Copyright © 2002 by Xi Zhou
All rights reserved

CLASSIFICATION OF MISSENSE MUTATIONS OF DISEASE GENES

by

Xi Zhou

Institute of Statistics and Desicion Science
Duke University

Date: _____

Approved:

_____
Edwin S. Iversen, Jr., Supervisor

_____
Giovanni Parmigiani, Supervisor

_____
Mike West

_____
Joellen M. Schildkraut

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Institute of Statistics and Desicion Science
in the Graduate School of
Duke University

2002

ABSTRACT

(Statistics)

CLASSIFICATION OF MISSENSE MUTATIONS OF DISEASE GENES

by

Xi Zhou

Institute of Statistics and Desicion Science
Duke University

Date: _____

Approved:

_____
Edwin S. Iversen, Jr., Supervisor

_____
Giovanni Parmigiani, Supervisor

_____
Mike West

_____
Joellen M. Schildkraut

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor of Philosophy in the
Institute of Statistics and Desicion Science in the Graduate School of
Duke University

2002

Abstract

Missense mutations of disease genes pose a challenging classification problem because of the uncertainty associated with their implications to the risk of disease. Assessing the risk implications is often complicated by small sample size and lack of an appropriate functional assay. For large genes such as BRCA1 and BRCA2, it is common to infer risk implications from pedigree data. It is typical to have a relatively small sample size for each mutation, and to only have pedigrees of individuals who are selected because of a high disease rate in the family. This selection mechanism is likely to overstate the mutation's contribution to risk of disease.

In this study, we develop a Bayesian hierarchical methodology which classifies missense mutations as deleterious or non-deleterious based on mutation specific penetrances estimated from pedigree data. We consider multiple competing genes and multiple phenotypes (e.g. cancer sites). The basis of our approach is to model the age–dependent mutation–specific penetrance functions by a hazard mixture of the phenocopy rate and penetrance of deleterious mutations. This permits us to take the age effect into account while accommodating limited sample size. We assume penetrances of known deleterious mutations and phenocopy rates are estimated in previous studies.

Using this mixture model for penetrance as the basis, we develop a Bayesian hierarchical approach to classify missense mutations. The mixture parameter above is a composite of the deleteriousness of the mutation and the selection bias. We compare these parameters to similarly estimated penetrances of known deleterious mutations and common polymorphisms. This allows us to separate the deleteriousness component from the bias component, and thus to reduce the effect of the selection bias inherited from the data collection mechanism, since pedigrees identified through

probands that are either negative (mostly with common polymorphisms) or positive (with known deleterious mutations or missense mutations) are collected based on the same sampling scheme. The model also takes into account the imperfect sensitivity of genotyping. Model parameters are estimated by using Markov Chain Monte Carlo methods. This approach is applied to the study of a sample of BRCA1 and BRCA2 missense mutations, using data collected at the Duke University Medical Center.

Contents

Abstract **iv**

List of Tables **viii**

List of Figures **ix**

Acknowledgements **xiii**

1 Introduction **1**

1.1 Probing the effects of a mutation through molecular methods 3

1.2 Characterizing the mutation through statistical methods 3

1.2.1 Likelihood calculation and ascertainment correction 5

1.2.2 Penetrance function 9

1.3 Characterizing missense mutations 9

1.4 Plan for the thesis 11

2 Data **12**

2.1 BRCA1 and BRCA2 12

2.2 The Duke missense mutation data set 15

3 Preliminary Model **22**

3.1 Model Specification 22

3.1.1 Conditional Likelihood 23

3.1.2 Penetrance Models 24

3.2 Model Application 26

3.2.1 Discrete Model 27

3.2.2 Continuous Model 29

3.3 Discussion 31

4 Hierachical classification model **35**

5 Simulation based validation **44**

5.1 Family history simulation 44

5.2 Likelihood calculation for simulated data 45

5.3 Validating the hierarchical classification model 47

6 Classifying missense mutations of BRCA genes **71**

6.1 Penetrance models and the likelihood calculation 72

6.2 Hierarchical Classification Results 74

7 Discussion **90**

A Full conditionals for Gibbs sampling **93**

Bibliography **95**

Biography **103**

List of Tables

2.1 Distribution of breast and ovarian cancer in probands categorized by their genetic test results. 16

2.2 Distribution of age of diagnosis of breast cancer in probands categorized by their genetic test results, where "positive" means positive with either missense mutations or known deleterious mutations. 17

2.3 Distribution of age of diagnosis of breast cancer in family members categorized by probands' genetic test results. 17

2.4 Number of families categorized by the number of breast cancer cases in the family and proband's genetic test result. 18

2.5 Number of families categorized by each increased number of breast cancer cases in the family and proband's genetic test results. 19

2.6 Total number of cases over total number of individuals among relatives in each age of diagnosis category categorized by probands' genetic test results. 19

2.7 Total number of cases over total number of individuals among relatives in each age of diagnosis category categorized by probands' genetic test results. 21

5.1 Proportion of incorrect classifications based on calculations for replicates of simulated data. 61

List of Figures

1.1 A family history of breast and ovarian cancers identified through a proband (indicated by the arrow). Label BC38 indicates that member 1 (the proband) was diagnosed with breast cancer at age 38, her current age or age at death (indicated by a diagonal slash) is 49; label OC57 indicates that one of her paternal aunt was diagnosed with ovarian cancer at age 57. Member 3 is free of both cancers. 6

3.1 Rescaled conditional likelihood plot. 28

3.2 Contour plots for joint marginal posterior based on different prior choices. 33

3.3 Shrinkage plot for models with different priors. x-axis is the mutation specific maximum likelihood estimate of the penetrance parameter γ_m. y axis is the posterior mean of γ_m. 34

4.1 Graphical illustration of the classification model. In this depiction, The K family histories are denoted by $f_i, i = 1, \ldots, K$. Each is identified through one proband who is tested for mutations at the site(s) of interest. Each proband's test results – negative, positive with a missense mutation – are indicated by circles, round-cornered rectangles and rectangles, respectively, enclosing the family history variable. Mutation-specific penetrances are plotted in the second row and are denoted by γ_{nn}, where the index nn identifies the mutation the proband carries. Values of nn ranging from m_1 through m_M correspond to the M unique missense mutations in the sample, those ranging from d_1 to d_N correspond to the N unique identified deleterious mutations, the index nd corresponds to the wild type genotype(s) and the index d corresponds to a common, unobserved deleterious genotype. Pedigrees of probands with mutation nn are connected to the appropriate mutation-specific penetrance parameter γ_{nn} by arrows. Dashed arrows indicate uncertainty about a proband's genotype. Mutations and their phenotypic effects, depicted at the bottom of the plot, are also connected by arrows. Uncertainty in the effects of the various missense mutations is denoted by dashed arrows. 39

4.2 Conditional structure of the hierarchical classification model 42

5.1 Plots of conditional likelihood of family histories given mutation specific penetrance versus mutation specific penetrance. Each colored curve represents the conditional likelihood for certain variant. And the conditional likelihoods are rescaled so that the x-axis and y-axis of each small box range from 0 to 1. Likelihoods for the wild type, deleterious, non-deleterious missense mutations and deleterious mutations are denoted by colors: light blue, blue, red, and orange, respectively. The grey vertical line in each small box shows the value of the penetrance of the particular mutation that was used to simulate the data. 49

5.2 Rescaled conditional likelihood of family histories given mutation specific penetrance vs. mutation specific penetrance ranging from 0 to 1. Each colored curve represents the conditional likelihood for certain variant. The total number of simulated family histories is 4000. Likelihoods for the wild type, deleterious, non-deleterious missense mutations and deleterious mutations are denoted by colors: light blue, blue, red, and orange, respectively. 53

5.3 Summary of interesting model parameters. x-axis represents the mutation index. Mutations 1-15 are non-deleterious missense mutations. Mutations 16-20 are deleterious missense mutations. And mutations 21-30 are known deleterious mutations. y-axis represents the value of the parameters. In each plot, dark lines represent mutation specific penetrances used to simulate the family histories, blue lines represents the maximum likelihood estimate of the mutation specific penetrance, red lines represent posterior mean of mutation specific penetrance estimated from the hierarchical classification model and green lines represent posterior probability that the mutation is deleterious. 58

5.4 Boxplot of the probability of deleteriousness of the missense mutations estimated from the replicates. Mutations denoted by "red" are non-deleterious missense mutations and mutations in "blue" are deleterious missense mutations". Plots (a)-(d) are results based on sample size 4000, (e)-(h) are based on sample of size 6000. Plots (a) and (e) are from based on population based data. Plots (b) and (f) are based on data ascertained through the progression ascertainment rule and the likelihoods are not corrected for ascertainment bias. Plots (c), (g), (d) and (f) are based on data ascertained through the affected probands ascertainment rule. For (c) and (g), the likelihoods are not corrected for ascertainment bias, while for (d) and (h), the likelihoods are corrected for ascertainment bias. 62

5.5 Summary of interesting model parameters for replicates of family histories. In each plot, x-axis represents the mutation index, y-axis represents the value of these parameters. In each plot, dark lines represent mutation specific penetrances used to simulate the family histories, blue lines represents the maximum likelihood estimate of the mutation specific penetrance, red lines represent posterior mean of mutation specific penetrance estimated from the hierarchical classification model and green lines represent posterior probability that the mutation is deleterious. 63

6.1 Penetrance curves of BRCA1 carriers under three penetrance models. Top plots are for penetrances of breast cancer. And bottom plots are for penetrances of ovarian cancer. 75

6.2 Penetrance curves of BRCA2 carriers under three penetrance models. Top plots are for penetrances of breast cancer. And bottom plots are for penetrances of ovarian cancer. 78

6.3 Histograms of the posterior samples of model parameters. Likelihoods calculation is based on penetrance model $\rho_m(a) = 1-(1-\rho(a))^{\frac{\gamma_m}{(1-\gamma_m)}}$. Prior distribution of model parameters are plotted in solid lines. . . . 82

6.4 Histograms of the posterior samples of model parameters. Likelihoods calculation is based on penetrance model $\rho_m(a) = (1-\gamma_m)\phi(a) + \gamma_m\rho^*(a)$. Prior distribution of model parameters are plotted in solid lines. 83

6.5 Histograms of the posterior samples of γ_{nd} and γ_d. Posterior means of these penetrance parameters are plotted in solid lines. 84

6.6 Posterior mean and 90% interval of γ_m's. Posterior probability of deleteriousness of missense mutations are denoted by "x". Estimates of missense mutations are in "green". Estimates of known deleterious mutations are in "red" color. And estimates for the two assumed polymorphisms from the negatives are in "blue". Penetrance model used here is $\rho_m(a) = 1-(1-\rho(a))^{\frac{\gamma_m}{(1-\gamma_m)}}$. 87

6.7 Posterior mean and 90% interval of γ_m's. Posterior probability of deleteriousness of missense mutations are denoted by "x". Estimates of missense mutations are in "green". Estimates of known deleterious mutations are in "red". And estimates for the two assumed polymorphisms from the negatives are in "blue". Penetrance model used here is $\rho_m(a) = (1 - \gamma_m)\phi(a) + \gamma_m \rho^*(a)$. 88

6.8 Posterior probability of deleteriousness of missense mutations. Missense mutations on BRCA1 gene are denoted in "light blue" and missense mutations on BRCA2 gene are denoted in "pink". Four mutations occur at the intron region of the gene are denoted by "*n". Actual mutations for each of them are as follows: *1, (5272+87)delT; *2, (5392+60)ins12; *3, G(7663+10)A; *4, A(4795-170)G. 89

Acknowledgements

Pursuing the PH. D. degree is a challenging journey, especially when a considerable part of it is spent thousands miles away from Duke. I am fortunate to have advisers Professor Giovanni Parmigiani and Professor Ed Iversen, who led me into the wonderful field of statistical genetics, provide me important guidance, uncanny insight and constant encouragement, even through weekly phone calls when I'm away from Duke. This thesis would not have been completed without their tremendous support. To them, I express my deepest gratitude.

I thank everyone in ISDS who have made my life there a most enjoyable one. Especially, I would like to thank Professor Mike West for his encouragement, help and allowing me to work remotely.

Special thanks goes to Professor Brani Vidakovic, who led me into ISDS through his encouragement and help and provided invaluable guidance and support during my first two years there.

I would also like to thank the Duke Cancer Center for providing the data. I thank Professor Joellen Schildkraut and Ms. Shelly Clark for useful discussions.

Finally, I extend my thanks to my husband, Yingkai Zhang, for his love, understanding, support and invaluable help; to my parents, for their love and encouragement; to my daughter, Minmin, for the happiness she brings to my life. I dedicate this thesis to them.

Chapter 1

Introduction

It has been known that the ultimate cause of many diseases, including cancer, is mutations on disease susceptibility genes. Recent rapid progress in molecular and computational technology has being accelerating the discoveries in every aspect of the genetics of human diseases. Advances in computational methods and technology improve our ability to identify disease susceptibility genes. Loci on the human genome that are associated with more than 20 different hereditary diseases have been identified (Fearon 1997), including some of major public health interest, such as the breast cancer susceptibility genes BRCA1 and BRCA2 (Miki et al. 1994; Wooster et al. 1995), colon cancer susceptibility gene APC (Syngal et al. 2000), and more recently, the schizophrenia susceptibility gene Neuregulin 1 (Stefansson et al. 2002). Advances in molecular technology have made isolation and cloning of increasing number of disease susceptibility genes available. Genetic tests have been available for members of high risk families of several inherited cancers, for example, retinoblastoma, polyposis coli, multiple endocrine neoplasia, and inherited breast and ovarian cancers, as a method for disease prevention and early diagnosis (Ponder 1997). While controversial, such tests are becoming increasingly common (Hoskins et al. 1995; Yan et al. 2000) with decreasing costs. However, how to accurately interpret the genetic tests

results remains a problem. Very often, a family history indicates increased risk, and a well known disease genes presents a mutation, but its significance is unknown. Decisions made on the basis of genetic tests are usually life-changing, and often involve radical preventive surgery, for example, mastectomy or oophorectomy for carriers of BRCA1 or BRCA2 mutations (Grann et al. 1998).

Genetic tests look for inherited mutations of disease susceptibility genes and classify them based on whether they are phenotype modifying. Types of mutations which significantly alter the gene sequence and lead to premature truncation of the gene product are phenotype-modifying. These mutations, such as frame-shift deletions, insertions, or nonsense mutations, are considered deleterious. There are also mutations, such as missense mutations, which only result in the substitution of one amino acid for another in the protein sequence and may or may not substantially changing the structure and functionality of the protein. These mutations are classified as of "unknown significance". As a whole, this type of mutation is relatively common. Their large number and the ambiguity of their effects on the protein product pose a major challenge for risk assessment (Cotton and Scriver 1998). Ignorance of the risk associated with this kind of mutations is not only restricted to understanding the molecular pathogenicity of these genes, but also to understanding an individual's disease risk, impeding the screening and clinical management of mutation carriers.

Currently, approaches to studying the effects of mutations usually proceed in two directions. The first direction is to use the molecular biological methods of a functional assay to probe the function of the mutation. The second is to use statistical methods to study the mutations' phenotype-modifying characteristics, mainly a carrier's risk to disease through population or family history studies. In this dissertation, statistical methods to characterize missense mutations through family history studies is will be studied in detail.

1.1 Probing the effects of a mutation through molecular methods

A biomedical approach is believed to be the ultimate test to decide whether a mutation is phenotype modifying (Cotton and Scriver 1998). This usually involves cloning the gene and sequencing and synthesizing probes for each mutation, then testing the functionality of the gene product through a functional assay. The advantage of this approach is that it may develop conclusive results based on a single patient sample. However, it requires that we are able to build an appropriate functional assay. Proteins encoded by these disease susceptibility genes may have a variety of functions in the cell. They may act as transmembrane receptors, cytoplasmic regulatory or structural proteins, transcription factors or regulators of transcription, cell cycle factors or DNA damage repair pathway proteins (Fearon 1997). To test the mutation effect on such large array of possibilities is likely to be costly and inefficient. This is especially true for large genes such as BRCA1 and BRCA2. To date, the exact function of BRCA1 and BRCA2 are still unclear (Welcsh et al. 2000). Although clones of these genes have been available for a while, a comprehensive functional study on the whole gene level remains a daunting task. Despite this, notable progress on parts of the protein have been made (Venkitaraman 2000; Hayes et al. 2000; Vallon-Christersson et al. 2001).

1.2 Characterizing the mutation through statistical methods

Most mutations of cancer genes do not determine the fate of the individual, although they substantially increase cancer risk. Mutation characterization through statistical methods therefore involves estimating the properties of a given mutation as it relates

to certain diseases, such as phenotype-genotype correlation and the mutation's population frequency. Typically, phenotype-genotype correlation is measured through penetrance functions, the cumulative probability distributions of developing cancer by age, conditional on a specific genetic variant. Otherwise, an odds ratio of disease between carriers and non-carriers can also provide a good assessment of a mutation's effect. Compared to the relatively expensive biomedical approach, statistical approaches are much cheaper as they utilize currently available epidemiological data. These data can be a sample of family histories, pairs of relatives or even unrelated individuals. They are usually collected in high risk clinics and are readily available for analysis.

Mutation characterization is a special kind of gene-characterization study. Instead of focusing on the deleterious mutations of the gene as a whole, it focuses on studying the effects of individual mutations, especially missense mutations because of the ambiguity of their effects on the protein product. Unlike deleterious mutations which result in the total lost of functionality of the protein product, each missense mutation may have different effect on the protein product and can not be treated the same. On the other hand, missense mutations are relatively common. On a disease gene, a significant proportion of the mutations are missense. This large number indicates that common methods for gene-characterization, though useful, are likely to be less efficient.

In the last decade, statistical methods have been successfully used to characterize several important disease genes, such as the BRCA1 and BRCA2 genes. Gene-characterization and other epidemiological studies which measure the association of disease with another covariate have much in common while the former focuses on heredity. Traditional epidemiological methods such as case-control or cohort studies usually serve as the foundation for risk assessment. Hereditary components in gene

characterization studies are usually built into a genetic model and examined through a case-control or cohort based analysis. To get accurate estimates, gene characterization usually needs to use efficient data, such as family histories, account for bias caused by sampling methods, such as different ascertainment criteria, and use an appropriate genetic model. Usually, these studies focused only on the disease gene and several deleterious mutations which occur relatively frequently in certain genetic groups (Struewing et al. 1997; Oddoux et al. 1996). A good example is characterization of BRCA1 gene. Before genetic tests were available, linkage and segregation analysis have been used (Claus et al. 1991; Easton et al. 1993; Easton et al. 1995) to characterize BRCA1. After genetic tests available, designs focused on using actual observed genotypes were proposed (Struewing et al. 1997; Gail et al. 1999).

Methods that directly characterize disease gene through penetrance estimation involves several important components, such as the specification of a genetic model, a penetrance function and ascertainment correction. The genetic model specifies how the disease gene segregates in the family. It usually includes the mode of inheritance. The penetrance function specifies a model for penetrance estimation as many diseases are age related. With these a likelihood of a given data set can be written. Often, data sets for these studies are collected by sampling heavily affected families. Thus, a likelihood based on this type of data needs ascertainment correction. Details of the likelihood calculation, ascertainment correction and penetrance function are described in the following sections.

1.2.1 Likelihood calculation and ascertainment correction

Family history is the most informative type of data for genetic analysis. It comprises disease status, age information (age of diagnosis for affected individuals, and current age or age at death for unaffected individuals) for each individuals in the family. For

members with genetic tests, genetic status is also included. Figure 1.1 illustrates a family history of breast and ovarian cancer identified through a proband (indicated by an arrow). In this illustration, circles denote female family members and squares denote male members. Individuals affected with disease are indicated by filled circles or squares. Deceased family members are indicated by a diagonal slash.

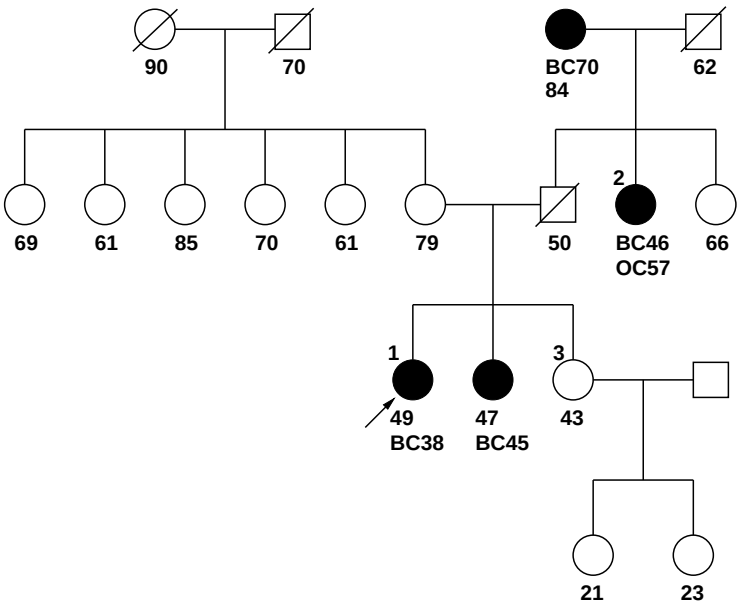


Figure 1.1: A family history of breast and ovarian cancers identified through a proband (indicated by the arrow). Label BC38 indicates that member 1 (the proband) was diagnosed with breast cancer at age 38, her current age or age at death (indicated by a diagonal slash) is 49; label OC57 indicates that one of her paternal aunt was diagnosed with ovarian cancer at age 57. Member 3 is free of both cancers.

Studies based on family histories often require calculation of the likelihood of observing phenotype data of such family histories. Let Y_{ij} denote the phenotype for the ith individual in the jth family. Usually, Y_{ij} includes both the disease status x_{ij} and age of onset a_{ij}. Under a random sampling scheme, the likelihood for the observed phenotypes of the family j (denoted as Y_j under a major gene model) can be written as

$$Pr(Y_j) = \sum_{g_j} Pr(Y_j|g_j)Pr(g_j),$$

where the sum extends over the possible genotypes g_j of the individuals in the jth family. This likelihood consists of two important components: the population frequency of different genotypes–prevalence q, which is needed for calculating $Pr(g_j)$; and the phenotype-genotype correlation–penetrance for calculating $Pr(Y_j|g_j)$. The calculation of the likelihood requires summation over all possible genotypes that individuals in the family history may have, which is usually computationally expensive. Under a general setting, algorithms have been developed for calculating this kind of likelihood and they rely on two independence structures. The first one depends on the conditional independent structure of genotypes in a family history, for example, individuals' genotypes are independent conditional on the genotypes of their parents. This structure allows the summation to be performed sequentially along the family history. The most famous such algorithm is developed by Elston and Stewart (1971) and extended by Lange and Elston (1975). Later it is best known as the "peeling algorithm" (Cannings et al. 1978). The other depends on the conditional independence of segregation indicators and involves direct observation or inference the segregation events in an experimental cross and then scoring the recombination events. The summation is performed sequentially along the chromosome. Lander and Green (1987) developed such algorithm. These algorithms have been used in both segregation and linkage analysis. A gene-characterization study usually involves only a single locus and is relatively easier to calculate.

As hereditary diseases are usually rare, simple random sampling can rarely provide enough affected individuals for statistical analysis. Data are usually collected based on "ascertainment" sampling. For example, family histories used for studying genetic diseases are usually ascertained through one or several affected individuals, called probands. Thus, the likelihood that a family history is observed depends on the phenotype or extent of disease in the family. Let A_j be the ascertainment event

that family j is ascertained. The likelihood of observing family j should be written as $P(Y_j|A_j)$ and $P(Y_j|A_j) = P(Y_jA_j)/P(A_j)$. It is clear that calculations without accounting for the sampling scheme may lead to biased estimates for genetic parameters. This problem has long been recognized. Fisher discussed the effects of ascertainment in estimating frequencies of disease genes (Fisher 1934). A large amount of literature has been devoted to methods for correcting ascertainment bias under different ascertainment sampling schemes. Several well discussed sampling schemes include single, multiple and complete ascertainment. These sampling schemes are traditionally aimed at family/sibship data (see (George and Elston 1991) for a review). Single ascertainment is for the sampling scheme under which the probability that a sibship being sampled is proportional to the number of affected individuals in the sibship. Complete ascertainment is for the sampling scheme under which each sibship with as least one affected individual has equal probability of being sampled. For multiple ascertainment, the probability that a sibship being sampled is a concave function of the number of affected individual in the sibship. Though more complicated, these methods have being extended to large pedigrees by conditioning the likelihood on pedigree structure (Elston and Sobel 1979; Ewens and Shute 1986; Shute and Ewens 1988). Other sampling schemes such as sequential sampling have been discussed by Cannings and Thompson (1978). Still, exact correction for ascertainment in most situations is difficult except in very special situations such as when the ascertainment is independent of phenotype and under single ascertainment. Furthermore, there are debates about whether the problem is intractable (Vieland and Hodge 1995; Elston 1995; Karunaratne and Elston 1998).

1.2.2 Penetrance function

Penetrance is an important component in the likelihood outlined above. In general, it measures the probability of developing disease given various risk factors including genetic effects. Under the major gene model, it measures the risk for disease given certain genotype. Often, it is age related and the penetrance curve is also known as the age of onset distribution. Methods developed to estimate the age of onset distribution often used survival analysis concepts(see (Chen et al. 1992) for a review). Various approaches have been used to model the age of onset distribution. Elston and George (1989) assume that the age of onset follows a logistic distribution. Abel and Bonney (1990) use a logistic regression to model the age dependent incidence for different age intervals. Gauderman et al. and Li et al. (Li and Thompson 1997; Gauderman and Thomas 1994) use a semi-parametric approach based on cox proportional hazard model to study the age of onset.

1.3 Characterizing missense mutations

Standard methods discussed earlier are useful when there are large enough data sets and the ascertainment bias is easy to correct. For missense mutations, directly estimating penetrance functions or risk ratios based on known schemes presented above is impractical. The difficulties are mainly due to limited sample size. Usually, the family history is moderate in size and the number of family histories for a single missense mutation is small. Complicating the situation, data are often not collected based on a single well formulated sampling scheme. Thus ascertainment bias is difficult to be accounted for. Therefore, for missense mutations, such analysis are rarely performed (Venkitaraman 2001). One way of making progress is by designing studies that obtain more informative data. Along this line, Petersen et al. developed a Bayesian approach to evaluate the mutation's specific risk and the probability that

a missense mutation is deleterious (Petersen et al. 1998). It is based on a design that requires testing the genotypes of the proband's affected relatives. The idea is to test the hypothesis that the mutation is disease causing against that the mutation is benign. If the mutation is not deleterious, the disease rate for carrier and non-carrier are the same which equals to the phenocopy rate. Otherwise, we should be able to observe significantly larger number of affected individuals being mutation carriers because the penetrance of the mutation would be similar as the penetrance of deleterious mutations. The Bayes factor in favor of the causality can be calculated. By incorporating prior knowledge of the disease causality, the probability of causality can also be provided. This study is currently the only approach designed to evaluate missense mutations risk effect. While informative and efficient in using limited testing resources, this method is restricted to families which have multiple cases and test results. If applied retrospectively to existing registries, this selection mechanism may also lead to bias. As a prospective approach it is difficult to implement quickly on a large scale.

Another difficulty presents in the missense mutation analysis is its large number. As missense mutations are one of the major types of mutations observed in the disease gene, studies analyzing them individually is not only inefficient but also may be subjected to greater bias due to random variation. However, methods to study them collectively may yield meaningful results because biologically, some missense mutations may share similar effects to the protein product and and model them as a collectively effectively uses limited information. This dissertation will focus on developing a Bayesian hierarchical approach to studying them collectively based on more general likelihood approaches. This method will be illustrated with the study of missense mutations of BRCA1/2 genes.

1.4 Plan for the thesis

In Chapter 2, family history of breast and ovarian cancers used in this thesis will be discussed. In Chapter 3, a simple hierarchical approach will be used to evaluate missense mutations' risk effects. Its limitations will be discussed. In Chapter 4, a hierarchical classification approach will be presented. This approach uses both families of probands who are tested positive and negative to calibrate the estimation, protecting the results from overestimation of the risk effects that would result from ascertainment bias. In Chapter 5, this approach will be tested through a simple simulated dataset where penetrances are assumed to be constant over age. How the approach performs for data ascertained under three different criteria will be discussed. In Chapter 6, we apply this model to study a sample of BRCA1 and BRCA2 missense mutations. The model discussed will be extended by assuming an age-dependent penetrance model with multiple cancer sites, and imperfect sensitivity of genotyping will be accounted for. The probability that each mutation is deleterious and the estimates of the mutation-specific penetrance will be provided. In chapter 7, we conclude with a discussion and outline of future work.

Chapter 2

Data

In this thesis, we develop methods to analyze unclassified mutations of disease genes. During the development, we use family histories of breast and ovarian cancers collected at the Duke University Cancer Center as an example. The application is focused on missense mutations of breast and ovarian cancer susceptibility genes BRCA1 and BRCA2. In this chapter, we will first provide some background on BRCA1 and BRCA2. Details of the Duke data will be described later.

2.1 BRCA1 and BRCA2

Breast and ovarian cancers are leading causes of cancer deaths among women in western society. Families with multiple breast and ovarian cancer cases have long been suspected to be at risk due to genetic factors. Linkage analysis by Hall et al (1990) provided the first convincing evidence for the existence of such factor by locating a breast cancer gene (later called BRCA1) to a marker on chromosome 17q21. BRCA1 was later identified by positional cloning by Miki and coworkers (1994). The existence of another breast cancer gene (BRCA2) was evidenced by localizing it to chromosome 13q (Wooster et al. 1995). BRCA2 was subsequently cloned in 1996 (Tavtigian et al. 1996). Structurally, both genes are large in size. BRCA1 is composed of 24 exons

and encodes a protein of 1863 amino acids. BRCA2 is composed of 27 exons and encodes a protein of 3418 amino acids. Proteins encoded by these two genes exhibit little resemblance to known proteins. This made it difficult to understand their exact function in the cell. It has long been suggested that they may act as tumor suppressors because of observed loss of heterozygosity in the wild type allele. Recent studies suggested their roles in DNA repair and transcriptional regulation, yet direct evidence is still needed (see (Welcsh and King 2001) for a review). For these two genes, mutations are found to be scattered throughout the coding sequence of both genes. Most of them are deleterious mutations which result in truncated proteins. A small proportion of them are missense mutations. On both genes, a total of over 500 distinct missense mutations have been reported to the on-line breast cancer mutation database (BIC 1997). Little is known about the implication of missense mutation in the pathogenesis of breast and ovarian cancer.

Genetic testing for BRCA1 and BRCA2 mutations was available soon after the cloning of these two genes. Methods used to detect mutations include direct DNA sequencing, single-strand conformation polymorphism assay, and others. While the first method rarely miss a mutation, it is labor intensive for large genes such as BRCA1 and BRCA2. Other methods are easier to perform and less costly, but may miss more mutations. Genetic test results can be positive or negative. Interpretation of these test results depends on the sensitivity, specificity of the testing method as well as the prevalence of the mutation.

Epidemiological characteristics of these two genes, such as prevalence and penetrance, were studied intensively in the last decade. Disease causing mutations of both genes are rare in the general population. An estimated prevalence of 0.0033 for all high risk dominant breast cancer susceptibility genes was provided by a segregation analysis of a large population based, case-control study (Claus et al. 1991).

Gene frequency of BRCA1 was indirectly estimated to be 0.0006 from a study by combining penetrance estimates of BRCA1 gene and population–based genetic epidemiology studies of British breast cancer patients (Ford and Easton 1995). Direct estimates of prevalences for individual BRCA1 and BRCA2 genes are not available due to limited sample of genetic tested individuals, large numbers of different mutations as well as imperfect genetic testing methods. However, estimated prevalences are available for several deleterious mutations commonly occurred among Ashkenazi Jews and Icelandic people (Roa et al. 1996; Johannesdottir et al. 1996).

Penetrances of these two genes are age dependent. The population based CASH study estimated the cumulative risk of breast cancer for carriers of the dominant breast cancer susceptibility genes to be 38% by age 50 years and 67% by age 70 (Claus et al. 1991). Gene specific penetrances were available after the localization of these two genes. Easton et al.(1993, 1995) used linkage analysis of breast and ovarian cancer families to estimate the penetrance of BRCA1. The cumulative risk of breast cancer for carriers of this gene is estimated to be 51% by age 50 years and 85% by age 70. Estimated cumulative risk of breast cancer for BRCA2 mutation carriers is 84% by age 70, which is similar to those of BRCA1 (Ford et al. 1998). Penetrances were also estimated based data of individuals tested with several deleterious mutations on BRCA1 and BRCA2 commonly observed among Ashkenazi Jews (Struewing et al. 1997). Carriers of BRCA1 and BRCA2 genes are also predisposed to increased risk of other cancers. For example, both genes confer increased risk for ovarian cancer and BRCA2 genes also increase the risk of male breast caner (see (Rahman and Stratton 1998) for a review).

2.2 The Duke missense mutation data set

Our data consists of 280 moderate sized family histories. Most of the family histories (277) are collected at the Duke University Cancer Center. We augment this data with 3 families of probands who tested positive with BRCA1 missense mutation R841W from Barker et al.'s study for comparison (Barker et al. 1996); these three families are believed to be ascertained in a similar way as those from Duke. Each of the family histories from Duke was ascertained through a proband who was tested for mutations on BRCA1 and BRCA2 genes. These probands were selected from a population of women with breast and/or ovarian cancer. The selection are largely based on rules related to personal history and family history of breast and ovarian cancers. The specific selection criteria are: individuals with very early onset of breast and/or ovarian cancer (< 30 years at diagnosis or < 40 if Jewish) regardless of family history; individuals with breast and/or ovarian cancer diagnosed between ages 31 and 55 and one first-degree relative with breast cancer diagnosed at or before age 55 or ovarian cancer diagnosed at any age; or individuals with breast and/or ovarian cancer diagnosed between ages 31 and 55 and at least two second-degree relatives with breast cancer diagnosed < 55 years of age or ovarian cancer at any age. However, sometimes the rules were not followed and there are families which do not meet any of the criteria listed above. Also these probands were not selected from affected individuals in the general population, but rather already identified to be high risk in some way. Such selection mechanism made it difficult or nearly impossible to write down explicitly the likelihood conditional on the ascertainment which is useful for unbiased risk assessment.

A typical family history includes the breast and ovarian cancer status of 1^{st} and 2^{nd} degree relatives of the proband as well as each individual's age and age(s) of diagnosis, if affected. An example of such family history is illustrated in Figure 1.1.

The majority of identified deleterious mutations are at BRCA1, while most of the identified missense mutations are at BRCA2. Among the 280 probands, 59 tested positive for known deleterious mutations on either BRCA1 or BRCA2 genes, with 41 carrying one of the 24 unique BRCA1 mutations identified in the sample and 18 carrying one of 15 unique BRCA2 mutations. Among probands tested positive for BRCA1 missense mutations, 13 of them carrying one of 6 unique missense mutations and another 3 each carrying two different missense mutations. Among 29 probands tested positive for BRCA2 missense mutations, 28 of them carrying one of 25 unique BRCA2 missense mutations and one probands carrying two BRCA2 missense mutations. The remaining 174 probands tested negative for any form of mutation at BRCA1 or BRCA2. None of the probands carrying two missense mutations have the same mutation as other probands. Table 2.1 shows the distribution of breast and ovarian cancers in probands categorized by their genetic test results. Note that there is no obvious distinction between genotypes based on the proband's affected status.

Table 2.1: Distribution of breast and ovarian cancer in probands categorized by their genetic test results.

	UV		Del		Neg
	BRCA1	BRCA2	BRCA1	BRCA2	
BC[a]	13	25	30	15	144
OVC[b]	1	2	4	1	9
Bi. BC[c]	3	5	13	4	19
BC & OVC	1	0	7	1	1

[a]Including bilateral cases but no OVC.

[b]OVC only.

[c]May contain individuals also with OVC.

Table 2.2 shows distribution of age of diagnosis for breast cancer in probands categorized by their genetic test results. Probands tested positive and who have had breast cancer have, on average, a slightly earlier age of diagnosis than those who

tested negative. The median age of diagnosis for those tested positive is 39.50 year old, while it is 43.50 for those who tested negative. When we break down the cases among those tested positive by the type of mutation as well as the location, we can see that those tested positive with deleterious mutations on BRCA1 seems to have an even earlier age of diagnosis with median 37.50 years old. This trend seems to be

Table 2.2: Distribution of age of diagnosis of breast cancer in probands categorized by their genetic test results, where "positive" means positive with either missense mutations or known deleterious mutations.

		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
UV	BRCA1	26.00	37.00	42.00	44.85	49.00	70.00
	BRCA2	29.00	36.00	42.00	43.40	52.00	63.00
Del	BRCA1	28.00	32.25	37.50	38.77	43.75	53.00
	BRCA2	29.00	37.00	41.00	41.67	47.00	54.00
	Positive	26.00	35.75	39.50	41.45	48.00	70.00
	Negative	19.00	37.75	43.50	42.88	48.00	65.00

preserved for family members of probands diagnosed with breast cancer. Table 2.3 tabulates the age at diagnosis of breast cancer for family members categorized by probands' genetic test results.

Table 2.3: Distribution of age of diagnosis of breast cancer in family members categorized by probands' genetic test results.

		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
UV	BRCA1	26.00	39.50	48.00	48.34	54.00	81.00
	BRCA2	24.00	40.00	48.50	49.04	55.00	82.00
Del	BRCA1	23.00	32.75	40.50	41.91	48.25	84.00
	BRCA2	26.00	39.00	47.00	48.65	58.00	76.00
	Positive	23.00	37.00	45.00	46.21	54.00	84.00
	Negative	16.00	40.00	48.00	48.92	56.00	90.00

Families categorized by number of breast cancer cases in the family and proband's

genetic test results are shown in Table 2.4. Most of the families in our study have less

Table 2.4: Number of families categorized by the number of breast cancer cases in the family and proband's genetic test result.

		3-	4+
UV	BRCA1	9	7
	BRCA2	23	6
Del	BRCA1	28	12
	BRCA2	14	4
	Positive	74(0.718)	29(0.282)
	Negative	135(0.785)	37(0.215)

than 3 breast cancer cases in the family history. For families of probands who tested positive, the percentage is about 71.8%. For those tested negative, the number is about 78.5%. There are slightly more families with probands tested positive with 4+ cases. When we break families of those tested positive by the type of the mutation as well as the location of the mutation, as shown in Table 2.4, wxe can see that a larger proportion of families of those tested positive with mutations at BRCA1 have 4 more cases than families with probands tested positive for mutations at BRCA2. Table 2.5 expanded Table 2.4 with each increased number of breast cancer cases.

Table 2.6 presents a summary of the family history data by affected status and age at diagnosis. Each cell gives the ratio of number affected to total number of family members for a specific age group, site (breast or ovarian cancer) and genetic test result across family histories. The table demonstrates that differences in family histories of probands with known deleterious mutations and those of mutation negatives lie not only in the number of cases but especially in the age at disease diagnosis. Table 2.7 expands Table 2.6 to show the number of cases and number of family members by missense mutation, age group and cancer site and demonstrates the difficulty in classifying individual missense mutations on the basis of simple summaries of extent

Table 2.5: Number of families categorized by each increased number of breast cancer cases in the family and proband's genetic test results.

		1	2	3	4	5	6	7+
UV	BRCA1	3	2	4	6	0	0	1
	BRCA2	4	11	8	3	3	0	0
Del	BRCA1	3	14	11	8	2	2	0
	BRCA2	2	4	8	0	1	2	1
	Positive	12	31	31	17	6	4	2
		(0.117)	(0.301)	(0.301)	(0.165)	(0.058)	(0.039)	(0.019)
	negative	26	56	53	22	13	1	1
		(0.151)	(0.326)	(0.308)	(0.127)	(0.076)	(0.006)	(0.006)

Table 2.6: Total number of cases over total number of individuals among relatives in each age of diagnosis category categorized by probands' genetic test results.

	Breast cancer				Ovarian cancer			
	(11,41]	(41,51]	(51,61]	(61,110]	(11,41]	(41,51]	(51,61]	(61,110]
del	25/89	21/34	13/35	2/82	2/76	5/35	7/38	4/88
neg	62/325	75/155	39/112	29/358	4/290	14/152	9/127	7/405
UV	28/146	38/69	19/61	19/158	6/129	4/59	5/69	1/177

of family history.

Observed difference in family disease phenotype between families of deleterious mutation positive probands, missense mutation positive probands and negative probands are subtle, depend on the number of cancers in the family, the types of cancers, the age at diagnosis of those cancers and, not reflected in the above tables, the exact relationships between affected and unaffected family members. Simple summaries of family history clearly cannot provide the necessary sensitivity to identify missense mutations with possible deleterious effects. What is needed is a low dimensional summary of family history capable of capturing the critical features in the family history and identifying mutations of different risk effects. In the following

Chapters, we develop statistical models to achieve this goal.

Table 2.7: Total number of cases over total number of individuals among relatives in each age of diagnosis category categorized by probands' genetic test results.

	Breast cancer				Ovarian cancer			
	(11,41]	(41,51]	(51,61]	(61,110]	(11,41]	(41,51]	(51,61]	(61,110]
N991D	0/2	3/4	1/3	0/5	0/2	0/3	0/3	0/6
G1788V	2/8	4/5	0/4	1/9	0/7	1/4	2/6	0/9
V2728I	0/4	2/2	1/2	2/5	0/4	0/1	0/2	0/6
K1690N	2/4	0/1	0/0	0/5	0/3	0/1	0/1	0/5
V1605I	2/10	1/2	0/4	0/3	0/9	0/3	0/4	0/3
D1902N	0/8	2/4	1/1	0/6	0/8	0/4	0/0	0/7
R1347G	1/8	5/6	1/3	1/13	0/8	0/3	0/4	0/15
Y179C	2/6	0/2	0/2	1/6	0/5	0/2	0/2	0/7
R2973C	0/4	1/2	1/3	1/9	0/4	0/2	0/3	0/9
T3013I	1/1	1/3	0/0	1/9	0/1	0/2	0/1	0/9
A2951T	0/2	4/6	0/0	3/10	1/2	0/3	0/1	0/12
G1529R	3/5	1/1	0/3	0/9	0/2	0/1	0/3	0/12
M1775R	4/9	3/4	0/2	0/1	0/6	1/5	0/3	0/2
R2034C	0/2	1/4	0/1	1/8	0/2	1/5	0/1	0/7
F2058I	0/2	0/0	2/3	2/6	0/2	0/0	0/2	0/7
del(97-98)	0/1	0/0	1/3	0/2	0/1	0/0	0/3	0/2
L1904V	2/9	2/5	1/5	0/5	0/8	0/4	0/7	0/5
T598A	0/9	0/1	3/4	0/6	0/9	0/1	0/4	0/6
K169R	1/6	0/0	0/0	0/3	0/5	0/1	0/0	0/3
S1140G	3/7	1/3	0/0	0/6	0/4	0/4	0/0	1/8
I1349T	0/10	2/2	1/2	0/3	0/10	0/0	0/4	0/3
E2856A	1/6	1/3	1/5	0/0	0/6	0/2	0/4	0/2
A2717S	1/3	0/0	0/0	1/5	0/2	0/0	1/2	0/4
S384F	0/3	1/2	1/2	1/2	2/3	1/2	1/1	0/3
M1652I	0/6	2/3	2/2	0/5	0/6	0/3	0/0	0/7
L2180F	2/7	0/3	1/5	0/10	0/6	0/3	0/6	0/10
R841W	1/4	1/1	1/2	4/7	3/4	0/0	1/2	0/8

Chapter 3

Preliminary Model

As we have seen in the previous chapter, sample size is typically small for individual mutations while their number in a database of high risk families is relatively large. This causes great difficulty for characterizing their risk effects individually. Biologically, some missense mutations may share similar effects on the protein product. Therefore, in this chapter, we explore the possibility of studying them together by using a Bayesian hierarchical model. To estimate the mutation's risk effect, the model borrows strength across different mutations, and the risk estimate associated with individual missense mutations is informed by other mutations. This model also allows for inclusion of prior information about each mutation's risk effect. However, it is only a preliminary model which is the building block for a more complex model described in the next chapter. This preliminary model has been applied to study a subgroup of missense mutations in our data set and its limitation will be discussed.

3.1 Model Specification

In this section, we describe a Bayesian hierarchical model to estimate mutation specific penetrances based on family history data. Two important components of the model are: the conditional likelihood of family histories based on dominant Mendelian

segregation and the penetrance function for risk assessment.

3.1.1 Conditional Likelihood

Let $x_{i,j}$ denote the disease status of jth individual in the ith family. It is a vector of 0's and 1's as the disease genes may predispose carriers to increased risks for several diseases. In our case, we consider both breast cancer, and ovarian cancer, as both are affected by the BRCA genes. Let s denote the type of disease with values 1, 2 and 3 corresponding to breast cancer, bilateral breast cancer and ovarian cancer, respectively. Age information, including age of diagnosis if affected, current age or age at death if unaffected, for the jth individual of family i is denoted as $a_{i,j}$. It is a vector of the same length as $x_{i,j}$. For a proband tested positive with mutation m, her/his genotype can be either $0m$ or mm, where allele 0 is the wild type and m takes values $1, \ldots, M$. We assume that the disease gene follow an autosomal dominant inheritance and the genotype for carrier of mutation m is written as m. This assumption is justified by the rare occurrence of missense mutations and the dominant nature of deleterious alleles. The genotype of proband i is denoted by $g_{0,i}$ and the genotype of the jth individual of the ith family is denoted by $g_{i,j}$. Under a major gene model, the likelihood of the observing family histories f_i can be obtained conditional on proband i's genotype $g_{0,i}$ and the collection of genetic parameters denoted as Θ, which includes mutation specific prevalences and and penetrances. It can be written as $P(f_i|g_{0,i}, \Theta)$. As the probabilities of getting disease for individuals of the same family are assumed to be independent conditional on their respective genotypes, this likelihood can be further written as

$$P(f_i|g_{0,i}, \Theta) = \sum_{\{g_{i,1},\ldots,g_{i,n_i}\}} \prod_{j=1}^{n_i} P(x_{i,j}, a_{i,j}|g_{i,j}, \Theta) P(g_{i,1}, \ldots, g_{i,n_i}|g_{0,i}, \Theta). \tag{3.1}$$

There are two building blocks of the likelihood, the probability of phenotype given

genotype and the probability of genotypes of family members given the genotype of the proband. The former is entirely described by the mutation specific penetrances and the latter depends on the mutation specific prevalences. In this analysis, we assume that the prevalences of different missense mutations are the same. We also assume that individuals in the ith family are either mutation free or carry the same mutation as proband i because of the rarity of these mutations. For different types of diseases considered in the model, it is assumed that the probability of having one disease is independent of the other conditional on the genotype and the mutation specific penetrance. Let $\rho_{m,s}(t)$ denote the penetrance of mutation m for disease s. Suppose individual j in the ith family carries mutation m and is diagnosed with disease 1 at age a_1 and then disease 3 at a_3 and died at age a_2, the probability of these phenotypes given his genotype can be written as the following

$$P(x_{i,k} = (1,0,1), a_{i,k} = (a_1, a_2, a_3) | g_{i,k} = m, \Theta) = \rho'_{m,1}(a_1)(1 - \rho_{m,2}(a_2))\rho'_{m,3}(a_3), \tag{3.2}$$

which is the product of the penetrances of the mutation at the ages of disease diagnosis if affected (denoted with $\rho'_{m,s}(t)$) and/or the probability of being disease free up to current age or age at death. The likelihood of observing all the ascertained family histories is the product of $P(f_i|g_{0,i}, \Theta)$.

3.1.2 Penetrance Models

As discussed in the second Chapter, it is impossible to estimate the age dependent penetrance of individual missense mutations directly because of the limited sample size. However, it is possible to estimate mutations specific risk effects by building a penetrance model on estimates of penetrance from previous studies. In a simple situation, we assume that a missense mutation is either deleterious or a benign polymorphism. Let $\rho_s(t)$ denote the penetrance of deleterious mutations for disease s and

$\phi_s(t)$ denote the phenocopy rate for disease s. Then, for missense mutation m, its penetrance for disease s can be written as

$$\rho_{m,s}(t) = \gamma_m \phi_s(t) + (1 - \gamma_m)\rho_s(t).$$

where γ_m is a dichotomous variable which takes values 0 and 1. Note that this penetrance model neglects the possibility of an intermediate value of γ_m. A more reasonable assumption is that the missense mutation may predispose carriers to increased risk of disease, but the increase is not as large as a deleterious mutation. In this case, the mutation specific penetrance can still be written as above but with γ_m as a weight parameter ranging from 0 to 1. We will refer the first penetrance model as the discrete model and the latter as the continuous model.

In both models, two assumptions are made. First, it is assumed that the dependence on age is described by the known penetrance functions and that γ_m is age independent. Second, it is assumed that the modifying effects to the penetrances by a given mutation are the same for different diseases. Thus, γ_m is the sole parameter describing the mutation specific risk effects. In some sense, it measures the collective risk effects by the same mutation to the different diseases under consideration in our study. If $\gamma_m = 1$, we have $\rho_{m,s}(t) = \rho_s(t)$ and it means that the mutation is probably a polymorphism. If $\gamma_m = 0$, it means that the mutation is as bad as disease causing mutations.

By incorporating the penetrance function, the conditional likelihood listed above is complete. This likelihood provides the marginal distribution for the observed phenotypes in the family histories and is a function of parameters $\Gamma = (\gamma_1, \ldots, \gamma_M)$. Mutations with similar risk effect should have similar γ values. Prior distributions will be discussed in the application to the family history of missense mutations of BRCA genes.

3.2 Model Application

In this section, the methods described above are applied to investigate the disease causality of missense mutations at BRCA1 and BRCA2. Using the discrete penetrance model, we studied 16 missense mutations which were available at an early stage in the study. Next, we applied the continuous model to study 26 missense mutations carried by 32 probands using an expanded dataset. This set of mutations is a subset of the missense mutations described in Chapter 2. Details of the application follow.

Two essential genetic parameters involved in the conditional likelihood are the prevalence and the penetrance. Prevalences of both BRCA1 and BRCA2 genes have been discussed in previous studies ((Ford and Easton 1995) about BRCA1, (Andersen 1996) about BRCA2). No estimates of prevalence are available for individual missense mutations. Thus, we use the published estimates and assume that the frequency of mutation occurrence for different mutations on the same gene is the same. Mutation specific penetrances for breast and ovarian cancers are the parameters under investigation. As described earlier, the penetrance models are built upon phenocopying rates and penetrances of BRCA genes from previous population based studies. Easton et al have studied the penetrance for a few disease causing mutations and the phenocopying rate in the general population (Easton et al. 1995; Easton et al. 1997). Ford et al.(1998) estimated the cumulative risk of breast cancer for women with a BRCA1 or a BRCA2 cancer-predisposing mutation from a sample of high risk families. Struewing et al. (1997) studied the cancer risk associated with several mutations among Ashkenazi Jews using a kin-cohort design. In this study, we take as our starting point the smooth parametric estimates distributed with the BRCAPRO carrier probability mode (Parmigiani et al. 1998; Iversen et al. 2000). These estimates are based on a meta-analysis of the data reported in Ford et al.(1998) and

Struewing et al. (1997) and are able to describe the penetrance in yearly intervals.

As we have seen in the previous section, to calculate the conditional likelihood, we must integrate over all possible genotypes of the relatives of the proband. This computationally expensive calculation is performed by a modified version of BRCAPRO. Details of the probability model implemented in BRCAPRO can be found in Berry et al. (1997) and Parmigiani et al. (1998). Briefly, BRCAPRO calculates the probability that an individual is a carrier of a deleterious BRCA1 or BRCA2 mutation given their family history of breast and ovarian cancer among first- and second-degree family members. BRCAPRO assumes an independent autosomal dominant mode of transmission for BRCA1 and BRCA2 and takes as inputs prevalence of deleterious BRCA1 and BRCA2 mutations and penetrances of breast and ovarian cancer among the two classes of mutation carriers and among carriers of benign polymorphisms. In this study, it is modified to calculate the likelihood of the observed family history given mutation prevalence and disease penetrance and the aforementioned genetic model. Plots of the likelihood as a function of γ_m are given in Figure 3.1. Each curve corresponds to one missense mutation. Each of these mutation specific likelihoods is normalized by the area underneath and multiplied by the same factor to compare them in the same plot. For each box, the x-axis corresponds to γ ranging from 0 to 1, while the y-axis corresponds to the scaled likelihood specified in Eq. 3.1. All plots are on the same scale.

3.2.1 Discrete Model

Recall that in the discrete penetrance model, γ_m only takes value either 0 or 1. Suppose the probability of $\gamma_m = 1$ is π, which can be regarded as the probability that the missense mutation is not deleterious or the proportion of the non–deleterious missense mutations in the missense mutation population. As no previous knowledge

h

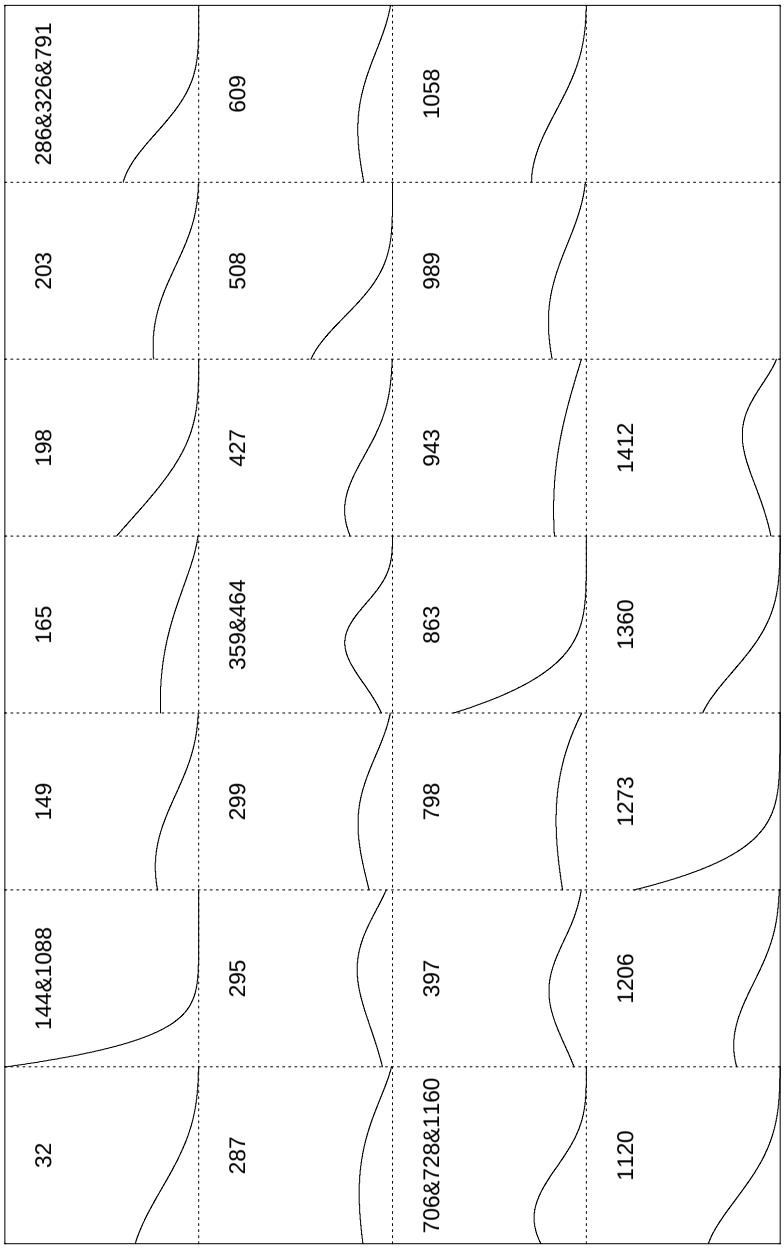


Figure 3.1: Rescaled conditional likelihood plot.

for π is available, a uniform distribution between 0 and 1 is used. Thus the prior can be specified as the following.

$$P(\gamma_m|\pi) = \pi I_{\{\gamma_m=1\}} + (1-\pi) I_{\{\gamma_m=0\}} \tag{3.3}$$

$$P(\pi) = 1 \qquad \pi \in [0,1] \tag{3.4}$$

By incorporating the likelihood, the posterior can be written as

$$P(\Gamma, \pi|data, g_0) \propto \prod_{m=1}^{M} [(\prod_i P(f_i|g_{0i}=m, \gamma_m)) P(\gamma_m|\pi)] P(\pi) \tag{3.5}$$

and the full conditionals can be simply obtained as the following

$$[\pi|others] \sim Beta(d, M-d) \tag{3.6}$$

$$[\gamma_m|others] \sim [\prod_i P(f_i|g_{0i}=m, \gamma_m)][\pi I_{\{\gamma_m=1\}} + (1-\pi) I_{\{\gamma_m=1\}}] \qquad m = 1, \ldots, M \tag{3.7}$$

Posterior samples of γ_m and π are obtained by using Gibbs sampling. The posterior mean and variance of parameter π are 0.0586 and 0.00303 respectively. Posterior means of γ_m's ranging from 0.000 to 0.0406. Thus all the missense mutations seem to be deleterious. This result is reasonable under this model specification. From the likelihood plot illustrated in 3.1, we can see that for most mutations, the probability of observing the family history given the penetrance of deleterious mutations is much larger than that given the phenocopy rate. This means that the penetrance of the deleterious mutations better explains the aggregation of diseases in the family history than the phenocopy rate. The reason for the similarity in the posterior estimates of the γ_m's is that the size of the data are still very limited and the estimates are shrunk toward a common value, the proportion of non-deleterious missense mutations.

3.2.2 Continuous Model

Under the continuous penetrance model, a natural choice for the prior of γ_m is a beta distribution with parameters (α_1, α_2). For the hyper-parameters α_1 and α_2,

we have several choices. Notice that in this Beta distribution, α_1 and α_2 can be interpreted as the number of benign and deleterious missense mutations, respectively. Hence, for each of these hyper-parameters we can assume Gamma distributions so that the mean of the distribution corresponds to the current observation (based on BIC database) and the variance corresponds to our certainty about this number (which can be specified based on the current observed value and the possible value out in the population which we are able to observe). For priors on beta distributions, one can specify hyper-priors based on a different parameterization, such as $\alpha_1/(\alpha_1+\alpha_2)$ and $\alpha_1+\alpha_2$, the mean percentage of the benign mutations and the total number of the missense mutations, respectively. We apply a prior resembles the Jefferey's for re-parameterized hyper-parameter $u=\alpha_1/(\alpha_1+\alpha_2)$ and a normal distribution with mean μ and variance σ^2 for the hyper-parameter $v=\log(\alpha_1+\alpha_2)$. Thus in this model the priors are

$$P(\gamma_m|\alpha_1,\alpha_2) = Beta(\alpha_1,\alpha_2), \tag{3.8}$$

$$[u] \sim 1/\sqrt{u(1-u)}, \tag{3.9}$$

$$[v] \sim Norm(\mu,\sigma^2). \tag{3.10}$$

And the posterior can be written as:

$$P(\underset{\sim}{\gamma_m},u,v|F,g_0) \propto \prod_{m=1}^{M}[(\prod_i P(f_i|g_{0i}=m,\gamma_m))P(\gamma_m|\alpha_1,\alpha_2)]f(u)f(v) \tag{3.11}$$

The marginal posterior of $[u,v]$ can be written as the following

$$[log(\frac{\alpha_1}{\alpha_2}), log(\alpha_1+\alpha_2)|F,g_0] \sim \prod_{m=1}^{M}\int_0^1 (\prod_i P(f_i|g_{0i}=m,\gamma_m))Beta(\gamma_m|\alpha_1,\alpha_2)\,d\gamma_m$$
$$Norm(v|\mu,\sigma^2)\alpha_1\alpha_2/(\alpha_1+\alpha_2)^2. \tag{3.12}$$

Notice that the conditional likelihood is actually a polynomial in γ_m. By using a polynomial function of γ_m to approximate the likelihood, we can obtain the analytical

form of the joint marginal posterior of $[u, v]$ based on the estimated coefficients of the polynomial. Then samples of u and v can be obtained jointly. Conditional on the sampled values of u and v, each of the γ_m can be sampled based the following distribution

$$[\gamma_m|F, g_0, \alpha_1, \alpha_2] \sim (\prod_i P(f_i|g_{0i} = m, \gamma_m))\gamma_m^{\alpha_1 - 1}(1 - \gamma_m)^{\alpha_2 - 1} \tag{3.13}$$

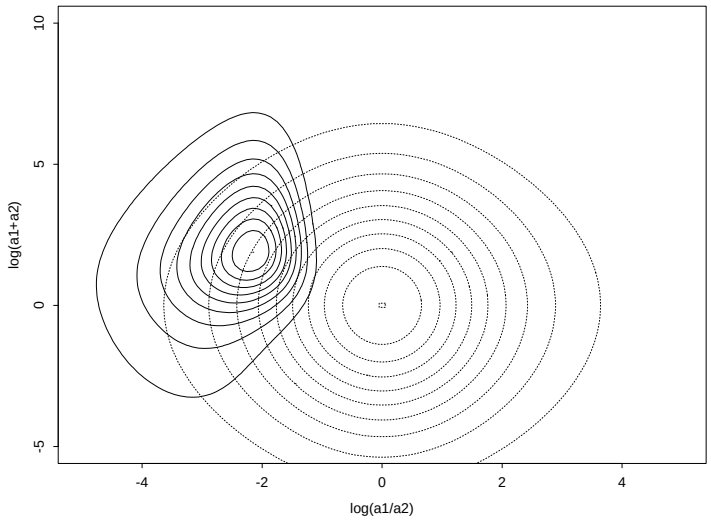
Choice of values for the parameters μ and σ depends on prior knowledge about the sample size. In this application, very little information is available, and we choose values so that the prior is relatively vague. For μ, we choose 0 so that the prior equivalent of the sample size is 1. For σ^2 we tried both 9, and 0.01. Contour plot of the joint prior distribution and marginal posterior of $[u, v]$ are plotted in Figure 3.2. We can see that the marginal posterior is sensitive to prior choices. By plotting the posterior mean of the γ_m's against their maximum likelihood estimates of γ_m's obtained from the conditional likelihood (see Figure 3.3), we can see that there is shrinkage of the posterior estimates. This indicates that information provided from the family history data is still limited. However, under different priors, the order of deleteriousness among these missense mutations are relatively stable.

3.3 Discussion

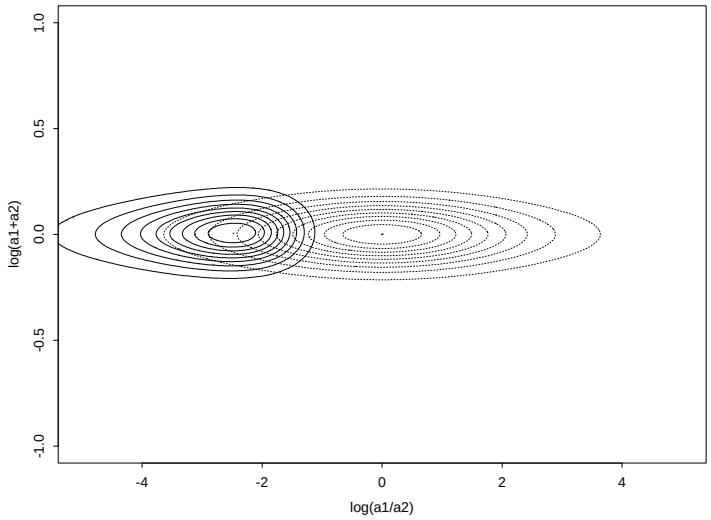
The Bayesian Hierarchical approach discussed in this chapter provide a useful way to analyze missense mutations' risk effects based on family history data, which would otherwise not be able to provide adequate support for analysis based on traditional epidemiological studies. To accommodate the sample size problem, penetrances of known deleterious mutations and phenocopy rates are used to build the penetrance models. Thus the mutation specific risk effect is estimated through one parameter without sacrificing the age dependence of the penetrance. The hierarchical structure

of the model allows us to study the missense mutations collectively and compare them systematically. It also ensures the sharing of information among the family histories of different missense mutations, as biologically these mutations may share similar effects to the function of the protein. The estimated mutation specific risk effect is the result taking into account not only the family histories related to the missense mutation, but also the distribution of mutation specific risk effects of all the mutations under study and the prior information.

However, several aspects of the model are still not very satisfactory. First, estimates of Γ are all clustered toward the deleterious end, a fact which is biologically implausible. One reason for this is that the application relies on family histories collected from high risk clinics. These families are usually ascertained because of the large aggregation of cases among family members. The model can not account for the bias caused by this ascertainment mechanism. Thus an estimate close to zero does not mean that the missense mutations is deleterious. But how to choose the cut-off for deleteriousness is difficult based on the method described in this chapter. The useful information provided by this application is thus the relative rank of deleteriousness of these missense mutations. Secondly, the penetrance model does not allow the mutation specific penetrance to go beyond the published estimates. This might be appropriate for population based data. But for our data, there is relative large number of cases in all the family. And this penetrance model has little power for discriminate these missense mutations. In the next chapter, a classification model will be discussed which takes into account the ascertainment mechanism.

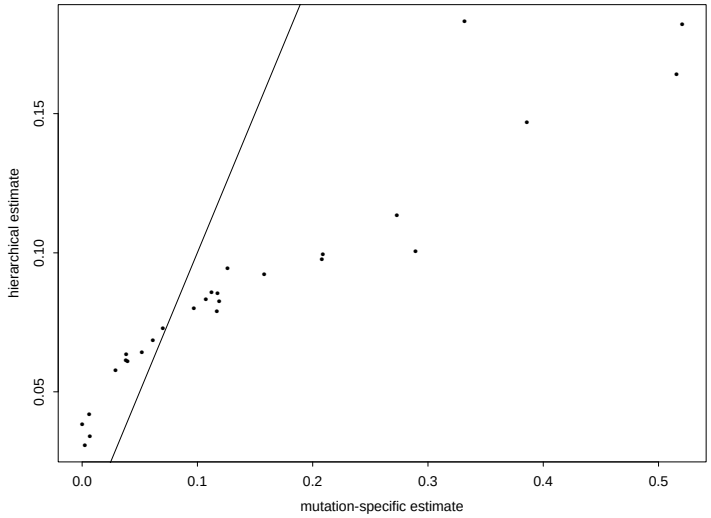


Prior: $Unif(u|0,1), N(v|0,9)$

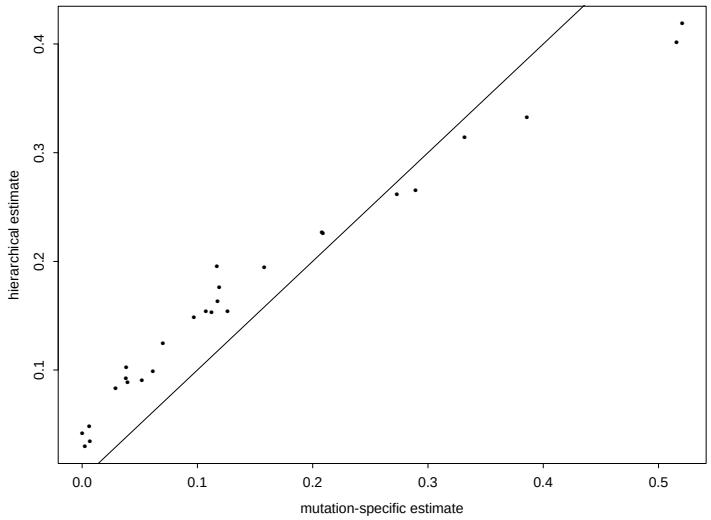


Prior: $Unif(u|0,1), N(v|0,0.01)$

Figure 3.2: Contour plots for joint marginal posterior based on different prior choices.



Prior: $Unif(u|0,1), N(v|0,9)$



Prior: $Unif(u|0,1), N(v|0,0.01)$

Figure 3.3: Shrinkage plot for models with different priors. x-axis is the mutation specific maximum likelihood estimate of the penetrance parameter γ_m. y axis is the posterior mean of γ_m.

Chapter 4

Hierachical classification model

As we have seen in the previous chapter, the hierarchical model ensures the study of the missense mutations' risk effects together and allows the possibility of comparing them systematically. Also, the one parameter penetrance model accommodates the limited sample size without sacrificing the age dependence of the penetrance. However, the results are still not satisfactory for the high risk data under investigation. Particularly, the set of family histories is collected in genetic counseling clinics and is highly selected for cancer related histories. Information about mutation specific parameters estimated from these families without ascertainment correction are often biased because individuals with high disease rate in the family is more likely to have genetic testing than those in the general population. Also, non-deleterious polymorphisms observed in high risk clinic may also exhibit large number of cases among relatives. Despite progress made in correcting ascertainment bias for the likelihood calculation, the actual ascertainment criteria are usually too complicated to be fully incorporated into the calculation. Also, to correct the ascertainment bias, one needs to estimate the parameters conditional on the ascertainment criteria. With limited information provided from family histories for each mutation, such correction further shrinks the effective sample size and results in greater uncertainty in the estimation of

these genetic parameters. On the other hand, to answer the question whether a missense mutation is deleterious or not, it is not necessary to have unbiased estimates of the genetic parameters. It is only required that we be able to discriminate the deleterious and non-deleterious variants through these parameters. Thus, instead of correcting the ascertainment bias in the likelihood calculation, we expand our study sample by using information from family histories of those tested positive with known deleterious mutations and those tested negative to help discriminate the deleterious and non-deleterious variants, and classify the missense mutation in a hierarchical setting.

With the expansion of the study sample, the accuracy of the genetic test becomes important as failure to correct for testing error is likely to result in biased estimates of mutation effect. In context of the classification, such bias may limit the model's ability to discriminate the deleterious and non-deleterious groups. Thus the testing error needs to be considered in the analysis.

In this chapter, we describe a Bayesian hierarchical classification model addressing these concerns. In this model, family histories of probands tested positive with deleterious mutations and family histories of probands tested negative are used to help classify missense mutations as deleterious or benign. The classification step is embedded in a Bayesian hierarchical analysis in which the mutation specific penetrance parameters of mutations with similar risk effects are modeled as arising from a common population whose characteristics are learned from data. And an allowance for genetic testing error is incorporated in the model.

The hierarchical classification model contain the following stages. First, the likelihood of family histories of certain mutation are calculated based on the genetic transmission model. Then, the variability of the genetic model parameters are modeled through a mixture of beta distributions with constraint. In the third stage, the

model incorporate the prior information about the parameters of the beta distributions.

In a genetic test, assay sensitivity – the probability that an assay find a mutation when one is present – is usually less than 1 and varies from method to method and locus to locus, while assay specificity, the probability that no mutation is detected when non is present is nearly 1 across methods and loci. Indeed, individuals who test positive with a mutation usually receive a confirmatory followup test, significantly reducing the possibility of false-positives. Accordingly, we assume that the genotype of individuals who test positive is that indicated by the test. And each unique mutation m is assumed to modify the penetrance through parameter γ_m. Individuals who test negative may be "true negatives" or "false negatives". We treat the genotype of negatives as a missing variable and multiply impute it. For true negatives, we assume that the aggregation of disease in the family is due to phenocopying at a rate captured in the parameter γ_{nd}, common to all true negatives in the study. We assume that false negatives carry a common deleterious mutation which is different from other known mutations in the positive or missense group and that they modify the penetrance through parameter γ_d. Furthermore, we assume that the rate of false negatives is $1 - \xi$.

Based on the information of mutation specific parameters γ's obtained from the family histories, the mutations are classified as deleterious or benign. To illustrate the general idea of the classification, we assume that there are total of K pedigrees collected. Each family history is denoted as f_i, where i is from 1 to K. Each pedigree is identified through one proband. Probands are tested with mutations on a disease susceptibility gene, in which I of them are tested negative, J-I of them are tested positive with M distinct missense mutations and K-J of them are tested positive with N distinct deleterious mutations. Figure 4.1 illustrates the general idea

of the classification procedure described above. In this figure, the K family histories are plotted in the top row and denoted by $f_i, i = 1, \ldots, K$. Each proband's test results – negative, positive with a missense mutation – are indicated by circles, round-cornered rectangles and rectangles, respectively, enclosing the family history variable. Mutation-specific penetrances are plotted in the second row and are denoted by γ_{nn}, where the index nn identifies the mutation the proband carries. Values of nn ranging from m_1 through m_M correspond to the M unique missense mutations in the sample, those ranging from d_1 to d_N correspond to the N unique identified deleterious mutations, the index nd corresponds to the wild type genotype(s) and the index d corresponds to a common, unobserved deleterious genotype. Pedigrees of probands with mutation nn are connected to the appropriate mutation-specific penetrance parameter γ_{nn} by arrows. Dashed arrows indicate uncertainty about a proband's genotype. Mutations and their phenotypic effects, depicted at the bottom of the plot, are also connected by arrows. Uncertainty in the effects of the various missense mutations is denoted by dashed arrows.

Let t_{0i} and g_{0i} denote the genetic test result and actual genotype of proband i, respectively. For a family identified through a proband who tested negative, the conditional likelihood can be written as

$$\begin{aligned} P(f_i|t_{0i},\Gamma,\xi) &= P(f_i|g_{0i}=m_{nd},\gamma_{nd})P(g_{0i}=m_{nd}|t_{0i},\xi) + \\ &\quad P(f_i|g_{0i}=m_d,\gamma_d)P(g_{0i}=m_d|t_{0i},\xi) \\ &= \xi P(f_i|g_{0i}=m_{nd},\gamma_{nd}) + (1-\xi)P(f_i|g_{0i}=m_d,\gamma_d), \end{aligned}$$

where Γ represents the collection of all the γ's. Let d_i denote the latent variable, with possible values 0 and 1, indicating whether the proband is a false-negative (0 for true negative). The joint distribution of the observed pedigree f_i and the unobserved indicator d_i conditional on the genetic parameters and the test results can be written

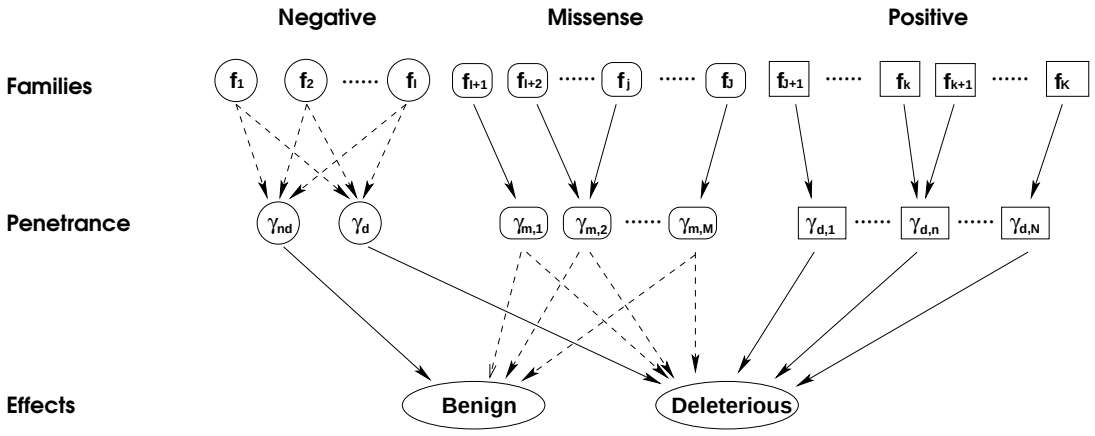


Figure 4.1: Graphical illustration of the classification model. In this depiction, The K family histories are denoted by $f_i, i = 1, \ldots, K$. Each is identified through one proband who is tested for mutations at the site(s) of interest. Each proband's test results – negative, positive with a missense mutation – are indicated by circles, round-cornered rectangles and rectangles, respectively, enclosing the family history variable. Mutation-specific penetrances are plotted in the second row and are denoted by γ_{nn}, where the index nn identifies the mutation the proband carries. Values of nn ranging from m_1 through m_M correspond to the M unique missense mutations in the sample, those ranging from d_1 to d_N correspond to the N unique identified deleterious mutations, the index nd corresponds to the wild type genotype(s) and the index d corresponds to a common, unobserved deleterious genotype. Pedigrees of probands with mutation nn are connected to the appropriate mutation-specific penetrance parameter γ_{nn} by arrows. Dashed arrows indicate uncertainty about a proband's genotype. Mutations and their phenotypic effects, depicted at the bottom of the plot, are also connected by arrows. Uncertainty in the effects of the various missense mutations is denoted by dashed arrows.

as

$$\begin{aligned} P(f_i, d_i | t_{0i} = 0, \Gamma, \xi) &= P(d_i | t_{0i} = 0, \xi) P(f_i | \Gamma, d_i) \\ &= [\xi P(f_i | g_{0i} = m_{nd}, \gamma_{nd})]^{(1-d_i)} [(1-\xi) P(f_i | g_{0i} = m_d, \gamma_d)]^{d_i} \end{aligned}$$

For family i identified through a proband who tested positive with mutation m (either missense or deleterious), as the testing result is considered accurate, the conditional likelihood can simply be written as

$$P(f_i | t_{0i} = m, \Gamma) = P(f_i | g_{0i} = m, \gamma_m)$$

Thus, let F denote the collection of all the families, and D_n denote the collection of latent indicators d_i for all the negative families, we can write down the joint

probability of family histories and latent variable D_N conditional on the genetic parameters as the following:

$$P(F, D_N|T_0, \Gamma, \xi) = \prod_i P(f_i, d_i|t_{0i} = 0, \gamma_{nd}, \gamma_d, \xi) \prod_j P(f_j|t_{0j} = m, \gamma_m)$$

where family index i and j represent families of probands tested negative and positive, respectively.

In the second stage of the hierarchical model, we assume that among the missense mutations, a proportion π are deleterious. As for each variant, the deleteriousness is captured through parameter γ_m, mutations of similar effects would have similar value of γ_m. Therefore, considering that in our model, γ's range from 0 to 1, we assume that the γ's for non-deleterious mutations are from a common Beta distribution with parameters α_1 and β_1. For deleterious mutations, we assume that the corresponding γ_m's are from a different Beta distribution with parameters α_2 and β_2. Thus for γ's of the negatives, γ_{nd} and γ_d, the priors are

$$\begin{aligned} P(\gamma_{nd}|\alpha_1, \beta_1, \alpha_2, \beta_2) &= Beta(\alpha_1, \beta_1), \\ P(\gamma_d|\alpha_1, \beta_1, \alpha_2, \beta_2) &= Beta(\alpha_2, \beta_2). \end{aligned}$$

For γ's of the missense mutations, the distribution which they are from are determined by their effects, whether they are deleterious or not. Let d_m denote the deleteriousness of the missense mutation. The prior distribution for γ_m conditional on d_m can be written as

$$P(\gamma_m|\alpha_1, \beta_1, \alpha_2, \beta_2, d_m) = Beta(\alpha_1, \beta_1)^{1-d_m} Beta(\alpha_2, \beta_2)^{d_m}.$$

By incorporating the prior information about the proportion of deleterious missense mutations π, the joint prior of γ_m and d_m is

$$P(\gamma_m, d_m|\alpha_1, \beta_1, \alpha_2, \beta_2, \pi) = [\pi Beta(\alpha_1, \beta_1)]^{1-d_m} [(1-\pi) Beta(\alpha_2, \beta_2)]^{d_m}.$$

For γ's of known deleterious mutations, it is assumed that they are from Beta(α_2, β_2).

To complete the model, priors and hyper-priors for these model parameters are specified for model parameters π, ξ, α_1, β_1, α_2, β_2. The choice of prior is based on prior evidence as well as computational convenience. Parameter ξ, the proportion of true negatives in those tested negative, is decided by the sensitivity and specificity of the test as well as the prevalence of mutations in the testing population based on Bayesian rule. As there are usually follow-up test for positive test results which significantly increase the specificity of the test, ξ is mainly be decided by the sensitivity of the test and the prevalence of mutations. Hence, in a specific application, choice of prior for xi should mainly decided by the sensitivity of the genetic test as well as prior knowledge about the prevalence of mutations in the testing population. For parameters π, as very often little information is known, choice of prior can be a relative vague proper prior. For parameters α_1, β_1, α_2, β_2, choice of hyper priors are usually based on prior knowledge as well as convenience for calculation. In our applications, we use uniform distributions between 1 and 100 as the hyper prior. To ensure identifiability, constraint need to be placed on these 4 model parameters. We use a relatively weak constraint: for any γ, the CDF of γ given α_1 and β_1 is uniformly greater than the CDF of γ given α_2 and β_2. This constraint can be written as

$$\int_0^x Beta(\gamma|\alpha_1,\beta_1)d\gamma \geq \int_0^x Beta(\gamma|\alpha_2,\beta_2)d\gamma, \quad \text{for any } x \in (0,1].$$

Let $C(\alpha_1,\beta_1,\alpha_2,\beta_2)$ denote the constraint and $\Theta = \{\Gamma, D_N, D_M, \alpha_1, \beta_1, \alpha_2, \beta_2, \xi, \pi\}$ denote the parameters in the model. Posterior can be written as

$$\begin{aligned} P(\Theta|F,T_0) \quad &\propto \quad P(F,D_N|T_0,\Gamma,\xi)P(\Gamma,D_M|\alpha_1,\beta_1,\alpha_2,\beta_2,\pi) \\ &\qquad f(\alpha_1)f(\beta_1)f(\alpha_2)f(\beta_2)I_{C(\alpha_1,\beta_1,\alpha_2,\beta_2)}f(\pi)f(\xi) \end{aligned}$$

Classification of individual missense mutations is carried out in context of that

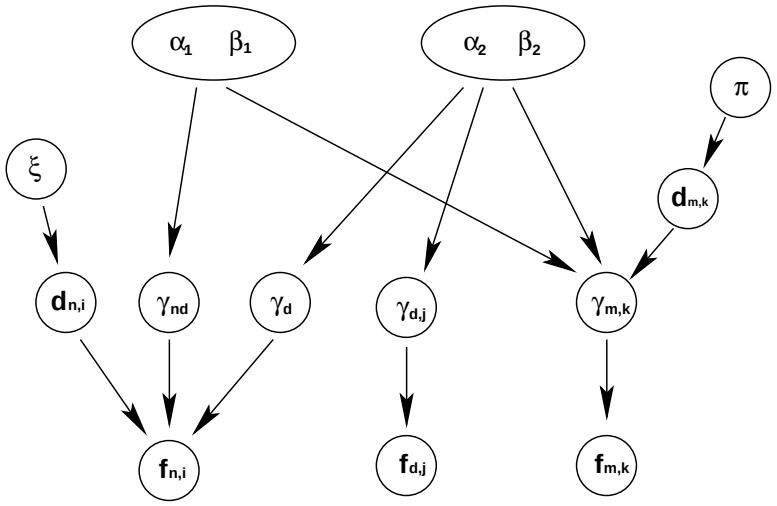


Figure 4.2: Conditional structure of the hierarchical classification model

mutation's d_m. In particular, inference focuses on the posterior distribution of d_m given F and T with missense mutation m being classified as deleterious if the posterior probability that $d_m = 1$ is large. This makes clear the ascertainment correction implicit in this approach, namely that classification is conditional on family history, F, and hence conditional on ascertainment as ascertainment is a function of F.

An illustration of the conditional structure of the model parameters are illustrated in Figure 4.2. For families identified through probands with different testing results, $f_{n,i}$ of the negatives, $f_{m,j}$ of those tested positive with missense mutations and $f_{p,k}$ of those tested positive with known deleterious mutations, the conditional structure is different. For each parameter, the full conditional can be written down. And parameters can be sampled iteratively conditioned on the sampled value of previous sampled parameters. The detail of the full conditionals are described in the Appendix.

This classification model only requires that the family histories are collected using the same ascertainment procedure. Operationally, this suggests that they are all recruited into the same high risk study or by the same high risk clinic. It is worth noting again that, using the methodology we describe, it is not necessary to correct the ascertainment bias by explicitly modeling the ascertainment procedure

in the likelihood calculation as the model implicitly corrects for ascertainment. The operating characteristics of this approach is evaluated in the context of a simulation study.

Chapter 5

Simulation based validation

We demonstrated in chapter 3 the preliminary model's ability to describe variability in disease association present in the family histories of missense mutation carriers. However, sample size is too limited to allow us to accurately estimate of mutation specific penetrance. Using the classification model, we are able to directly estimate the probability of deleterious for each mutation. Given high risk ascertainment, it is not immediately clear that estimates of this probability are accurate. In this chapter, we assess classification accuracy, how the accuracy of penetrance estimation affects the accuracy of classification, how sensitive our method is to different ascertainment rules and how well our method performs in a larger sample. These assessments are made by using data simulated under three ascertainment rules: population based sampling, affected proband sampling and progressive ascertainment. In the following sections, we describe data simulation, the likelihood calculation and an application of the classification model to the simulated data sets.

5.1 Family history simulation

In this simulation study, family histories are simulated given a simple study design and a stylized genetic model. In particular, family histories consist of two parents and

an offspring, and we consider a single autosomal dominant disease gene associated with a single disease. We assume that the allele frequency (prevalence) for each mutation of the disease gene is the same and that penetrances of the mutations of the disease susceptibility gene are constant over age.

Family histories are generated in two steps. First, the genotypes of family member are generated. Genotypes of the parents are generated independently based on the prevalence of each mutation. Genotypes of the children are generated based on the genotypes of the parents under independent segregation of alleles. Secondly, the phenotype of each family member is simulated based on the genotype of the individual. For individuals whose genotype is $0X$ or XX, where 0 denotes the wild type allele and X a certain mutation, we assume the probability of getting disease is the penetrance of mutation X, ρ_X. For individuals whose genotype is XY, where Y is a different mutation, we assume the probability of getting disease is the larger of ρ_X and ρ_Y.

In particular, we consider a population segregating a total of 30 different mutations. Among them, 20 are missense mutations including 15 non-deleterious ones and 5 deleterious ones, and 10 mutations are known deleterious mutations. The prevalence for each mutation is assumed to be 0.002. Penetrances of these mutations are assumed to arise from two distinct Beta distributions. Non-deleterious mutation penetrances are generated from a Beta(2,18) distribution. Deleterious mutation penetrances are generated from a Beta(8,2) distribution.

5.2 Likelihood calculation for simulated data

Mirroring the setting of a gene characterization study, we assume that the family histories are ascertained through the proband who has had the genetic test and the genotype of the proband is known. In the simulation study, the proband for each

family is assumed to be the child and the genetic test is assumed to be accurate. Genotypes of parents can be inferred based on the observed genotype of the child and the allele frequency of the possible mutations. The disease status for each family member is observed. Together with information about allele frequency, the likelihood of the family history given mutation specific penetrance parameter can be calculated.

Let x_i denote the disease status of the i^{th} family member, and g_1 the proband's genetic test results. The probability of observing a given family history conditional on the proband's genotype is

$$P(x_1, x_2, x_3|g_1 = m) = \frac{\sum_{g_2,g_3} P(x_1|g_1)P(x_2|g_2)P(x_3|g_3)P(g_1 = m|g_2, g_3)P(g_2, g_3)}{\sum_{g_2,g_3} P(g_1|g_2, g_3)P(g_2, g_3)}. \tag{5.1}$$

Where $P(x_i|g_i = m) = \rho_m^{x_i}(1 - \rho_m)^{(1-x_i)}$. There is a very small possibility that some probands carry two different mutations. To simplify the simulation, families with probands carrying two distinct mutations are discarded.

Given parents' genotypes g_2, g_3, the conditional probability distribution of the child's genotype $P(g_1|g_2, g_3)$ can be easily calculated using Mendel's laws. However, considering a disease gene with m variants and excluding genotypes of the form XY, the number of possible genotypes is quite large, $1+2m$. The probability $P(g_1 = i|g_2, g_3)$ where i is a particular genotype of the proband, can be written as a $(1+2m)\times(1+2m)$ matrix. Hence the conditional likelihood is a complicated polynomial of the penetrances of all the different mutation variants. With a relatively low population frequency for each mutation, much of the information provided by the likelihood is for the penetrance of the mutation the proband carries, while little information can be obtained for the penetrances of other mutations. Thus, in our calculation, for each family history, the probability $P(x_1, x_2, x_3|g_1 = m)$ is approximated by a polynomial in ρ_m, the penetrance of the mutation carried by the proband.

5.3 Validating the hierarchical classification model

In practice, data are usually collected under a (possibly complex) set of ascertainment rules. Using the likelihood shown above to estimate penetrance is likely to yield biased estimates as it is no longer the correct likelihood. In many cases, the specific ascertainment rules are not very clear making it difficult to correct the likelihood. While unbiased penetrance estimates are difficult to obtain, it may still be possible to infer whether a mutation is deleterious or not through the classification procedure described earlier. In this simulation study, we assess performance of the hierarchical classification model for data ascertained under three different ascertainment rules:

- **Population Based:** Use all the family histories generated in the simulation.
- **Progressive Ascertainment Rule:** Among families without disease, sample 10% of them. Among families with 1 diseased individual, sample 50% of them. Among families with 2 diseased cases, sample 70% of them. And among families with 3 diseased cases, sample 90% of them.
- **Affected Proband:** Only ascertain families in which the proband is diagnosed with disease.

For data ascertained via an affected proband, the ascertainment bias is easy to correct for by explicitly using a modified likelihood

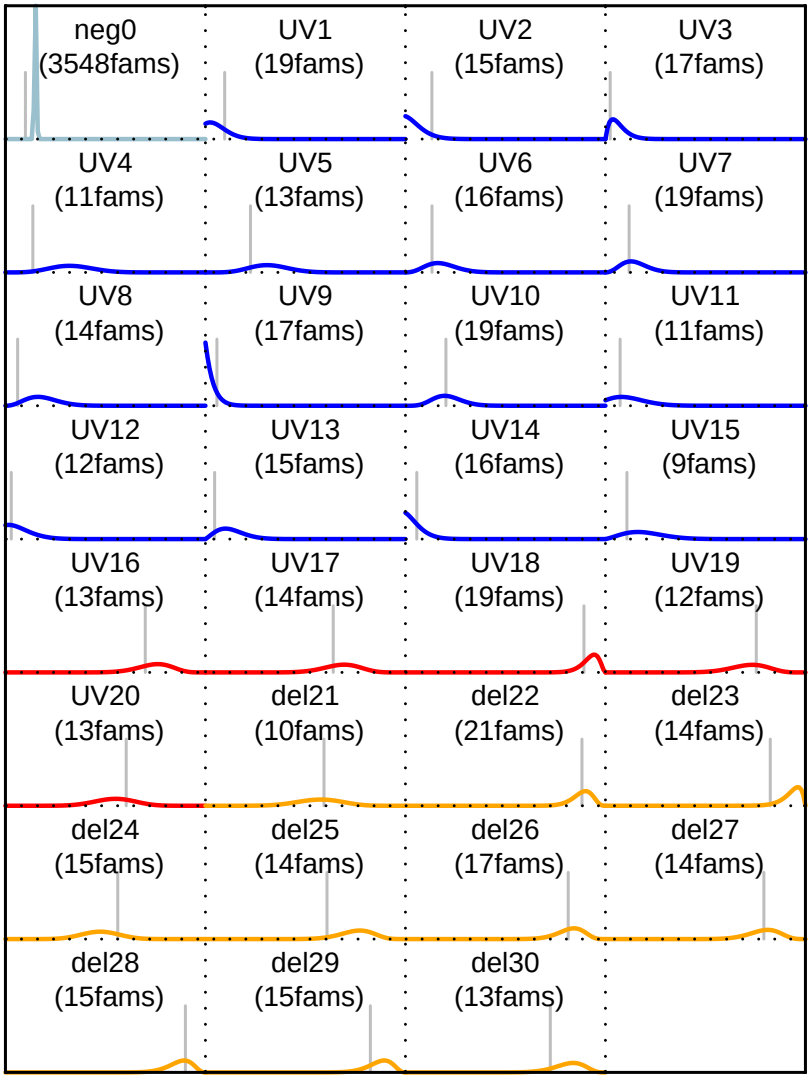
$$P(x_2, x_3|g_1 = m, x_1 = 1) = \frac{\sum_{g_2,g_3} P(x_2|g_2)P(x_3|g_3)P(g_1 = m|g_2, g_3)P(g_2, g_3)}{\sum_{g_2,g_3} P(g_1 = m|g_2, g_3)P(g_2, g_3)}. \quad (5.2)$$

Thus, we also compare the performance of classification for calculations based on the ascertainment corrected (Eq. 5.2) and uncorrected (Eq. 5.1) likelihoods. Furthermore, we assess the performance of our classification model for data of different sample size.

The design of the simulation study is as follows: We simulate 10 replicates of family histories with sample size 4000 and 10 replicates with sample size 6000. For each set of family histories, families are ascertained through the three ascertainment rules stated above. Next, we calculate the likelihood given in Equation 5.1. For data ascertained based on ascertainment via an affected proband, we also calculate the likelihood with ascertainment correction given in Equation 5.2. Finally, we fit the hierarchical classification model to the ascertained data sets using uniform(1,100) hyper-priors for the parameters of the two Beta distributions. To ensure identifiability, a relatively loose prior constraint was imposed so that the cumulative probability of the Beta distribution of the non-deleterious mutations is uniformly greater than that of the deleterious mutations. As there is no testing error simulated in our sample, the parameter ξ is set to 1. For the parameter π, which specifies the proportion of deleterious missense mutations, a beta distribution with parameters (2,2) is used.

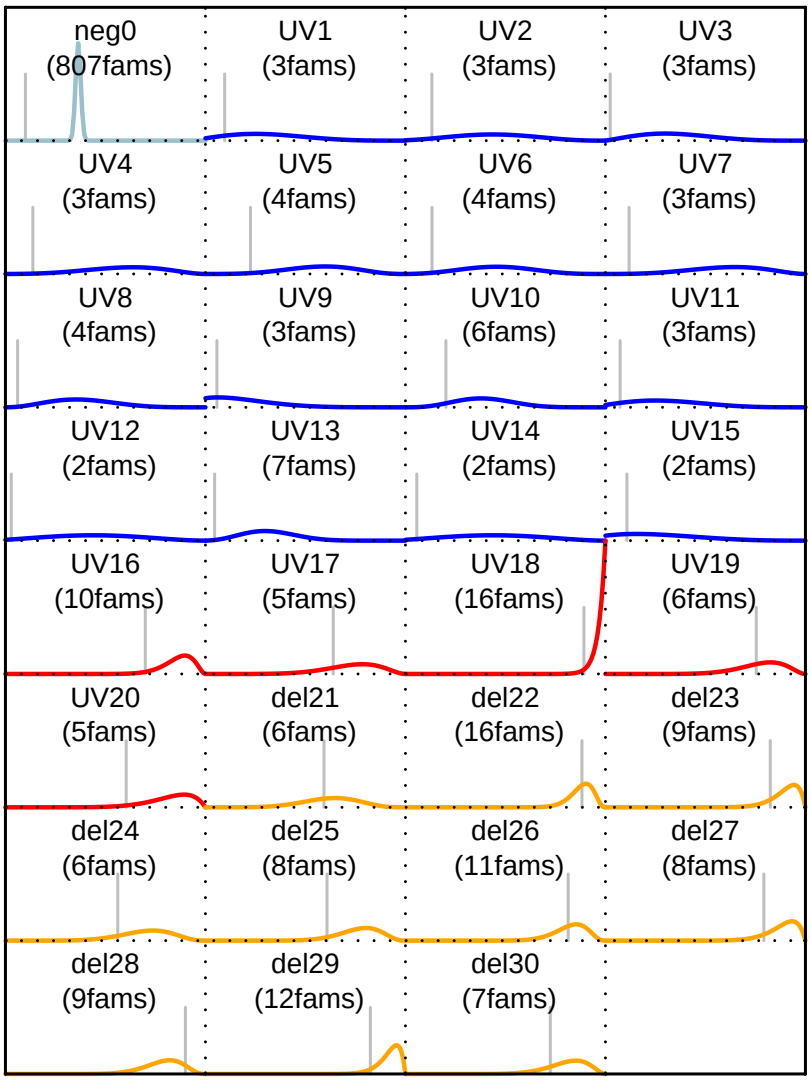
One replicate of a sample size 4000 data set serves to illustrate the detail of the calculation. Figure 5.1 plots the rescaled conditional likelihood of family histories for each mutation given mutation specific penetrance versus mutation specific penetrance ρ_m under three ascertainment criteria. Likelihoods for the wild type, deleterious, non-deleterious missense mutations and deleterious mutations are denoted by colors: light blue, blue, red, and orange, respectively. The grey vertical line in each small box shows the value of the penetrance of the particular mutation that was used to simulate the data.

Our goal is to classify the missense mutations to the correct classes based on these likelihood information. In Figure 5.2, likelihoods are plotted together to show the effect on the likelihood of γ_m under these different ascertainment rules. The coloring scheme is the same as in Figure 5.1.

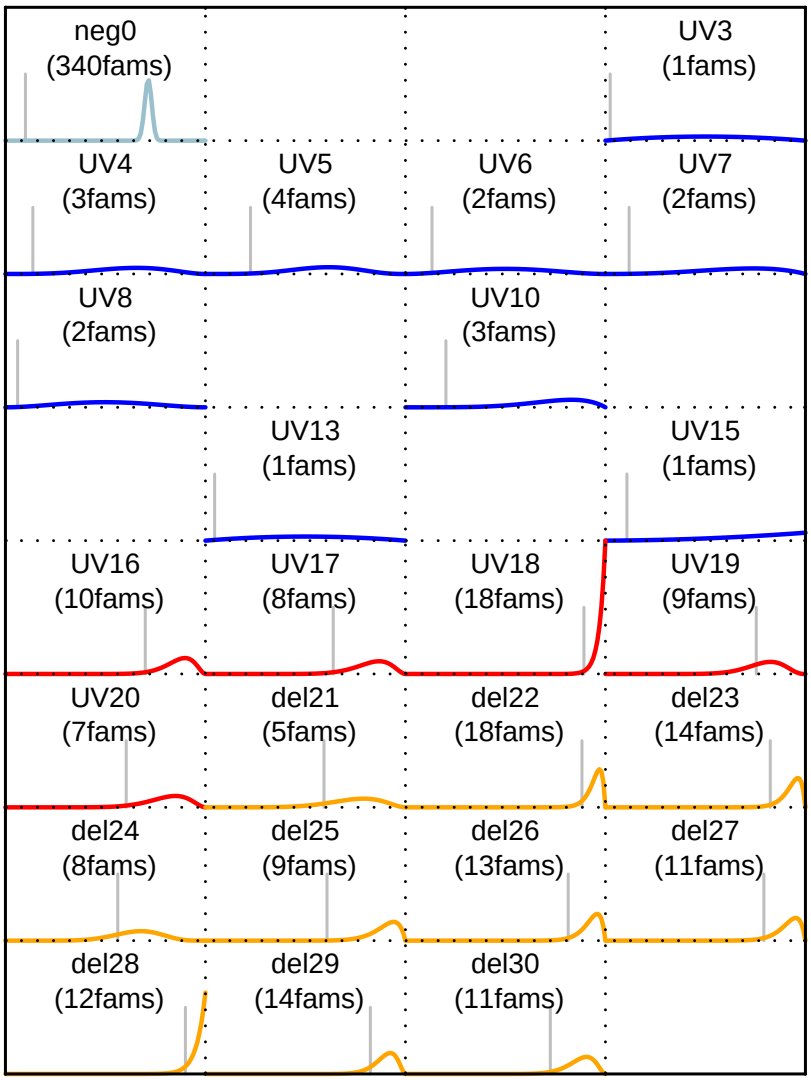


(a) Rescaled likelihood plots. Family histories are population based.

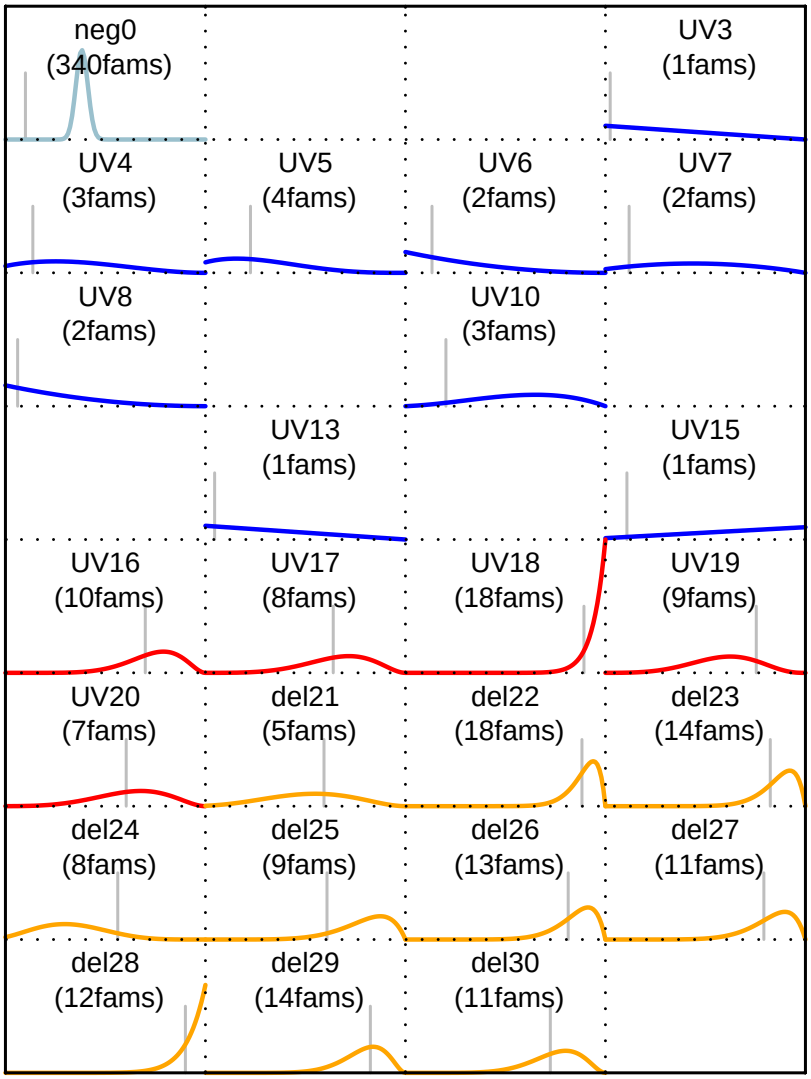
Figure 5.1: Plots of conditional likelihood of family histories given mutation specific penetrance versus mutation specific penetrance. Each colored curve represents the conditional likelihood for certain variant. And the conditional likelihoods are rescaled so that the x-axis and y-axis of each small box range from 0 to 1. Likelihoods for the wild type, deleterious, non-deleterious missense mutations and deleterious mutations are denoted by colors: light blue, blue, red, and orange, respectively. The grey vertical line in each small box shows the value of the penetrance of the particular mutation that was used to simulate the data.



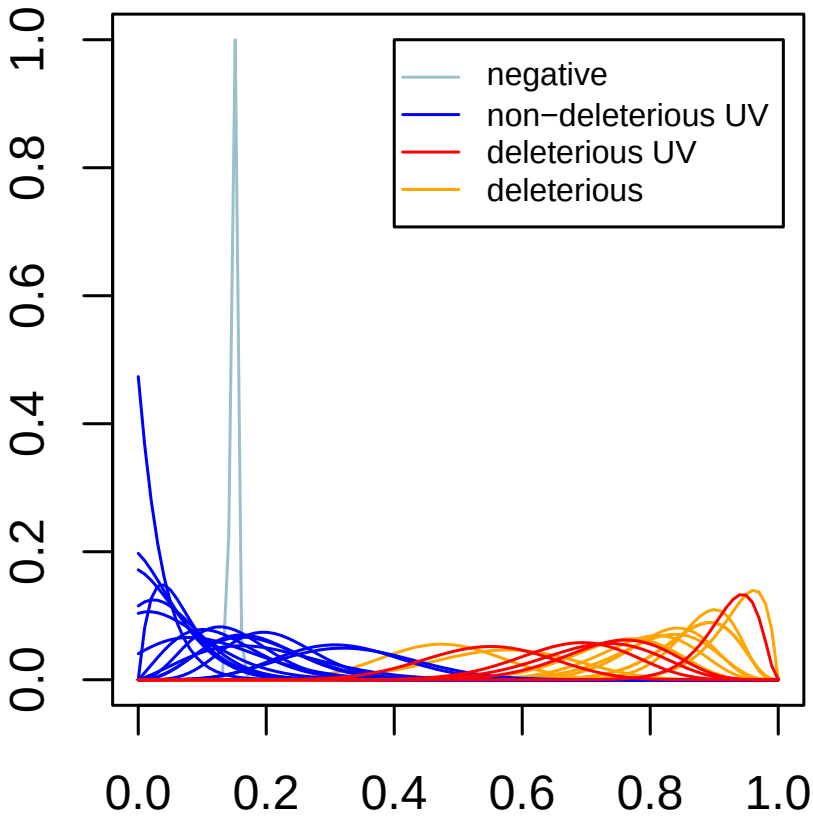
(b) Figure 5.1 continued. Rescaled likelihood plots. Family histories are ascertained through progressive ascertainment rule. Likelihoods are calculated without correcting ascertainment bias.



(c) Figure 5.1 continued. Rescaled likelihood plots. Family histories are ascertained through affected proband ascertainment rule. Likelihoods are calculated without correcting ascertainment bias.

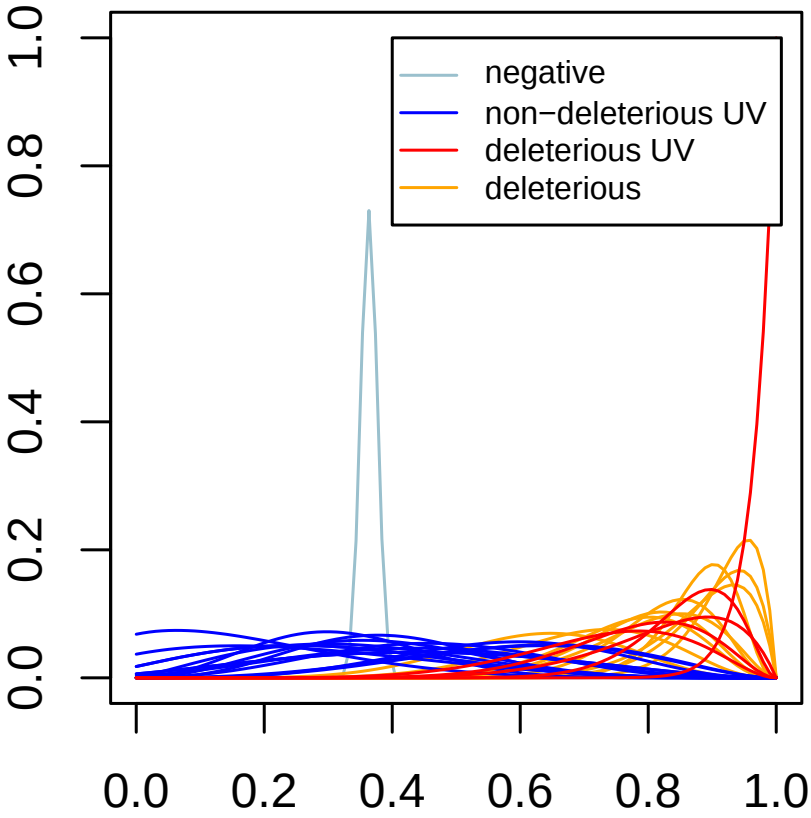


(d) Figure 5.1 continued. Rescaled likelihood plots. Family histories are ascertained through affected proband ascertainment rule. Likelihoods are calculated with ascertainment correction.

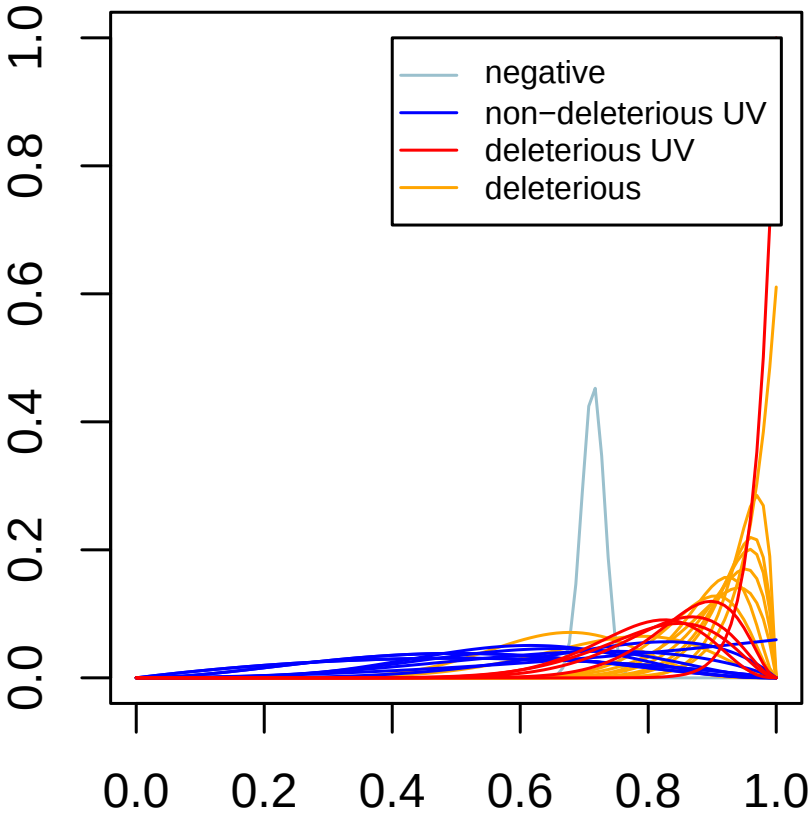


(a) Figure 5.2 (continued). Rescaled likelihood plots. Family histories are population based.

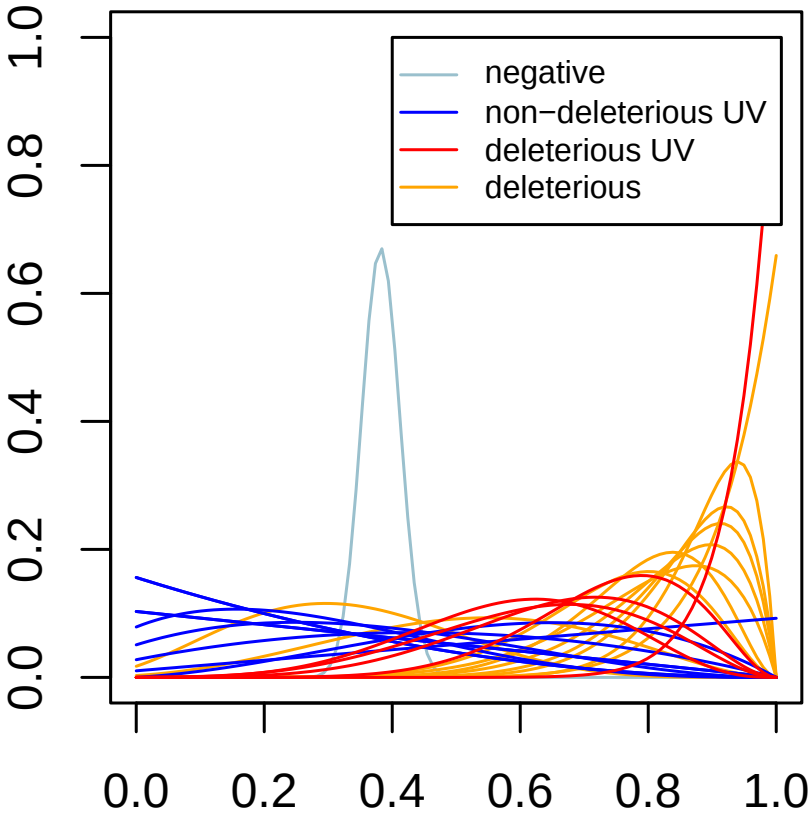
Figure 5.2: Rescaled conditional likelihood of family histories given mutation specific penetrance vs. mutation specific penetrance ranging from 0 to 1. Each colored curve represents the conditional likelihood for certain variant. The total number of simulated family histories is 4000. Likelihoods for the wild type, deleterious, non-deleterious missense mutations and deleterious mutations are denoted by colors: light blue, blue, red, and orange, respectively.



(b) Figure 5.2 continued. Rescaled likelihood plots. Family histories are ascertained through progressive ascertainment rule. Likelihoods are calculated without correcting ascertainment bias.



(c) Figure 5.2 continued. Rescaled likelihood plots. Family histories are ascertained through affected proband ascertainment rule. Likelihoods are calculated without correcting ascertainment bias.



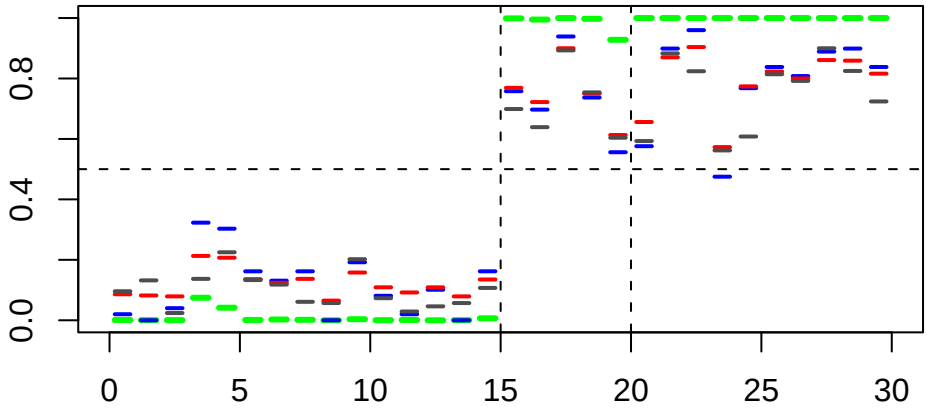
(d) Figure 5.2 continued. Rescaled likelihood plots. Family histories are ascertained through affected proband ascertainment rule. Likelihoods are calculated with ascertainment correction.

Notice that, for population based data, even though we approximated the likelihood calculation, the conditional likelihoods shows very good agreement with the penetrance of the mutation used to sample the data in the sense that the modes are very close to the 'true' values. Hence, the likelihood approximation appears to contribute little bias to population-based estimates of penetrance. From the likelihood of the negative families, however, we can see some effect of the approximation used in the likelihood calculation. The likelihood of the negatives is maximized at a slightly larger value than the phenocopy rate 0.1 set in the simulation. The reason for this is that, despite the fact that the probands tested negative, these families may still have individuals carrying a mutation with higher penetrance.

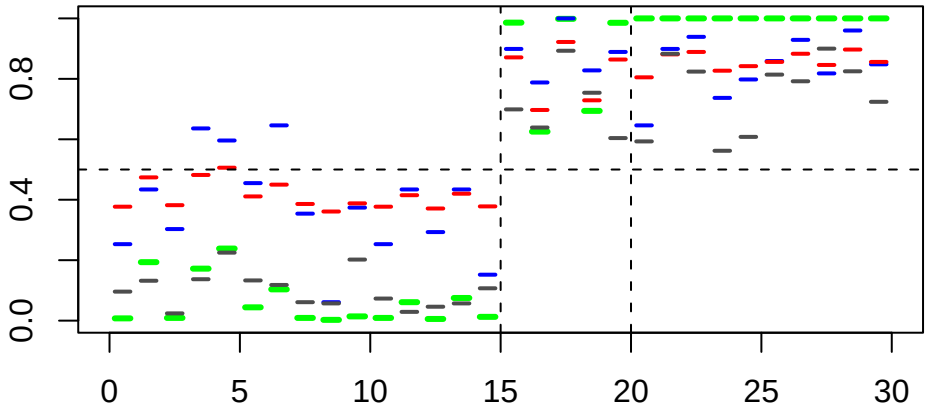
Fewer families are sampled using the progressive ascertainment rule. Most of the likelihoods calculated maximize at values greater than the penetrance used to sample the data. This parallels the situation we encountered in using data from high risk clinics. Even fewer families and mutations are sampled under the affected proband scheme. Under this scheme, the likelihoods show greater uncertainty about the value of the associated mutation specific penetrance, whether including an ascertainment correction or not. Estimates based on the uncorrected likelihoods show greater bias than their ascertainment corrected analogues.

Summaries of the estimated posterior means of the penetrance parameters and the classification result (the probability that a mutation is deleterious) are plotted in Figure 5.3.

From Figure 5.3, we can see that, for population based data (Panel (a)), the estimated posterior means of penetrance (red line) are fairly close to the maximum likelihood estimates (blue line) and the true penetrances (dark line), and the posterior probability that the mutation is deleterious (green line) is a good classifier. In particular, thresholding this probability at $1/2$ correctly classifies all 20 missense mutations.

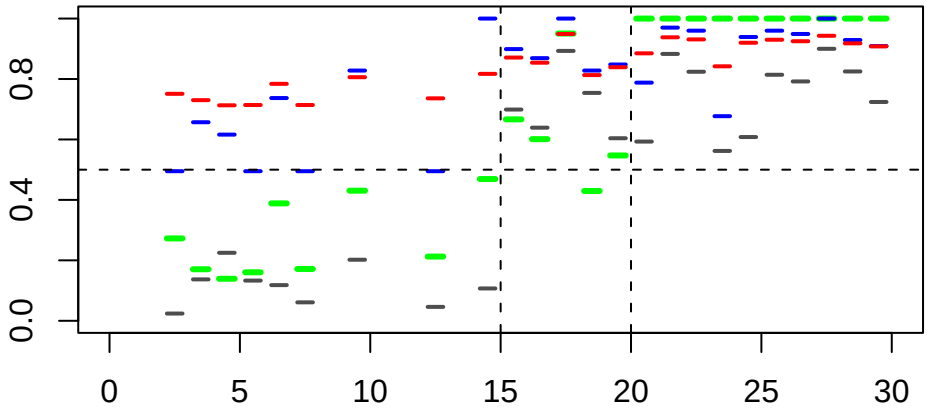


(a) Simulation results. Family histories are population based.

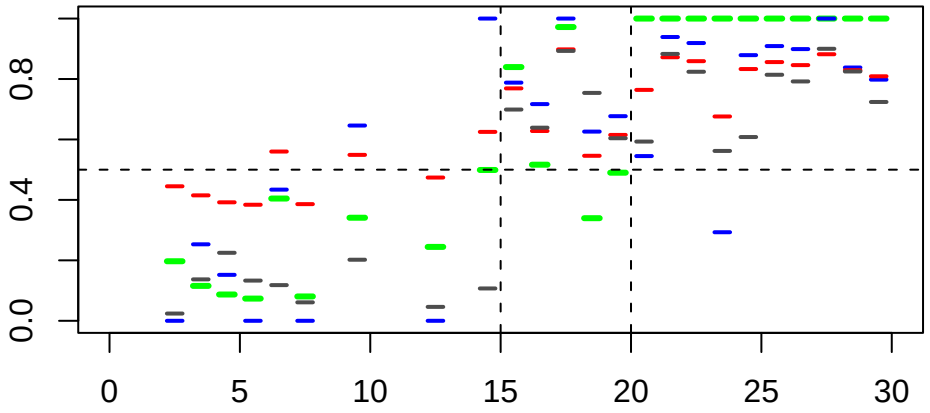


(b) Simulation results. Family histories are ascertained through progressive ascertainment rule. Likelihoods are calculated without correcting ascertainment bias.

Figure 5.3: Summary of interesting model parameters. x-axis represents the mutation index. Mutations 1-15 are non-deleterious missense mutations. Mutations 16-20 are deleterious missense mutations. And mutations 21-30 are known deleterious mutations. y-axis represents the value of the parameters. In each plot, dark lines represent mutation specific penetrances used to simulate the family histories, blue lines represents the maximum likelihood estimate of the mutation specific penetrance, red lines represent posterior mean of mutation specific penetrance estimated from the hierarchical classification model and green lines represent posterior probability that the mutation is deleterious.



(c) Figure 5.3 continued. Simulation results. Family histories are ascertained through affected proband ascertainment rule. Likelihoods are calculated without correcting ascertainment bias.



(d) Figure 5.3 continued. Simulation results. Family histories are ascertained through affected proband ascertainment rule. Likelihoods are calculated with ascertainment correction.

For data ascertained using the progressive ascertainment procedure (Panel (b)), posterior and maximum likelihood estimates of penetrance are clearly biased. One reason for this is that fewer families are sampled. However, the posterior probability that the mutation is deleterious remains an accurate classifier. Here, thresholding at 1/2 again leads to perfect classification. For data ascertained because of an affected proband, the plot (Panel (c)) shows greater bias in the maximum likelihood and posterior estimates of penetrance. But here, too, the posterior probability that the mutation is deleterious remains a good classifier, reflecting the true class of the mutation in all but one case. The ascertainment corrected likelihood for affected proband data leads to less biased estimates of penetrance but has only a minor impact on mutation classification.

We repeat the same analysis for another 9 replicated sets of family histories with each set having 4000 families, as well as 10 replicates of family histories with each set having 6000 families. The results are similar to those evidenced in Figure 5.3. Figure 5.4 displays boxplot of the probability of deleteriousness across replicates for each missense mutation. Classification results for all the replicates are presented in Figures 5.5(a)-5.5(h).

If we classify mutations with posterior probability of deleteriousness greater than 0.5 as deleterious, we can calculate the proportion of incorrect classifications based on the results obtained from these replicates. Table 5.3 summarizes misclassification rates. If sampling is population based, classification is very accurate. Further more, under progressive and affected proband ascertainment, error rates are higher but the classification is still good, and there is little difference between the corrected and uncorrected likelihoods. In particular, the classification seems to be slightly better using uncorrected likelihoods. This is likely because the ascertainment correction results in less information and thus greater uncertainty in penetrance estimates. Note

also that the classification model's performance improves as the size of the data set increases for all modes of ascertainment considered.

Table 5.1: Proportion of incorrect classifications based on calculations for replicates of simulated data.

sample size	popul.based	asc.rule2	asc.rule3 lik.no.correction	asc.rule3 lik.corrected
4000	0.005	0.046	0.125	0.153
6000	0.005	0.026	0.070	0.094

In summary, the simulation study shows that: (1). in general, the hierarchical classification model accurately classifies missense mutations; (2). the model does so even when the data are not population based and when the likelihood does not implicitly correct for mode of ascertainment; (3). the latent classification variable is robust to ascertainment bias even while the penetrance parameter is not; (4). classification accuracy increases with sample size.

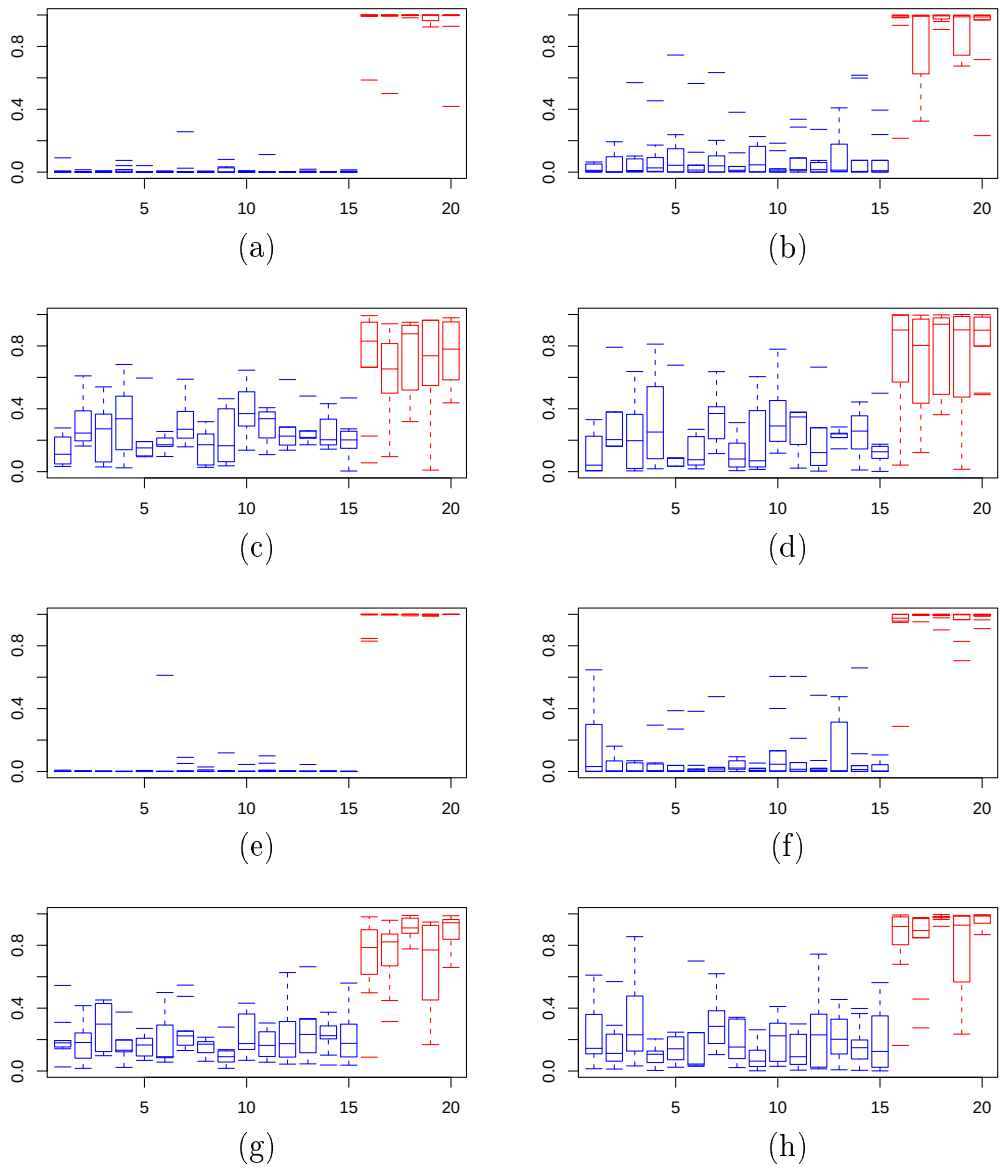
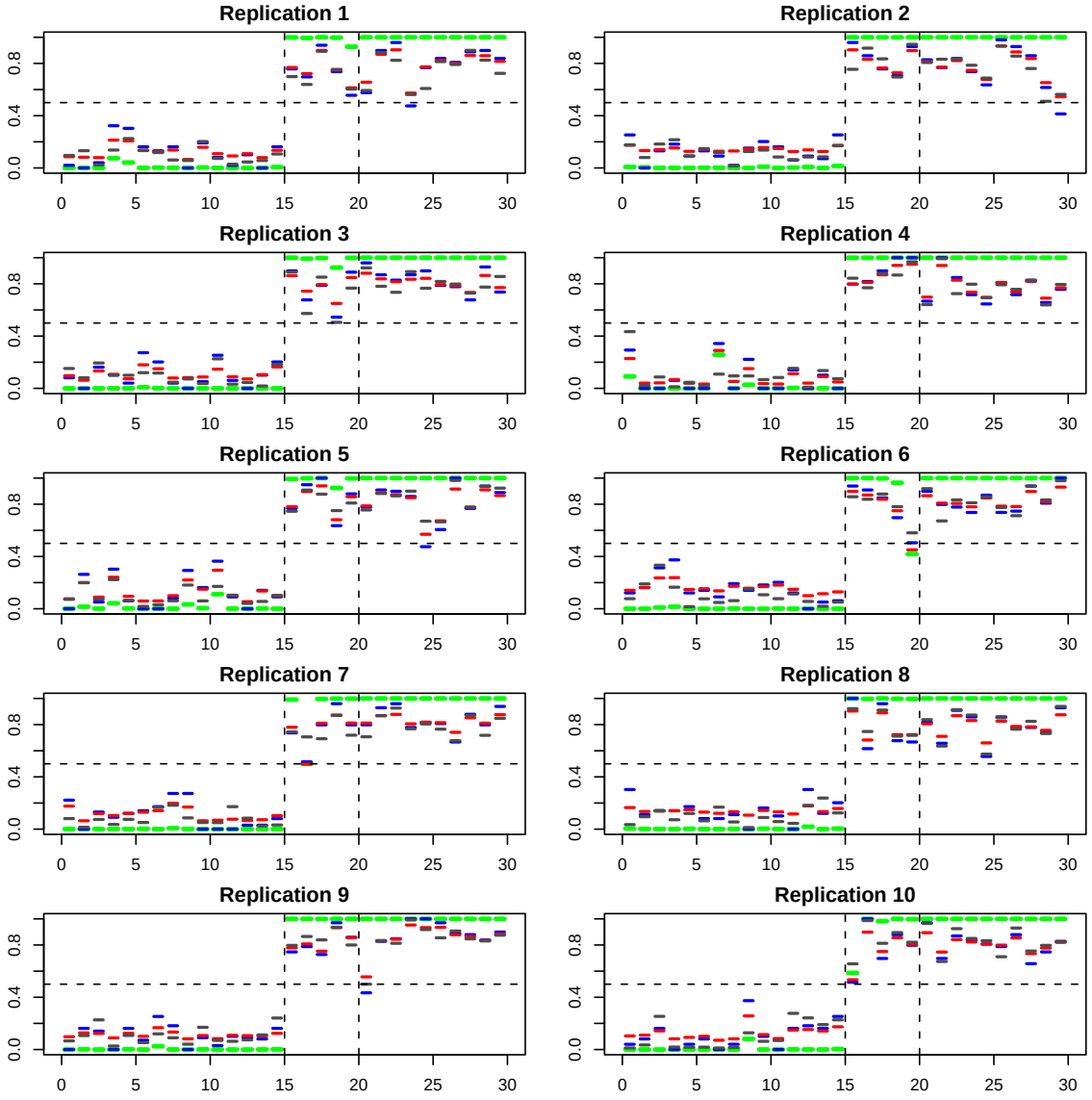
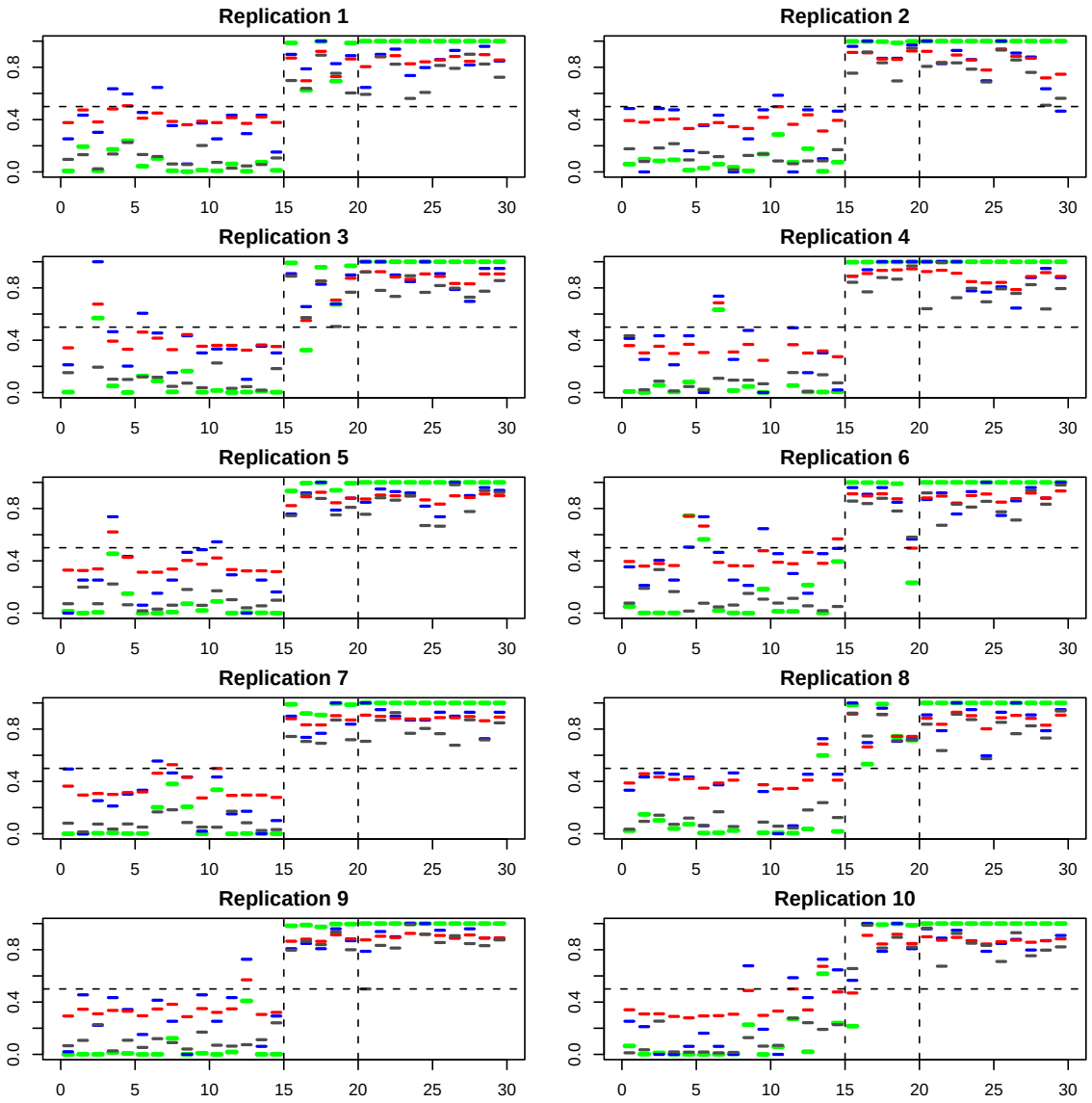


Figure 5.4: Boxplot of the probability of deleteriousness of the missense mutations estimated from the replicates. Mutations denoted by "red" are non-deleterious missense mutations and mutations in "blue" are deleterious missense mutations". Plots (a)-(d) are results based on sample size 4000, (e)-(h) are based on sample of size 6000. Plots (a) and (e) are from based on population based data. Plots (b) and (f) are based on data ascertained through the progression ascertainment rule and the likelihoods are not corrected for ascertainment bias. Plots (c), (g), (d) and (f) are based on data ascertained through the affected probands ascertainment rule. For (c) and (g), the likelihoods are not corrected for ascertainment bias, while for (d) and (h), the likelihoods are corrected for ascertainment bias.

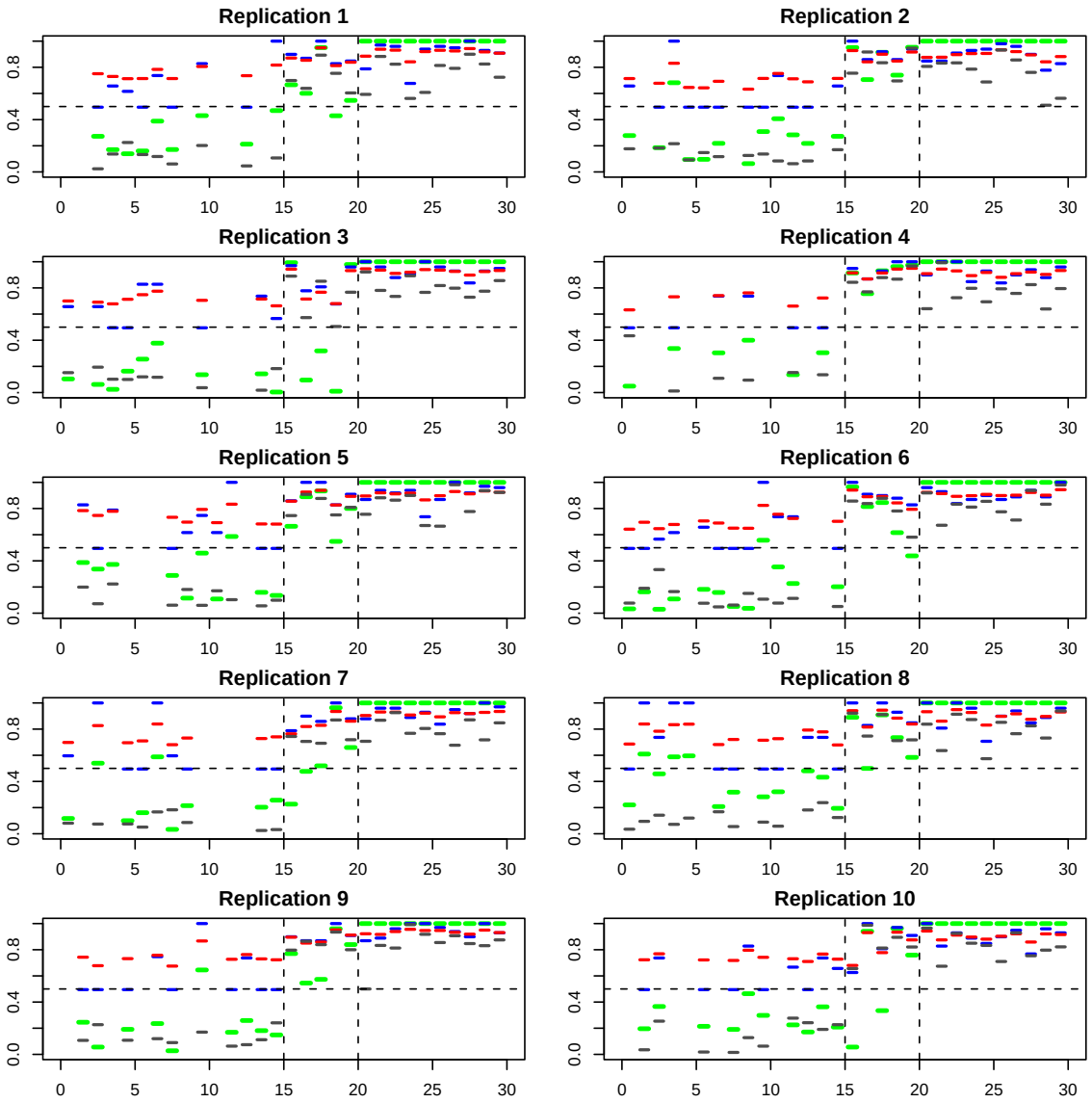


(a) Sample size: 4000. Data are population based.

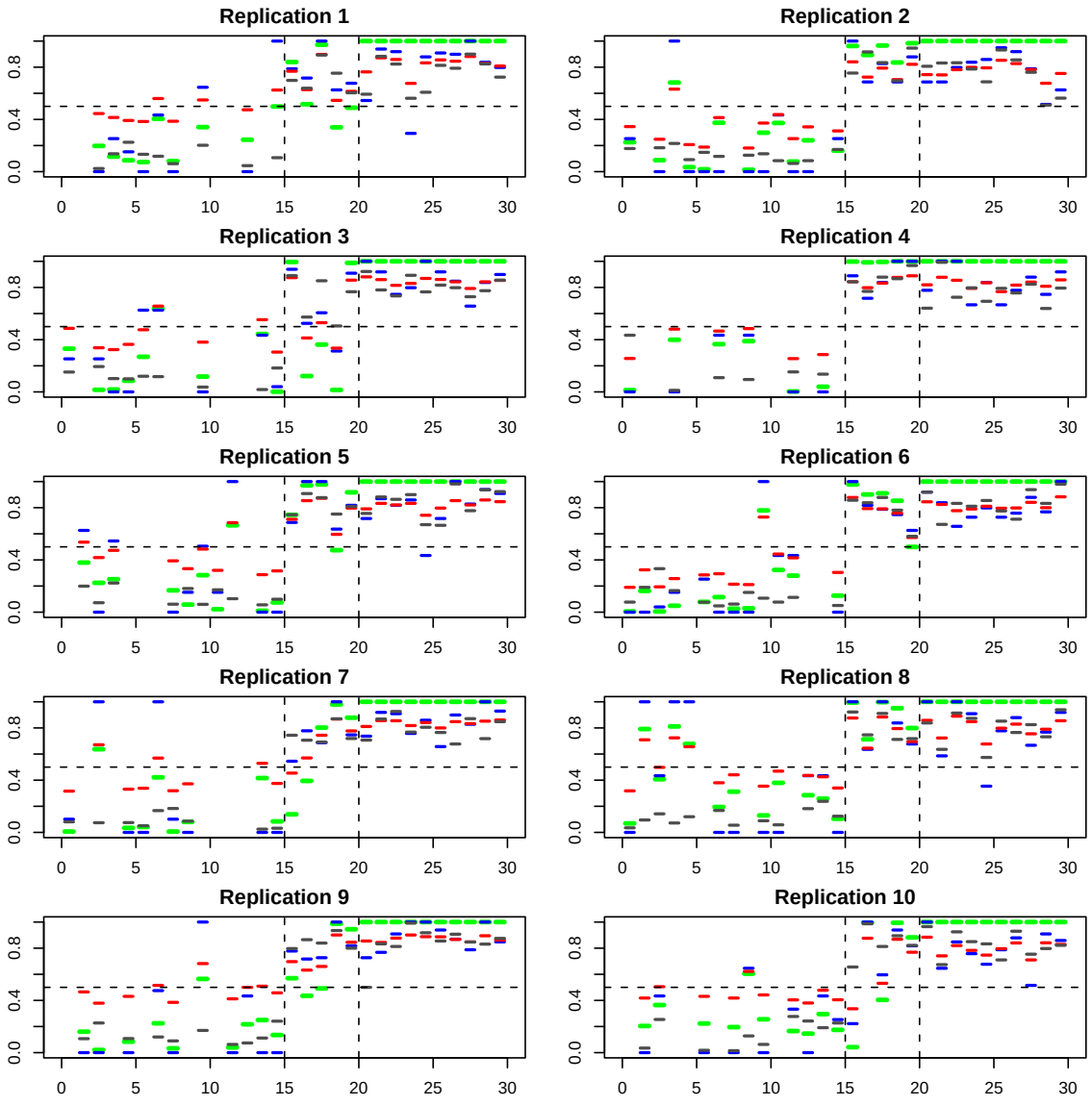
Figure 5.5: Summary of interesting model parameters for replicates of family histories. In each plot, x-axis represents the mutation index, y-axis represents the value of these parameters. In each plot, dark lines represent mutation specific penetrances used to simulate the family histories, blue lines represents the maximum likelihood estimate of the mutation specific penetrance, red lines represent posterior mean of mutation specific penetrance estimated from the hierarchical classification model and green lines represent posterior probability that the mutation is deleterious.



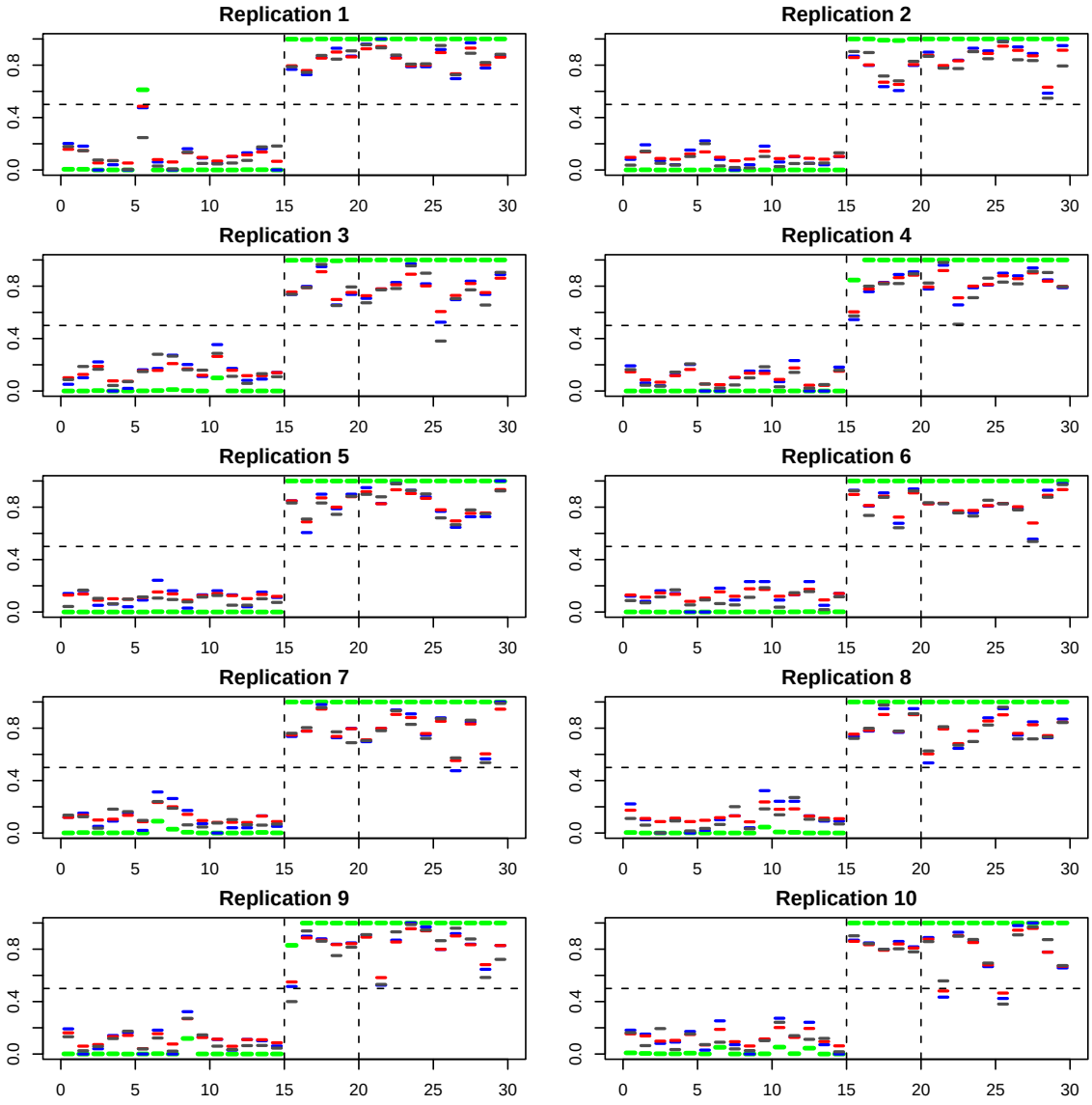
(b) Figure 5.5 continued. Sample size: 4000. Data are ascertained through progressive ascertainment rule. Likelihoods are calculated ascertainment correction.



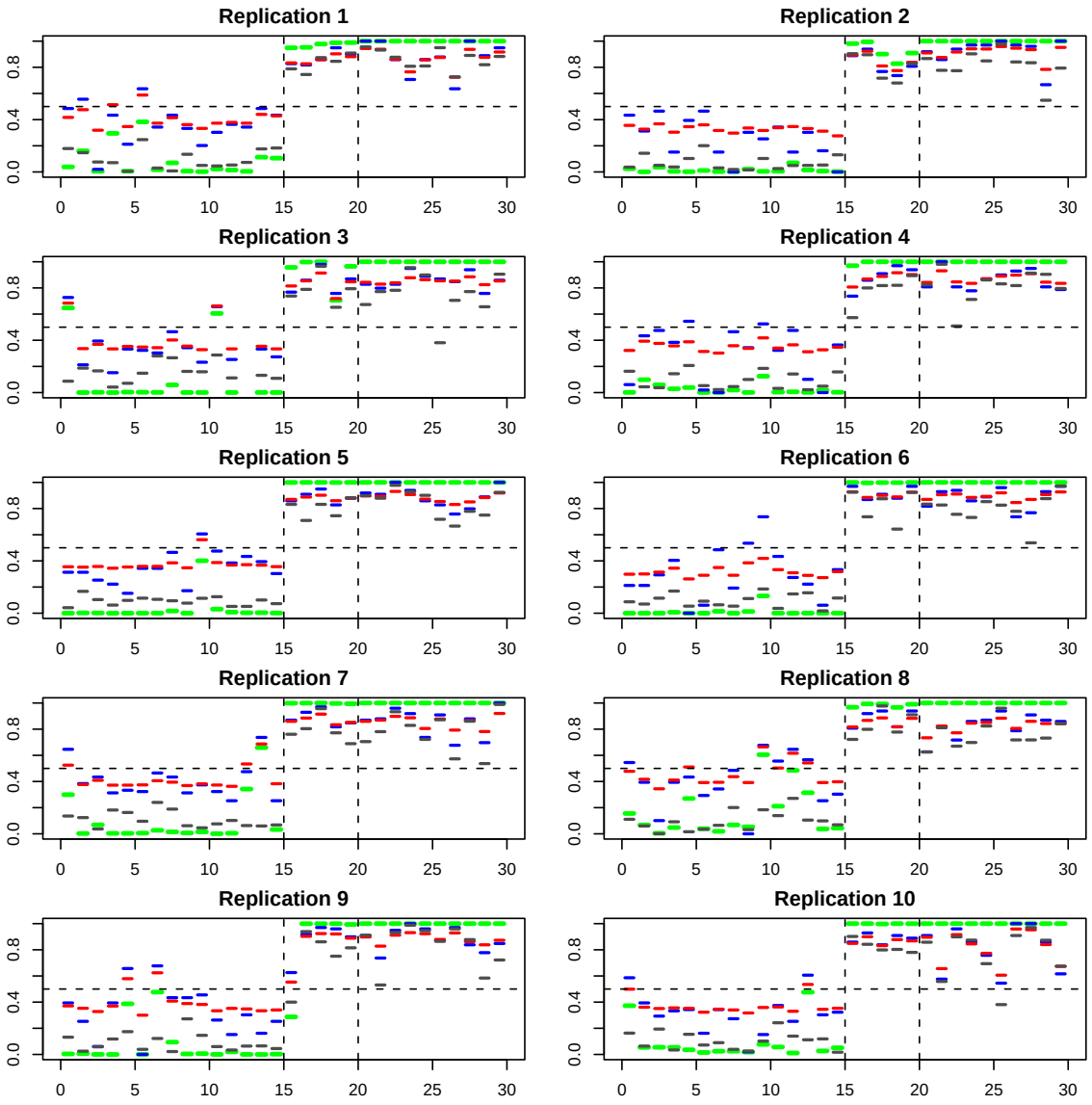
(c) Figure 5.5 continued. Sample size: 4000. Data are ascertained through affected proband ascertainment rule. Likelihoods are calculated without ascertainment correction.



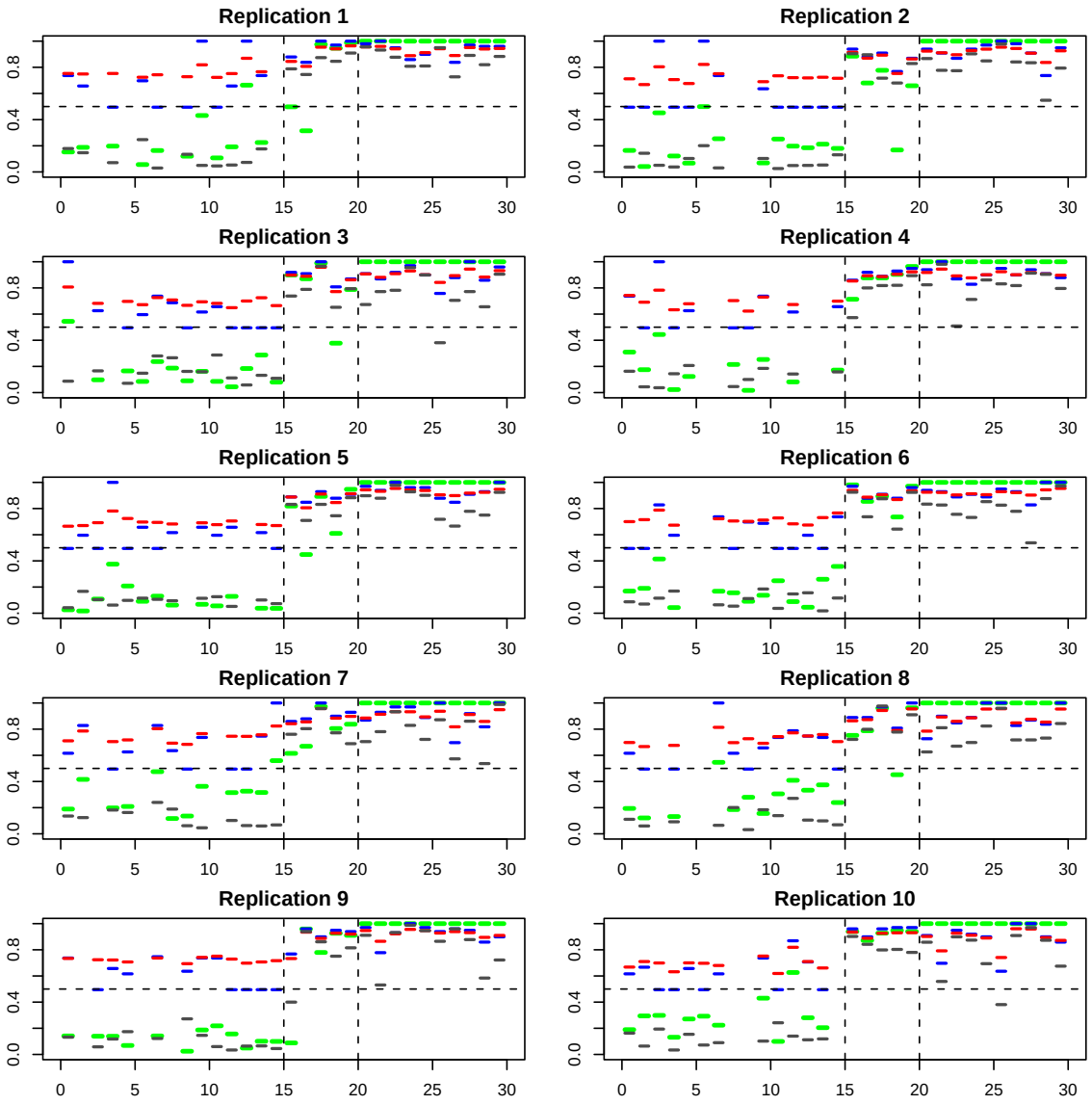
(d) Figure 5.5 continued. Sample size: 4000. Data are ascertained through affected proband ascertainment rule. Likelihoods are calculated with ascertainment correction.



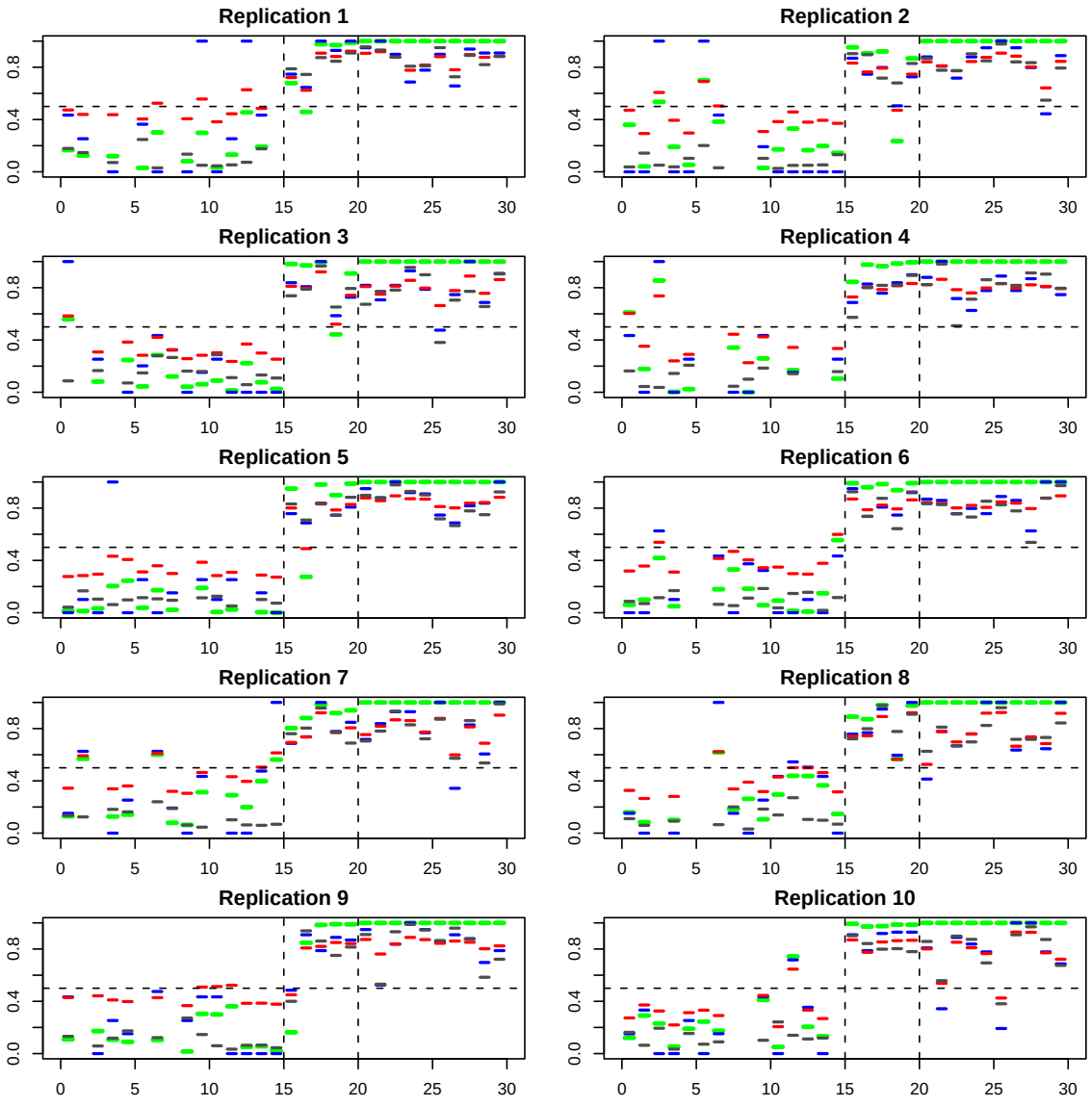
(e) Figure 5.5 continued. Sample size: 6000. Data are population based.



(f) Figure 5.5 continued. Sample size: 6000. Data are ascertained through progressive ascertainment rule. Likelihoods are calculated without ascertainment correction.



(g) Figure 5.5 continued. Sample size: 6000. Data are ascertained through affected proband ascertainment rule. Likelihoods are calculated without ascertainment correction.



(h) Figure 5.5 continued. Sample size: 6000. Data are ascertained through affected proband ascertainment rule. Likelihoods are calculated with ascertainment correction.

Chapter 6

Classifying missense mutations of BRCA genes

As demonstrated in the simulation study, the hierarchical classification model performs reasonably well in the simulated data sets. The latent variable of class reveals to be a more robust measure of missense mutations' deleteriousness than the estimated penetrance parameter, especially when the data are ascertained through certain rules. In this chapter, we apply this hierarchical classification model to study the deleteriousness of missense mutations of breast cancer susceptibility genes BRCA1 and BRCA2.

Comparing with the simulation study, complications arises from both the likelihood calculation and the classification step. For the likelihood calculation, the complications are mainly from the fact that the genetic model is more complicated. First, instead of studying mutations on one particular locus, we study mutations on two breast cancer susceptibility genes. Secondly, instead of studying one disease type, we study both breast cancer and ovarian cancer. Thirdly, instead of assuming constant penetrance over age, penetrances are assumed to be age dependent. And the size of the family histories are usually larger. Further complications come from the testing error which needed to be evaluated in this application.

In this chapter, we first discuss penetrance models essential to the likelihood calculation. Then the hierarchical classification results based on two different penetrance models will be presented.

6.1 Penetrance models and the likelihood calculation

From previous section, we can see that the data do show some difference in terms of number of cases and age of diagnosis for families of probands with different genetic test results. However, to be able to detect the difference at the level of individual mutations requires the model to be able to recover the subtle difference exhibit in the limited family histories associated with each mutation. As described earlier, to do that, we use a full likelihood model which counts in not only the number of cases in the family, the age of diagnosis, but also the relationship among cases and the size of the family history. And the differences in family histories of specific mutations are expected to be exhibited through the estimated penetrances.

As family histories in our study are all moderate in size, specifying a fully parameterized penetrance model is likely to overwhelm the data and miss the main differences shown in the families. Hence, a one-parameter penetrance model would be ideal if it is able to capture the main differences shown in the family histories and at the same time being able to capture the age dependent nature of these inherited diseases. In this application, we consider three penetrance models with different focuses.

First, similar to the continuous penetrance model discussed in Chapter 3, it is assumed that the mutation specific penetrance is the weighted average of known penetrance of deleterious mutations and the phenocopying rate of the disease estimated from previous studies. Let $\rho_{m,s}(a)$ denote the mutation specific penetrance

of disease s for mutation m, $\rho_s(a)$ denote the penetrance of known deleterious mutations of disease s estimated from previous studies, $\phi_s(a)$ denote the phenocopying rate estimated from previous studies. The model can be written as:

$$\rho_{m,s}(a) = (1 - \gamma_m)\phi_s(a) + \gamma_m \rho_s(a), \tag{6.1}$$

where $\gamma_m \in [0, 1]$ is the mutation specific weight. It describes how mutation modifies the penetrance of disease and is expected to capture the degree of deleteriousness of the mutation. With this penetrance model, the parameter γ_m has an meaningful interpretation. When $\gamma_m = 0$, the mutation is indistinguishable from a benign polymorphism; when $\gamma_m = 1$, it is indistinguishable from known deleterious mutations. Parameter γ_m itself is assumed to be independent of age, and cancer site to accommodate limited sample size. However, the overall mutation specific penetrance is age dependent.

In penetrance model 6.1, the mutation specific penetrance are assumed to be at most as bad as the known deleterious penetrance or as benign as the phenocopy rate. It may be best to be used when the size of the family histories are relatively large where the variation of the penetrance is relatively small. To be able to consider the possibility that the mutation specific penetrance may be even worse than the penetrance of known deleterious mutations, we specify the second penetrance model. In this model, instead of using known penetrance of deleterious mutations estimated from population studies as the upper limit of the mutation specific penetrances, we use the asymptotic limit of penetrance estimated based on the model described in Parmigiani et al. (1998) and Iversen et al. (2000) as the upper limit. The asymptotic limit is obtained by assuming that every BRCA1 carrier will ultimately be affected with breast cancer, and penetrance for ovarian cancer and of BRCA2 mutations are adjusted accordingly. The penetrance model is written as

$$\rho_{m,s}(a) = (1 - \gamma_m)\phi_s(a) + \gamma_m \rho^*_s(a), \tag{6.2}$$

where parameter $\rho_s^*(a)$ is the asymptotic penetrance. This penetrance model allow the mutation specific penetrance as bad as the asymptotic penetrance limit of known deleterious mutations when $\gamma_m = 1$.

Although in penetrance model 6.2, more possibility are considered through the penetrance specification, the mutation specific penetrance is still confined to known penetrances. A greater flexibility can be achieved through our penetrance model 6.3, which is written as the following:

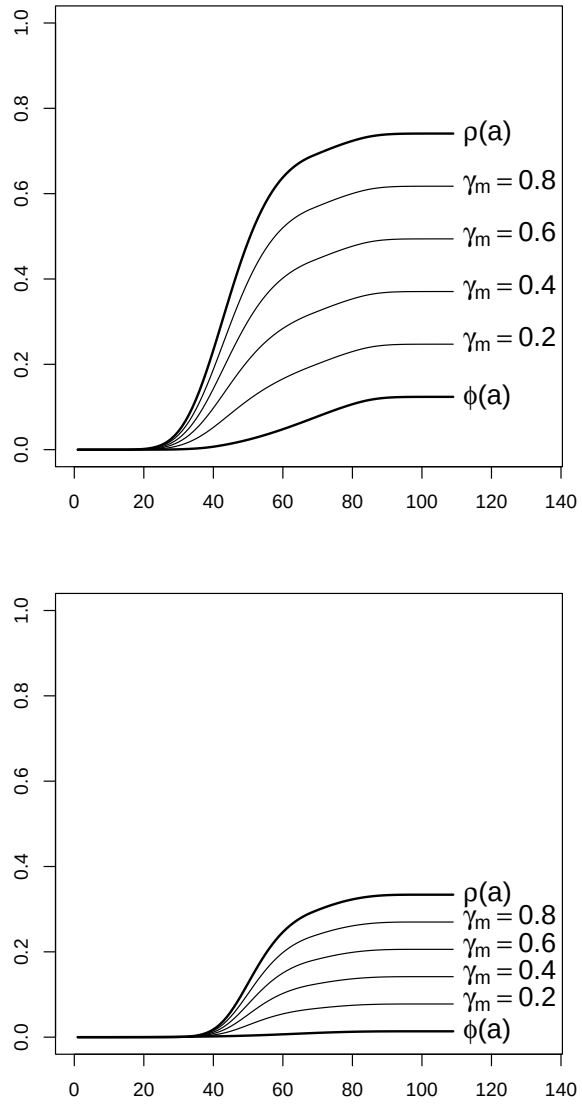
$$\rho_{m,s}(a) = 1 - (1 - \rho_s(a))^{\frac{\gamma_m}{(1-\gamma_m)}}. \tag{6.3}$$

Under this penetrance model, only known penetrance of the deleterious mutations are used. With different values of γ_m, the allowed effect of the mutation can be lethal with mutation specific penetrance $\rho_{m,s}(a) = 0$ when $\gamma_m = 0$ or can be very protective with $\rho_{m,s}(a) = 1$ when $\gamma_m = 1$.

Figure 6.1 and 6.2 shows the mutation specific penetrance curves of breast and ovarian cancers for BRCA1 and BRCA2 mutation carriers under these three different penetrance models. Because of the high risk character exhibit in our data, using penetrance model 6.1 is likely to have difficulty in discriminating the negative and positive families. In the application follows, we focus on the latter two penetrance models.

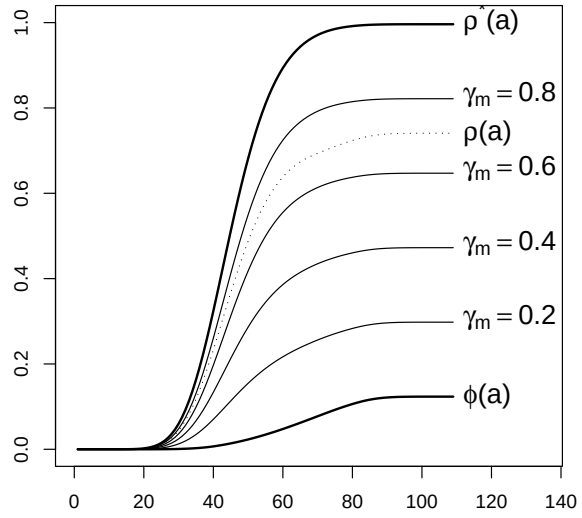
6.2 Hierarchical Classification Results

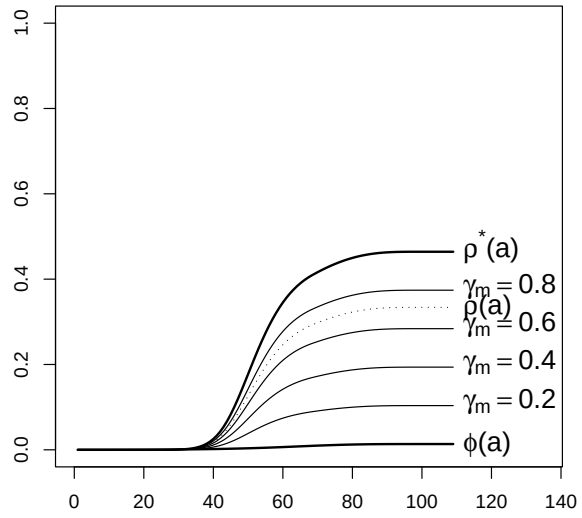
The conditional likelihoods are incorporated into the hierarchical classification model described in Chapter 4. Vague yet proper priors are used for parameters π and ξ. For parameter π we used beta distribution with parameters $(2, 2)$. For parameter ξ, the proportion of false negatives of the genetic test, we chose the prior following the discussion in Chapter 4. Specifically, we assume that the sensitivity of the genetic



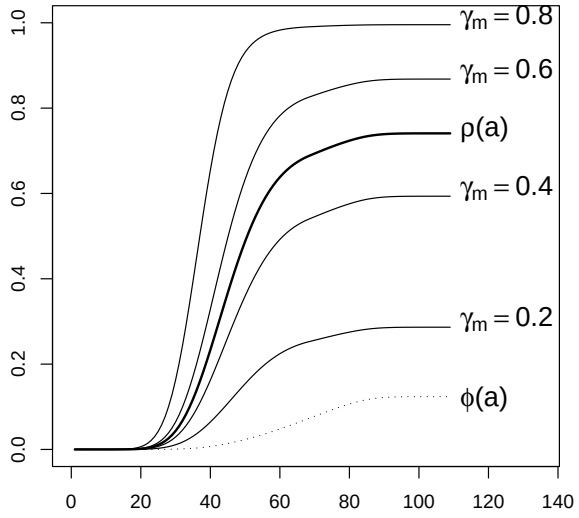
(a) $\rho_m(a) = (1 - \gamma_m)\phi(a) + \gamma_m \rho(a)$

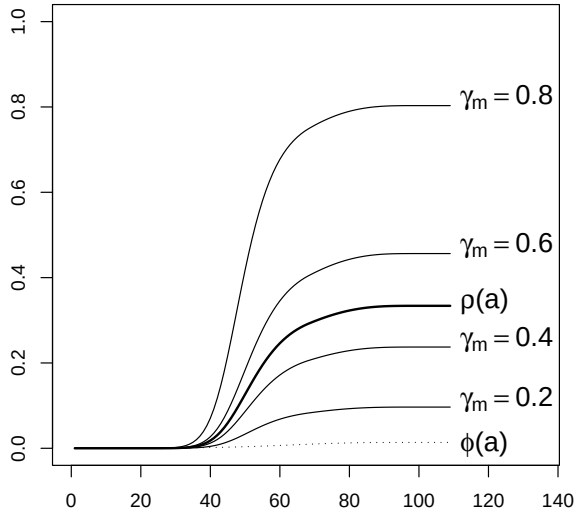
Figure 6.1: Penetrance curves of BRCA1 carriers under three penetrance models. Top plots are for penetrances of breast cancer. And bottom plots are for penetrances of ovarian cancer.



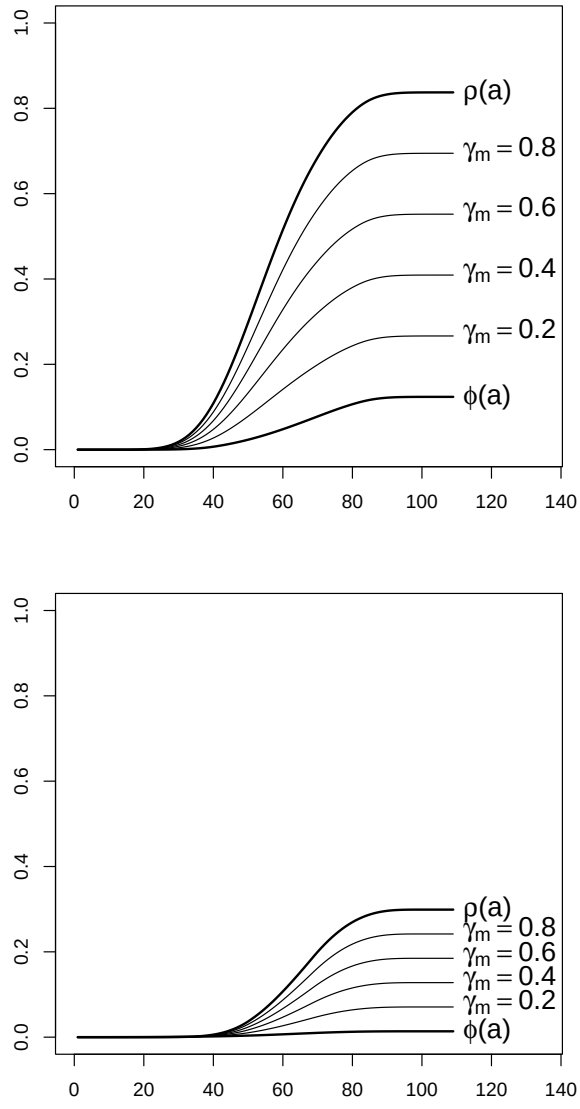


(b) $\rho_m(a) = (1 - \gamma_m)\phi(a) + \gamma_m \rho^*(a)$



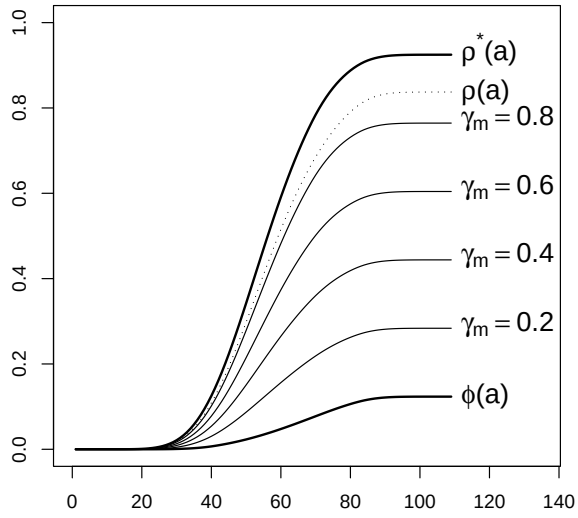


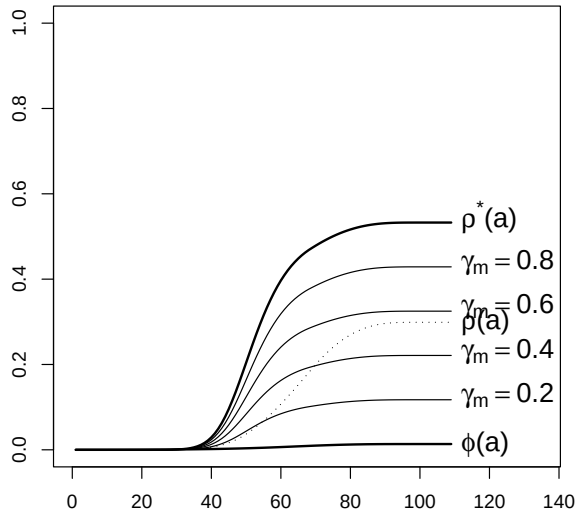
(c) $\rho_m(a) = 1 - (1 - \rho(a))^{\frac{\gamma_m}{(1-\gamma_m)}}$



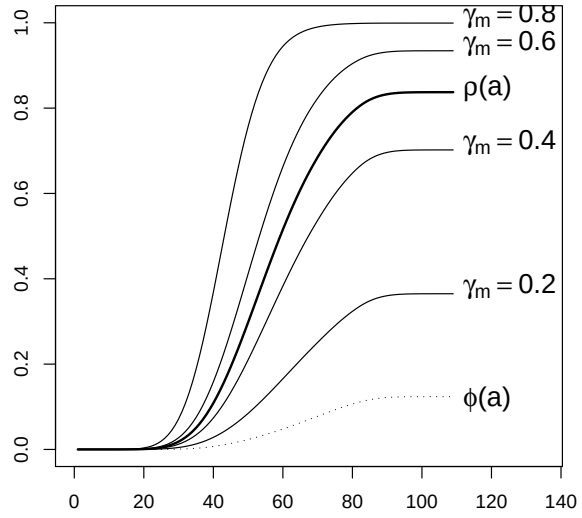
(a) $\rho_m(a) = (1 - \gamma_m)\phi(a) + \gamma_m \rho(a)$

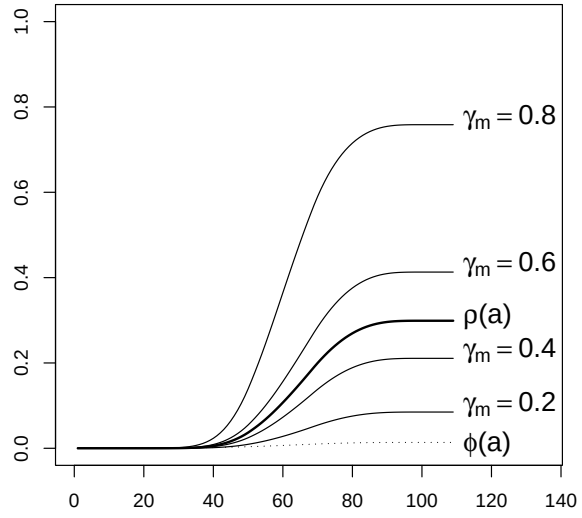
Figure 6.2: Penetrance curves of BRCA2 carriers under three penetrance models. Top plots are for penetrances of breast cancer. And bottom plots are for penetrances of ovarian cancer.





(b) $\rho_m(a) = (1 - \gamma_m)\phi(a) + \gamma_m \rho^*(a)$





(c) $\rho_m(a) = 1 - (1 - \rho(a))^{\frac{\gamma_m}{(1-\gamma_m)}}$

test (SSCP) was about 65% to 75% and the prevalence of mutations in the sample is between 30% and 50%. Using Bayesian rule, a relatively conservative choice for ξ would be around 0.7. Hence, a beta distribution with parameters $(7, 3)$ was used. For hyper parameters α_1 β_1 and α_2 β_2, we used uniform distributions ranging from 1 to 100 as hyper prior. For likelihoods based on each of the penetrance models, marginal posterior of model parameters are sampled using the Gibbs sampling scheme outlined in the Chapter 4.

We sampled the posterior distribution of model parameters using the Gibbs sampling scheme outlined in Chapter 4 given each of the three penetrance parameterization described earlier. For each, we carried out 200000 iterations, retaining every 10th. For each parameter, we obtain 20000 samples. Trace plots of the posterior samples appeared stationary and the chains passed the Heidelberger and Welch test for stationarity (Heidelberger and Welch 1983). Furthermore, the samples are sufficient to estimate the 2.5 percentile within the accuracy of 0.5% with probability 95% based on Raftery and Lewis diagnostic (Raftery and Lewis 1996).

Histograms of model parameters show that the data can provide some information and improve our understanding of the model parameters, especially in the case of using penetrance model 6.3. Difference between marginal posterior of parameters $(\alpha_1,\ \beta_1)$ and $(\alpha_2,\ \beta_2)$ shows that there is difference between the penetrance non-deleterious polymorphisms and the penetrances of deleterious mutations, and the model is able to capture the difference. The histograms of model parameters sampled with the likelihoods based on the survival form of the penetrance model along with their priors are plotted in Figure 6.3. Similarly, histograms of model parameters sampled with the likelihoods based on the weighted average form of the penetrance model along with their priors are plotted in Figure 6.4. From Figure 6.4, we can see that less information available to estimate parameters α_1 and α_2. As we will

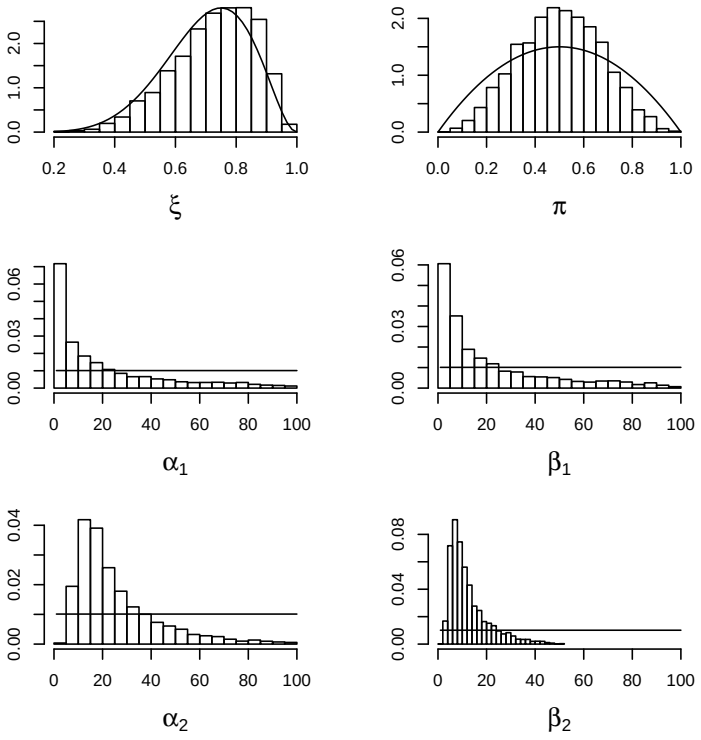


Figure 6.3: Histograms of the posterior samples of model parameters. Likelihoods calculation is based on penetrance model $\rho_m(a) = 1 - (1 - \rho(a))^{\frac{\gamma_m}{(1-\gamma_m)}}$. Prior distribution of model parameters are plotted in solid lines.

see later, this reflect the limitation of the weighted form of the penetrance model. The variation of the penetrances based on our sample is greater than this penetrance allows, especially at the deleterious end. And it limits the model's ability to discern the deleterious and non-deleterious mutations.

The hierarchical model also estimates the posterior of penetrance parameters γ_m for deleterious and missense mutations. For those tested negative, as we assumed that they may either carry a common non-deleterious polymorphism or a common deleterious mutation, the penetrance parameters for these two assumed variants, denoted by γ_{nd} and γ_d, respectively, are also estimated. Figure 6.5 shows the histograms of the posterior samples of γ_{nd} and γ_d from the model using likelihoods calculated using penetrance model 6.3 as well as the one using penetrance model 6.2. Posterior

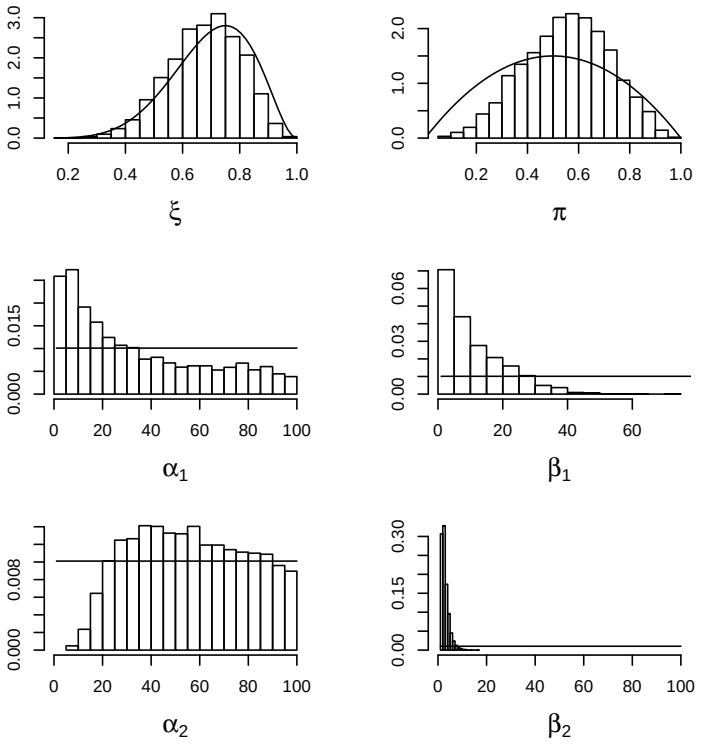


Figure 6.4: Histograms of the posterior samples of model parameters. Likelihoods calculation is based on penetrance model $\rho_m(a) = (1 - \gamma_m)\phi(a) + \gamma_m\rho^*(a)$. Prior distribution of model parameters are plotted in solid lines.

means are denoted by the vertical lines in the plots. From this plot, we can see that penetrance model 6.3 has greater discriminate power over penetrance model 6.2. In the second case, the plot shows that even with the asymptotic extreme as the upper limit, the model still has difficulty to separate the difference in deleteriousness shown in the family history at the deleterious end. The reason for this is that the family histories are ascertained in the high risk clinic, and have much higher disease rate and earlier age of diagnosis than the general population. Survival form of the penetrance shows greater promise to separate the negative families and the deleterious families. However, one need to be cautious while interpreting these penetrance estimates. As the main purpose of the model is to classify the missense mutations, the difference shown in the family histories of the negatives and those known deleterious mutations

are sufficient. There is no need to compare them with the population data as long as they are ascertained the same way. The simulation study has shown that even with ascertainment bias, the model was able to perform well in classification. However, the penetrance estimates are biased when the data are ascertained in favor of families with large number of cases and early age of diagnosis and should be treated as a relative measure. Posterior mean of γ_m's and the 90% interval are plotted in

$$\rho_m(a) = (1-\gamma_m)\phi(a) + \gamma_m \rho^*(a) \quad \rho_m(a) = 1 - (1-\rho(a))^{\frac{\gamma_m}{(1-\gamma_m)}}$$

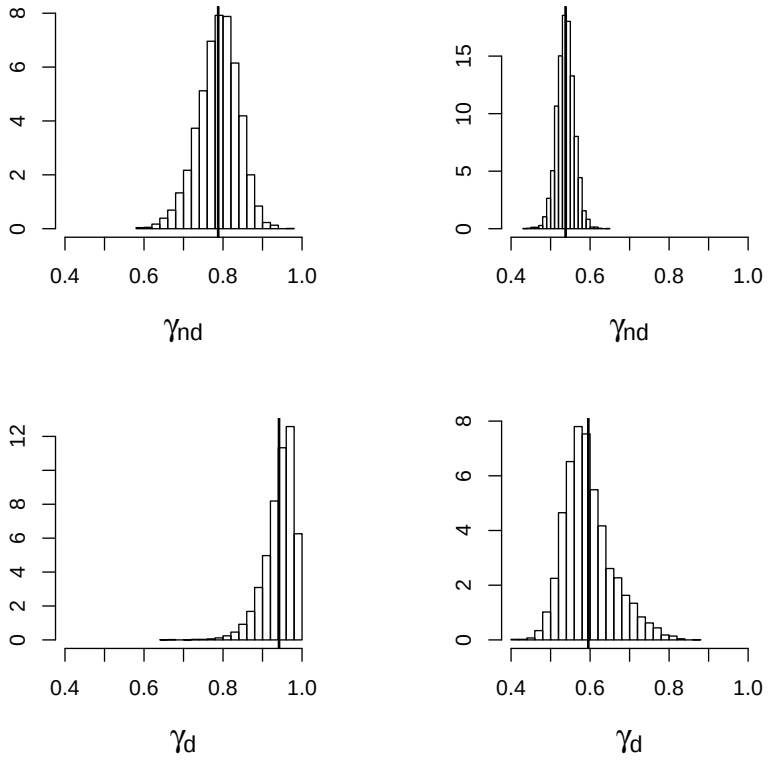


Figure 6.5: Histograms of the posterior samples of γ_{nd} and γ_d. Posterior means of these penetrance parameters are plotted in solid lines.

figure 6.6 and figure 6.7.

Posterior Probability of deleteriousness of missense mutations are plotted in figure 6.8. Mutations denoted as "*n" are unclassified variants located at the intron region of the gene. The probability of deleteriousness estimated in this sample ranges

from 0.197 to 0.86 for estimates based on the survival form of the penetrance model and it ranges from 0.168 to 0.737 for estimates based on the weighted average form of the penetrance. The estimates show good correlation under these two different penetrance models. The differences between ranks under these two penetrance models are within 5 for 25 of the 35 missense mutations. However, there do have some differences in ranking. The largest rank differences are found in mutations for which the family histories have relatively large number of cases while the cases having relatively late age of diagnosis. Mutations for which the families with a relatively large number of cases are usually ranked higher regardless of the age of diagnosis in the classification using penetrance model6.2. While the difference in age of diagnosis are better captured by the calculation based on the survival form of the penetrance model. For example, for mutations such as I481F, V1605I and mutation set M1652I and (5272+87)delT, each has one family history available and all three have comparable size. In each of these family histories, there are 4 cases (3 breast cancer and 1 bilateral breast cancer cases). The notable difference among these three family histories is the age of diagnosis of these cases. For mutation I481F, the ages of diagnosis are(47,47, 50, 59). For mutation V1605I, the ages of diagnosis are (36, 36, 38, 47). For mutation set M1625I and (5272+87)delT, the ages of diagnosis are (42, 50, 55,59). The ranks for these three mutations are 6,7,8 respectively, under the classification using "fp" penetrance model. While they rank 7, 2, 29, respectively under the classification using the survival form of the penetrance model.

After this work have been completed, we were notified by Ms. Shelly Clark in the Duke University Cancer Center that two of the missense mutations (C64Y and M1775R) in our data had been classified as deleterious missense mutations and missense mutation S1140G is actually a polymorphism. In our data, mutations S1140G and C64Y were identified in the same individual, thus the posterior probability of

deleterious 0.822 estimated from our model is likely the effects of C64Y. Mutation M1775R is identified with another intronic variation in the same proband, the estimated posterior probability for this combination is 0.86. Because intronic region do not directly affect the protein product of the gene, it is likely that this high probability is because of mutation M1775R.

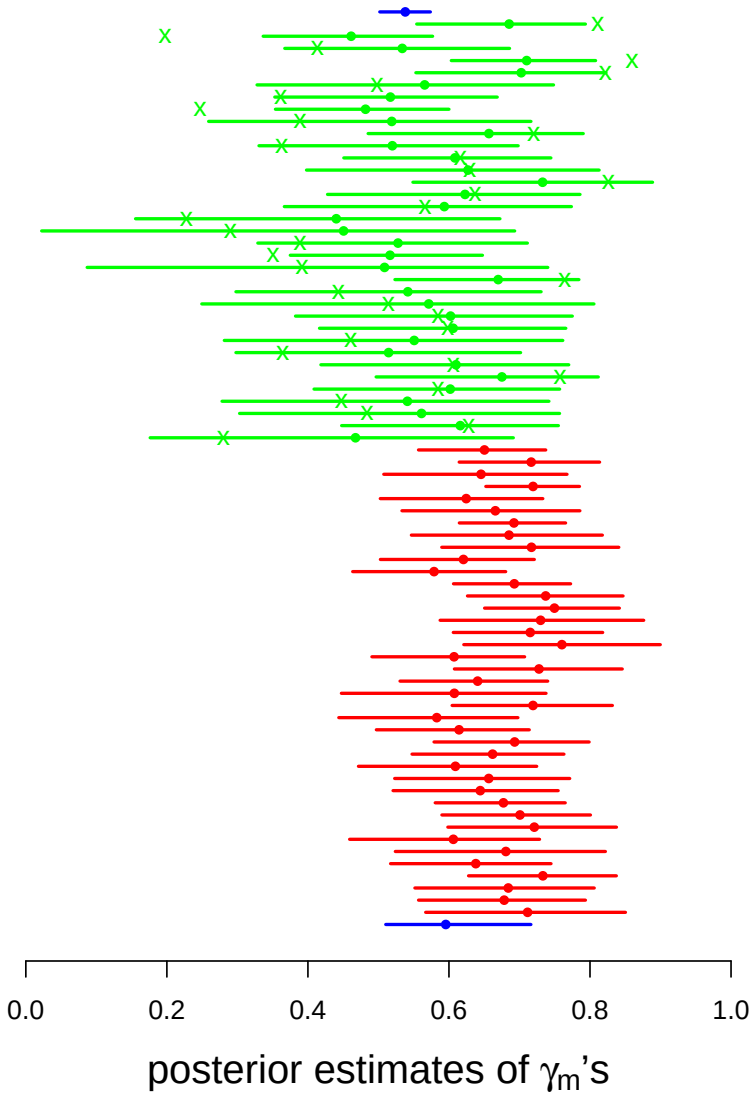


Figure 6.6: Posterior mean and 90% interval of γ_m's. Posterior probability of deleteriousness of missense mutations are denoted by "x". Estimates of missense mutations are in "green". Estimates of known deleterious mutations are in "red" color. And estimates for the two assumed polymorphisms from the negatives are in "blue". Penetrance model used here is $\rho_m(a) = 1 - (1 - \rho(a))^{\frac{\gamma_m}{(1-\gamma_m)}}$.

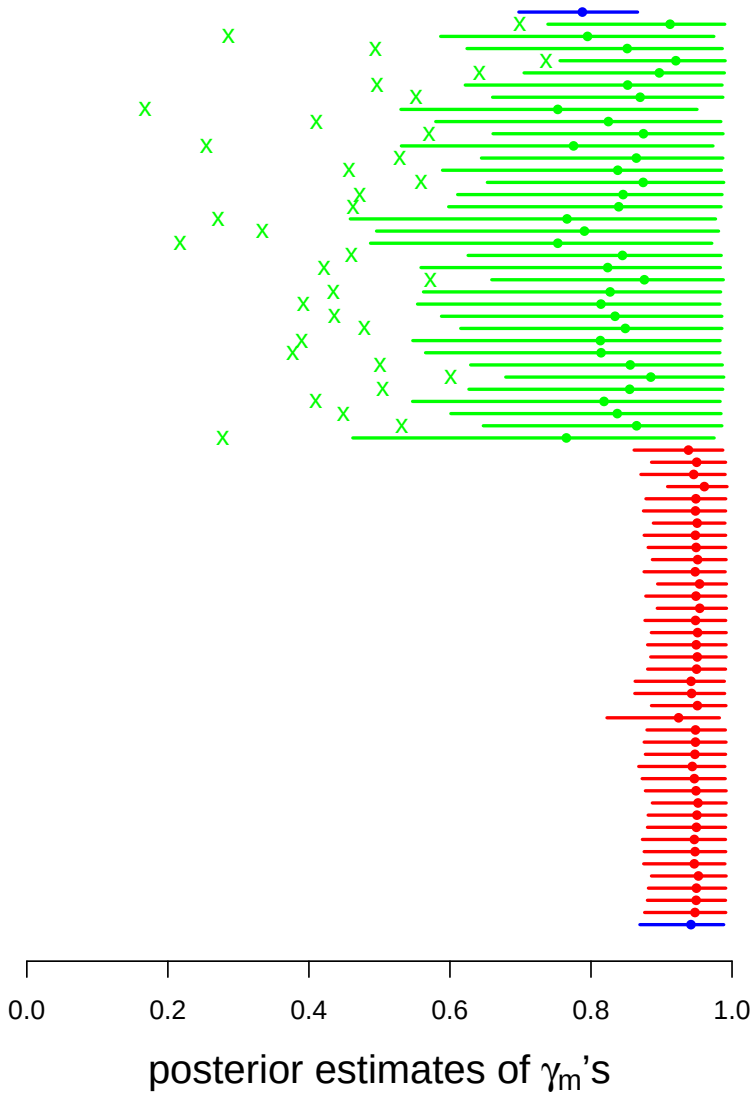


Figure 6.7: Posterior mean and 90% interval of γ_m's. Posterior probability of deleteriousness of missense mutations are denoted by "x". Estimates of missense mutations are in "green". Estimates of known deleterious mutations are in "red". And estimates for the two assumed polymorphisms from the negatives are in "blue". Penetrance model used here is $\rho_m(a) = (1 - \gamma_m)\phi(a) + \gamma_m \rho^*(a)$.

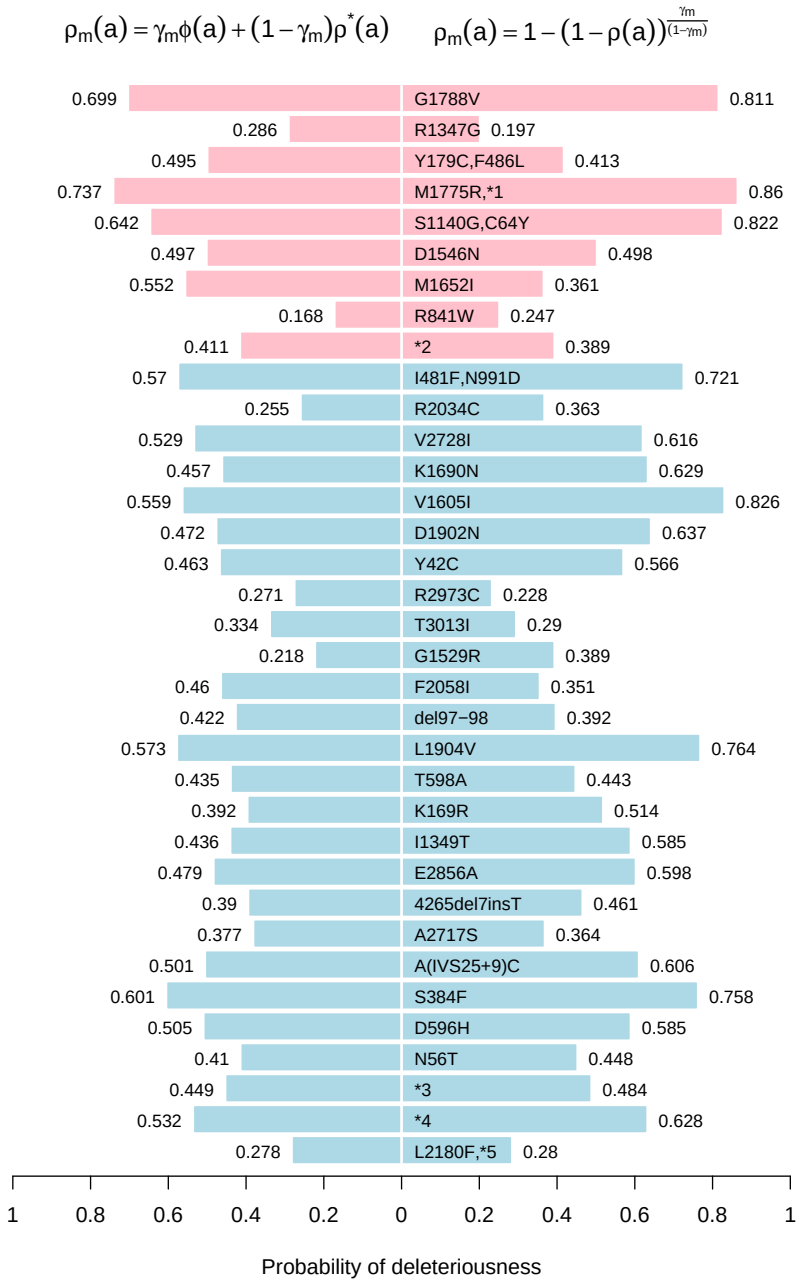


Figure 6.8: Posterior probability of deleteriousness of missense mutations. Missense mutations on BRCA1 gene are denoted in "light blue" and missense mutations on BRCA2 gene are denoted in "pink". Four mutations occur at the intron region of the gene are denoted by "*n". Actual mutations for each of them are as follows: *1, (5272+87)delT; *2, (5392+60)ins12; *3, G(7663+10)A; *4, A(4795-170)G.

Chapter 7

Discussion

The story of the functional genomics of disease genes in human cancer is still in its infancy. Yet important clinical decisions already rely on our understanding of the role of genetic variants of these genes. To make progress, we will have to rely at least partially on information coming from highly selected subgroups of individuals and families at high risk. In this study, we present statistical methods for classifying genetic variants of the BRCA genes using data from high risk families.

In this study, we have developed a Bayesian hierarchical method to study disease causality of missense mutations. This method provides a framework for evaluating the deleteriousness of the mutations all at once and allows for systematic comparison of the evidence of causality from observed family histories. This allows us to present results which are useful for both genetic counselors and molecular biologists. Such an analysis can potentially provide more insights about gene function through a family history study approach.

In the past decades, extensive family history data has been collected for the purpose of studying the association between disease and the genes has been studied. However, various studies have suggested that even for mutation carriers, there is variation in the susceptibility to disease and in the age of disease onset. The reason

for this variation may come from genetic, or environmental factors or both. There have been numerous studies focused on the environmental modifiers of the Mendelian disease genes. Few evaluated the variation of phenotypic implications caused by variants of the same gene. One of the main reasons for that is the lack of a sufficient sample for each mutation, and, the fact that current data collected from high risk clinics may lead to overestimating the effects of mutation. However, this data is informative. The method we describe can provide meaningful information on disease causality of the missense mutations because it compares these mutations to those that significantly change the function of the protein and to those that do not (benign polymorphisms). Also, the comparison is made on a homogeneous group of data collected in the same high risk clinic.

Our methods assumes that polymorphisms are not disease associated. If this assumption is not valid, the method we describe may underestimate the disease association of the missense mutations. With the availability of more pedigrees on both positive carriers and carriers of known non-deleterious polymorphisms, estimation will be more accurate. Another factor which may affect our model estimation is that our penetrance model is built on currently available penetrance and phenocopying rates. Although these estimates are the best that are available, improved estimation of these parameters is likely improve the reliability of our method.

The method described in this thesis focused on the issue of classification in presence of ascertainment bias using limited sample size. The availability of a large number of pedigrees for each missense mutation will enable more accurate estimation of the risk of disease for each mutation. Several of our model assumptions can be generalized when sample size are larger. For example, instead of using one parameter γ_m for both breast cancer as well as ovarian cancer, we can use two separate parameters. Also the model can be extended to study gene-gene interactions and

gene-environment interactions.

Appendix A

Full conditionals for Gibbs sampling

Given other parameters and the data, the distribution of ξ depends on the proportion of true non-deleterious mutations in the negative families and the prior information

$$[\xi|\Theta^{(s)}_{-\xi}, data] \sim \xi^{\sum_i I_{\{d_i^{(s)}=0\}}}(1-\xi)^{\sum_i I_{\{d_i^{(s)}=1\}}} f(\xi), \quad i = 1, \ldots, 60.$$

Similarly, the full conditional of π depends on the proportion of non-deleterious mutations in the missense group and the prior

$$[\pi|\Theta^{(s)}_{-\pi}, data] \sim \pi^{\sum_m I_{\{d_m^{(s)}=0\}}}(1-\pi)^{\sum_m I_{\{d_m^{(s)}=1\}}} f(\pi), \quad m = 61, \ldots, 76.$$

For latent variable d_m of missense mutation m, the full conditional possible situations can be written as

$$[d_m|\Theta^{(s)}_{-d_m}, data] \sim Bern(p_1)$$

where p_1 is the probability of d_m being equal to 1 and can be written as

$$p_1 = \frac{(1-\pi^{(s)})Beta(\gamma_m^{(s)}|\alpha_1^{(s)}, \beta_1^{(s)})}{\pi^{(s)}Beta(\gamma_m^{(s)}|\alpha_0^{(s)}, \beta_0^{(s)}) + (1-\pi^{(s)})Beta(\gamma_m^{(s)}|\alpha_1^{(s)}, \beta_1^{(s)})} \quad m = 61, \ldots, 76.$$

The full conditional of latent variable d_i for the negatives can be written as

$$[d_i|\Theta^{(s)}_{-d_i}, data] \sim Bern(p_2)$$

where d_i equals to 1 and p_2 is the probability of proband i has a false- negative test result and can be written as

$$p_2 = \frac{(1-\xi)^{(s)} P(f_i|g_{0i}, \gamma_0^{(s)})}{\xi^{(s)} P(f_i|g_{00}, \gamma_0^{(s)}) + (1-\xi^{(s)}) P(f_i|g_{0i}, \gamma_{m=i}^{(s)})} \quad i = 1, \ldots, 60.$$

The full conditionals for the γ's are different for different mutations. For the assumed non-deleterious mutation "0",it is

$$[\gamma_0|\Theta_{-\gamma_0}^{(s)}, data] \sim \begin{cases} [\prod_{i\in\{d_i^{(s)}=0\}} P(f_i|g_{00}, \gamma_0)] Beta(\gamma_0|\alpha_0^{(s)}, \beta_0^{(s)}) & \text{if } \sum_{i=1}^{60} I_{\{d_i^{(s)}=0\}} > 0 \\ Beta(\gamma_0|\alpha_0^{(s)}, \beta_0^{(s)}) & \text{if } \sum_{i=1}^{60} I_{\{d_i^{(s)}=0\}} = 0 \end{cases}$$

For the assumed deleterious mutations of the negatives may carry, it is

$$[\gamma_{m=i}|\Theta_{-\gamma_{m_i}^{(s)}}, data] \sim \begin{cases} P(f_i|g_{0i}, \gamma_{m=i}) Beta(\gamma_{m=i}|\alpha_1^{(s)}, \beta_1^{(s)}) & \text{if } d_i^{(s)} = 1 \\ Beta(\gamma_{m=i}|\alpha_1^{(s)}, \beta_1^{(s)}) & \text{if } d_i^{(s)} = 0 \end{cases}$$

where $i = 1, \ldots, 60$. For the missense mutations ($m = 61, \ldots, 76$), we have

$$[\gamma_m|\Theta_{-\gamma_m}^{(s)}, data] \sim P(f_m|g_{0m}, \gamma_m)[Beta(\gamma_m|\alpha_0^{(s)}, \beta_0^{(s)})]^{(1-d_m^{(s)})}[Beta(\gamma_m|\alpha_1^{(s)}, \beta_1^{(s)})]^{d_m^{(s)}}.$$

For the mutations in the positive group($m = 77, \ldots, 92$), we have

$$[\gamma_m|\Theta_{-\gamma_m}^{(s)}, data] \sim P(f_m|g_{0m}, \gamma_m)[Beta(\gamma_m|\alpha_1^{(s)}, \beta_1^{(s)})].$$

The full conditionals of the α's and β's can also be written out as the following

$$[\alpha_0|\Theta_{-\alpha_0}^{(s)}, data] \propto f(\alpha_0) \frac{\Gamma(\alpha_0 + \beta_0^{(s)})}{\Gamma(\alpha_0)} (\gamma_0^{(s)})^{\alpha_0 - 1} [\prod_{m:\{d_m^{(s)}=0\}} \frac{\Gamma(\alpha_0 + \beta_0^{(s)})}{\Gamma(\alpha_0)} (\gamma_m^{(s)})^{\alpha_0 - 1}]$$

$$[\beta_0|\Theta_{-\beta_0}^{(s)}, data] \propto f(\beta_0) \frac{\Gamma(\alpha_0^{(s)} + \beta_0)}{\Gamma(\beta_0)} (1-\gamma_0^{(s)})^{\beta_0 - 1} [\prod_{m:\{d_m^{(s)}=0\}} \frac{\Gamma(\alpha_0^{(s)} + \beta_0)}{\Gamma(\beta_0)} (1-\gamma_m^{(s)})^{\beta_0 - 1}]$$

$$[\alpha_1|\Theta^{(s)}_{-\alpha_1}, data] \propto f(\alpha_1)[\prod_{m:\{77,\ldots,92 ord_m^{(s)}=1\}} \frac{\Gamma(\alpha_1+\beta_1^{(s)})}{\Gamma(\alpha_1)}(\gamma_m^{(s)})^{\alpha_1-1}]$$

$$[\beta_1|\Theta^{(s)}_{-\beta_1}, data] \propto f(\beta_1)[\prod_{m:\{77,\ldots,92 ord_m^{(s)}=1\}} \frac{\Gamma(\alpha_1^{(s)}+\beta_1)}{\Gamma(\beta_1)}(1-\gamma_m^{(s)})^{\beta_1-1}]$$

Bibliography

Abel, L. and G. E. Bonney (1990). A time-dependent logistic hazard function for modeling variable age of onset in analysis of familial diseases. *Genet. Epidemiol. 7*(6), 391–407.

Andersen, T. (1996). Genetic heterogeneity in breast cancer susceptibility. *Acta. Oncol. 35(4)*, 407.

Barker, D. F., E. R. Almeida, G. Casey, P. R. Fain, S.-Y. Liao, I. Masunaka, B. Noble, T. Kurosaki, and H. Anton-Culver (1996). BRCA1 R841W: A strong candidate for a common mutation with moderate phenotype. *Genet. Epidemiol. 13*, 595–604.

Berry, D. A., G. Parmigiani, J. Sanchez, J. Schildkraut, and E. Winer (1997). Probability of carrying a mutation of breast-ovarian cancer gene BRCA1 based on family history. *J Natl Cancer Inst 89*, 227–238.

BIC (1997). National institutes of health, breast cancer information core: An open access on-line breast cancer mutation data base. http://www.nhgri.nih.gov/Intramural_research/Lab_transfer/Bic/.

Cannings, C., E. A. Thompson, and M. H. Skolnick (1978). Probability functions on complex pedigrees. *Adv. Appl. Probab. 10*(1), 26–61.

Chen, W. J., S. V. Faraone, and M. T. Tsuang (1992). Estimating age at onset distributions - a review of methods and issues. *Psychiatr. Genet. 2*(4), 219–238.

Claus, E. B., N. Risch, and W. D. Thompson (1991). Genetic analysis of breast cancer in the cancer and steroid hormon study. *Am. J. Hum. Genet. 48*, 232–242.

Cotton, R. G. H. and C. R. Scriver (1998). Proof of "disease causing" mutation.

Hum. Mutat. 12(1), 1–3.

Easton, D. F., D. T. Bishop, D. Ford, G. P. Crockford, and T. B. C. L. Consortium (1993). Genetic linkage analysis in familial breast and ovarian cancer: Results from 214 families. *Am J Human Genetics 52*, 678–701.

Easton, D. F., D. Ford, and D. T. *et al.* . Bishop (1995). Breast and ovarian-cancer incidence in BRCA1-mutation carriers. *Am. J. Hum. Genet. 56*, 265–271.

Easton, D. F., L. Steele, et al. (1997). Cancer risks in two large breast cancer families linked to BRCA2 on chromosome 13q12-13. *Am. J. Hum. Genet. 61*, 120–128.

Elston, R. C. (1995). Ťwixt cup and lip: How intractable is the ascertainment problem? *American Journal of Human Genetics 56*, 15–17.

Elston, R. C. and V. T. George (1989). Age of onset, age at examination, and other covariates in the analysis of family data. *Genet. Epidemiol. 6*(1), 217–220.

Elston, R. C. and E. Sobel (1979). Sampling considerations in the gathering and analysis of pedigree data. *American Journal of Human Genetics 31*, 62–69.

Elston, R. C. and J. Stewart (1971). A general model for the genetic analysis of pedigree data. *Human Heredity 21*, 523–542.

Ewens, W. J. and N. C. E. Shute (1986). A resolution of the ascertainment sampling problem .1. theory. *Theor. Popul. Biol. 30*(3), 388–412.

Fearon, E. R. (1997). Human cancer syndromes: clues to the origin and nature of cancer. *Science 278*(5340), 1043–1050.

Fisher, R. A. (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals of Eugenics 6*, 13–25.

Ford, D. and D. F. Easton (1995). The genetics of breast and ovarian cancer. *Br. J. Cancer 72*, 805.

Ford, D., D. F. Easton, M. Stratton, S. Narod, D. Goldgar, P. Devilee, D. T. Bishop, et al. (1998). Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. *Am. J. Hum. Genet 62*, 676–689.

Gail, M. H., D. Pee, J. Benichou, and R. Carroll (1999). Designing studies to estimate the penetrance of an identified autosomal dominant mutation: cohort, case-control, and genotyped-proband designs. *Genet. Epidemiol. 16*(1), 15–39.

Gauderman, W. J. and D. C. Thomas (1994). Censored survival models for genetic epidemiology - a gibbs sampling approach. *Genet. Epidemiol. 11*(2), 171–188.

George, V. T. and R. C. Elston (1991). Ascertainment: An overview of the classical segregation analysis model for independent sibships. *Biometrics Journal 33*, 741–753.

Grann, V. R., K. S. Panageas, W. Whang, K. H. Antman, and A. I. Neugut (1998). Decision analysis of prophylactic mastectomy and oophorectomy in brca1-positive or brca2-positive patients. *J. Clin. Oncol. 16*(3), 979–985.

Hall, J. M., M. K. Lee, B. Newman, J. E. Morrow, L. A. Anderson, B. Huey, and M. C. King (1990). Linkage of early-onset familial breast cancer to chromosome 17q12. *Science 250*, 1684–1689.

Hayes, F., C. Cayanan, D. Barilla, and A. N. A. Monteiro (2000). Functional assay for brca1: Mutagenesis of the cooh-terminal region reveals critical residues for transcription activation. *Cancer Res. 60*(9), 2411–2418.

Heidelberger, P. and P. Welch (1983). Simulation run length control in the presence of an initial transient. *Operations Research 31*, 1109–1144.

Hoskins, K. F., J. E. Stopfer, K. A. Calzone, S. D. Merajver, T. R. Rebbeck, J. E. Garber, and B. L. Weber (1995). Assessment and counseling for women with a

family history of breast cancer a guide for clinicians. *JAMA 273*, 577–585.

Iversen, Jr, E. S., G. Parmigiani, D. A. Berry, and J. Schildkraut (2000). Genetic susceptibility and survival: Application to breast cancer. *Journal of the American Statistical Association 95*, 28–42.

Johannesdottir, G., J. Gudmundsson, J. T. Bergthorsson, A. Arason, B. A. Agnarsson, G. Eiriksdottir, O. T. Johannsson, A. Borg, S. Ingvarsson, D. F. Easton, V. Egilsson, and R. B. Barkardottir (1996). High prevalence of the 999del5 mutation in icelandic breast and ovarian cancer patients. *Cancer Res. 56*(16), 3663–3665.

Karunaratne, P. M. and R. C. Elston (1998). Likelihood calculation conditional on observed pedigree structure. *American Journal of Human Genetics 62*, 738–739.

Lander, E. S. and P. Green (1987). Construction of multilocus genetic-linkage maps in humans. In *Proc. Natl. Acad. Sci. U. S. A.*, pp. 2363–2367. Published as Proc. Natl. Acad. Sci. U. S. A., volume 84, number 8.

Lange, K. and R. C. Elston (1975). Extension to pedigree analysis. *Human Heredity 25*, 95–105.

Li, H. Z. and E. Thompson (1997). Semiparametric estimation of major gene and family-specific random effects for age of onset. *Biometrics 53*(1), 282–293.

Miki, Y., J. Swenson, D. Shattuck-Eidens, P. A. Futreal, K. Harshman, S. Tavtigian, et al. (1994). A strong candidate for the breast and ovarian cancer susceptibility: gene BRCA1. *Science 266*, 66–71.

Oddoux, C., J. P. Struewing, C. M. Clayton, et al. (1996). The carrier frequency of the BRCA2 6174delT mutation among ashkenazi jewish individuals is approximately 1%. *Nat. Genet. 14*, 188–190.

Parmigiani, G., D. A. Berry, and O. Aguilar (1998). Determining carrier probabilities for breast cancer susceptibility genes BRCA1 and BRCA2. *Am. J. Hum. Genet. 62*, 145–158.

Petersen, G. M., G. Parmigiani, and D. Thomas (1998). Missense mutations in disease genes: A Bayesian approach to evaluate causality. *American Journal of Human Genetics 62(6)*, 1516–1524.

Ponder, B. (1997). Genetic testing for cancer risk. *Science 278*(5340), 1050–1054.

Raftery, A. E. and S. M. Lewis (1996). Implementing mcmc. In W. R. Gilks, S. Richardson, and D. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*, London, pp. 115–127. Chapman and Hall.

Rahman, N. and M. R. Stratton (1998). The genetics of breast cancer susceptibility. *Annu. Rev. Genet. 32*, 95–121.

Roa, B. B., A. A. Boyd, K. Volcik, and C. S. Richards (1996). Ashkenazi Jewish population frequencies for common mutations in BRCA1 and BRCA2. *Nature Genetics 14*, 185–187.

Shute, N. C. E. and W. J. Ewens (1988). A resolution of the ascertainment sampling problem .2. generalizations and numerical results. *Am. J. Hum. Genet. 43*(4), 374–386.

Stefansson, H., E. Sigurdsson, S. V, et al. (2002). Neuregulin 1 and susceptibility to schizophrenia. *Am. J. Hum. Genet. 71*, 877–892.

Struewing, J. P., P. Hartge, S. Wacholder, S. M. Baker, M. Berlin, M. Mcadams, M. M. Timmerman, L. C. Brody, and M. A. Tucker (1997). The risk of cancer associated with specific mutations of brca1 and brca2 among ashkenazi jews. *N. Engl. J. Med. 336*(20), 1401–1408.

Syngal, S., D. Schrag, M. Falchuk, N. Tung, F. A. Farraye, D. Chung, M. Wright,

A. Whetsell, G. Miller, and J. E. Garber (2000). Phenotypic characteristics associated with the APC gene I1307K mutation in Ashkenazi Jewish patients with colorectal polyps. *JAMA 284*, 857–860.

Tavtigian, S. V., J. Simard, J. Rommens, F. Couch, D. Shattuckeidens, S. Neuhausen, S. Merajver, S. Thorlacius, K. Offit, D. Stoppalyonnet, C. Belanger, R. Bell, S. Berry, R. Bogden, Q. Chen, T. Davis, M. Dumont, C. Frye, T. Hattier, S. Jammulapati, T. Janecki, P. Jiang, R. Kehrer, J. F. Leblanc, J. T. Mitchell, J. Mcarthurmorrison, K. Nguyen, Y. Peng, C. Samson, M. Schroeder, S. C. Snyder, L. Steele, M. Stringfellow, C. Stroup, B. Swedlund, J. Swensen, D. Teng, A. Thomas, T. Tran, T. Tran, M. Tranchant, J. Weaverfeldhaus, A. K. C. Wong, H. Shizuya, J. E. Eyfjord, L. Cannonalbright, F. Labrie, M. H. Skolnick, B. Weber, A. Kamb, and D. E. Goldgar (1996). The complete brca2 gene and mutations in chromosome 13q-linked kindreds. *Nature Genet. 12*(3), 333–337.

Vallon-Christersson, J., C. Cayanan, K. Haraldsson, N. Loman, J. T. Bergthorsson, K. Brondum-Nielsen, A. M. Gerdes, P. Moller, U. Kristoffersson, H. Olsson, A. Borg, and A. N. A. Monteiro (2001). Functional analysis of brca1 c-terminal missense mutations identified in breast and ovarian cancer families. *Hum. Mol. Genet. 10*(4), 353–360.

Venkitaraman, A. R. (2000). The breast cancer susceptibility gene, brca2: at the crossroads between dna replication and recombination? *Philos. Trans. R. Soc. Lond. Ser. B-Biol. Sci. 355*(1394), 191–198.

Venkitaraman, A. R. (2001). Functions of brca1 and brca2 in the biological response to dna damage. *J. Cell Sci. 114*(20), 3591–3598.

Vieland, V. J. and S. E. Hodge (1995). Inherent intractability of the ascertainment problem for pedigree data - a general likelihood framework. *Am. J. Hum.*

Genet. 56(1), 33–43.

Welcsh, P. L. and M. C. King (2001). Brca1 and brca2 and the genetics of breast and ovarian cancer. *Hum. Mol. Genet. 10*(7), 705–713.

Welcsh, P. L., K. N. Owens, and M. C. King (2000). Insights into the functions of brca1 and brca2. *Trends Genet. 16*(2), 69–74.

Wooster, R., G. Bignell, J. Lancaster, S. Swift, S. Seal, and J. Mangion (1995). Identification of the breast cancer susceptibility gene BRCA2. *Nature 378*, 789–92.

Yan, H., K. W. Kinzler, and B. Vogelstein (2000). Genetic testing — present and future. *Science 289*, 1890–1892.

Biography

Xi (Kathy) Zhou was born on August 14, 1973, in Chengdu, China. She obtained her B.S. degree in Astrophysics from Nanjing University in July, 1996. She began her graduate study at Duke University in the fall of 1998 and obtained M.S. in statistics from Duke University in May, 2000.