomes intractable" [15]. The objective of this paper is to find a score-like function of the observed data which yields efficient unbiased penetrance estimates for every plausible ascertainment scheme and underlying pedigree structure.

The following chapter will contain the following:

1. a characterization of the unbiased scores,
2. a characterization of the standard error of the unbiased scores, and
3. methods to optimize the standard error of the unbiased scores.

Chapter 2

Methodology

2.1 Estimating Penetrance Under Non-parametric Assumptions on the Ascertainment Scheme

The objective of this chapter is to outline an approach for devising efficient and unbiased score statistics for the setting where the ascertainment process is incompletely specified. Section 2.1 discusses a method to find unbiased and efficient penetrance estimates when non-parametric assumptions on the ascertainment scheme are made. Section 2.2 discusses a method to find unbiased and efficient penetrance estimates when parametric assumptions on the ascertainment scheme are made. Section 2.3 details the process of solving for the efficient, unbiased penetrance estimate using the score equation methods derived in Sections 2.1 and 2.2. Section 2.4 summarizes the algorithm.

2.1.1 Deriving the Form of the Constrained Optimization Problem

Let $\mathcal{X}_i$ denote the set of conceivable pedigree structures for the i^{th} family, i from 1 to n, and let $X_i \in \mathcal{X}_i$ denote the family's actual pedigree structure. For every pedigree structure x in $\mathcal{X}_i$, let $\mathcal{G}_x$ denote the possible genotype configurations of the members of the pedigree, and let $\mathcal{Y}_x$ denote the possible phenotype configurations of the members of the pedigree. Let $G_i \in \mathcal{G}_{X_i}$ denote the actual genotype configuration of the members of the pedigree, and let $Y_i \in \mathcal{Y}_{X_i}$ denote the actual phenotype configuration of the members of the

pedigree. In this dissertation, consider the case in which the conditional distribution of the G_i and Y_i, given the X_i is fully parameterized by the population penetrance parameter π. The support for the complete data for the i^{th} family may be denoted by $\mathcal{C}_i$, where

$$\mathcal{C}_i = \{(x, g, y) : x \in \mathcal{X}_i, g \in \mathcal{G}_x, y \in \mathcal{Y}_x\}$$

However, complete information on the full pedigree structure and pedigree members' genotypes and phenotypic is rarely available. The observed information is a coarsened version of the complete data. The observed pedigree structure, genotype information, and phenotype information is shaped by the ascertainment scheme. The ascertainment schemes are modeled as maps from the set of complete data to subsets of the sample space of the complete data.

Let $\mathcal{Z}_i$ be the set of all conceivable ascertainment schemes for family i, and let Z_i denote the actual ascertainment scheme used for family i. More formally, $\mathcal{Z}$ may be defined as the set of conceivable mappings from the set of complete data $\mathcal{C}$ into the power set of $\mathcal{C}$. In practice, some information pertaining to the ascertainment scheme may be available to the analyst. Such information reflects knowledge (if any) of the proband sampling frame, aspects of the sampling scheme, and other factors which make certain potential subjects and their corresponding genotypic and phenotypic information observable. This additional information (the presence and nature of which may be dependent upon the complete information for family i), narrows the possible mappings Z that could have led to the observed data. Define $\mathcal{U}_i$ as the set of all conceivable maps from $\mathcal{Z}_i \times \mathcal{C}_i$ into the power set of $\mathcal{Z}_i$, and let U_i denote the mapping that characterizes the process of reporting additional information. Lastly, let $\mathcal{W}_i$ denote the set of

$$\mathcal{Z}_i \times \bigcup_{x \in \mathcal{X}_i} \{x\} \times \mathcal{G}_x \times \mathcal{Y}_x,$$

and let V_i denote the external information known about the ascertainment processes in the ascertainment of the i^{th} family, $U_i(Z_i, X_i, G_i, Y_i)$. The information which is observed for the i^{th} family is (W_i, V_i), where

$$W_i = Z_i(X_i, G_i, Y_i)$$

and

$$V_i = U_i(Z_i, X_i, G_i, Y_i).$$

There may potentially exist multiple configurations of the ascertainment schemes, available external information on the ascertainment scheme, and underlying complete data which yield identical events in $\mathcal{W}_i$ and corresponding conditional probabilities. In other words, there may exist complete data configurations (x_j, g_j, y_j) and $(x_{j'}, g_{j'}, y_{j'})$, external information mappings u_k and $u_{k'}$, and ascertainment scheme $z_l, z_{l'}$ such that

$$(z_k(x_j, g_j, y_j), u_l(z_j, x_j, g_j, y_j)) = (z_{k'}(x_{j'}, g_{j'}, y_{j'}), u_{l'}(z_{j'}, x_{j'}, g_{j'}, y_{j'}))$$

for some $j, j' \in \{1, \ldots, |\mathcal{C}_i|\}$, $k, k' \in \{1, \ldots, |\mathcal{Z}_i|\}$, and $l, l' \in \{1, \ldots, |\mathcal{U}_i|\}$. The indices which satisfy the above condition create a set of equivalence classes of observable events (w, v). Thus, the observed information actually provides information about the equivalence class for the underlying pedigree structure and ascertainment scheme which could have resulted in the observed information. Let $\mathcal{H}_i$ denote the set of equivalence classes defined by the above equation for the i^{th} family.

Define $q_{x,z,u}(w, v)$ to be the conditional probability mass function of (W_i, V_i), given the underlying pedigree structure $X_i = x$, ascertainment procedure $Z_i = z$, and informative mapping $U_i = u$. More specifically,

$$q_{x,z,u}(w, v) = \sum_{g,y:z(x,g,y)=w,u(z,x,g,y)=v} p^{\pi}(g, y|x),$$

where dependence on π is suppressed in the notation.

Define t_i for $i = 1, \ldots, n$ to be a function from $\mathcal{W}_i \times \mathcal{V}_i$ into $\mathbb{R}$, and consider test statistics of the form

$$\sum_{i=1}^{n} t_i(W_i, V_i; \pi).$$

Without assumptions on the distribution of the Z_i, for unbiasedness of the test statistic, it

is necessary that for all i, for all $x \in \mathcal{X}_i$, $z \in \mathcal{Z}_i$, and $u \in \mathcal{U}_i$,

$$\begin{aligned} & \mathbb{E}^{\pi}\{t_i(z(x, G_i, Y_i), u(z, x, G_i, Y_i); \pi)\} \\ = & \sum_{w \in \mathcal{W}_i, v \in \mathcal{V}_i} t_i(w, v; \pi) q_{x,z,u}(w, v) \\ = & \ 0 \end{aligned}$$

The choice of the t_i to maximize power subject to the unbiasedness conditions is not determined absent further assumptions on the X_i,Z_i, and U_i, assumptions such as a prior on their distribution. Make a working assumption of independence between the (X_i, Z_i, U_i) pairs, and let $\gamma^i_{X,Z,U}$ denote the joint probability mass function corresponding to a working prior on the (X_i, Z_i, U_i). For $w \in \mathcal{W}_i$ and $v \in \mathcal{V}_i$, define $\tilde{q}_i(w, v)$ to be

$$\sum_{(x,z,u) \in \mathcal{X}_i \times \mathcal{Z}_i \times \mathcal{U}_i} \gamma^i_{x,z,u} q_{x,z,u}(w, v)$$

.

The variance of the test statistic, σ^2, may be expressed as

$$\begin{aligned} & \mathbb{E}\left\{ \sum_{i=1}^{n} \sum_{w \in \mathcal{W}_i, v \in \mathcal{V}_i} t_i^2(w, v; \pi) q_{X_i,Z_i,U_i}(w, v) - \left[\sum_{w \in \mathcal{W}_i, v \in \mathcal{V}_i} t_i(w, v; \pi) q_{X_i,Z_i,U_i}(w, v) \right]^2 \right\} \\ = & \sum_{i=1}^{n} \sum_{w \in \mathcal{W}_i, v \in \mathcal{V}_i} \sum_{(x,z,u) \in \mathcal{X}_i \times \mathcal{Z}_i \times \mathcal{U}_i} t_i^2(w, v; \pi) \gamma^i_{x,z,u} q_{x,z,u}(w, v) \\ & - \left[\sum_{w \in \mathcal{W}_i, v \in \mathcal{V}_i} \sum_{(x,z,u) \in \mathcal{X}_i \times \mathcal{Z}_i \times \mathcal{U}_i} t_i(w, v; \pi) \gamma^i_{x,z,u} q_{x,z,u}(w, v) \right]^2 \\ = & \sum_{i=1}^{n} \sum_{w \in \mathcal{W}_i, v \in \mathcal{V}_i} t_i^2(w, v; \pi) \tilde{q}_i(w, v) - \left[\sum_{w \in \mathcal{W}_i, v \in \mathcal{V}_i} t_i(w, v; \pi) \tilde{q}_i(w, v) \right]^2, \end{aligned}$$

suggesting an estimate of the variance of the test statistic,

$$\widehat{\sigma}^2 \ = \ \sum_{i=1}^{n} t_i^2(W_i, V_i; \pi)\big|_{\pi = \hat{\pi}} .$$

Under broad regularity conditions, the asymptotic variance of $\hat{\pi}$ may be well approximated by the variance of the score of $\hat{\pi}$ divided by the square expected slope. (See

the appendix.) We seek to replace the score terms corresponding to the observed data (W, V) with the score for a function of the data t such that the variance of π may be minimized subject to unbiasedness conditions on $t(w, v; \pi)$. The variance of $\hat{\pi}$ can be minimized by minimizing the variance of t subject to a constraint on the square expected value of the slope and the constraints necessary for unbiasedness.

The derivative with respect to π of the expectation of the test statistic is

$$\begin{aligned} & \frac{\partial}{\partial \pi} \sum_{i=1}^{n} \sum_{(x,z,u) \in \mathcal{X}_i \times \mathcal{Z}_i \times \mathcal{U}_i} \gamma^i_{x,z,u} \sum_{w \in \mathcal{W}_i, v \in \mathcal{V}_i} t_i(w, v; \pi) q_{x,z,u}(w, v) \\ = & \sum_{i=1}^{n} \sum_{w \in \mathcal{W}_i, v \in \mathcal{V}_i} t_i(w, v; \pi) \sum_{(x,z,u) \in \mathcal{X}_i \times \mathcal{Z}_i \times \mathcal{U}_i} \gamma^i_{x,z,u} q_{x,z,u}(w, v) S_{x,z,u}(w, v) \end{aligned}$$

where the conditional score $S_{x,z,u}(w, v)$ is the derivative with respect to π of the natural log of $q_{x,z,u}(w, v)$.

To first order, if the γ^i are correctly specified, then the optimal power for an unbiased test statistic is achieved by taking t to maximize

$$\begin{aligned} & \sum_{i=1}^{n} \sum_{w \in \mathcal{W}_i, v \in \mathcal{V}_i} \sum_{(x,z,u) \in \mathcal{X}_i \times \mathcal{Z}_i \times \mathcal{U}_i} t_i^2(w, v; \pi) \gamma^i_{x,z,u} q_{x,z,u}(w, v) \\ & - \left[\sum_{w \in \mathcal{W}_i, v \in \mathcal{V}_i} \sum_{(x,z,u) \in \mathcal{X}_i \times \mathcal{Z}_i \times \mathcal{U}_i} t_i(w, v; \pi) \gamma^i_{x,z,u} q_{x,z,u}(w, v) \right]^2 \\ = & \sum_{i=1}^{n} \sum_{w \in \mathcal{W}_i, v \in \mathcal{V}_i} t_i^2(w, v; \pi) \tilde{q}_i(w, v) - \left[\sum_{w \in \mathcal{W}_i, v \in \mathcal{V}_i} t_i(w, v; \pi) \tilde{q}_i(w, v) \right]^2 \end{aligned}$$

subject to the unbiasedness constraint and a constraint on

$$\left[\sum_{i=1}^{n} \sum_{w \in \mathcal{W}_i, v \in \mathcal{V}_i} t_i(w, v; \pi) \sum_{(x,z,u) \in \mathcal{X}_i \times \mathcal{Z}_i \times \mathcal{U}_i} \gamma^i_{x,z,u} q_{x,z,u}(w, v) S_{x,z,u}(w, v) \right]^2$$

For each family i, let γ_i be a vector of prior probabilities on each $(x, z, u) \in \mathcal{X}_i \times \mathcal{Z}_i \times \mathcal{U}_i$ for the i^{th} family. Additionally, define Q_i to be a matrix of the conditional probabilities $q_{x,z,u}(w, v)$ for the i^{th} family that is row-indexed by the (x, z, u) and column-indexed by

$(w,v)_i = \left((w,v)_{i,1}, ..., (w,v)_{i,N_i}\right)^T$. Since certain configurations of ascertainment scheme and underlying complete pedigree structure may yield identical events and corresponding conditional probabilities, Q_i may feature duplicate rows. To avoid singularity issues, Q_i may be instead be row-indexed by the equivalence classes in $\mathcal{H}_i$. Defining Q_i in this manner requires that the γ^i be redefined as the vector of prior probabilities over each element in the equivalence classes in $\mathcal{H}_i$. Using matrix notation, the variance of the test statistic may be expressed as

$$\sum_{i=1}^{n} t_i^T M_i t_i$$

where M_i is a matrix corresponding to the i^{th} family whose entries are

$$M_i[k,k] = \tilde{q}_i\left((w,v)_k\right) - \left[\tilde{q}_i\left((w,v)_k\right)\right]^2$$

(for $k \in \{1, \ldots, N_i\}$) on the diagonals, and

$$M_i[k,l] = -\left[\tilde{q}_i\left((w,v)_k\right)\right] \times \left[\tilde{q}_i\left((w,v)_l\right)\right]$$

(for $k \neq l \in \{1, \ldots, N_i\}$).

Furthermore, defining $\nabla\tilde{q}_i$ to be the derivative with respect to π of $\tilde{q}_i$, the squared derivative with respect to π of the expectation of the test statistic may be expressed as

$$\left[\sum_{i=1}^{n} t_i^T \nabla_\pi \tilde{q}_i\right]^2$$

Using matrix notation, the unbiasedness conditions may be equivalently expressed as $Q_i t_i = \mathbf{0}$ for all i.

It follows that the variance of the penetrance estimate $\hat{\pi}$ is approximately

$$\mathbb{V}(\widehat{\pi}) = \frac{\sum_{i=1}^{n} t_i^T M_i t_i}{\left[\sum_{i=1}^{n} t_i^T \nabla_\pi \tilde{q}_i\right]^2}$$

The function t which yields optimal power by minimizing the variance of $\widehat{\pi}$ (subject to the unbiasedness conditions) under the null hypothesis can be found by minimizing

$$\sum_{i=1}^{n} t_i^T M_i t_i,$$

subject to the condition that for all π and i in $1, \ldots, n$, t_i is unbiased under all (x, z, u) configurations, that is, that

$$Q_i t_i = \mathbf{0}$$

for all i, and subject to a constraint on the denominator of the variance of $\hat{\pi}$,

$$\left[\sum_{i=1}^{n} t_i^T \nabla_\pi \tilde{q}_i\right]^2$$

or, equivalently, by maximizing

$$\left[\sum_{i=1}^{n} t_i^T \nabla_\pi \tilde{q}_i\right]^2$$

subject to $Q_i t_i = \mathbf{0}$ for all i and subject to a constraint on the denominator of the variance of $\hat{\pi}$,

$$\sum_{i=1}^{n} t_i^T M_i t_i.$$

Here, we indicate the calculations for the former. In this dissertation, t will be found by the

$$\sum_{i=1}^{n} t_i^T M_i t_i,$$

subject to $Q_i t_i = \mathbf{0}$ for all i and subject to a constraint on the denominator of the variance of $\hat{\pi}$,

$$\left[\sum_{i=1}^{n} t_i^T \nabla_\pi \tilde{q}_i\right]^2.$$

It should be noted that constraining the denominator is equivalent to constraining $\sum_{i=1}^{n} t_i^T \nabla_\pi \tilde{q}_i$ to a constant. Also, the function of the empty set event for each family i is set to 0 since the family was ascertained and, thus, some amount of information was observed. Hence, the optimization is done by omitting the column of the Q_i matrices corresponding to the empty set and proceeding accordingly to solve for the remaining entries of the t_i vectors.

The solution to this constrained optimization problem is a family-by-family weighted function of the Q_i matrices for each family i. The weights correspond to each row in the Q_i matrix for all i. The next section details how to calculate the appropriate weights for the (x, z, u) configurations for each family included in the study.

2.1.2 Finding the Appropriate Weights for Constrained Optimization Problem

The previous section discussed how the solution to the constrained minimization problem may be found by minimizing the variance of the test statistic $\sum_{i=1}^{n} t(W_i, V_i; \pi)$, expressible as $\sum_{i=1}^{n} t_i^T M_i t_i$, subject to the unbiasedness constraint and a constraint on the gradient of the expectation of the test statistic, $\sum_{i=1}^{n} t_i^T \nabla_\pi \tilde{q}_i$. This section will detail how to find the weights needed to achieve unbiasedness and efficiency of the test statistic.

Suppose that for any family i, $(X, Z, U)_{i,j}$ is independent of $(X, Z, U)_{i',j'}$ for all $(X, Z, U)_i$ in $2^{\mathcal{X}_i \times \mathcal{Z}_i \times \mathcal{U}_i}$ and $(X, Z, U)_{i'}$ in $2^{\mathcal{X}_{i'} \times \mathcal{Z}_{i'} \times \mathcal{U}_{i'}}$, $\forall i, i' \in \{1, \ldots, n\}$.

Taking the test statistic to be $\sum_{i=1}^{n} t_i(W_i, V_i; \pi)$, the t which minimizes the variance of the test statistic subject to the unbiasedness conditions is found by separately minimizing the variance of the test statistic with respect to t_i for each i while constraining $\sum_{i=1}^{n} t_i^T \nabla_\pi \tilde{q}_i$. The solution to this minimization problem is $t_{\text{sol}} = (t_{\text{sol},1}, \ldots, t_{\text{sol},n})$, where

$$t_{\text{sol},i} = \frac{1}{2} \lambda_0 M_i^{-1} \left[\nabla \tilde{q}_i - Q_i^T \lambda_i \right]$$

where λ_i is a vector of weights corresponding to each row of the Q_i matrix and is given by

$$\left(Q_i M_i^{-1} Q_i^T \right)^{-1} Q_i M_i^{-1} \nabla \tilde{q}_i,$$

$$\forall i \in \{1, \ldots, n\}$$

and constant λ_0 is a free parameter in $\mathbb{R} \setminus \{0\}$.

Note:

1. The entry of each t_i corresponding to the empty set must be set to zero *after* carrying out the above calculation.

2. Due to the assumption that noncarriers cannot be affected and the way in which $\mathcal{W}_i$ is defined, the Q_i matrix will likely possess zero-columns, resulting in the singularity

of matrix M_i. In the event of singularity of M_i, the pseudoinverse of M_i (i.e. M_i^+) may be used in place of M_i^{-1}. (This will result in the function being defined as 0 at the w events corresponding to the zero-columns.) Neither the test statistic variance nor the bias will be affected by using the pseudoinverse in place of the actual inverse.

Proof. The solution to this problem may be found through the use of Lagrange multipliers. The derivative of $\sum_{i=1}^{n} t_i^T M_i t_i$ with respect to t_f (for $f \in 1, \ldots, n$) is $2\tilde{M}_f t_f$. Let λ_0 be a Lagrange multiplier for the $\sum_{i=1}^{n} t_i^T \nabla_\pi \tilde{q}_i$ constraint, and let $\lambda_f^* = (\lambda_{f,1}^*, \ldots, \lambda_{f,L_f}^*)$ be a vector of Lagrange multipliers associated with each $t_i^T q_{i,j} = 0$ constraint for family f. The solution to the minimization problem must satisfy the following equation:
$\nabla_{t_f} \sum_{i=1}^{n} t_i^T M_i t_i = \lambda_0 \nabla_{t_f} \sum_{i=1}^{n} t_i^T \nabla_\pi \tilde{q}_i + \lambda_{f,1}^* \nabla_{t_f} t_f^T q_{f,1} + \ldots + \lambda_{f,L_f}^* \nabla_{t_f} t_f^T q_{f,L_f}$
The above equation simplifies to

$$2M_f t_f = \lambda_0 \nabla_\pi \tilde{q}_f + \sum_{j=1}^{L_f} \lambda_{f,j}^* q_{f,j}$$

$$\Leftrightarrow \quad 2M_f t_f = \lambda_0 \nabla_\pi \tilde{q}_f + Q_f^T \lambda_f^*.$$

Solving for t_f yields the following:

$$t_f = \frac{1}{2} M_f^{-1} \left[\lambda_0 \nabla \tilde{q}_f + Q_f^T \lambda_f^* \right].$$

Substituting this expression for t_f into the $Q_f t_f = 0$ constraint yields the following solution for λ_f^*:

$$\lambda_f^* = -\lambda_0 (Q_f M_f^{-1} Q_f^T)^{-1} Q_f M_f^{-1} \nabla_\pi \tilde{q}_f^T$$

Thus, the solution for t_i is given by

$$t_f = \frac{1}{2} \lambda_0 M_f^{-1} \left[\nabla_\pi \tilde{q}_f - Q_f^T \lambda_f \right]$$

$$\forall f \in \{1, \ldots, n\}$$

where $\lambda_f = \left(Q_f M_f^{-1} Q_f^T \right)^{-1} Q_f M_f^{-1} \nabla_\pi \tilde{q}_f^T$. □

2.2 Estimating Penetrance Under Parametric Assumptions on the Ascertainment Scheme

2.2.1 Deriving the Form of the Constrained Optimization Problem

Define X_i, G_i, Y_i, Z_i, and U_i as in Section 2.1. Similarly, let W_i denote the observed phenotype and genotype information and V_i be the observed external mapping information for pedigree i. Let $\mathcal{Z}_i$ denote the set of all conceivable ascertainment schemes, i.e. maps from

$$\bigcup_{x \in \mathcal{X}_i} \{x\} \times \mathcal{G}_x \times \mathcal{Y}_x$$

into $\mathcal{W}_i$, and let $\mathcal{U}$ be the set of all possible maps from $\mathcal{Z} \times \mathcal{C}$ into the power set of the set of all conceivable mappings $\mathcal{Z}$.

The objective remains to find a function of observed data $(w, v)_1, \ldots, (w, v)_n$ which yields a test statistic from which unbiased and efficient penetrance estimates may be derived. The unbiasedness condition may be relaxed if it is deemed appropriate to posit certain assumptions on the Z_i. Suppose that it is assumed that for an index set Θ and for maps from Θ into $\mathcal{Z}_i$, $\theta \longrightarrow z_i^\theta$, that there is some $\theta \in \Theta$ such that for all i, $Z_i = z_i^\theta$. The parameter θ will be assumed to be $\sqrt{n}$-estimable. Let $p(\theta)$ be the prior probability on θ, and let $\alpha_{x,u}^i$ to be the i^{th} family's prior probability on underlying complete pedigree x and underlying external information mapping u. As in the previous chapters, define $p_i^\pi(g, y|x)$ to be the conditional probability mass function for G_i and Y_i, given X_i. Let $q_{x,u}^\theta(w, v)$ be the probability of observing w and v for underlying pedigree x, ascertainment scheme z^θ, and external information mapping u, i.e.

$$q_{x,u}^\theta(w, v) = \sum_{g,y: z^\theta(x,g,y)=w, u(z,x,g,y)=v} p^\pi(g, y|x),$$

where dependence on π is suppressed in the notation.

Let $\hat{\theta}$ be an unbiased estimate of θ. Under these parametric assumptions on the distribution of Z, the sufficient conditions for approximate unbiasedness of t are that for a consistent estimator $\hat{\theta} \in \Theta$, for all $x_i \in \mathcal{X}_i$, and for all $u \in \mathcal{U}_i$,

$$\begin{aligned} & \sum_{i=1}^{n} \mathbb{E}^{\pi}\{t_i(z^{\hat{\theta}}(x_i, G_i, Y_i), u_i(z^{\hat{\theta}}, x_i, G_i, Y_i); \pi)\} \\ = & \sum_{i=1}^{n} \sum_{w \in \mathcal{W}_i, v \in \mathcal{V}_i} t_i(w, v; \pi) q^{\hat{\theta}}_{x,u}(w, v) \\ = & \ 0. \end{aligned}$$

For each family i, define $Q_i(\hat{\theta})$ to be a matrix of the conditional probabilities $q^{\hat{\theta}}_{x,u}(w, v)$ for the i^{th} family that is row-indexed by the (x, u) configurations and column-indexed by $(w, v)_i = ((w, v)_{i,1}, ..., (w, v)_{i,N_i})^T$. Additionally, define t_i to be $(t_i(w, v; \pi)_1, ..., t_i(w, v; \pi)_{N_i}))^T$ as in previous sections. Using matrix notation, the unbiasedness conditions may be re-expressed as follows:

$$\sum_{i=1}^{n} t_i^T q^{\hat{\theta}}_{i,j_i} = 0$$

for all $j_i \in 1, \ldots, L_i$ and for all $\hat{\theta} \in \Theta$.

Under the parametric assumption posited on Z and the working prior on the X_i and U_i, the variance of the test statistic $\sum_{i=1}^{n} t(w_i, v_i; \pi)$ may be expressed as

$$\sum_{i=1}^{n} \left\{ \sum_{(w,v) \in \mathcal{W}_i \times \mathcal{V}_i} t_i^2(w, v; \pi) \left(\sum_{(x,u) \in \mathcal{X}_i \times \mathcal{U}_i} \alpha^i_{x,u} q^{\hat{\theta}}_{x,u}(w, v) \right) - \left[\sum_{(w,v) \in \mathcal{W}_i \times \mathcal{V}_i} t_i(w, v; \pi) \left(\sum_{(x,u) \in \mathcal{X}_i \times \mathcal{U}_\rangle} \alpha^i_{x,u} q^{\hat{\theta}}_{x,u}(w, v) \right) \right]^2 \right\}$$

Using matrix notation, the variance of the test statistic may be expressed as

$$\sum_{i=1}^{n} t_i^T B_i t_i$$

where B_i is a matrix corresponding to the i^{th} family whose entries are

$$B_i[k,k] = \sum_{(x,u)\in\mathcal{X}_i\times\mathcal{U}_i} \alpha^i_{x,u} q^{\hat{\theta}}_{x,u}((w,v)_k) - \left[\sum_{(x,u)\in\mathcal{X}_i\times\mathcal{U}_i} \alpha^i_{x,u} q^{\hat{\theta}}_{x,u}((w,v)_k)\right]^2$$

(for $k = 1, \ldots, N_i$) on the diagonals, and

$$B_i[k,l] = -\left[\sum_{(x,u)\in\mathcal{X}_i\times\mathcal{U}_i} \alpha^i_{x,u} q^{\hat{\theta}}_{x,u}((w,v)_k)\right] \times \left[\sum_{(x,u)\in\mathcal{X}_i\times\mathcal{U}_i} \alpha^i_{x,u} q^{\hat{\theta}}_{x,u}((w,v)_k)\right]$$

(for $k \neq l \in \{1, \ldots, N_i\}$) on the off-diagonals.

Under the parametric assumptions posited on Z and the assumed prior on the X_i and U_i, the derivative with respect to π of the expectation of the test statistic may be expressed as

$$\begin{aligned}
& \nabla_\pi \mathbb{E}^\pi \sum_{i=1}^n t_i(W_i, V_i; \pi) \\
\approx\ & \nabla_\pi \mathbb{E}_{X_i,U_i,\theta} \mathbb{E}^\pi \left\{ \sum_{i=1}^n [t_i(W_i, V_i; \pi) | X_i, U_i, \theta] \right\} \Bigg|_{\theta=\hat{\theta}} \\
=\ & \nabla_\pi \sum_{i=1}^n \sum_{(w,v)\in\mathcal{W}_i\times\mathcal{V}_i} t_i(w, v; \pi) \sum_{(x,u)\in\mathcal{X}_i\times\mathcal{U}_i} \alpha^i_{x,u} q^{\hat{\theta}}_{x,u}(w, v) \\
=\ & \sum_{i=1}^n \sum_{(w,v)\in\mathcal{W}_i\times\mathcal{V}_i} t_i(w, v; \pi) \left[\sum_{(x,u)\in\mathcal{X}_i\times\mathcal{U}_i} \alpha^i_{x,u} \nabla_\pi q^{\hat{\theta}}_{x,u}(w, v)\right] \\
=\ & \sum_{i=1}^n \sum_{(w,v)\in\mathcal{W}_i\times\mathcal{V}_i} t_i(w, v; \pi) \left[\sum_{(x,u)\in\mathcal{X}_i\times\mathcal{U}_i} \alpha^i_{x,u} q^{\hat{\theta}}_{x,u}(w, v) S^{\hat{\theta}}_{x,u}(w, v)\right]
\end{aligned}$$

where the score $S^{\hat{\theta}}_{x,u}(w, v)$ is the derivative with respect to π of the natural log of $q^{\hat{\theta}}_{x,u}(w, v)$.

Let $\breve{q}_i(w, v) = \sum_{x\in\mathcal{X}_i, u\in\mathcal{U}_i} \alpha^i_{x,u} q^{\hat{\theta}}_{x,u}(w, v)$, and let $\breve{q}_i$ be the vector $(\breve{q}_i(w, v)_1, \ldots, \breve{q}_i(w, v)_{N_i})^T$. Denote $\nabla_\pi \breve{q}_i$ as the element-wise derivative with respect to π of the $\breve{q}_i$ vector. The derivative with respect to π of the expectation of the test statistic may be alternatively expressed as $\sum_{i=1}^n t_i^T \nabla \breve{q}_i$.

Using matrix notation, the variance of the penetrance estimate $\hat{\pi}$ may be expressed as

$$\mathbb{V}(\widehat{\pi}) \approx \frac{\sum_{i=1}^{n} t_i^T B_i t_i}{\left[\sum_{i=1}^{n} t_i^T \nabla_\pi \breve{q}_i\right]^2}$$

The objective of this chapter is to find function $t = (t_1, \ldots, t_n)$ such that, under the parametric assumptions on Z_i and correct specification of the $\alpha^i_{X,U}$, the power for the test statistic is maximized subject to the unbiasedness constraint. To first order, if the $\alpha^i_{x,u}$ are correctly specified, then the optimal power for an unbiased test statistic may be achieved by taking t to minimize

$$\sum_{i=1}^{n} t_i^T B_i t_i$$

subject to the unbiasedness conditions and a constraint on

$$\sum_{i=1}^{n} t_i^T \nabla_\pi \breve{q}_i$$

It should be noted that the function of the empty set event for each family i is set to 0 since the family was ascertained and, thus, some amount of information was observed. Hence, the optimization is done by omitting the column of the $Q_i(\theta)$ matrices corresponding to the empty set and proceeding accordingly to solve for the remaining entries of the t_i vectors.

Define t as the vector $(t_1|\ldots|t_n)$, $\breve{q}$ as the vector $(\breve{q}_1|\ldots|\breve{q}_n)$, and block-diagonal matrix B as

$$\begin{pmatrix} B_1 & & & \\ & B_2 & & 0 \\ & & B_3 & \\ & 0 & & \ddots \\ & & & & B_n \end{pmatrix}$$

Let $Q(\hat{\theta})$ be a matrix composed of all combinations of rows from each $Q_i(\hat{\theta})$ matrix (omitting the column of Q_i corresponding to the empty set). More specifically, the columns of $Q(\hat{\theta})$ are composed of all elements of

$$\{q_{1,1}(\hat{\theta}), \ldots, q_{1,L_1}(\hat{\theta})\} \times \ldots \times \{q_{n,1}(\hat{\theta}), \ldots, q_{n,L_n}(\hat{\theta})\},$$

where $q_{i,j}(\hat{\theta})$ represents the j^{th} row of the $Q_i(\hat{\theta})$ matrix whose column corresponding to the empty set is omitted.

It is convenient to re-express the aforementioned minimization problem as finding t to minimize

$$t^T B t$$

subject to a constraint on $t^T \nabla \breve{q}$ and subject to the condition

$$Q(\hat{\theta})t = \mathbf{0}$$

for all $\sqrt{n}$-consistent $\hat{\theta} \in \Theta$.

The solution to this constrained minimization is a weighted function of the $Q(\hat{\theta})$ and B matrices. It should be noted that this solution is only approximate and may result in a loss of information due to the estimation of θ. Nonetheless, the solution provides a reasonable bound by which penetrance estimates may be derived in the setting where parametric assumptions can be made on Z. The specific form of the solution to this constrained minimization problem is found in the next section.

2.2.2 Finding the Appropriate Weights for Constrained Optimization Problem

The previous section discussed how the solution to the constrained minimization problem may be found by minimizing the variance of the test statistic $\sum_{i=1}^{n} t(W_i, V_i; \pi)$, expressible as $t^T Bt$, subject to the unbiasedness constraint and a constraint on the gradient of the expectation of the test statistic, $t^T \nabla \breve{q}$. This section will detail how to find the weights needed to achieve unbiasedness and efficiency of the test statistic.

Suppose that for any family i, $(X,U)_{i,j}$ is independent of $(X,U)_{i,k}$ for all (X,U) in $\mathcal{X}_i \times \mathcal{U}_i$ and that θ is $\sqrt{n}$ estimable and identifiable. Under parametric assumptions on the ascertainment scheme and taking the test statistic to be $\sum_{i=1}^{n} t_i(W_i, V_i; \pi)$, the t which minimizes the variance of the test statistic subject to the unbiasedness conditions is of the form

$$t = \frac{1}{2}(B)^{-1}\left[\lambda_0 \nabla \breve{q} + Q(\hat{\theta})\lambda(\hat{\theta})\right],$$

where the constant λ_0 is a nonzero free parameter and vector of weights $\lambda(\hat{\theta})$ is given by

$$\left(Q(\hat{\theta})B^{-1}Q(\hat{\theta})^T\right)^{-1} Q(\hat{\theta})B^{-1}\nabla_\pi \breve{q}$$

Depending on the form of the B matrix, B may be singular. In the event of singularity of B, the pseudoinverse of B (i.e. B^+) may be used in place of B^{-1}.

Proof. The solution to this problem may be found through the use of Lagrange multipliers. The derivative of $t^T Bt$ with respect to t is $2Bt$. Let λ_0 be a Lagrange multiplier for the $t^T \nabla_\pi \breve{q}$ constraint, and let $\lambda(\hat{\theta})$ be a vector of Lagrange multipliers associated with the $Q(\hat{\theta})^T t = \mathbf{0}$ constraint for $\hat{\theta} \in \Theta$. The solution to the minimization problem must satisfy the following equation:

$$\nabla_t t^T Bt = \lambda_0 \nabla_t (t^T \nabla_\pi \breve{q}) + \nabla_t Q(\hat{\theta})^T t \; \lambda^*(\hat{\theta})$$

The above equation simplifies as follows:

$$2Bt = \lambda_0 \nabla_\pi \breve{q} + Q(\hat{\theta})\lambda^*(\hat{\theta})$$

Solving for t yields the following:

$$t = \frac{1}{2} B^{-1} \left[\lambda_0 \nabla_\pi \breve{q} + Q(\hat{\theta})\lambda^*(\hat{\theta}) \right].$$

Substituting this expression for t into the $Q(\hat{\theta})^T t = \mathbf{0}$ constraint yields the following solution for weight vector $\lambda^*(\hat{\theta})$:

$$-\lambda_0 \left(Q(\hat{\theta}) B^{-1} Q(\hat{\theta})^T \right)^{-1} Q(\hat{\theta}) B^{-1} \nabla_\pi \breve{q}.$$

Thus, the solution for t is

$$t = \frac{1}{2} \lambda_0 B^{-1} \left[\nabla_\pi \breve{q} - Q(\hat{\theta})^T \lambda(\hat{\theta}) \right],$$

where $\lambda(\hat{\theta}) = \left(Q(\hat{\theta}) B^{-1} Q(\hat{\theta})^T \right)^{-1} Q(\hat{\theta}) B^{-1} \nabla_\pi \breve{q}^T$.

□

2.3 Solving for the Penetrance Estimate

2.3.1 Non-parametric Assumptions on Z

Let $n_i = |\mathcal{W}_i \times \mathcal{V}_i|$. Let $(w, v)^i_*$ represent the outcome observed for family i, and let D_i be an $(n_i - 1) \times (n_i - 1)$ matrix whose only nonzero entry is a one corresponding to the $(w, v)^{i\text{th}}_*$ diagonal entry. (Recall, the entry corresponding to the empty set event is omitted in the calculation of the t vector since the t entry corresponding to the empty set event is set to zero.) Then the optimal function of the observed outcome (optimal in the sense that the function of the data that yields the minimum variance among the set of all unbiased estimators) is given by the following:

$$t((w, v; \pi)^i_*) = \left\| \frac{1}{2} \lambda_0 D_i M_i^{-1} \left[\nabla_\pi \tilde{q}_i - Q_i^T (Q_i M_i^{-1} Q_i^T)^{-1} Q_i M_i^{-1} \nabla_\pi \tilde{q}_i \right] \right\|_2$$

If information $((w,v)_*^1,\ldots,(w,v)_*^n)$ is observed for families $i,\ldots,n$, then the "observed" test statistic is

$$\sum_{i=1}^{n} t((w,v;\pi)_*^i) = \sum_{i=1}^{n} \left\| \frac{1}{2}\lambda_0 D_i \left[M_i^{-1}\nabla_\pi \tilde{q}_i - Q_i^T (Q_i M_i^{-1} Q_i^T)^{-1} Q_i M_i^{-1} \nabla_\pi \tilde{q}_i \right] \right\|_2$$

Since under the null hypothesis (for the correct value of penetrance π_o), the test statistic has expectation 0, the penetrance estimate $\hat{\pi}$ is the solution to the equation $\sum_{i=1}^{n} t_i((w,v;\pi)_*^i) = 0$. In other words,

$$\hat{\pi} = \arg_\pi \sum_{i=1}^{n} \left\| \frac{1}{2}\lambda_0 D_i M_i^{-1} \left[\nabla_\pi \tilde{q}_i - Q_i^T (Q_i M_i^{-1} Q_i^T)^{-1} Q_i M_i^{-1} \nabla_\pi \tilde{q}_i \right] \right\|_2 = 0$$

2.3.2 Parametric Assumptions on Z

Let $(w,v)_*^1,\ldots,(w,v)_*^n$ represent the outcome observed for families $1,\ldots,n$, and let D be an $(\sum_{i=1}^{n} n_i - 1) \times (\sum_{i=1}^{n} n_i - 1)$ matrix whose only nonzero entries contain a one corresponding to the $(w,v)_*^1,\ldots,(w,v)_*^n$ diagonal entries. Assume that all events corresponding to the empty set in the t vector have been removed. Then the test statistic $\sum_{i=1}^{n} t_i((w,v;\pi)_*^i)$ is given by the following:

$$\left\| \frac{1}{2}\lambda_0 D B^{-1} \left[\nabla_\pi \breve{q} - Q(\hat{\theta})^T \left(Q(\hat{\theta}) B^{-1} Q(\hat{\theta})^T \right)^{-1} Q(\hat{\theta}) B^{-1} \nabla_\pi \breve{q}^T \right] \right\|_2 .$$

Since under the null hypothesis (for the correct value of penetrance π_o), the test statistic has expectation 0, the penetrance estimate $\hat{\pi}$ is the solution to the equation $\sum_{i=1}^{n} t_i((w,v;\pi)_*^i) = 0$. Thus,

$$\hat{\pi} \approx \arg_\pi \left\| \frac{1}{2}\lambda_0 D B^{-1} \left[\nabla_\pi \breve{q} - Q(\hat{\theta})^T \left(Q(\hat{\theta}) B^{-1} Q(\hat{\theta})^T \right)^{-1} Q(\hat{\theta}) B^{-1} \nabla_\pi \breve{q}^T \right] \right\|_2 = 0.$$

In both cases, the penetrance estimates can be found by using numerical solvers. In the following examples, the penetrance estimates were peformed using R software.

2.4 Summary of Algorithm to Find Efficient Penetrance Estimates

2.4.1 Non-parametric Assumptions on Z

This section contains a summary of the algorithm for non-parametric assumptions on the ascertainment mappings $Z_i, \ldots, Z_n$. To carry out the algorithm, do steps 1-8 for all families in the study and then proceed with steps 9 and 10 to solve for the penetrance estimate.

1. Calculate Q_i matrix, based on the set of all plausible pedigree structures $\mathcal{X}_i$, mappings $\mathcal{Z}_i$, and external information mappings $\mathcal{U}_i$. Omit the Q_i column corresponding to the empty set (i.e. the event of family i not being ascertained), and call the resulting matrix Q_i^*.

2. Calculate $\tilde{q}_i$ based on Q_i^* and the vector of prior probabilities on the ascertainment scheme and pedigree combinations as follows: $\tilde{q}_i = \gamma_i^T Q_i^*$

3. Calculate the M_i matrix based on $\tilde{q}_i$.

4. Calculate $\lambda = (Q_i M_i^+ Q_i^T)^+ Q_i M_i^+ \nabla_\pi \tilde{q}_i$

5. Calculate $t_i = \frac{1}{2}\lambda_0 M_i^+ \left[\nabla \tilde{q}_i - Q_i^T \lambda_i\right]$

6. Suppose $(\tilde{w}, \tilde{v})_i$ corresponds to the event of family i being unobserved. Complete the t_i vector by setting $t_i((\tilde{w}, \tilde{v})_i)$ equal to 0.

7. Find the t_i entry corresponding to the information that was actually observed, $(w, v)_i^*$. This will be a [very complicated] function of the penetrance parameter π.

8. Calculate the test statistic $\sum_{i=1}^{n} t((w, v; \pi)_i^*)$. This, too, will be a complicated function of the penetrance parameter.

9. Use numeric methods to find the π-value which satisfies the equation $\sum_{i=1}^{n} t((w, v; \pi)_i^*) = 0$. This is the penetrance estimate.

2.4.2 Parametric Assumptions on Z

This section contains a summary of the algorithm for parametric assumptions on the ascertainment mappings Z. However, they may be adapted to the parametric setting in the case in which the Z_i are parameterized by θ by estimating θ, constructing the $Q^{\hat{\theta}}$ matrix, and proceeding accordingly. To carry out the algorithm, first estimate θ as outlined in step 1 and then carrying out steps 2-4 for all families in the study and then proceed with steps 5-11 to solve for the penetrance estimate.

1. Calculate $Q_i(\hat{\theta})$ matrix, based on the set of all plausible pedigree structures $\mathcal{X}_i$, external information mappings $\mathcal{U}_i$, and $\hat{\theta}$. Omit the $Q_i(\hat{\theta})$ column corresponding to the empty set (i.e. the event of family i not being ascertained), and call the resulting matrix $Q_i(\hat{\theta})^*$.

2. Calculate $\tilde{q}_i^{\hat{\theta}}$ based on $Q_i(\hat{\theta})^*$ and the vector of prior probabilities on the ascertainment scheme and pedigree combinations.

3. Calculate the B_i matrix based on $\tilde{q}_i$.

4. Create aggregated vector $\breve{q} = (\breve{q}_1|\ldots|\breve{q}_n)$ and block-diagonal matrix B, where the diagonals consist of matrices $B_1, \ldots, B_n$.

5. Create $Q(\hat{\theta})$, a matrix composed of all possible combinations of rows from each $Q_i(\hat{\theta})^*$ matrix.

6. Calculate $\lambda = (Q(\hat{\theta})B^+Q(\hat{\theta})^T)^+Q(\hat{\theta})B^+\nabla_\pi \breve{q}^{\hat{\theta}}$

7. Calculate $t = \frac{1}{2}\lambda_0 B^+ \left[\nabla \breve{q}^{\hat{\theta}} - Q(\hat{\theta})^T\lambda(\hat{\theta})\right]$

8. Find the t entries corresponding to the information that was actually observed: $t(w, v; \pi)_1^*, \ldots, t(w, v; \pi)_n^*$. Each entry will be a [very complicated] function of the penetrance π.

9. Calculate the test statistic $\sum_{i=1}^{n} t((w, v; \pi)_i^*)$. This, too, will be a complicated function of the penetrance parameter.

10. Use numeric methods to find the π-value which satisfies the equation $\sum_{i=1}^{n} t((w, v; \pi)_i^*) = 0$. This is the penetrance estimate.

Chapter 3

Applications

3.1 An Illustrative Single-Family Example

In this simple case, assume that an affected and an unaffected subject are observed. It is believed that the two subjects comprise the entirety of the underlying complete pedigree structure, a sibling pair. This sibling pair structure is the only element of the complete pedigree structure set $\mathcal{X}$. The observed subjects will be referred to as "Subject 1" and "Subject 2." The set of complete observable information is summarized by the following table:

Subject 1, Subject 2 Geno.	**Subject 1, Subject 2 Pheno.**	$(\mathbf{w}, \mathbf{v})$
Carrier, Carrier	Affected, Affected	$(w, v)_1$
Carrier, Carrier	Affected, Unaffected	$(w, v)_2$
Carrier, Carrier	Unaffected, Affected	$(w, v)_3$
Carrier, Carrier	Unaffected, Unaffected	$(w, v)_4$
Carrier, Not Carrier	Affected, Affected	$(w, v)_5$
Carrier, Not Carrier	Affected, Unaffected	$(w, v)_6$
Carrier, Not Carrier	Unaffected, Affected	$(w, v)_7$
Carrier, Not Carrier	Unaffected, Unaffected	$(w, v)_8$
Not Carrier, Carrier	Affected, Affected	$(w, v)_9$
Not Carrier, Carrier	Affected, Unaffected	$(w, v)_{10}$
Not Carrier, Carrier	Unaffected, Affected	$(w, v)_{11}$
Not Carrier, Carrier	Unaffected, Unaffected	$(w, v)_{12}$
Not Carrier, Not Carrier	Affected, Affected	$(w, v)_{13}$
Not Carrier, Not Carrier	Affected, Unaffected	$(w, v)_{14}$
Not Carrier, Not Carrier	Unaffected, Affected	$(w, v)_{15}$
Not Carrier, Not Carrier	Unaffected, Unaffected	$(w, v)_{16}$
Carrier, –	Affected, –	$(w, v)_{17}$
Carrier, –	Unaffected, –	$(w, v)_{18}$
Not Carrier, –	Affected, –	$(w, v)_{19}$
Not Carrier, –	Unaffected, –	$(w, v)_{20}$
–, Carrier	– , Affected	$(w, v)_{21}$
–, Carrier	– , Unaffected	$(w, v)_{22}$
–, Not Carrier	– , Affected	$(w, v)_{23}$
–, Not Carrier	– , Unaffected	$(w, v)_{24}$
$\varnothing$	–, –	$(w, v)_{25}$

Table 3.1: Observable Genotypic and Phenotypic Outcomes for Family Size 2

Note: "–" means that the subject's information wasn't observed.

A non-parametric approach on the ascertainment scheme will be taken in this example.

The probability of each subject being a carrier was taken to be 0.5. Suppose that it is believed that three ascertainment schemes were possible when the subjects were observed. The first possible scheme is full ascertainment (sampling everyone in the pedigree regardless of phenotype and/or genotype); the second possible scheme involves sampling the entire pedigree if it contains at least two carriers; and the third possible scheme is sampling the entire pedigree if at least one of the subjects is a carrier. These three ascertainment schemes comprise the set of plausible ascertainment schemes $\mathcal{Z}$. Though subsets of the complete pedigree are theoretically possible based on the way $\mathcal{W}$ is defined, under the presumed possible pedigree schemes, the entire pedigree is observable. In this particular case, no information is known about the proband status of the subjects, though it may be presumed that Subject 1 is the proband since he/she was the only affected subject observed. Under non-parametric assumptions on Z, the events in Table 3.1 result in the following transposed conditional probability (Q) matrix:

W	Sibship, Full	Sibship, ≥ 1 Carrier	Sibship, Complete
CC AA	$\frac{1}{4}\pi^2$	$\frac{1}{4}\pi^2$	$\frac{1}{4}\pi^2$
CC AU	$\frac{1}{4}\pi(1-\pi)$	$\frac{1}{4}\pi(1-\pi)$	$\frac{1}{4}\pi(1-\pi)$
CC UA	$\frac{1}{4}\pi(1-\pi)$	$\frac{1}{4}\pi(1-\pi)$	$\frac{1}{4}\pi(1-\pi)$
CC UU	$\frac{1}{4}(1-\pi)^2$	$\frac{1}{4}(1-\pi)^2$	0
CN AA	0	0	0
CN AU	$\frac{1}{4}\pi$	$\frac{1}{4}\pi$	$\frac{1}{4}\pi$
CN UA	0	0	0
CN UU	$\frac{1}{4}(1-\pi)$	$\frac{1}{4}(1-\pi)$	0
NC AA	0	0	0
NC AU	0	0	0
NC UA	$\frac{1}{4}\pi$	$\frac{1}{4}\pi$	$\frac{1}{4}\pi$
NC UU	$\frac{1}{4}(1-\pi)$	$\frac{1}{4}(1-\pi)$	0
NN AA	0	0	0
NN AU	0	0	0
NN UA	0	0	0
NN UU	$\frac{1}{4}$	0	0
Subject 1 - C A	0	0	0
Subject 1 - C U	0	0	0
Subject 1 - N A	0	0	0
Subject 1 - N U	0	0	0
Subject 2 - C A	0	0	0
Subject 2 - C U	0	0	0
Subject 2 - N A	0	0	0
Subject 2 - N U	0	0	0
$\varnothing$	0	$\frac{1}{4}$	$1-\frac{1}{4}(2-\pi)^2$

Table 3.2: Transposed Q matrix for Sibship of Size 2

where the first column corresponds to complete ascertainment of the full sibling pair, the second column corresponds to the "affecteds only" sampling and the full sibling pair, the

third column corresponds to just observing the first sibling under complete ascertainment, and the fourth column corresponds to just observing sibling 1 only if he/she is affected. In this table, "C" denotes "Carrier", "A" denotes "Affected", "N" denotes "Not a Carrier," and "U" denotes "Unaffected".

Each row of the Q matrix represents the conditional probability of each possible outcome based on a configuration of the underlying pedigree, ascertainment scheme, and informative mapping. Under all three plausibles schemes considered in this example, the full pedigree is sampled if the ascertainment criteria are satisfied. Thus, events $(w, v)_{17}$ through $(w, v)_{25}$ would not occur under these schemes. The probabilities for each event follow the likelihood for full ascertainment outlined in Section 1.5. Under the second plausible ascertainment scheme, the full sibling pair is only observed if at least one of the siblings is a carrier. Events not satisfying this requirement have probability mass 0 under this scheme. Thus, all events in which both subjects are noncarriers have probability mass zero under this ascertainment scheme, and the empty set has mass equal to the probability that both subjects are noncarriers. Under complete ascertainment, the events in which at least one of the subjects is affected have non-zero mass (with the exception of the empty set event). Regardless of the plausible ascertainment scheme, events which violate the assumption that noncarriers cannot be affected with the phenotype are assigned probability mass zero.

An even prior will be taken over all three possible (x, z, u) configurations. Thus, $\gamma = (0.\bar{3}, 0.\bar{3}, 0.\bar{3})$. Note that in some cases, different ascertainment schemes and underlying complete pedigree structure configurations may yield the same conditional probabilities, resulting in identical rows in the Q matrix. In such a case, the resulting M matrix will be singular. A quick fix for such a case is to eliminate all but one of the duplicate rows and to combine the prior probabilities corresponding to all of the eliminated duplicate rows. Furthermore, rows in Q which have probability mass 1 on the empty set entry and prior probability 0 may be omitted from the Q matrix since impossible (x, z, u) configurations will contribute no information in the estimation of π.

The even prior taken over the three possible (x, z, u) configurations yields a $\tilde{q}$ vector of $(0.\bar{3}, 0.\bar{3}, 0.\bar{3}) \times Q$. With the M matrix being defined as in previous chapters (where the diagonals are equal to $\tilde{q}(w, v)_k - \tilde{q}^2(w, v)_k$ and the off-diagonals are equal to $-\tilde{q}(w, v)_k \tilde{q}(w, v)_l$), M is still be a singular matrix, as it contains several zero-columns, most which correspond to phenotypic-genotypic configurations in which at least one non-carrier is affected with the genotype. (Per conventional assumptions, this has probability 0 of occurring.) Thus, the problem must be redefined to minimize $t^T M^+ t$ subject to a constraint on $t^T \nabla_\pi \tilde{q}$ and a subject to $Qt = \mathbf{0}$, where M^+ is the pseudoinverse of M. Using the pseudoinverse allows for a loosening of the regularity conditions on M and $Q^T M^{-1} Q$ being non-singular. This method also yields identical results if one were to remove the zero-rows from the Q matrix and the corresponding entries in the t and $\tilde{q}$.

In this case, event $(w, v)_2$ is observed. Since only one "family" is observed, the test statistic is simply $t\big((w, v)_2\big)$. As such, the penetrance estimate $\hat{\pi}$ in this case is the value of π which satisfies $t\big((w, v)_2\big) = 0$. The π value (constrained to be between 0 and 1) which yields the zero root is approximately 0.31.

The following vector is the result under the settings $\pi = 0.31$ and λ_0 (the free parameter) equal to -1:

Table 3.3: t vector for illustrative example for $\pi = \hat{\pi}$

	$t(W,V)$
$t\big((w,v)_1\big)$	4.66
$t\big((w,v)_2\big)$	0
$t\big((w,v)_3\big)$	0
$t\big((w,v)_4\big)$	-2.90
$t\big((w,v)_5\big)$	0
$t\big((w,v)_6\big)$	1.451
$t\big((w,v)_7\big)$	0
$t\big((w,v)_8\big)$	-1.45
$t\big((w,v)_9\big)$	0
$t\big((w,v)_{10}\big)$	0
$t\big((w,v)_{11}\big)$	1.45
$t\big((w,v)_{12}\big)$	-1.45
$t\big((w,v)_{13}\big)$	0
$t\big((w,v)_{14}\big)$	0
$t\big((w,v)_{15}\big)$	0
$t\big((w,v)_{16}\big)$	0
$t\big((w,v)_{17}\big)$	0
$t\big((w,v)_{18}\big)$	0
$t\big((w,v)_{19}\big)$	0
$t\big((w,v)_{20}\big)$	0
$t\big((w,v)_{21}\big)$	0
$t\big((w,v)_{22}\big)$	0
$t\big((w,v)_{23}\big)$	0
$t\big((w,v)_{24}\big)$	0
$t\big((w,v)_{25}\big)$	0

Due to the fact that in this example two of the three plausible ascertainment schemes are fairly similar in terms of the conditional probabilities that they yield, the penetrance is not overly sensitive to the prior. However, as can be seen in Table 3.4, the penetrance estimate does noticeably decrease as the prior probability on the complete ascertainment scheme configuration increases. Typically, the extent to which the penetrance estimate is significantly affected by the prior depends on the set of possible ascertainment schemes, set of underlying complete pedigrees, and informative mappings for each family.

Table 3.4: Penetrance Estimate as a Function of Prior Probabilities - Sibship of Size 2

Prior probabilities $\gamma = (\gamma_1, \gamma_2, \gamma_3)$	Penetrance Estimate - $\hat{\pi}$
$(0, 0, 1)$	0
$(0, 0.25, 0.75)$	0
$(0, 0.5, 0.5)$	0.19
$(0, 0.75, 0.25)$	0.385
$(0, 1, 0)$	0.5
$(0.25, 0, 0.75)$	0
$(0.25, 0.25, 0.5)$	0.153
$(0.25, 0.5, 0.25)$	0.375
$(0.25, 0.75, 0)$	0.5
$(0.5, 0, 0.5)$	0.105
$(0.5, 0.25, 0.25)$	0.364
$(0.5, 0.5, 0)$	0.5
$(0.75, 0, 0.25)$	0.35
$(0.75, 0.25, 0)$	0.5
$(1, 0, 0)$	0.5

In general, if the $\mathcal{X}_i \times \mathcal{Z}_i \times \mathcal{U}_i$ yield fairly "similar" vectors in the Q_i matrices, then the penetrance estimate will not vary greatly according to the prior distributions placed on the (x, z, u) configurations. Consider a case in which the underlying ascertainment scheme

is believed to be either affecteds only ascertainment or full ascertainment. When the family size is relatively small, affecteds only ascertainment is known to be highly inefficient even in cases in which the underlying pedigree structure is completely known. While the efficiency of the penetrance estimate increases with a larger sample size, the affecteds only ascertainment scheme always yields less efficient penetrance estimates than the full ascertainment scheme. The difference in efficiency of the two ascertainment schemes is more pronounced for small sample sizes. Thus, in the case in which the sample size is small, the choice in prior is more likely to play a significant role in the variance of the penetrance estimate.

In this case, the pedigree structure is small, so information is limited and precision is low, but more information is typically available for larger pedigrees. Figure 3.1 demonstrates how the variance of $\hat{\pi}$ decreases with increasing family size under full ascertainment and affecteds only ascertainment.

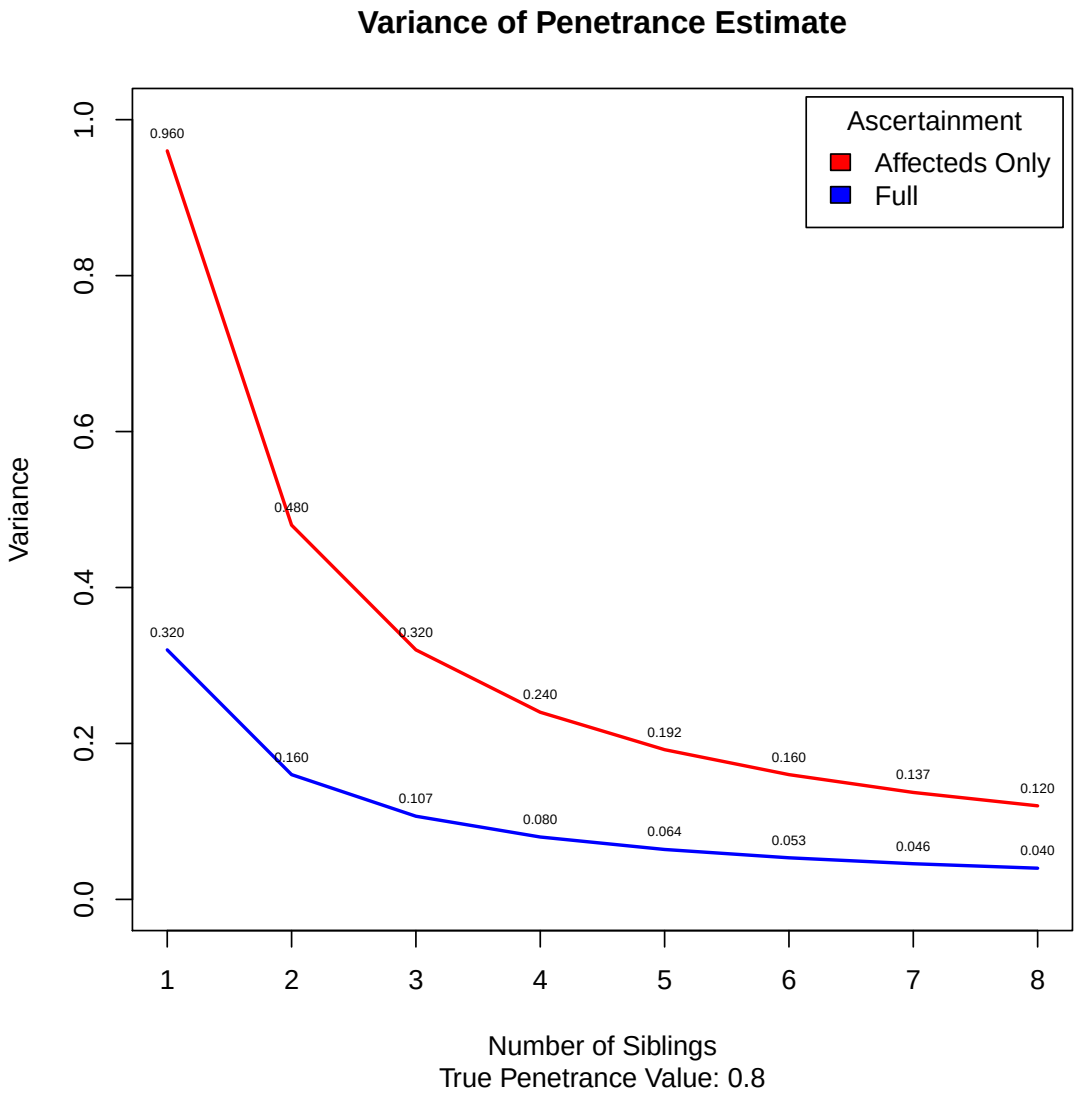


Figure 3.1: Comparison of Variance of $\hat{\pi}$ under Affecteds Only Ascertainment and Full Ascertainment as a Function of Family Size

3.2 Impact of Pedigree Structure on Efficiency of Peneterance Estimate under Full Ascertainment

In this section, five pedigree structures, each with 4 pedigree members, are considered. The difference in the corresponding Q matrices for each structure occurs due to the

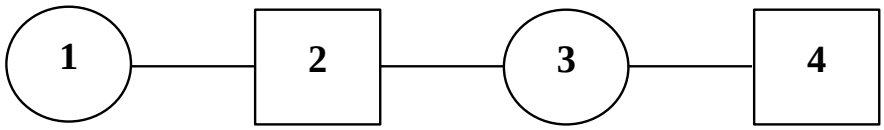
differing number of outcomes that adhere to the assumption that children of parents who are both noncarriers cannot be carriers. A pedigree is not ascertainable if these assumptions are violated. (For example, for Structure 2, any outcome in which Subjects 1 and 2 are unaffected noncarriers and either Subject 3 or Subject 4 is a carrier would not be observable.) The constraints on the possible observable outcomes generated by the pedigree structure result in more probability mass being placed on the entries corresponding to the empty set in the Q matrices.

In these simulation studies, complete ascertainment is assumed, and Subject 1 is assumed to be the proband. Under complete ascertainment and a true penetrance value of 0.80, the approximate variance of the penetrance is 0.107 for all pedigree structures considered in this section. This demonstrates how, for a fixed family size, penetrance value, ascertainment scheme which brings in the entire pedigree subject to preset inclusion criteria, and informative mapping, the variance of the penetrance $\hat{\pi}$ associated with a given pedigree is invariant of the pedigree structure.

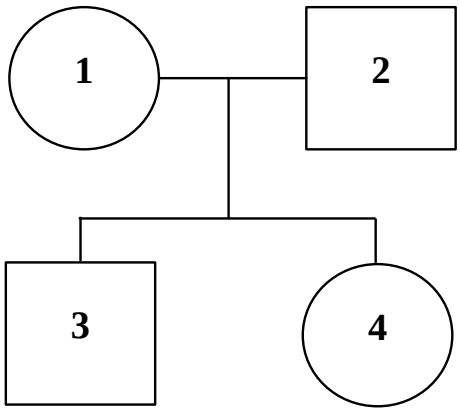
However, the variance of the penetrance estimate under sequential ascertainment schemes which bring in subjects based on phenotype status of subjects already observed is not invariant of the pedigree structure. Consider the case in which Subject 1 is still known to be the proband, and the ascertainment scheme of sequentially bringing in offspring and parents of affected members already ascertained. Under this scheme and knowledge of Subject 1 being the proband, only Subject 1 is observable for Structures 1 and 3. This results in an infinite penetrance estimate variance (indicating no information). For Structure 2, Subjects 3 and 4 are observable, but Subject 2 is not. This results in a penetrance estimate of 0.16. For Structure 4, all subjects are observable by Subject 1 being affected since all of Subject 1's offspring may enter the study under this scheme. This yields a penetrance estimate variance of 0.107. Structure 5 has seven possible outcomes observable. By Subject 1 being the proband, Subject 2 will be observed. If Subject 2 is affected, Subject 3 will be observed, and if Subject 3 is affected, then Subject 4 will be observed. This yields a penetrance estimate of 0.42. Thus, the variance can be

greatly affected by sampling schemes which do not bring in the entire pedigree (subject to preset inclusion criteria).

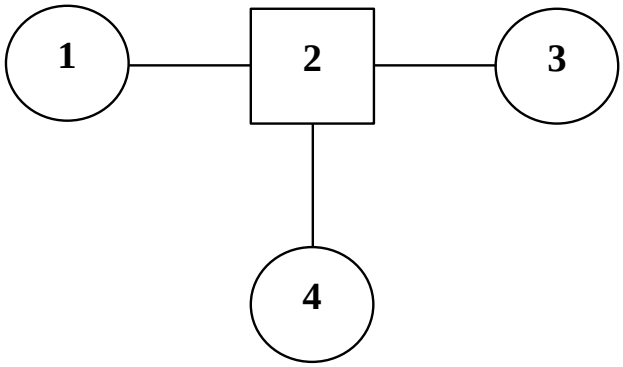
Structure 1:



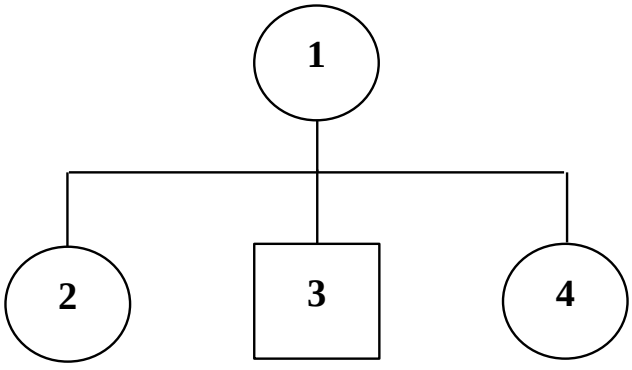
Structure 2:



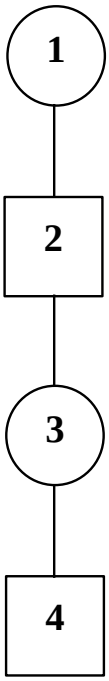
Structure 3:



Structure 4:



Structure 5:



3.3 Application to Pedigree Data

This study features seven pedigrees–labeled as Families A through G. Figures 3.2-3.8 show the observed pedigree, observed phenotype, and observed genotype for each family. Note that only affected status of the subjects is shown. In Family A, Subject 1 is an unaffected carrier, and Subject 2 is an unaffected noncarrier. In Family B, Subject 1 is an unaffected carrier, and Subject 701 is an unaffected noncarrier. In Family C, Subject 3 is an unaffected carrier, and Subjects 2, 4, and 401 are unaffected noncarriers. In Family D, Subjects 1 and 4 are unaffected carriers, and Subject 2 is an unaffected noncarrier. In Family E, Subjects 5 and 6 are unaffected carriers, and Subjects 1 and 4 are unaffected noncarriers. In Family F, Subject 5 is an unaffected carrier, and Subject 2 is an unaffected noncarrier. Lastly, in Family G, Subject 1 is an unaffected carrier, and Subjects 2 and 3 areunaffected noncarriers.

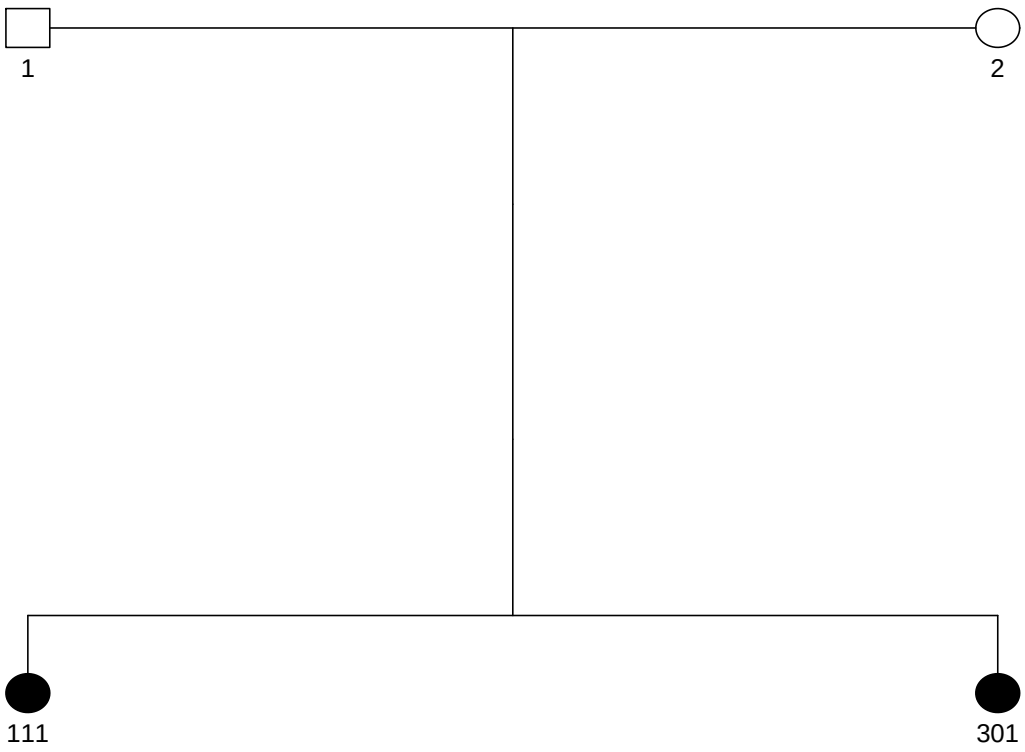


Figure 3.2: Family A Pedigree Structure

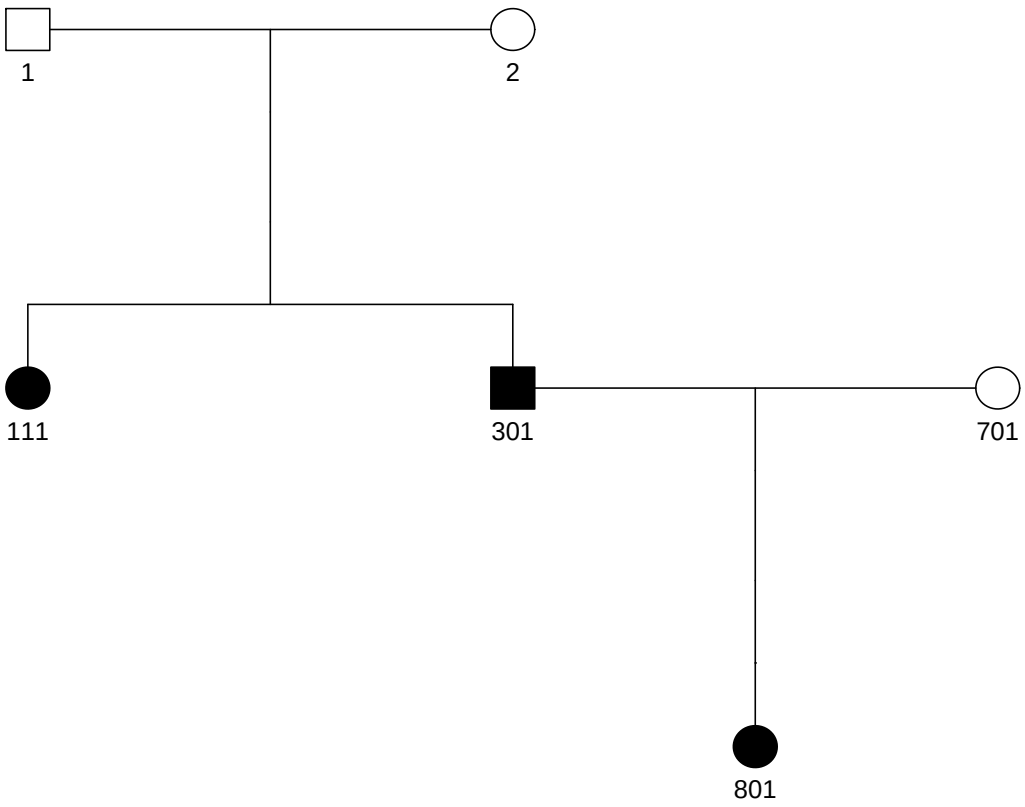


Figure 3.3: Family B Pedigree Structure

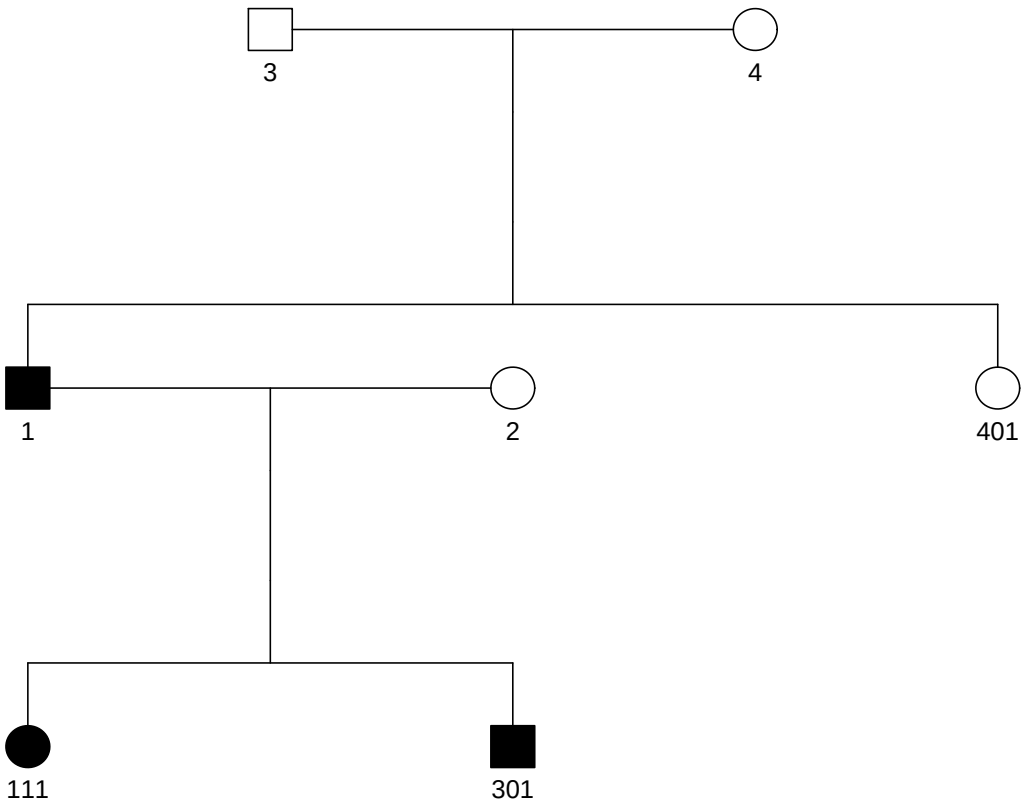


Figure 3.4: Family C Pedigree Structure

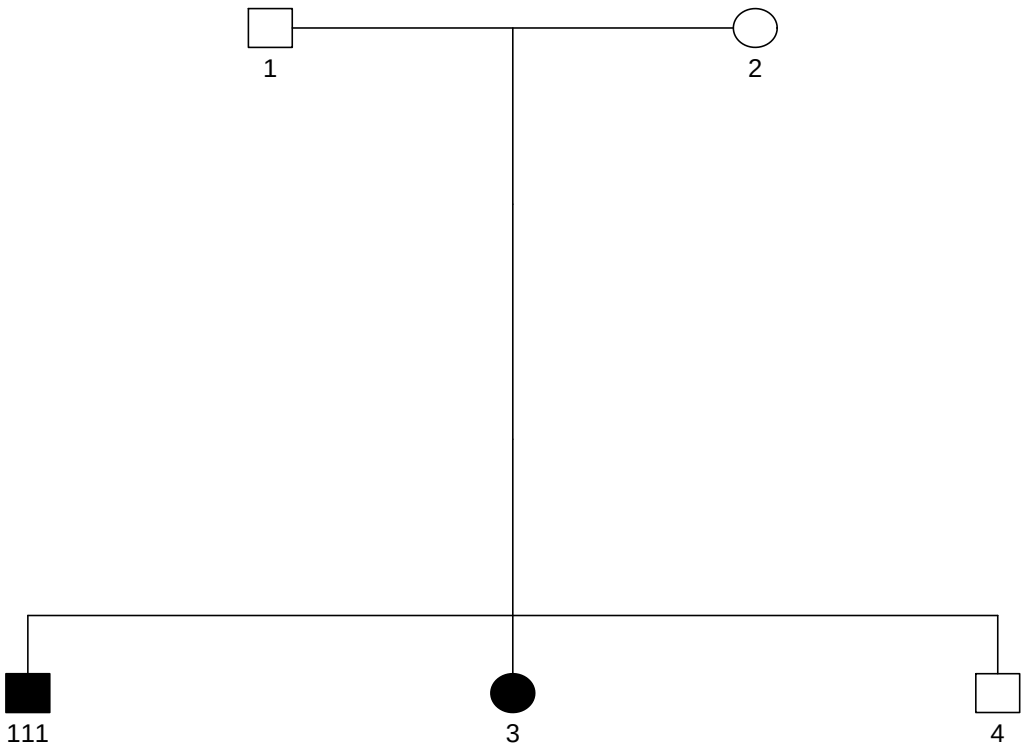


Figure 3.5: Family C Pedigree Structure

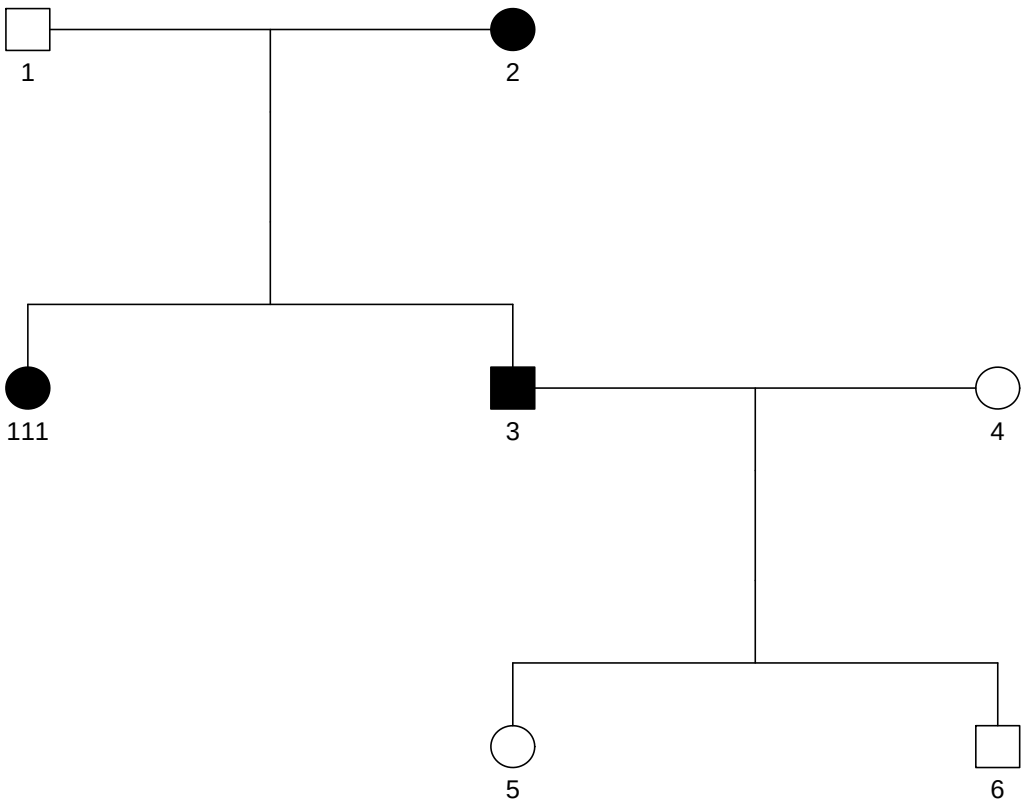


Figure 3.6: Family D Pedigree Structure

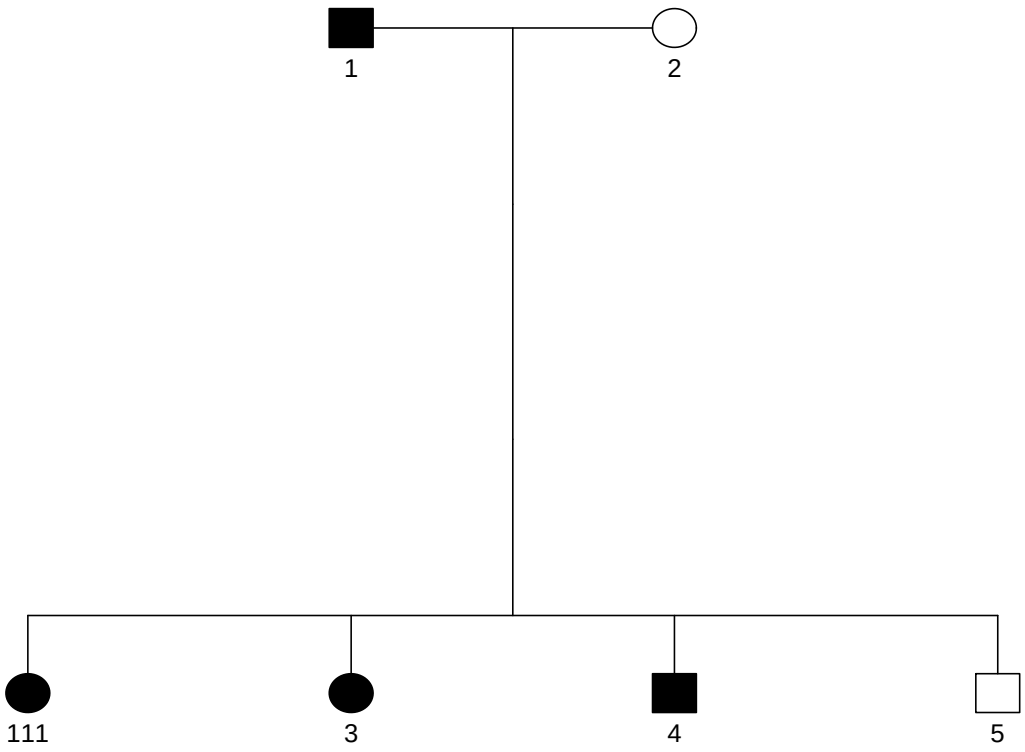


Figure 3.7: Family E Pedigree Structure

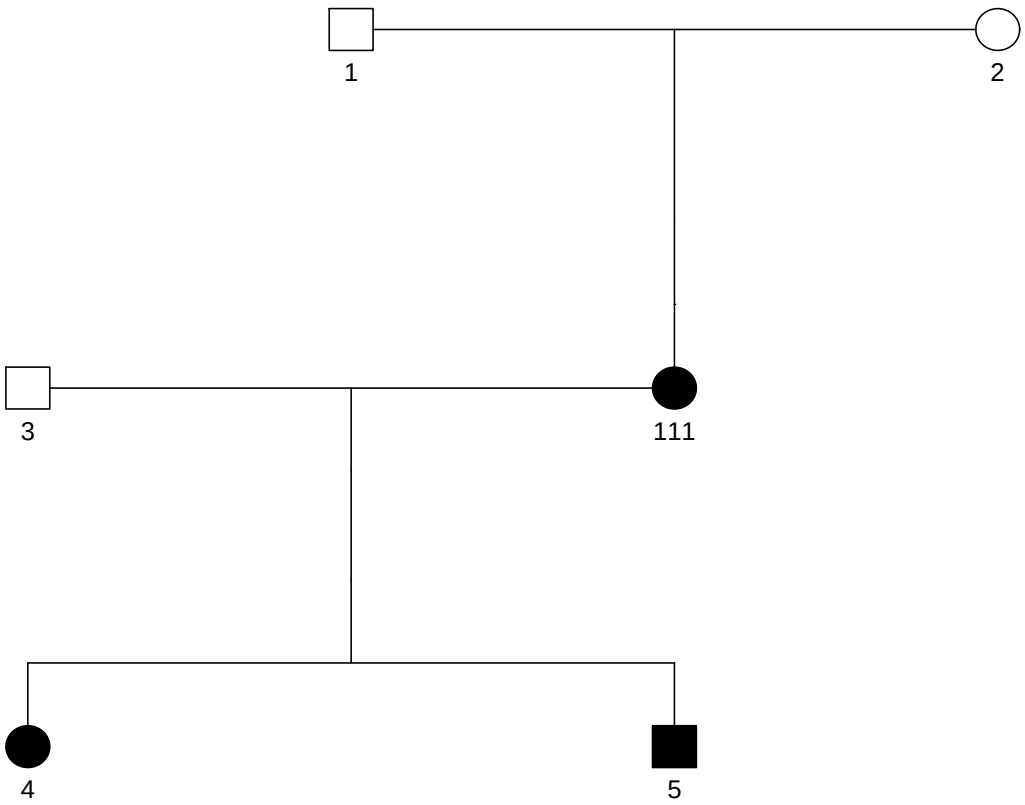


Figure 3.8: Family F Pedigree Structure

In this study, there is sufficient reason to believe that the observed pedigree structures are indeed the complete pedigree structures. These pedigrees were ascertained via one of two possible forms of complete ascertainment. The first plausible ascertainment scheme is complete ascertainment if at least one member of the pedigree is affected. The second plausible ascertainment scheme is that the entire family was eligible to enter the study if at least two siblings in the pedigree are affected. Thus, $\mathcal{X}_i$ is of dimensionality 1, $\mathcal{Z}_i$ is of dimensionality 2, and $\mathcal{U}_i$ is of dimensionality 1, for each $i \in \{1, \ldots, 7\}$.

Table 3.5: Penetrance Estimates over a Range of Prior Probabilities on the Ascertainment Scheme

Prior (≥ 1 Affected, ≥ 2 Affected)	$\hat{\pi}$	Var($\hat{\pi}$)
(1,0)	0.6687	0.011
(0.9,0.1)	0.5949	0.013
(0.8,0.2)	0.5415	0.015
(0.7,0.3)	0.5050	0.016
(0.6,0.4)	0.4801	0.016
(0.5,0.5)	0.4627	0.016
(0.4,0.6)	0.4501	0.016
(0.3,0.7)	0.4405	0.016
(0.2,0.8)	0.4332	0.016
(0.1,0.9)	0.4273	0.016
(0,1)	0.4225	0.016

Table 3.2 details the range of penetrance estimates for assumed priors on the ascertainment scheme and underlying pedigree. If an even prior is assumed on both (x, z, u) schemes, then the method yields a penetrance estimate of approximately 46.3%. The variance of this penetrance estimate is approximately 1.6%. In this case, the estimates are fairly invariant to the prior placed on the ascertainment scheme.

Figure 3.9 displays the test statistic as a function of penetrance. A vertical reference line is drawn at the penetrance estimate, which is the penetrance value that satisfies the equation $\sum_{i=1}^{n} t_i\big((w, v)_i\big) = 0$.

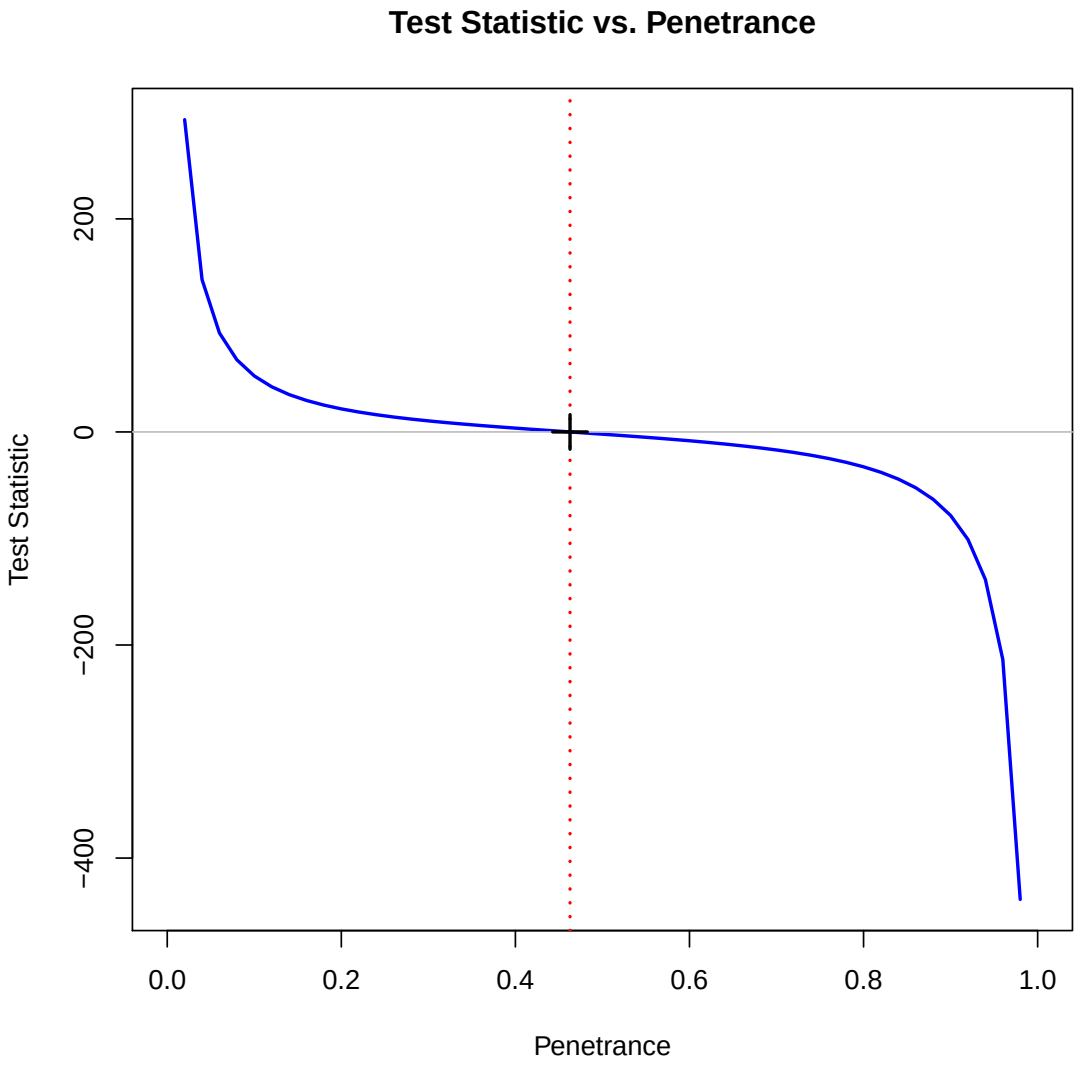


Figure 3.9: Score Test Statistic vs. Penetrance for Familial Data

Chapter 4

Conclusion

Unless the underlying pedigree structure is known for each family brought into a penetrance study, the problem of constructing a likelihood for the observed genotypic and phenotypic information becomes difficult, but not intractable. This paper details how one may designate a set of potential underlying pedigree structures and incorporate knowledge of the proband class, the familial sampling frame, and other factors which can determine what portion of the complete data for each observed family in order to generate efficient and unbiased penetrance estimates.

The method presented in this paper has clear connections to score tests. The function t may be seen as a score-like function which reflects a mixture distribution over adequately conditioned likelihoods for each family (under proper specification of the prior distributions on the pedigree structure and ascertainment scheme). The resulting solution for $t(W, V)$ ultimately leads to an efficient and unbiased score equation for the true penetrance value for the mixture likelihood.

In accordance to Mendelian laws, offsprings' genotypic and phenotypic configurations are dependent upon their parents' genotypes and phenotypes. If both parents are unaffected noncarriers, then all offspring of the parents must, too, be unaffected noncarriers. Configurations which violate this requirement have zero probability of occurring. This results in a greater amount of probability mass being placed on Q matrix elements

corresponding to the empty set. For a fixed family size, penetrance value, ascertainment scheme which brings the entire family into the study subject to the inclusion criteria, and informative mapping, the variance of the penetrance estimate is invariant of the structure of the pedigree. However, if sequential sampling is implemented, the variance may be significantly affected by the types of ascertainment schemes in the $\mathcal{Z}_i$.

In general, very little information is available from small to moderately-sized pedigrees. As shown in Figure 3.1, under full ascertainment, there is still little information available, especially in cases in which the family size is less than 3. The variance may also be sensitive to the prior placed on the underlying pedigree structure and ascertainment scheme to the extent that the mappings are different. If the (x, z, u) configurations yield fairly similar vectors in the corresponding Q matrix, then the variance of $\hat{\pi}$ will not significantly fluctuate according to the prior. The advantage to using the method presented in the paper is that one may still derive unbiased penetrance estimates despite relative uncertainty of the underlying complete pedigree structure, complete phenotypic and genotypic information, and mapping scheme. However, under sequential ascertainment schemes, particularly PD schemes, the variance of the penetrance can significantly be affected by the pedigree structure.

It should be noted that this method is not resistant to cases in which two portions of the same pedigree are brought into the study without the researchers' knowledge. This could potentially lead to a loss in efficiency, as there are cases in which a single family with many members can provide more information than a two or more families with few members per pedigree. (This assertion largely depends on the structure of the large single-family pedigree structure.)

Further extensions of this method would require finding a solution to making this method less computationally intensive. Additionally, in cases in which the study contains large pedigrees (i.e. pedigrees with 10 or more members in the underlying complete pedigrees), the dimensionality of the Q matrix can become too large for many software packages to

handle the inversions required to find the proper weights associated with each family's possible pedigree and ascertainment scheme configuration.

Bibliography

[1] Carayol, Jerome and Catherine Bonaiti-Pellie. "Estimating Penetrance From Family Data Using a Retrospective Likelihood When Ascertainment Depends on Genotype and Age of Onset." *Genetic Epidemiology*. 27 (2004): 109-117.

[2] Cannings, C. and E.A. Thompson. "Ascertainment in the sequential sampling of pedigrees." *Clinical Genetics*. 12 (1977): 208-212.

[3] Bonaiti, Bernard, Valerie Bonadona, Herve Perdry, et al. "Estimating penetrance from multiple case families with predisposing mutations: extension of the 'genotype-restricted likelihood' (GRL) method." *European Journal of Human Genetics*. (2010): 1-7.

[4] Ottman, Ruth, Karina Berenson, and Christie Barker-Cummings. Recruitment of Families in Genetic Studies of Epilepsy. *Epilepsia*. (2005).

[5] Hodge, Susan E. "Conditioning on Subsets of the Data: Applications to Ascertainment and Other Genetic Problems". *American Journal of Human Genetics*. 43 (1988): 364-373.

[6] Elston, Robert C. "INVITED EDITORIAL: 'Twixt Cup and Lip: How Intractable Is the Ascertainment Problem?". *American Journal of Human Genetics*. 56 (1995): 15-17.

[7] Elston, R.C. and E. Sobel. "Sampling Considerations in the Gathering and Analysis of Pedigree Data". *American Journal of Human Genetics*. 31 (1979): 62-69.

[8] Shute, Nereda C.E. and W.J. Ewens. "A resolution of the ascertainment sampling problem. I. Theory.". *Theoretical Population Biology*. 30(3) (1986): 388-412.

[9] Shute, Nereda C.E. and W.J. Ewens. "A Resolution of the Ascertainment Sampling Problem. II. Generalizations and Numerical Results". *American Journal on Human Genetics*. 43 (1988): 374-386.

[10] Shute, Nereda C.E. and W.J. Ewens. "A Resolution of the Ascertainment Sampling Problem. III. Pedigees". *American Journal on Human Genetics*. 43 (1988): 387-395.

[11] Vieland, Veronica J. and Susan E. Hodge. "Inherent Intractability of the Ascertainment Problem for Pedigree Data: A General Likelihood Framework". *American Journal of Human Genetics*. 56 (1995): 33-43.

[12] Robert C. Elston et al (eds.). *Statistical Human Genetics: Methods and Protocols,* Methods in Molecular Biology, vol. 850. Springer Science+Business Media, LLC 2012.

[13] Ottman, Ruth and Mervyn Susser. "Data collection strategies in genetic epidemiology: The epilepsy family study of Columbia University". *Journal of Clinical Epidemiology* 45(7) (1992) 721-727.

[14] Ottman, Ruth, Berenson, Karina, and Christie Barker-Cummings. "Recruitment of Families for Genetic Studies of Epilepsy". *Epilepsia* 46(2) (2005) 290-297.

[15] Ewens, Warren and Robert C. Elston. "Assumptions for Different Ascertainment Models in Human Genetics." *Statistical Human Genetics: Methods and Protocols*. Ed. Robert C. Elston et al. (2012) 187-209.

[16] Stene, Jon. "Sampling Considerations in the Gathering and Analysis of Pedigree Data". *Biometrics*. 33(3) (1977): 523-527.

Appendix A

APPENDIX

A.1 Brief Discussion of Theory of Estimating Equations

Let $W_i = Z_i(X_i, G_i, Y_i)$ and $V_i = U_i(Z_i, X_i, G_i, Y_i)$ be the observed information for family i. Furthermore, let $P_x\{z(x, G_i, Y_i) \in w, u(z, x, G_i, Y_i) \in v\}$ be the probability that the observed data would be (w, v) for complete pedigree structure x, ascertainment scheme z, and informative mapping u. Then the score, $S_i(\pi)$, be the derivative with respect to π (the parameter of interest) of the natural log of $P_x\{z(x, G_i, Y_i) \in w, u(z, x, G_i, Y_i) \in v\} \times \gamma_{x,z,u}$, where $\gamma_{x,z,u}$ is the prior probability on (x, z, u). Under the assumption of independence, the score for families $1, \ldots, n$ is simply the sum of the scores: $\sum_{i=1}^{n} S_i(\pi)$. Denote the maximum likelihood estimate of π as $\hat{\pi}$ and the "true" penetrance value as π_o.

A Taylor expansion of $S_i(\hat{\pi})$ yields the following:

$$S_i(\hat{\pi}) = S_i(\pi_o) + \nabla_\pi S_i(\pi_o) \cdot (\hat{\pi} - \pi_o) + \frac{\nabla_\pi^2 S_i(\pi)}{2!} \cdot (\hat{\pi} - \pi)^2 + \ldots$$

This may be rearranged as follows:

$$\nabla_\pi S_i(\pi_o) = \frac{S_i(\hat{\pi}) - S_i(\pi_o)}{\hat{\pi} - \pi_o} - \frac{1}{\hat{\pi} - \pi_o} \sum_{k=2}^{\infty} \frac{\nabla_\pi^{(k)} S_i(\pi_o)}{k!} (\hat{\pi} - \pi_o)^k$$

Assuming that the higher degree terms are neglegible and that slope of the score behaves as a constant function of π and as a random variable near π_o, the slope of the $\sum_{i=1}^{n} S_i(\pi)$ versus π curve may be approximated by

$$\mathbb{E}\left\{\frac{\partial}{\partial\pi}\sum_{i=1}^{n} S_i(\pi_o)\right\}$$

This leads to the following approximation for $\hat{\pi}$:

$$\hat{\pi} \approx \frac{\sum_{i=1}^{n} S_i(\hat{\pi}) - \sum_{i=1}^{n} S_i(\pi_o)}{\mathbb{E}\left\{\frac{\partial}{\partial\pi}\sum_{i=1}^{n} S_i(\pi_o)\right\}} + \pi_o$$

Thus, the variance of $\hat{\pi}$ is approximately

$$\frac{\mathbb{V}\left\{\sum_{i=1}^{n} S_i(\hat{\pi})\right\}}{\left[\mathbb{E}\left\{\frac{\partial}{\partial\pi}\sum_{i=1}^{n} S_i(\pi_o)\right\}\right]^2}$$

A.2 Efficiency of the Penetrance Estimate

Consider the problem of testing $\pi = \pi_o$ versus the alternative $\pi \neq \pi_o$. Upon proper specification of the likelihood and regularity conditions, an efficient score test may be used. Let (W_i, V_i) be the observed data for the i^{th} family in the study. The likelihood $f(w_i, v_i)$ of observing (w_i, v_i) for the i^{th} family may also be further parameterized by parameters other than the parameter of interest π. Let λ denote this set of nuisance parameters. Under this setup, the full likelihood for families $1, \ldots, n$ may be expressed as

$$f_{\pi,\lambda}((w,v)_1, \ldots, (w,v)_n) = \prod_{i=1}^{n} f_{\pi,\lambda}((w,v)_i).$$

Let $S_i(\pi)$ be the derivative with respect to π of log-likelihood of the i^{th} family. This is also known as the score for π for the i^{th} family. The total score for π is $S(\pi) = \sum_{i=1}^{n} S_i(\pi)$.

Let

1. $I_{\pi\pi} = -\mathbb{E}\frac{\partial^2}{\partial\pi^2}\log f_{\pi,\lambda}(w_1,\ldots,w_n)$,
2. $I_{\pi\lambda} = I_{\lambda,\pi} = -\mathbb{E}\frac{\partial^2}{\partial\pi\partial\lambda}\log f_{\pi,\lambda}(w_1,\ldots,w_n)$, and
3. $I_{\lambda\lambda} = -\mathbb{E}\frac{\partial^2}{\partial\lambda^2}\log f_{\pi,\lambda}(w_1,\ldots,w_n)$

Under the null hypothesis $\pi = \pi_o$, the expectation of the total score evaluated at π_o is 0, and its variance is the adjusted Fisher information,

$$I(\pi_o) = I_{\pi\pi} - I_{\pi\lambda}I_{\lambda\lambda}^{-1}I_{\lambda\pi}\Big|_{\pi=\pi_o}$$

The above variance is known as the "Cramer-Rao" lower bound and is the minimum variance attainable among all unbiased estimators of the parameter of interest, π.

Under the regularity conditions, the following hold and are equivalent:

1. $\frac{S(\pi_o)}{\sqrt{I(\pi_o)}} \sim N(0,1)$
2. $\frac{S^2(\pi_o)}{I(\pi_o)} \sim \chi_1^2$

The above expressions create a framework by which efficient estimating equations (and subsequent hypothesis tests and confidence intervals) may be constructed. However, care must be exercised in expressing the likelihood so as to reflect the pertinent information on how the family was brought into the study. The score expression will only be unbiased for zero if the likelihood expression sufficiently accounts for the ascertainment process. If a score expression is unbiased for zero, then the $\hat{\pi}$ which solves the score equation is unbiased for π. A likelihood expression cannot be constructed without knowledge of the underlying pedigree structure and/or ascertainment scheme. However, if a range of

plausible underlying complete pedigree structures and ascertainment schemes can be defined, efficient and unbiased score equations may be calculated (which can subsequently yield to unbiased and efficient penetrance estimates).