

Classification of missense mutations of disease genes

Xi ZHOU, Edwin S. IVERSEN, JR., Giovanni PARMIGIANI

An accurate estimate of disease risk is critical to clinical management of individuals with a missense mutation of a disease susceptibility gene. In practice, it is a challenge to accurately assess a missense mutation's contribution to disease susceptibility as sufficient sample size or an appropriate functional assay are often not available. For cancer susceptibility genes, measures of disease association will in the foreseeable future be based on pedigree data collected in clinics specializing in individuals at high risk of genetic susceptibility to cancer. Absent a correction, this sampling mechanism is likely to lead to an overstatement of the mutation's contribution to disease. In this study, we propose a Bayesian hierarchical model to study the disease causality of missense mutations given pedigree data collected in the high risk setting. The model classifies missense mutations as deleterious or non-deleterious. The hierarchical structure of the model makes the systematic comparison of the effects of different missense mutations possible, allowing us to study them as a group instead of one at a time. In addition to the missense pedigrees, the model utilizes a group of pedigrees identified through probands tested positive for known deleterious mutations and a group of test-negative pedigrees, both obtained from the same clinic, to calibrate the classification and control for potential ascertainment bias. We apply this model to study missense mutations of breast-ovarian susceptibility genes BRCA1 and BRCA2, using data collected at the Duke University Medical Center in Durham, North Carolina.

1. INTRODUCTION

Cancer is caused, in part, by genetic changes (mutations) in DNA. While many of these changes are somatic, and occur during one's lifetime, some are inherited. Loci on the human genome that are associated with inherited susceptibility to cancer have been identified for several cancer sites, including some of major public health interest, such as the breast cancer susceptibility genes BRCA1 and BRCA2 (Miki, Swenson, Shattuck-Eidens, Futreal, Harshman, Tavtigian, et al. 1994; Wooster, Bignell, Lancaster, Swift, Seal, and Mangion 1995), the colon cancer susceptibility gene APC (Syngal, Schrag, Falchuk, Tung, Farraye, Chung, Wright, Whetsell, Miller, and Garber 2000) and the hereditary non-polyposis colorectal cancer (HNPCC) genes (Lynch and de la Chapelle 1999). Genetic tests for inherited mutations of cancer genes have been developed and, while controversial, are becoming increasingly common (Hoskins, Stopfer, Calzone, Merajver, Rebbeck, Garber, and Weber 1995; Yan, Kinzler, and Vogelstein 2000). Genetic tests look for mutations of disease susceptibility genes. If one is found, it is classified as deleterious (or disease causing), wild type, or as a variant of unknown significance (VUS) based on whether it is phenotype-modifying. It is well established that mutations that

This work was funded in part by the NCI through the Specialized Program of Research Excellence (SPORE) in Breast Cancer at Duke University, P50 CA68438 and through the Cancer Genetics Network at Duke University, U24 CA78157, and at The Johns Hopkins University, U24 CA78148.

DRAFT October 10, 2002

D R A F T

October 10, 2002, 12:38pm

D R A F T

lead to premature truncation of the gene product, such as frame-shift deletions, insertions, or nonsense mutations, are regarded as deleterious. However, for missense mutations, which are mutations that results in the substitution of one amino acid for another in the sequence, it is difficult to predict whether the change will affect the protein product and its function, and ultimately modify cancer risk (Cotton and Scriver 1998).

Missense mutations, as a group, are relatively common. For breast and ovarian cancer susceptibility genes BRCA1 and BRCA2, to date, over 500 distinct missense mutations have been identified (BIC 1997). But, because of their large number and the ambiguity of their effects on the protein product, limited progress has been made so far in classifying individual missense mutations as deleterious or not. A large number of families with high incidence of cancer harbor missense mutations of unknown significance and present to genetic counseling clinics for testing and advice. In this setting, a very serious dilemma arises: family history indicates increased risk, similar in magnitude to that exhibited by known deleterious mutations, but its significance is unknown. Decisions made on the basis of genetic tests are usually life-changing, and often involve radical preventive surgery, such as mastectomy or oophorectomy, exacerbating the dilemma (Grann, Panageas, Whang, Antman, and Neugut 1998).

Currently, approaches to studying the effects of mutations usually proceed in one of two directions. The first is to use molecular biological methods of functional assay to probe the function of the mutation. Such biological analysis is believed to be the ultimate test to decide whether a mutation is phenotype modifying. However, even with the availability of the clone of a gene, sequencing and synthesizing probes for each potential mutation is still very difficult. For large genes, such as BRCA1 and BRCA2, despite the notable progress made to date on understanding the function of parts of the gene product, a functional assay of the whole protein has yet to be developed (Venkitaraman 2000; Hayes, Cayanan, Barilla, and Monteiro 2000; Vallon-Christersson, Cayanan, Haraldsson, Loman, Bergthorsson, Brondum-Nielsen, Gerdes, Moller, Kristoffersson, Olsson, Borg, and Monteiro 2001). The second direction is to empirically evaluate the phenotype-genotype correlation based on epidemiological data; for example, a sample of family histories. Evidence of association from this type of study usually comes from an estimate of the difference in risk of disease between mutation carriers and non-carriers. Mutations are often aggregated, and analyses that are specific to individual variants are rare and usually confined to variants with relatively high frequency in certain genetic groups (Struewing, Hartge, and Wacholder 1997; Oddoux, Struewing, and Clayton 1996). For missense mutations, such analyses are rarely performed because of limited sample size (Venkitaraman 2001). One way of making progress in this direction 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, Parmigiani, and Thomas 1998). Their approach is based on a design that requires testing the genotypes of the proband's affected

relatives. 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.

With the decreasing cost of genotyping, the number of mutations of unknown significance is likely to increase, while understanding the effects of these mutations in a timely fashion remains critical. Efforts to develop networks of institutions sharing information about high risk families (such as the NCI's Cancer Genetics Network) are underway. These efforts will generate a large amount of data, but it will be common for most of the mutations to have only a small number of family histories. Progress on quantitative methodologies for integrating this kind of data at the scale of full genomic variability is still lacking. The goal of this work is to develop and illustrate a classification methodology that can contribute to a more systematic, accurate and timely use of available data on inherited susceptibility due to missense mutations.

In this study we describe a general methodology for using data accrued in genetic counseling clinics to classify missense mutations as deleterious or not. The typical layout of the data is as the following: an individual, often referred to as the proband, is tested for mutations on the gene/genes. The result of the test comprises whether the alleles found are common polymorphisms (normal) or are mutated, and which mutations are found. It is possible that no mutations are found even when some are present. In addition, a pedigree with history of relevant cancer for the proband's family members is ascertained via an interview with the proband, and sometimes verified through medical records. No other family members are typically tested. If they were, the methodology developed could still be applied, and the data would be more informative. So the case considered here is not only the most common but also the most challenging.

Most mutations of cancer genes do not determine the fate of the individual, although they substantially increase cancer risk. Genetic effects are therefore characterized by penetrance functions, that is cumulative probability distributions of developing cancer by age, conditional on a specific genetic variant. Operationally, we define a variant to be deleterious if the associated penetrance is increased compared to the penetrance of normal polymorphisms (also called phenocopy rate). Family histories provide some evidence about the association between disease and genotype and can be used for classifying effects of missense mutations.

In practice, there are, however, a number of concerns, which motivate this study: (1) the number of family histories available for each missense mutation is typically small, making it difficult to estimate a mutation-specific age-dependent penetrance function; (2) the set of family histories collected in genetic counseling clinics is highly selected for cancer related histories, creating the potential for substantial overestimation of penetrance; (3) there is biological foundation

for considering the ensemble of missense mutations of a gene as likely to be similar in effect, although the degree of similarity is unknown; and (4) mutations may be missed by testing.

We address these challenges as follows. We use data from the literature to estimate penetrance functions for common polymorphisms and known deleterious mutations, and introduce a single parameter penetrance model to describe missense mutation-specific penetrance as a function of those estimates. To address ascertainment bias, we compare cancer histories of families with missense mutations to those derived from families with known deleterious mutations, on one hand, and to those derived from negative families, ascertained using the same mechanism, on the other. The classification step is embedded in a Bayesian hierarchical analysis in which the missense mutation-specific penetrance parameters are modeled as arising from a common population whose characteristics are learned from the data. Finally, an allowance for genetic testing error is incorporated in the model.

We have organized the paper as follows. In the next section we describe the structure of our model in detail. In Section 3, we study the operating characteristics of our hierarchical classification approach in context of a simple simulation experiment where penetrance is assumed to be constant over age and the data are generated under one of three ascertainment rules. In Section 4, we apply the model to study a sample of BRCA1 and BRCA2 missense mutations obtained from a study of women at high risk of breast cancer conducted at Duke University's Comprehensive Cancer Center. We provide estimates of the probability that each observed missense mutation is deleterious and estimate the mutation-specific penetrance functions of each. We end the paper with a discussion.

2. METHOD

Family studies provide evidence about the association between phenotype (for example, incidence of disease, age at diagnosis, etc.) and genotype through the disease- and genotype-specific penetrance functions. Typically only a small subset of each study family, often only the proband, is genotyped. At the analysis stage, genotypes of the remaining family members are treated as random variables governed by, in the case of known mode of transmission, a specified genetic model for inheritance of the disease genes (see, for example, Elston and Stewart 1971). For each family member, the likelihood of the observed phenotype is calculated from the relevant penetrance function(s). Hence, inference about the disease association of specific genotypes is a statistical question of penetrance estimation.

There is a rich literature on estimation of penetrance curves (also known as age at onset curves). In it, these functions have been modeled in a number of ways and are sometimes modulated by covariates other than the genotype of interest, for example, by relevant environmental factors. Elston 1973 suggests a normal cumulative distribution function tempered with an asymptote to model age at onset, while Elston and George 1989 use a logistic distribution

incorporating covariates. Abel and Bonney 1990 describe a logistic hazard function to model penetrance as a function of genotype and other covariates. Gauderman and Thomas 1994 describe a methodology for estimating penetrance using a proportional hazards model and Li and Thompson 1997 extend this model to include a family-specific random effect to capture other sources of familial aggregation.

2.1 Ascertainment Bias

The method of sampling, or ascertainment, of families is critical to penetrance estimation. Disease associated mutations of genes, including BRCA1 and BRCA2, are rare and expensive to detect. As a result, most available data on tested individuals is generated in clinics or research programs specializing in familial disease where referrals are made on the basis of features in the family history related to penetrance of the disease. The literature on ascertainment bias in family studies is extensive (Fisher 1934; Elston 1973; Elston and Sobel 1979; Elston and Bonney 1984; Ewens and Shute 1986; Sawyer 1990; Dawson 1994; Rabinowitz 1996; Bonney 1998), and includes work on the specific problems of segregation analysis (Morton 1959; Boehnke and Greenberg 1984; George and Elston 1991; Vieland and Hodge 1995; Elston 1995) and gene characterization (Winter 1980). One approach to the ascertainment problem is to design studies incorporating population-based sampling schemes (Thomas 1999). Unfortunately, such designs are not practical for the study of specific missense allelotypes because individual variants are exceedingly rare.

Because family data is often collected from high risk subpopulations using a variety of complex, sometimes informal, protocols it is often difficult or impossible to condition on ascertainment. Estimates of mutation-specific parameters derived from this type of data without an ascertainment correction will likely be biased because individuals with high disease rates in their family are more likely to have genetic testing than those in the general population. Likewise, non-deleterious polymorphisms observed in the high risk setting may also exhibit large numbers of cases among relatives. Our focus is on the question of whether individual missense mutations are deleterious or not. To answer this question, it is not necessary to have unbiased estimates of the genetic parameters, it is only necessary that we be able to discriminate the deleterious and non-deleterious variants through these parameters. Instead of formally correcting for the ascertainment mechanism in family-specific contributions to the likelihood, we expand the data set to incorporate family histories of individuals tested positive with known deleterious mutations and those tested negative to form comparison groups against which we classify missense mutations with the same ascertainment. Hence the classification procedure is itself implicitly conditional on ascertainment and, importantly, does not require specification of the method of ascertainment or the extremely difficult conditional modeling of family-specific contributions to the likelihood. The

only requirement is that all three groups of individuals — mutation negatives, deleterious mutation positives and missense mutation positives — are ascertained in the same way.

2.2 Genetic Assay Error

Genotype is measured using biological assays, or 'tests,' with imperfect operating characteristics. Failure to correct for assay error is likely to result in biased estimates of mutation effect. In context of the classification, the bias may limit the model's ability to discriminate the deleterious and non-deleterious groups. Assay sensitivity — the probability that an assay finds a mutation when one is present — is less than 1 and varies from method to method and loci to loci, while assay specificity, the probability that no mutation is detected when none is present, is nearly 1 across methods and loci. Indeed, individuals who test positive for 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.

3. MODEL

The hierarchical classification model we propose contains the following stages. First, the likelihood of family histories associated with each mutation are calculated based on a genetic transmission model, assumed known. Next, the variability of mutation-specific genetic model parameters are modeled through a constrained mixture of beta distributions. In the third stage, the model incorporates prior information about the parameters of the beta distributions.

Each unique mutation, m , is assumed to modify penetrance through a scalar mutation-specific 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 cancer cases in other family members are 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 missense mutations are classified as deleterious or benign. To illustrate the procedure, 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. Proband is tested for mutations on a disease susceptibility gene, of whom we assume I tested negative, $J - I$ tested positive with M distinct missense mutations and $K - J$ tested positive with N distinct deleterious mutations. Figure 1 illustrates the general structure of the classification procedure described above. In this figure, the

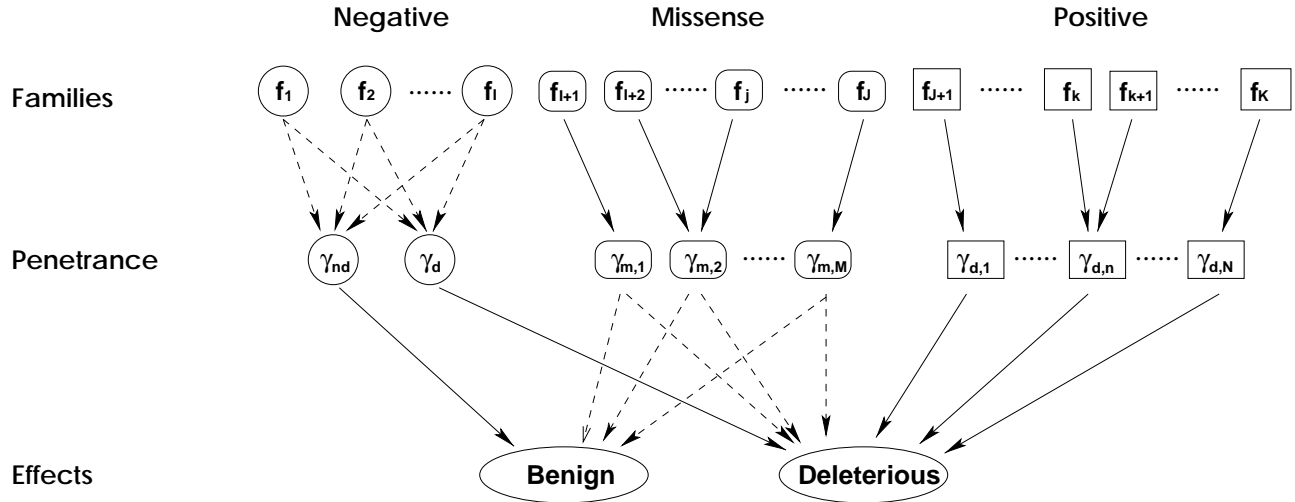


Figure 1. Graphical illustration of the classification model. In this depiction, the K family histories are denoted by $f_i, i = 1, \dots, K$. Each is identified through one proband who is tested for mutations at the site(s) of interest. Each proband's test results — negative, positive with a missense mutation or positive with a deleterious mutation — are indicated by circles, round-cornered rectangles and rectangles, respectively. Mutation-specific penetrances 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 through d_N correspond to the N unique identified deleterious mutations, the index nd corresponds to the wildtype genotype(s) and the index d corresponds to a common, unobserved deleterious genotype. Pedigrees of probands with mutation nn are connected to γ_{nn} by arrows. Dashed arrows indicate uncertainty about a proband's genotype. Mutations and their effects are also connected by arrows. Uncertainty in the effects of the various missense mutations is denoted by dashed arrows.

K family histories are plotted in the top row and denoted by $f_i, i = 1, \dots, K$. Each proband's test results — negative, positive with a missense mutation or positive with a deleterious 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 through d_N correspond to the N unique identified deleterious mutations, the index nd corresponds to the wildtype 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, and let m_{nd} denote the wildtype (non-deleterious) genotype and m_d denote the unobserved deleterious genotype. For a family identified through a proband who tested negative, the conditional likelihood can be written as

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

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 test results can be written as

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

For family i , identified through a proband who tested positive with mutation m , the assay result is considered accurate and the conditional likelihood can simply be written as

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

Let F denote the collection of all the families and D_n denote the collection of latent indicators d_i for all negative families. We can write the joint probability of family histories and the 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 indices i and j represent families of probands tested negative and positive, respectively.

At the second stage of the hierarchical model, we assume that among missense mutations, a proportion, π , of them are deleterious. The degree to which each variant is deleterious is captured through its parameter γ_m . Mutations of similar phenotypic effect will have similar values of their γ_m . We model γ 's as falling in the 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 Beta distribution with parameters α_2 and β_2 . Thus the priors are

$$P(\gamma_{nd} | \alpha_1, \beta_1, \alpha_2, \beta_2) = \text{Beta}(\alpha_1, \beta_1),$$

$$P(\gamma_d | \alpha_1, \beta_1, \alpha_2, \beta_2) = \text{Beta}(\alpha_2, \beta_2).$$

The distribution that gives rise to the missense mutation γ 's is determined by whether they are deleterious or not. Let d_m denote the deleteriousness of missense mutation m . The prior distribution for γ_m conditional on d_m is

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

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

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

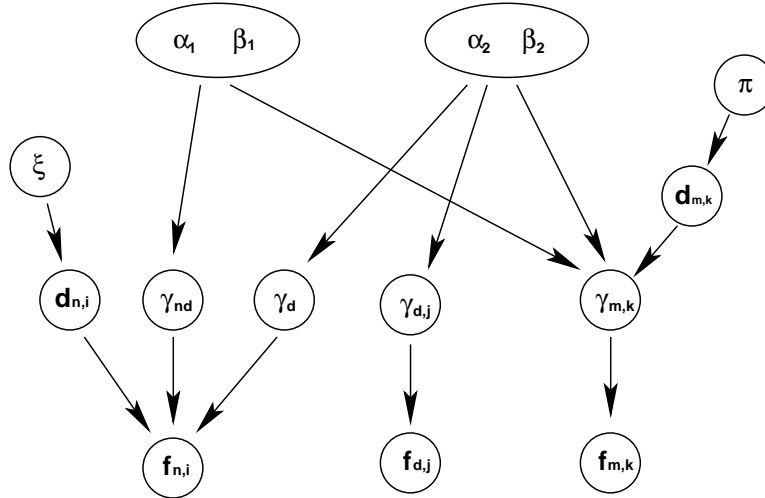


Figure 2. Conditional dependence structure of the hierarchical classification model

The γ 's of known deleterious mutations are assumed to arise from a $\text{Beta}(\alpha_2, \beta_2)$ distribution.

To complete the model, we specify priors and hyper-priors for the 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 feasibility. The 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 tested population using Bayes' rule. Since we assume that test specificity is 1, the prior for ξ should be determined by what is known about the sensitivity of the genetic test as well as prior knowledge about the prevalence of mutations in the tested population. We choose a relatively vague proper prior for π . For the parameters α_1 , β_1 , α_2 , β_2 , the choice of hyper priors is based on prior knowledge as well as computational convenience. In our application, we use uniform distributions between 1 and 100 as the hyper prior. To ensure identifiability we place a weak constraint on these 4 model parameters: for any γ , the CDF of γ given α_1 and β_1 must be uniformly greater than the CDF of γ given α_2 and β_2 . This constraint can be written as

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

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. The posterior can be written as

$$P(\Theta|F, T_0) \propto P(F, D_N|T_0, \Gamma, \xi)P(\Gamma, d_M|\alpha_1, \beta_1, \alpha_2, \beta_2, \pi)f(\alpha_1)f(\beta_1)f(\alpha_2)f(\beta_2)I_{C(\alpha_1, \beta_1, \alpha_2, \beta_2)}f(\pi)f(\xi)$$

An illustration of the conditional dependence structure of the model parameters is given in Figure 2. For families identified through probands with different testing results, $f_{n,i}$ for negatives, $f_{m,j}$ for those tested positive with missense mutations and $f_{p,k}$ for 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. Details of the full conditional distributions are given in the Appendix.

The classification model we describe requires that all family histories used in the analysis 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 for potential ascertainment bias by explicitly modeling the ascertainment procedure in the likelihood calculation as the model implicitly corrects for ascertainment. In the next section we evaluate the operating characteristics of this approach in context of a simulation study.

4. SIMULATION-BASED VALIDATION

In this section, we assess the performance of our hierarchical approach using data simulated under three different ascertainment rules: population-based sampling, affected proband sampling and high risk sampling. Details are given later. We are interested in four issues: the accuracy of classification; the accuracy of penetrance estimates and the effect of that on classification accuracy; the sensitivity to different ascertainment rules; and the performance of the model in large samples. In what follows, we describe simulation of the data, calculation of the likelihood and application of the classification model to the simulated data sets.

4.1 Family history simulation

We simulate three member nuclear pedigrees consisting of parents and one offspring and consider only one heritable disease. This disease is assumed to be associated with a single disease susceptibility gene with an autosomal dominant mode of inheritance. Further, we assume that the allele frequency (prevalence) for each mutated variant of the disease gene is the same and that mutation-specific penetrances are constant over age.

The family histories are sampled in two steps. First, the genotypes of family members are generated. Genotypes of the parents are generated independently based on the prevalence of each mutation, then genotypes of the children are generated based on the genotypes of their parents under independent segregation of alleles. Second, the phenotype of each family member is simulated based on the genotype of the individual. Let 0 denote the wild type allele and X a specific mutated allele. The probability that individuals whose genotype is $0X$ or XX have the disease is equal to the penetrance of mutation X , denoted ρ_X . For individuals whose genotype is XY , where Y is a different mutation from X , their probability of getting disease is assumed to be the larger of ρ_X and ρ_Y .

We consider a population segregating a total of 30 unique mutations. Among them, 20 are missense mutations of which 15 are non-deleterious and 5 are deleterious. The remaining 10 mutations are known deleterious mutations. The

prevalence for each mutation is taken to be 0.002. Penetrances of these mutations are drawn from two distinct Beta distributions. Penetrances of non-deleterious mutations are generated from a Beta(2,18) distribution, while those of deleterious mutations are generated from Beta(8,2).

4.2 Likelihood calculation for simulated data

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

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

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

Where $P(x_i | g_i = m) = \rho_m^{x_i} (1 - \rho_m)^{(1-x_i)}$.

The conditional probability $P(g_1 | g_2, g_3)$ of the child's genotype given the parental genotypes can be calculated from Mendel's laws. The number of possible genotypes at a disease gene with M variants, ignoring genotypes of the form XY , is $1 + 2m$. Hence, $P(g_1 = i | g_2, g_3)$ can be written as a $(1 + 2M) \times (1 + 2M)$ matrix. The contribution of a family to the likelihood is $P(x_1, x_2, x_3 | g_1 = m)$. Because of the marginalization of unknown parental genotypes, this is a complicated polynomial involving the penetrances of all mutations. When, as is the case here, the population frequency of each mutation is low, most of the information provided by the likelihood above is about the penetrance of the mutation that the proband carries. Therefore, we approximate the family-specific contribution to the likelihood by a polynomial in the penetrance ρ_m of the mutation carried by the proband.

4.3 Model validation

In this simulation study, we confine our attention to three different ascertainment schemes:

- **Population-Based:** Randomly sample family histories from the general population.
- **Affected Proband:** Only ascertain families in which the proband is diagnosed with disease.

- **High-Risk:** Sample families without disease with probability 0.1, families with 1 diseased member with probability 0.5, families with 2 diseased cases with probability 0.7 and families with 3 affected individuals with probability 0.9.

These schemes are stylized versions of study designs used in gene characterization studies. In practice, participants in family studies are usually collected under more complex set of ascertainment rules, most closely resembling the high risk scheme.

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

$$P(x_2, x_3 | g_1 = m, x_1 = 1) = \frac{\sum_{g_2, g_3} P(x_2 | g_2) P(x_3 | g_3) P(g_1 = m | g_2, g_3) P(g_2, g_3)}{\sum_{g_2, g_3} P(g_1 = m | g_2, g_3) P(g_2, g_3)}. \quad (2)$$

Thus, for this mode of ascertainment, we also compare classification performance based on the ascertainment corrected and uncorrected likelihoods. Furthermore, we assess performance of our classification model for data sets of different sample size.

The design of the simulation study is as follows: We simulate 10 replicate data sets of family histories with sample size 4000 and 10 replicates with sample size 6000. For each set, families are ascertained through each of the three rules stated above. Next, we calculate the likelihood given in Equation 1 for each mutation. For data ascertained given an affected proband, we calculate the ascertainment corrected likelihood given in Equation 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, we impose the relatively loose prior constraint that the cumulative probability of the Beta distribution of the non-deleterious mutations is uniformly greater than that of the deleterious mutations. As we do not simulate testing error, we set ξ to 1. We place a Beta(2,2) prior on π , the proportion of deleterious missense mutations.

One replicate of the sample size 4000 data sets serves to illustrate details of the calculation. Figure 3 plots mutation-specific likelihoods for the penetrance parameter by ascertainment rule. Likelihoods for the wild type, non-deleterious missense mutation, deleterious missense mutation and known deleterious mutation penetrance parameters are denoted by colors: light blue, blue, red, and orange, respectively.

The conditional likelihoods for mutation penetrances for the population-based data pictured in Figure 3 show very good correlation with the penetrance values (not plotted) 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. The likelihood of the negative families is, however, maximized at a slightly larger value than

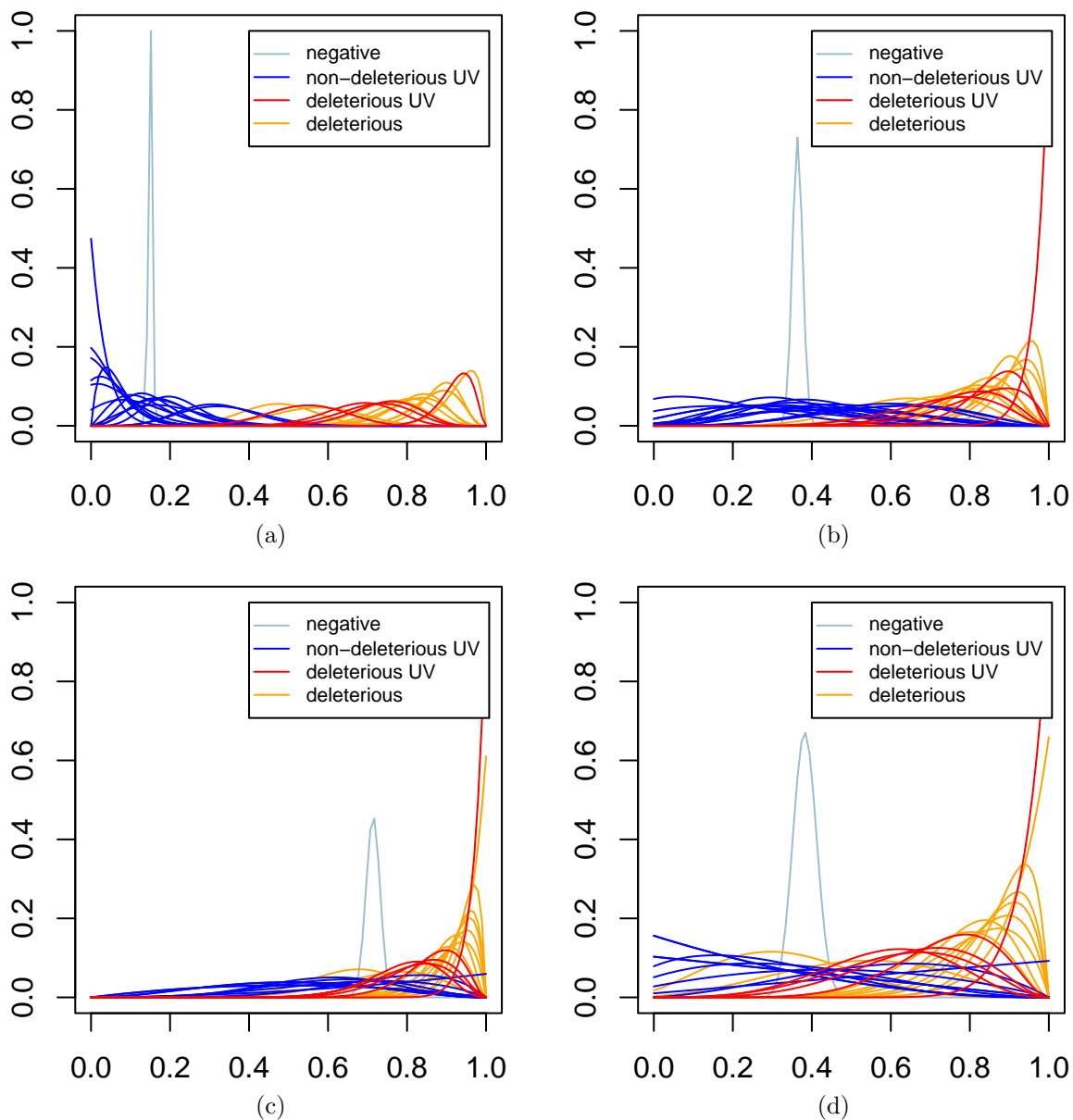


Figure 3. Rescaled conditional likelihood of family histories given mutation-specific penetrance vs. mutation-specific penetrance. The total number of simulated family histories is 4000. In panel (a), family histories are population based; in panel (b), they are ascertained through the high-risk rule; in panel (c), they are ascertained through disease affected probands and the likelihoods are calculated without correcting for ascertainment; in panel (d), they are ascertained through affected probands and the likelihoods are calculated conditional on ascertainment.

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 a parent with a mutation. Under the high-risk ascertainment rule, only a subset of the 4000 families are sampled. Of the mutations represented in the sample, most of the associated likelihoods are maximized at values greater than the penetrance used to sample the data. This parallels the situation we encounter in using data from high-risk clinics without correcting for ascertainment. 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.

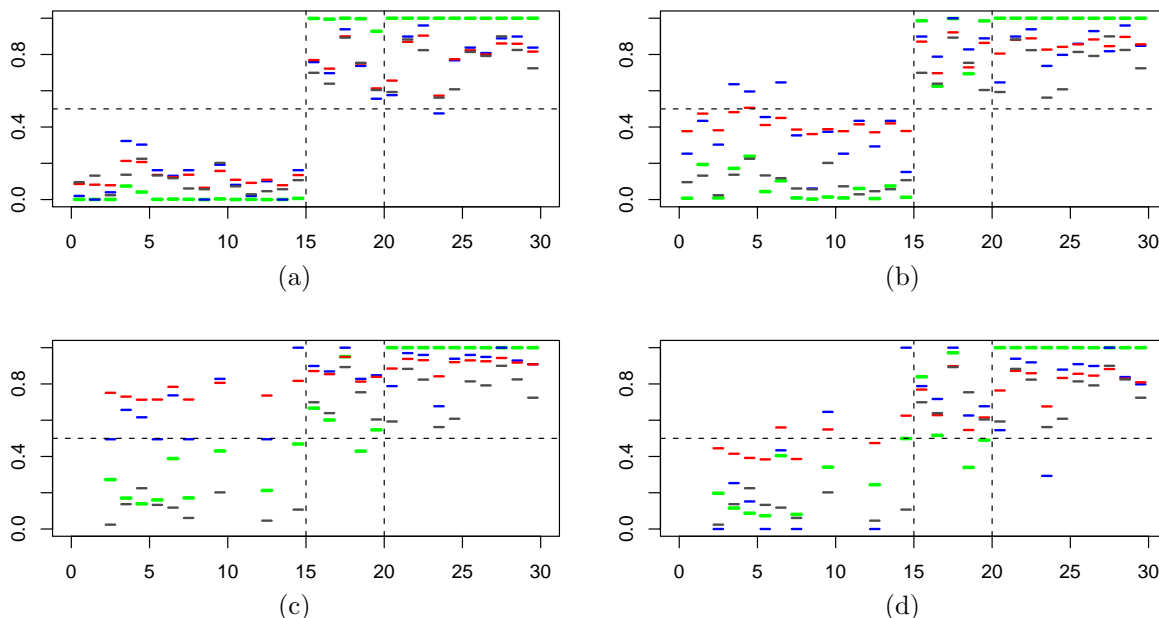


Figure 4. Summary of posterior estimates of model parameters from the replicate summarized in Figure 3. The horizontal axis represents mutation index. Mutations 1-15 are non-deleterious missense mutations, mutations 16-20 are deleterious missense mutations and mutations 21-30 are known deleterious mutations. The vertical 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 penetrances, red lines represent posterior mean of mutation-specific penetrance estimated from the hierarchical classification model and green lines represent the posterior probability that the mutation is deleterious. Panel (a) summarizes the population based analysis; panel (b), the high-risk analysis; panel (c), the uncorrected affected proband analysis; and panel (d), the corrected affected proband analysis.

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 4. Note that, for population-based data (Panel (a)), the estimated posterior means of penetrance (red lines) are fairly close to the maximum likelihood estimates (blue lines) and the true penetrances (dark lines), and that the posterior probability that the mutation is deleterious (green lines) is a good classifier. In particular, thresholding this probability at $1/2$ correctly classifies all 20 missense mutations. For the high-risk data (Panel (b)), posterior and maximum likelihood estimates of penetrance are clearly biased. 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 show similar trends to those evidenced in Figure 4. Figure 5 displays boxplots of the probability of deleteriousness across replicates for each missense mutation.

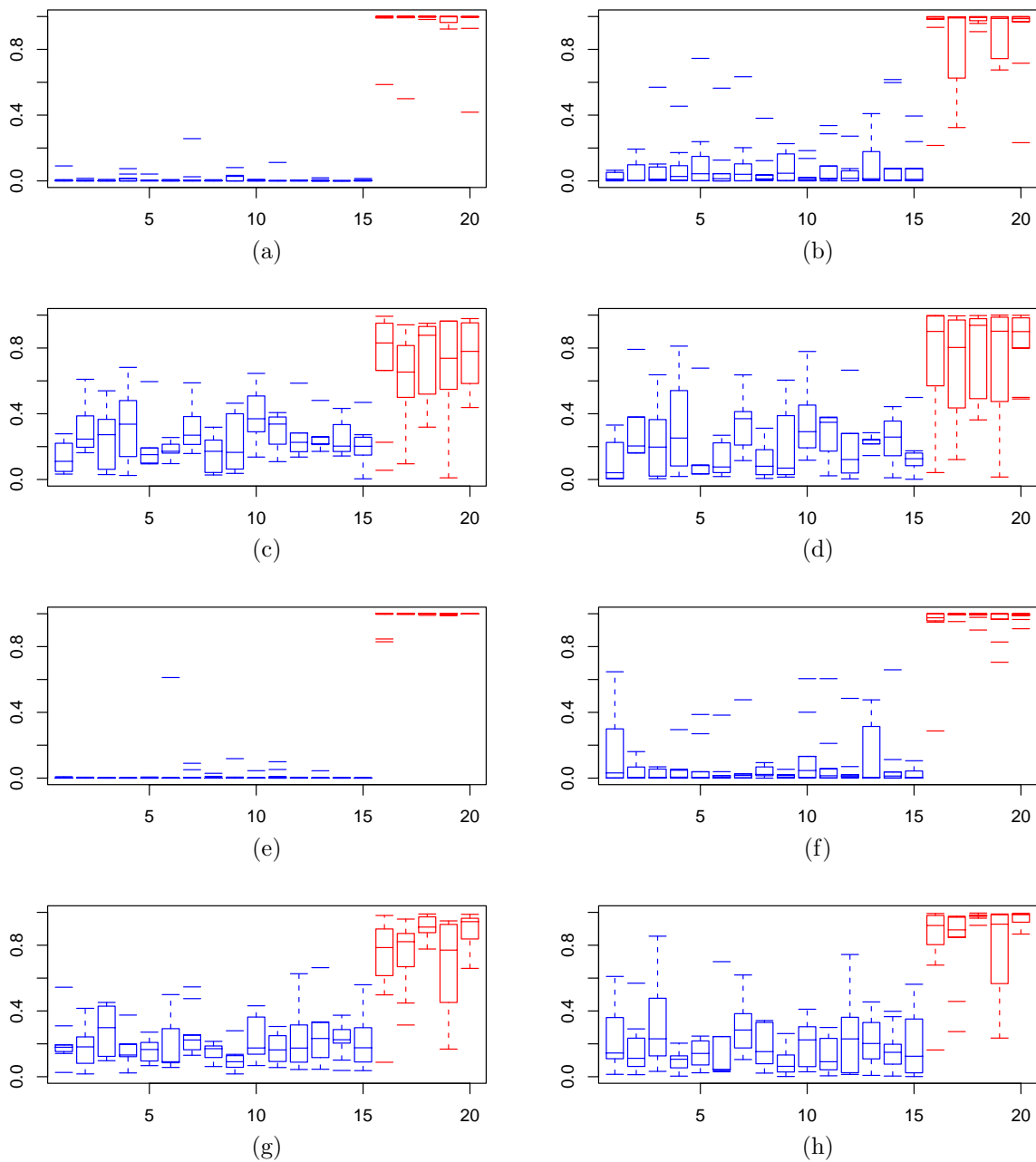


Figure 5. Boxplots of the probability of deleteriousness of the missense mutations estimated across replicates. Mutations denoted by blue are non-deleterious missense mutations, while mutations in red are deleterious missense mutations. Plots (a)-(d) are results based on samples of size 4000, while panels (e)-(h) are based on samples of size 6000. Plots (a) and (e) summarize the population-based samples, plots (b) and (f) summarize the high-risk samples and plots (c), (g), (d) and (f) summarize the affected proband samples. In plots (c) and (g), the likelihoods are not corrected for ascertainment mechanism, while for (d) and (f), they are.

Suppose we classify as deleterious those missense mutations with a posterior probability of deleteriousness greater than 0.5. We summarize the resulting misclassification rates in Table 1. If sampling is population-based, classification is very accurate. Furthermore, under high-risk 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 the uncorrected likelihood. This is likely because the ascertainment cor-

rection 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 1. Proportion of incorrect classifications across simulation replicates.

sample size	population-based	high-risk	affected proband no likelihood correction	affected proband likelihood correction
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; and, not surprisingly, (4), classification accuracy increases with sample size.

5. APPLICATION TO MISSENSE MUTATIONS AT BRCA1 AND BRCA2

In this section, we apply the hierarchical classification model to study a data set of missense mutations of the breast cancer susceptibility genes BRCA1 and BRCA2 identified (primarily) at Duke University in context of a study of women at high risk of breast and ovarian cancer. This analysis involves complexities not present in the simulation study. In particular, the genetic model is of two loci, each predisposing carriers to elevated risk of two diseases (breast and ovarian cancer) with associated age dependent penetrance functions. Furthermore, individual family histories are typically much larger and more complex and genetic assay error complicates classification. In what follows, we describe the data set, detail a penetrance model and present results of an analysis based on this model.

5.1 Description of Study Sample

Our data consists of a total of 280 moderate sized family histories. Most of the families (277) were collected at the Duke University Cancer Center. Each of these family histories was ascertained through a proband who was tested for mutations at BRCA1 and BRCA2. 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, Almeida, Casey, Fain, Liao, Masunaka, Noble, Kurosaki, and Anton-Culver 1996); these three families are believed to be ascertained in a similar way as those from Duke. Family histories include the breast and ovarian cancer status of all 1st and 2nd degree relatives of the proband as well as each individual's age and age(s) at diagnosis, if affected. 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 at either BRCA1 or BRCA2, with 41 carrying one of the 24 unique

BRCA1 mutations identified in the sample and 18 carrying one of 15 unique BRCA2 mutations. A further 16 probands tested positive for one of 9 unique BRCA1 missense mutations and 29 tested positive for one of 26 unique BRCA2 missense mutations. The remaining 174 probands tested negative for any form of mutation at BRCA1 or BRCA2. Table 2 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. Distribution of breast and ovarian cancer in probands categorized by their genetic test results.

	UV		Del		Neg
	BRCA1	BRCA2	BRCA1	BRCA2	
Number	16	29	41	18	174
Breast Cancer ¹	13	25	30	15	144
Ovarian Cancer ²	1	2	4	1	9
Bilateral B.C. ³	3	5	13	4	19
B.C. & Ov.C.	1	0	7	1	1

Table 3 shows the distribution of age at diagnosis of breast cancer in affected probands categorized by genetic test result. Probands who tested positive and who have had breast cancer have, on average, a slightly earlier age at diagnosis

Table 3. Distribution of age at diagnosis of breast cancer in probands categorized by genetic test result, where "positive" denotes positive with either a missense or known deleterious mutation.

		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

than those who tested negative. The median age at diagnosis among those who tested positive is 39.50 year old, while it is 43.50 for those who tested negative. When we stratify by type of mutation as well as location, carriers of deleterious mutations at BRCA1 appear to have an earlier age of diagnosis, on average, with a sample median of 37.50 years old. This trend is reflected in breast cancer affected family members of probands. Table 4 tabulates the age at diagnosis of breast cancer for family members categorized by the proband's genetic test result.

Table 4. Distribution of age at diagnosis of breast cancer in family members categorized by proband's 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 result are shown in Table 5. Most family histories contain fewer than 3 breast cancer cases. Among families of mutation positive probands,

Table 5. Family histories categorized by the number of cases in the family and the proband’s genetic test result.

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

this percentage is about 72.4%, while among those who tested negative, it is about 78.4%. Among mutation positive families, BRCA1 families appear to have, on average, more breast cancer than BRCA2 families.

Table 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 7 expands on Table 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 of family history.

Table 6. Total number of cases over total number of individuals among relatives by age at diagnosis and proband’s genetic test result. “del” denotes deleterious, “neg” negative and “UV” unclassified variant (missense mutation).

	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

Observed differences in familial disease phenotype between families of deleterious mutation positive individuals, missense mutation positive individuals and negative individuals are subtle, depend on the number of cancers in the family, the types of cancers, the ages 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 classify the disease causality of missense mutations in this kind of data set. What is needed is a low dimensional summary of family history to use as a classification variable, one that captures the critical features in the family history. In the next section we propose such a measure.

5.2 Penetrance Model and Likelihood Calculation

Table 7. Total number of cases over total number of individuals among relatives by age of diagnosis and proband's genetic test result for those individuals who tested positive with a missense mutation.

	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

While the raw data evidence differences in family histories according to the genetic test result of the proband, a genetic model is needed to recover and fully exploit the subtle differences exhibited in the limited family histories associated with each mutation. To accomplish this discrimination, we use a full likelihood based approach which takes into account the exact structure of the pedigree, the ages of its members, the breast and ovarian cancer status of its members and the ages of diagnosis of affected members. Disease incidence and age at diagnosis enter the model through site- and mutation-specific penetrance functions. As family histories in our study are all moderate in size, specifying a fully parameterized mutation-specific penetrance model is likely to overfit the data and miss the main differences between family histories associated with individual mutations. As an alternative, we implement a simple one parameter penetrance model.

A variety of studies have focused on penetrance of breast and ovarian cancers among carriers of deleterious BRCA1 and BRCA2 mutations (Easton et al. 1995; Easton et al. 1997; Ford et al. 1998; Struewing et al. 1997; Fodor et al. 1998). This literature serves as the foundation for a simple penetrance model. We take as our starting point the smooth parametric estimates distributed with the BRCAPRO carrier probability model (Parmigiani 2002; Parmigiani et al. 1998) and described in 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.

Let $\rho_{m,s}(a)$ denote the mutation specific penetrance of disease s at age a for mutation m , let $\rho_{s,l}(a)$ denote the penetrance of disease s at age a among known deleterious mutations at locus l , that is $BRCA1$, estimated from previous studies and let $\phi_s(a)$ denote the phenocopy rate estimated from previous studies. A natural choice for the penetrance model is:

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

In this parameterization, the baseline hazard is that associated with carriers of benign polymorphisms and $\gamma_m/(1-\gamma_m)$ is the hazard ratio associated with mutation m . When $\gamma_m = 0$ the mutation is protective of disease ($\rho_{m,s}(a) = 0$ for all a), when $\gamma_m = 0.5$ it is indistinguishable from a benign polymorphism and when $\gamma_m = 1$ it is 'lethal' ($\rho_{m,s}(a) = 1$ for all a). Although parameter γ_m is assumed to be independent of age and cancer site to accommodate limited sample size, the overall mutation specific penetrance is age dependent. This parameterization of penetrance indexes a family of penetrance curves by a one dimensional variable with support $[0,1]$. Numerous other families of curves may be obtained in a similar fashion (Zhou 2002), but sensitivity analysis of the classification model to this choice is beyond the scope of this paper.

In Figure 6, we plot the implied penetrance curves of $BRCA1$ and $BRCA2$ mutations for breast and ovarian cancers over a range of values of γ_m under the penetrance model described above. A single mutation specific value of γ_m simultaneously indexes breast and ovarian cancer penetrance curves for the loci ($BRCA1$ or $BRCA2$) at which the mutation was identified.

The conditional likelihood based on any of these penetrance models can be calculated as described in the simulation section. A factor that complicates this calculation for the high risk breast cancer data is that family histories are large, consisting of all first- and second-degree relatives of the proband, and vary in structure from family to family. Computing the likelihood requires summing over all possible combinations of genotype among family members conditional on the proband's genotype. This calculation is performed through a modified version of the software application BRCAPRO. Details of the probability model implemented in BRCAPRO can be found in Berry, Parmigiani, Sanchez, Schildkraut, and Winer 1997 and Parmigiani, Berry, and Aguilar 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 penetrance of breast and ovarian cancer among the two classes of mutation carriers and among carriers of benign

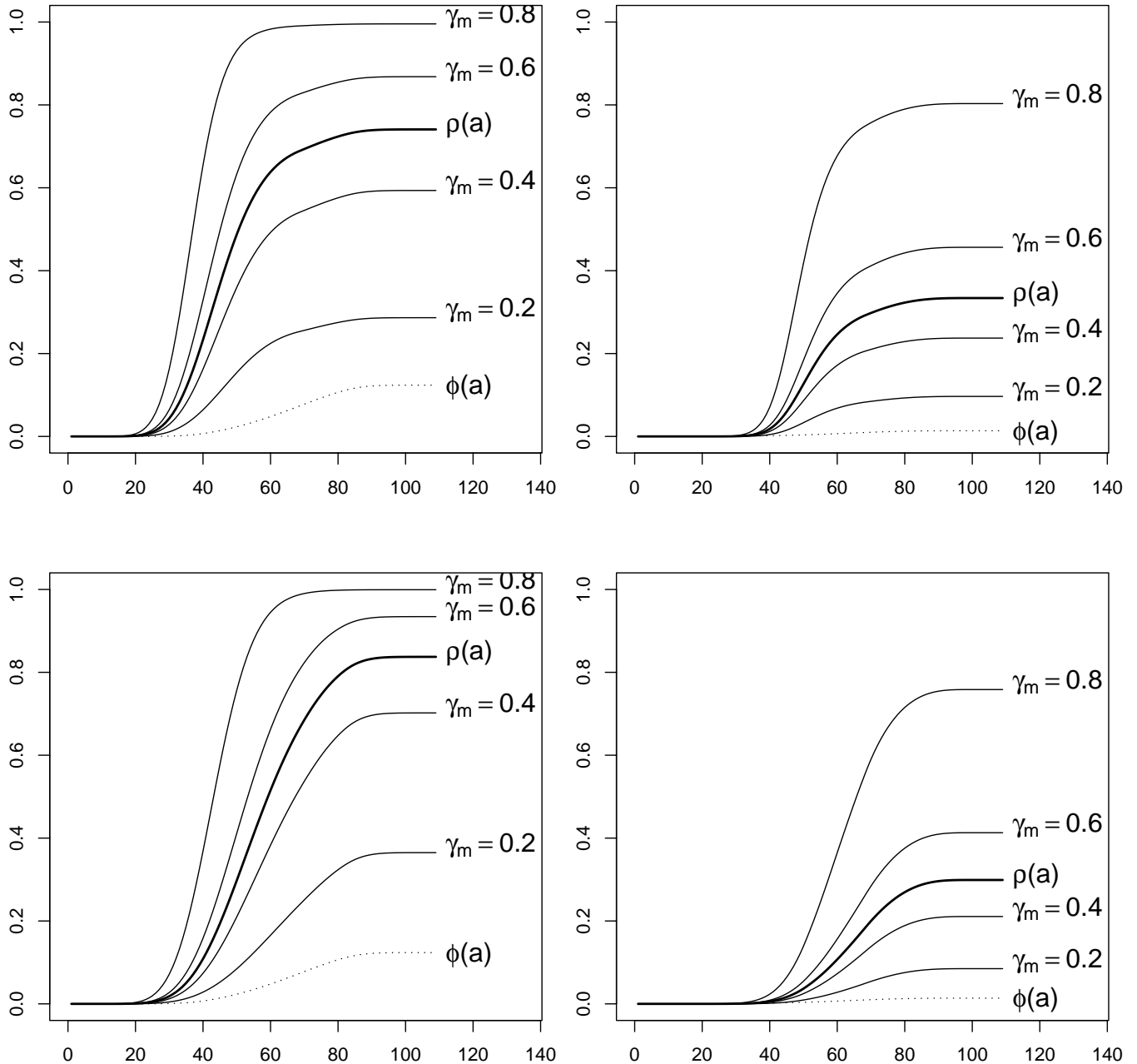


Figure 6. Penetrance of breast and ovarian cancers for carriers of *BRCA1* and *BRCA2* mutations under the assumed penetrance model. The first row plots penetrance curves of breast (left panel) and ovarian (right panel) cancer for *BRCA1* mutation carriers. The second row plots penetrance curves of breast (left panel) and ovarian (right panel) cancers for *BRCA2* mutation carriers.

polymorphisms. We modified the software to calculate the likelihood of the observed family history given mutation prevalence and disease penetrance and the aforementioned genetic model.

5.3 Hierarchical Classification Results

We construct the hierarchical classification model as described in Section 3. Vague proper priors were used for model parameters π and ξ . We used beta distribution with parameters (2, 2) for π and, for ξ , the proportion of false negatives of the genetic test, we chose the prior following the discussion in Section 3. Specifically, we assumed that the sensitivity

of the genetic test (SSCP) was about 65% to 75% and that the prevalence of mutations in the sample is between 30% and 50%. Using Bayes' rule, a relatively conservative choice for ξ would be around 0.7. Hence, we chose a beta distribution with parameters (7,3) as the prior on ξ . Finally, we placed Uniform(1,100) hyper priors on the hyper parameters α_1 , β_1 , α_2 and β_2 .

We sampled the posterior distribution of model parameters using the Gibbs sampling scheme outlined in the Section 3 given each of the three penetrance parameterizations. For each, we carried out 200000 iterations, retaining every 10th. 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 were declared sufficient to estimate the 2.5 percentile of any of the marginal posteriors within an accuracy of 0.5% with probability 95% based on the Raftery and Lewis diagnostic (Raftery and Lewis 1996).

Histograms of marginal posterior samples of model parameters are plotted in Figure 7 with their respective priors. The data are very weakly informative for π and virtually uninformative for ξ . This reflects the fact that it is difficult to locate the anchor populations (carriers of benign polymorphisms and carriers of deleterious mutations) given data obtained through high risk ascertainment. The data is, however, informative for parameters α_1 , β_1 , α_2 , and β_2 of the anchor population beta distributions. We estimate the posterior mean of the distribution of non-deleterious γ 's to be $E(\alpha_1/(\alpha_1 + \beta_1) \mid \text{Data}) = 0.44$ with 95% equal-tailed interval (0.22,0.57) and estimate the posterior mean of the distribution of deleterious γ 's to be $E(\alpha_2/(\alpha_2 + \beta_2) \mid \text{Data}) = 0.67$ with 95% equal-tailed interval (0.62,0.71). Figure 8 plots the beta distributions associated with estimates of posterior expectations of the classification distribution parameters α_1 , β_1 , α_2 , and β_2 . Note that the model's survival parameterization captures a full range variability in mutation specific penetrance parameters among both carriers of benign and deleterious variants.

Figure 9 plots histograms of the posterior samples of γ_{nd} and γ_d , penetrance parameters for carriers of non-deleterious (nd) and of deleterious (d) mutations among individuals who tested mutation negative at BRCA1 or BRCA2. Posterior means are denoted by the vertical lines in the plots. This study's focus is on inference for the missense mutations, whose individual effects are couched in terms of the mutation specific penetrance parameters γ_m for each mutation m . Estimated posterior means of the γ_m 's and associated 90% equal-tailed posterior intervals are plotted in Figure 10. Estimates associated with missense mutations are plotted in green, estimates of known deleterious mutations are plotted in red and estimates associated with the two assumed polymorphisms among the negatives are plotted in blue. In addition, an estimate of the posterior probability of disease causality of each missense mutation is denoted by an "x". Posterior estimates of mutation specific γ_m 's vary widely from missense mutation to missense mutation with some

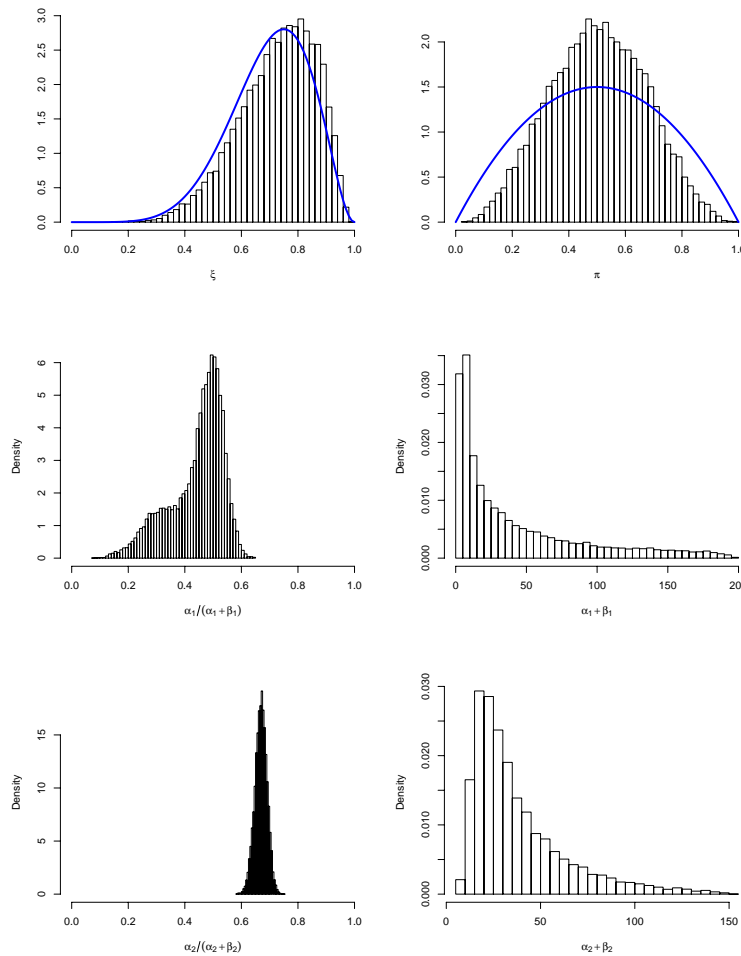


Figure 7. Histograms of posterior samples of model parameters. Prior distributions on model parameters are plotted in solid lines.

showing strong evidence of being from the population of deleterious mutations (see, for example, the four cases near the top of the plot with posterior probability of disease causality of 0.8 or above).

The posterior probabilities of disease causality (deleteriousness) of the missense mutations are plotted in Figure 11. Mutations labeled unknown are unclassified variants located at the intronic region of the gene. The probabilities of disease causality estimated in this sample range from 0.197 to 0.860. Mutations M1775R, S1140G and G1788V at BRCA1 and mutations V1605I and L1904V at BRCA2 have the strongest association with disease. The association is much weaker for R1347G at BRCA1 and R2973C at BRCA2. Mutation R841W, discussed by Barker, Almeida, Casey, Fain, Liao, Masunaka, Noble, Kurosaki, and Anton-Culver 1996 is less strongly associated with disease than most of the mutations in our study.

6. DISCUSSION

In this study, we have developed a Bayesian hierarchical method to study disease causality of missense mutations. This method provides a framework for simultaneously evaluating the disease association of a group of mutations and allows

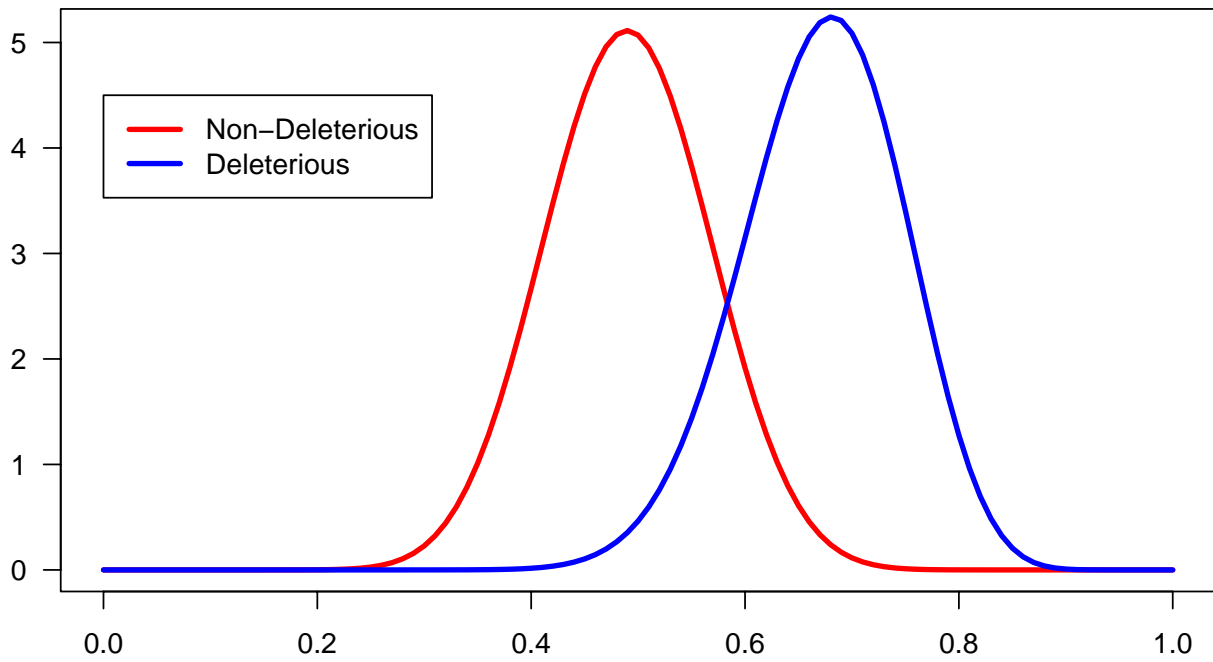


Figure 8. Beta distributions associated with estimates of posterior expectations of the classification distribution parameters α_1 , β_1 , α_2 , and β_2 . Note that the model captures a full range variability in mutation specific penetrance parameters among both carriers of benign and deleterious variants.

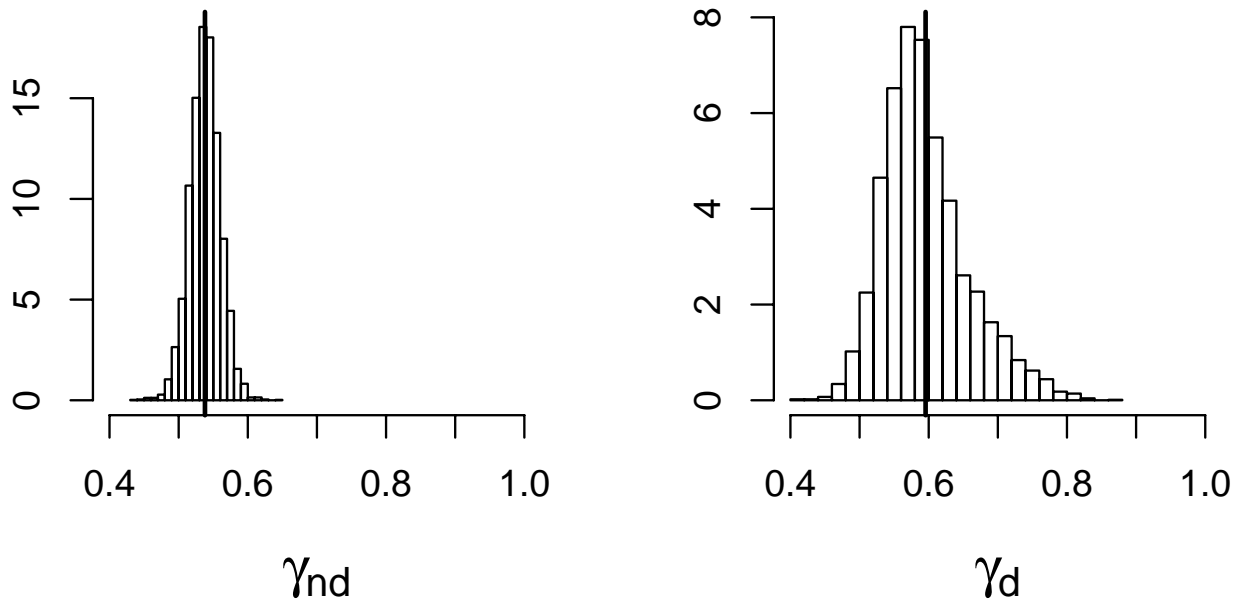


Figure 9. Histograms of the posterior samples of γ_{nd} and γ_d . Posterior means of these penetrance parameters are plotted in solid lines.

for a systematic comparison of the evidence of causality from observed family histories. This systematic comparison allows us to present results which are useful to both genetic counselors and molecular biologists, while providing insights into gene function through an analysis of high risk family history data.

Family studies of Mendelian disease are increasingly common. Many have suggested that among mutation carriers, there is variation in disease susceptibility and age at disease onset. Possible reasons for this kind of variation may include

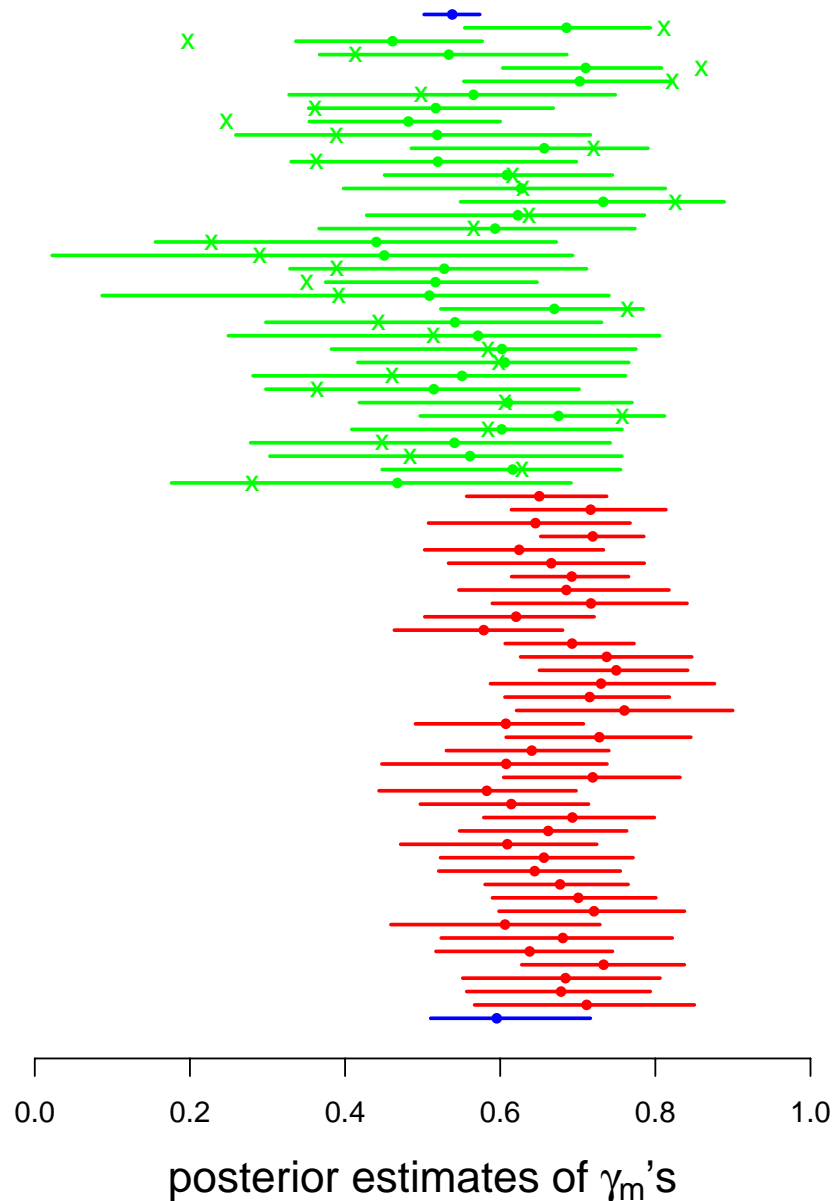


Figure 10. Posterior means and 90% equal-tailed intervals of the γ_m 's. An estimate of the posterior probability of disease causality of each missense mutation is denoted by an "x". Estimates associated with missense mutations are plotted in green, estimates of known deleterious mutations are plotted in red and estimates associated with the two assumed polymorphisms — nondeleterious (top) and deleterious (bottom) — among the negatives are plotted in blue.

genetic factors, environmental factors or both. While there have been numerous studies focused on environmental modifiers of Mendelian disease, few have been aimed at evaluating variability in phenotypic exemplification associated with different variants at a common locus. The primary reason for this is the lack of sufficiently large samples for specific variants and biases induced by data collection in high risk clinics. Our method builds a framework for analysis of this kind of data set.

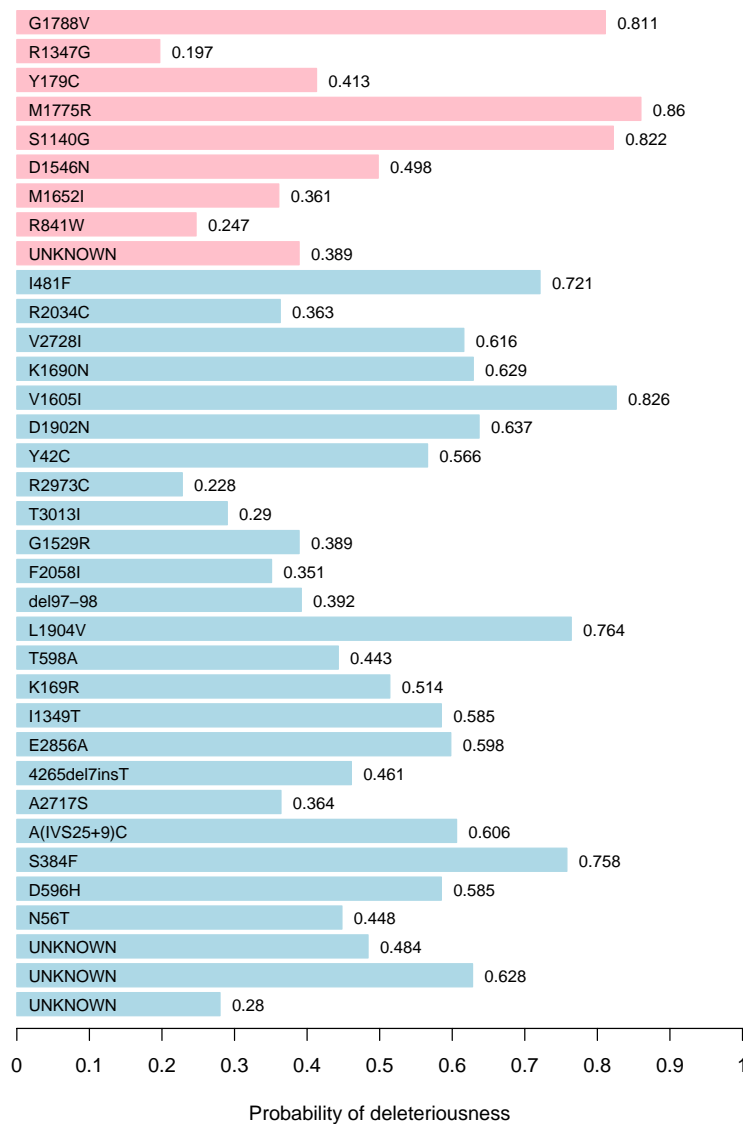


Figure 11. Posterior probabilities of deleteriousness for the missense mutations. Missense mutations at *BRCA1* are plotted in light blue and those at *BRCA2* are plotted in pink.

This method provides a meaningful evaluation of the disease causality of group of missense mutations because it compares them to mutations with a known association to disease and to those with no functional change in the protein, i.e. common polymorphisms. This evaluation is model based and is made with respect to a homogeneous (with respect to ascertainment) group of data. In our analysis, we assume that all polymorphisms of the disease gene(s) under study are not associated with disease. If this assumption is violated, our method may underestimate the disease association of the missense mutations if one or more polymorphism is indeed associated with increased risk of disease. The penetrance model we use for classification is simple, involving only one parameter, and is based on currently available

estimates of penetrance and the phenocopy rate. More realistic models of penetrance would involve more flexible, higher dimensional families of penetrance curves, but would require larger data sets for estimation. Substituting a more sophisticated family of penetrance curves would, however, require only minimal changes to the model. Improved estimates penetrance and phenocopy rate may also improve the reliability of the method.

There are a number of opportunities for expanding the study of missense mutations at BRCA1 and BRCA2 to encompass larger sets of missense mutations. The National Cancer Institute's Cancer Genetics Network, for example, maintains several large family history data sets with structure similar to data analyzed here. These data sets comprise samples collected at multiple centers, under different modes of ascertainment. The model we describe could be applied to this type of data set after a modification to hierarchically model the multiple centers and modes of ascertainment. The encouraging results we obtain on a relatively small data set suggests that our method will perform extremely well on large multi-center data sets, providing critically important information on disease risk to genetic counselors, molecular biologists and, most importantly, to the carriers of missense mutations themselves.

References

- Abel, L. and G. E. Bonney (1990). A time-dependent logistic hazard function for modeling variable age of onset in analysis of familial diseases. *Genetic Epidemiology* 7, 391–407.
- 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/.
- Boehnke, M. and D. A. Greenberg (1984). The effects of conditioning on probands to correct for multiple ascertainment. *American Journal of Human Genetics* 36, 1298–1308.
- Bonney, G. E. (1998). Ascertainment corrections based on smaller family units. *American Journal of Human Genetics* 63, 1202–1215.
- Cotton, R. G. H. and C. R. Scriver (1998). Proof of "disease causing" mutation. *Hum. Mutat.* 12(1), 1–3.
- Dawson, D. V. (1994). Ascertainment models incorporating effects of variable age of onset. *American Journal of Medical Genetics* 53, 340–347.

- Easton, D. F., D. Ford, D. T. Bishop, and The Breast Cancer Linkage Consortium (1995). Breast and ovarian cancer incidence in BRCA1-mutation carriers. *Am J Human Genetics* 56, 265–71.
- Easton, D. F., L. Steele, P. Fields, W. Ormiston, D. Averill, P. A. Daly, R. McManus, S. L. Neuhausen, D. Ford, R. Wooster, L. A. Cannon-Albright, M. R. Stratton, and D. E. Goldgar (1997). Cancer risks in two large breast cancer families linked to BRCA2 on chromosome 13q12-13. *Am J Human Genetics* 61, 120–128.
- Elston, R. C. (1973). Ascertainment and age of onset in pedigree analysis. *Human Heredity* 23, 105–112.
- Elston, R. C. (1995). \hat{T} wixt cup and lip: How intractable is the ascertainment problem? *American Journal of Human Genetics* 56, 15–17.
- Elston, R. C. and G. E. Bonney (1984). Sampling considerations in the design and analysis of family studies. In D. C. Rao, R. C. Elston, L. H. Kuller, M. Feinleib, C. Carter, and R. Havlik (Eds.), *Genetic Epidemiology of Coronary Heart Disease — Past, Present, and Future*. New York: Alan R. Liss, Inc.
- Elston, R. C. and V. T. George (1989). Age of onset, age at examination and other covariates in the analysis of family data. *Genetic Epidemiology* 6, 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. i. theory. *Theoretical Population Biology* 30, 388–412.
- Fisher, R. A. (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals of Eugenics* 6, 13–25.
- Fodor, F. H., A. Weston, I. J. Bleiweiss, M. M. McCurdy, L D a nd Walsh, P. I. Tartter, S. T. Brower, and C. M. Eng (1998). Frequency and carrier risk associated with common BRCA1 and BRCA2 mutations in Ashkenazi Jewish breast cancer patients. *American Journal of Human Genetics* 63, 45–51.
- 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.
- Gauderman, W. J. and D. C. Thomas (1994). Censored survival models for genetic epidemiology: a Gibbs sampling approach. *Genetic Epidemiology* 11, 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, March). Decision analysis of prophylactic mastectomy and oophorectomy in brca1-positive or brca2-positive patients. *J. Clin. Oncol.* 16(3), 979–985.
- Hayes, F., C. Cayanan, D. Barilla, and A. N. A. Monteiro (2000, MAY 1). 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. Berry, and J. Schildkraut (2000). Genetic susceptibility and survival: application to breast cancer. *Journal of the American Statistical Association* 95, 28–42.
- Li, H. and E. Thompson (1997). Semiparametric estimation of major gene and family-specific random effects for age of onset. *Biometrics* 53, 282–293.
- Lynch, H. T. and A. de la Chapelle (1999). Genetic susceptibility to non-polyposis colorectal cancer. *J Med Genet* 36(11), 801–818.
- 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.
- Morton, Newton, E. (1959). Genetic tests under incomplete ascertainment. *American Journal of Human Genetics* 11, 1–16.
- Oddoux, C., J. P. Struewing, and C. M. *et al.* . Clayton (1996). The carrier frequency of the BRCA2 6174delT mutation among ashkenazi jewish individuals is approximately 1%. *Nat. Genet.* 14, 188–190.
- Parmigiani, G. (visited 10/2002). BRCAPRO website. <http://astor.som.jhmi.edu/brcapro>.
- Parmigiani, G., D. A. Berry, and O. Aguilar (1998). Determining carrier probabilities for breast cancer susceptibility genes BRCA1 and BRCA2. *American Journal of Human Genetics* 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.
- Rabinowitz, D. (1996). A pseudolikelihood approach to correcting for ascertainment bias in family studies. *American Journal of Human Genetics* 59, 726–730.
- Raftery, A. E. and S. M. Lewis (1996). Implementing MCMC. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*, London, pp. 115–127. Chapman and Hall.

- Sawyer, S. (1990). Maximum likelihood estimators for incorrect models, with an application to ascertainment bias for continuous characters. *Theoretical Population Biology* 38, 351–366.
- 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. *New England Journal of Medicine* 336, 1401.
- Struewing, J. P., P. Hartge, and S. *et al.* Wacholder (1997). The risk of cancer associated with specific mutations of brca1 and brca2 among ashkenazi jews. *N. Engl. J. Med.* 336, 1401–8.
- 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.
- Thomas, D. C. (1999). Design of gene characterization studies: an overview. *Journal of the National Cancer Institute Monographs* 26, 17–23.
- Vallon-Christersson, J., C. Cayan, 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, FEB 15). 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, FEB 29). 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, October). 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. *American Journal of Human Genetics* 56, 33–43.
- Winter, R. M. (1980). The estimation of phenotype distributions from pedigree data. *American Journal of Medical Genetics* 7, 537–542.
- 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.
- Zhou, X. (2002, December). *Disease Causality of Missense Mutations*. Ph. D. thesis, Duke University, Durham, NC 27708.

APPENDIX

Given other model parameters and the data, the distribution of ξ depends on the proportion of true non-deleterious mutations

in the negative families and the prior:

$$[\xi|\Theta_{-\xi}^{(s)}, data] \sim \xi^{\sum_i (1-d_i^{(s)})} (1-\xi)^{\sum_i d_i^{(s)}} f(\xi).$$

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

$$[\pi|\Theta_{-\pi}^{(s)}, data] \sim \pi^{\sum_m (1-d_m^{(s)})} (1-\pi)^{\sum_m d_m^{(s)}} f(\pi).$$

The full conditional for the latent variable d_m associated with missense mutation m can be written as

$$[d_m|\Theta_{-d_m}^{(s)}, data] \sim \text{Bern}(p = \frac{(1-\pi^{(s)})\text{Beta}(\gamma_m^{(s)}|\alpha_2^{(s)}, \beta_2^{(s)})}{\pi^{(s)}\text{Beta}(\gamma_m^{(s)}|\alpha_1^{(s)}, \beta_1^{(s)}) + (1-\pi^{(s)})\text{Beta}(\gamma_m^{(s)}|\alpha_2^{(s)}, \beta_2^{(s)})}).$$

where p is the probability of d_m being equal to 1. The full conditional of the latent variable d_i for the negatives can be written

$$[d_i|\Theta_{-d_i}^{(s)}, data] \sim \text{Bern}(p = \frac{(1-\xi^{(s)})P(f_i|g_{0i} = m_d, \gamma_d^{(s)})}{\xi^{(s)}P(f_i|g_{0i} = m_{nd}, \gamma_{nd}^{(s)}) + (1-\xi^{(s)})P(f_i|g_{0i} = m_d, \gamma_d^{(s)})}).$$

where p is the probability that proband i has a false-negative test result and d_i equals to 1.

The full conditionals for the γ 's depend on mutation. For the assumed non-deleterious mutations “ m_{nd} ”, the full conditional of γ_{nd} is

$$[\gamma_{nd}|\Theta_{-\gamma_{nd}}^{(s)}, data] \sim \prod_i P(f_i|g_{0i} = m_{nd}, \gamma_{nd})^{1-d_i} \text{Beta}(\gamma_{nd}|\alpha_1^{(s)}, \beta_1^{(s)}).$$

For the assumed deleterious mutation, m_d , that negatives may carry the full conditional is

$$[\gamma_d|\Theta_{-\gamma_d}^{(s)}, data] \sim \prod_i P(f_i|g_{0i} = m_d, \gamma_d)^{d_i} \text{Beta}(\gamma_d|\alpha_2^{(s)}, \beta_2^{(s)}).$$

For the missense mutations, we have

$$[\gamma_m|\Theta_{-\gamma_m}^{(s)}, data] \sim P(f_m|g_{0m}, \gamma_m)[\text{Beta}(\gamma_m|\alpha_1^{(s)}, \beta_1^{(s)})]^{(1-d_m^{(s)})} [\text{Beta}(\gamma_m|\alpha_2^{(s)}, \beta_2^{(s)})]^{d_m^{(s)}}.$$

For mutations in the positive group, we have

$$[\gamma_m|\Theta_{-\gamma_m}^{(s)}, data] \sim P(f_m|g_{0m}, \gamma_m)[\text{Beta}(\gamma_m|\alpha_2^{(s)}, \beta_2^{(s)})].$$

To sample from the marginal posterior of α 's and β 's, first sample each parameter according to the full conditionals, retaining that satisfy the constraint described in Section 3. The full conditionals can be written

$$[\alpha_1|\Theta_{-\alpha_1}^{(s)}, data] \propto f(\alpha_1) \frac{\Gamma(\alpha_1 + \beta_1^{(s)})}{\Gamma(\alpha_1)} (\gamma_{nd}^{(s)})^{\alpha_1-1} \left[\prod_{m:\{d_m^{(s)}=0\}} \frac{\Gamma(\alpha_1 + \beta_1^{(s)})}{\Gamma(\alpha_1)} (\gamma_m^{(s)})^{\alpha_1-1} \right],$$

$$[\beta_1|\Theta_{-\beta_1}^{(s)}, data] \propto f(\beta_1) \frac{\Gamma(\alpha_1^{(s)} + \beta_1)}{\Gamma(\beta_1)} (1 - \gamma_{nd}^{(s)})^{\beta_1-1} \left[\prod_{m:\{d_m^{(s)}=0\}} \frac{\Gamma(\alpha_1^{(s)} + \beta_1)}{\Gamma(\beta_1)} (1 - \gamma_m^{(s)})^{\beta_1-1} \right],$$

$$[\beta_2 | \Theta_{-\beta_2}^{(s)}, data] \propto f(\beta_2) \frac{\Gamma(\alpha_2^{(s)} + \beta_2)}{\Gamma(\beta_2)} (1 - \gamma_d^{(s)})^{\beta_2 - 1} \left[\prod_{m: \{d_m^{(s)} = 1\}} \frac{\Gamma(\alpha_2^{(s)} + \beta_2)}{\Gamma(\beta_2)} (1 - \gamma_m^{(s)})^{\beta_2 - 1} \right].$$

where $m : \{d_m^{(s)} = 1\}$ includes both deleterious missense mutations and known deleterious mutations.