Published in final edited form as:

J Am Stat Assoc. 2005 ; 100: 399–409.

Population-Calibrated Gene Characterization: Estimating Age at Onset Distributions Associated With Cancer Genes

Edwin S. Iversen Jr. and Sining Chen

Edwin S. Iversen, Jr. is Research Assistant Professor, Department of Biostatistics and Bioinformatics and Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708 (E-mail: iversen@stat.duke.edu). Sining Chen is Postdoctoral Fellow, Oncology Biostatistics, Johns Hopkins University, Baltimore, MD 21205 (E-mail: sining@jhu.edu).

Abstract

Phenotypic characterization of rare disease genes poses a significant statistical challenge, but the need to do so is clear. Clinical management of patients carrying a disease gene depends crucially on an accurate characterization of the genetically predisposed disease, including its likelihood of occurrence among mutation carriers, natural history, and response to treatment. We propose a formal yet practical method for controlling for bias due to ignoring ascertainment, defined as the sampling mechanism, when quantifying the association between genotype and disease using data on high-risk families. The approach is more statistically efficient than conditioning on the variables used in sampling. In it, the likelihood is adjusted by a factor that is a function of sampling weights in strata defined by those variables. It requires that these variables and the sampling probabilities in the strata they define either are known or can be estimated. The latter requires a second, population-based dataset. As an example, we derive ascertainment-corrected estimates of penetrance for the breast cancer susceptibility genes BRCA1 and BRCA2. The Bayesian analysis that we use incorporates a modified segregation model and prior data on penetrance derived from the literature. Markov chain Monte Carlo methods are used for inference.

Keywords

Age at onset; Ascertainment correction; Breast cancer susceptibility genes; Model-based poststratification; Penetrance

1. INTRODUCTION

Characterization of disease genes poses a significant challenge. Because Mendelian disease genes are rare, reliable, population-based data on which to characterize the predisposed phenotype is difficult to obtain. However, the need to do so is clear; clinical management of such patients depends crucially on knowing whether the predisposed disease phenotype is fundamentally different from the sporadic in its rate of incidence, age at onset distribution, natural history, or response to treatment. For example, the decision to prophylactically treat a disease-free gene carrier may depend on her probability of subsequently developing the disease.

A key quantity associated with Mendelian traits is penetrance of the gene(s), defined as the distribution of phenotypes associated with a particular genotype. This article focuses on estimation of penetrance curves, defined as the cumulative distribution function (cdf) of the age at disease onset distribution among individuals with the disease genotype. With simple study designs, penetrance may be estimated empirically or, as is necessary with full pedigree data, modeled by parameterizing the appropriate genetic likelihood in terms of a parametric or semiparametric function for age at onset.

Individuals in family studies are often sampled, or ascertained, on the basis of the extent of disease evident in the family. Careful study design may reduce the influence of ascertainment on estimates of penetrance (for an overview, see Thomas 1999). Because carefully designed and executed population-based gene characterization studies are difficult and expensive to implement, and because it is in high-risk populations that rare Mendelian genes first come to light, there will be a persistent need to estimate penetrance using high-risk family data. Such analyses must account for the sampling mechanism by which families are ascertained.

In practice, the variables used in making ascertainment decisions are often unknown, vary from clinic to clinic, and depend on whether the individual is a patient in a clinic or part of formal study. Furthermore, these variables are typically combined to form complex ascertainment rules. High-risk *clinic* rules are more complex and not always adhered to.

In this article we propose a formal, yet practical method for controlling for ascertainment bias that is more statistically efficient than conditioning on the variables used in sampling. In it, the likelihood is adjusted by a factor that is a function of sampling weights in strata defined by those variables. It requires that these variables and the associated sampling probabilities are known. If these probabilities are not known, they may be estimated if a second, population-based dataset that includes the variables is available. If it is not known which variables were used in ascertainment, they may be inferred through an analysis to identify differences between the high-risk and population-based datasets. This may be accomplished by, for example, searching among variables summarizing family history for those that predict sample membership.

The article is organized as follows. In Section 2 we describe ascertainment-corrected models for family data, including our general modeling approach. We detail a simulation experiment comparing three competing approaches in Section 3. In Section 5 we describe our analysis of the high-risk and population-based datasets, summarized in Section 4, to estimate breast and ovarian cancer penetrance associated with the breast cancer susceptibility genes BRCA1 and BRCA2. We conclude with a discussion in Section 6.

2. MODELS AND NOTATION

In what follows, we consider ascertainment conditional models for family data. Let θ denote parameters describing disease penetrance and phenocopy rate and let π denote prevalence of mutated alleles of the disease gene(s) under study; notation and development in the case of multiple genes that define a syndrome involving multiple diseases is the same. Family studies collect phenotype data (e.g., disease status and age or age at disease diagnosis) on members of each sampled individual's family. We call this "family history" or "pedigree" data (following Elston and Stewart 1971) and denote it by $\mathbf{F}$. In the spirit of Ewens and Shute (1986), we partition $\mathbf{F}$ into two parts: $\mathbf{F}_1$, denoting the subset of variables in $\mathbf{F}$ on which the ascertainment decision is made and $\mathbf{F}_2$ denoting what remains; hence $\mathbf{F} = (\mathbf{F}_1, \mathbf{F}_2)$. For example, $\mathbf{F}_1$ may be the number of disease-affected first-degree family members. Let $\mathbf{T}$ denote genetic test result (s) for family members tested for disease gene mutations.

In the case where data are sampled at random from the target population, ascertainment is independent of θ , π , and $\mathbf{F}$, and the appropriate likelihood is proportional to the joint sampling model $\Pr(\mathbf{F}, \mathbf{T} | \theta, \pi)$. This is Elston and Stewart's general model for pedigree data (Elston and Stewart 1971), with a modification to accommodate genetic test results.

The models that we describe have in common that they assume the event at which a family is ascertained, A , is conditionally independent of model parameters given the ascertainment variable(s), that is, that

$$\Pr(A | \mathbf{F}_1, \theta, \pi) = \Pr(A | \mathbf{F}_1). \quad (1)$$

We focus on the ascertainment-corrected joint (ACJ) model, $\Pr(\mathbf{F}, \mathbf{T} | A, \boldsymbol{\theta}, \boldsymbol{\pi})$ (Kraft and Thomas 2000; Hodge 1988). Assuming (1), we have

$$\Pr(\mathbf{F}, \mathbf{T} | A, \boldsymbol{\theta}, \boldsymbol{\pi}) \propto \frac{\Pr(\mathbf{F}, \mathbf{T} | \boldsymbol{\theta}, \boldsymbol{\pi})}{\Pr(A | \boldsymbol{\theta}, \boldsymbol{\pi})}. \quad (2)$$

We model $\Pr(A | \boldsymbol{\theta}, \boldsymbol{\pi})$, once again assuming (1), by introducing $\mathbf{F}_1$ via

$$\Pr(A | \boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_f \Pr(A | \mathbf{F}_1 = f) \Pr(\mathbf{F}_1 = f | \boldsymbol{\theta}, \boldsymbol{\pi}).$$

Here $\mathbf{F}_1$ is a discrete variable, and the sum is over all possible values f of $\mathbf{F}_1$. The ACJ model, as defined by the foregoing two expressions, differs from the model of Kraft and Thomas (2000) in its factorization of $\mathbf{F}$ as $(\mathbf{F}_1, \mathbf{F}_2)$ and the simplifications resulting from application of (1).

When ascertainment is via a simple and strictly adhered-to sampling plan, the stratum-specific ascertainment probabilities, $\Pr(A | \mathbf{F}_1 = f)$, will be known. In the case where the variables used in ascertainment are known but the rule is not, these probabilities must be modeled. Note that

$$\begin{aligned} \Pr(A | \boldsymbol{\theta}, \boldsymbol{\pi}) &\propto \sum_f \frac{\Pr(\mathbf{F}_1 = f | A)}{\Pr(\mathbf{F}_1 = f)} \Pr(\mathbf{F}_1 = f | \boldsymbol{\theta}, \boldsymbol{\pi}) \\ &= E_{(\mathbf{F}_1 | \boldsymbol{\theta}, \boldsymbol{\pi})} \left(\frac{\Pr(\mathbf{F}_1 | A)}{\Pr(\mathbf{F}_1)} \right). \end{aligned} \quad (3)$$

Here $\Pr(\mathbf{F}_1 = f | A)$ is the density of the sampling variable in the high-risk dataset, and $\Pr(\mathbf{F}_1 = f)$ is its population-based analog. Their ratio will be finite and positive, because $\Pr(\mathbf{F}_1)$ places mass everywhere that $\Pr(\mathbf{F}_1 | A)$ does. Hence, given population-based data on the sampling variable $\mathbf{F}_1$, it is possible to estimate $\Pr(A | \boldsymbol{\theta}, \boldsymbol{\pi})$ without making explicit the connection between the particular ascertainment rule used (unknown) and model parameters, as long as unaccounted-for differences in population substructure between the two datasets is ignorable.

When ascertainment depends on $\mathbf{F}_1$, the *empirical* joint distribution of family histories and test results $(\mathbf{F}, \mathbf{T})$ will have sampling weights defined by the ratio $\Pr(\mathbf{F}_1) / \Pr(\mathbf{F}_1 | A)$ evaluated at the sampled $\mathbf{F}_1$. Hence $[E_{(\mathbf{F}_1 | \boldsymbol{\theta}, \boldsymbol{\pi})} (\Pr(\mathbf{F}_1 | A) / \Pr(\mathbf{F}_1))]^{-1}$ can be viewed as a model-based sampling weight. Simpler model-based approaches to analysis of ascertained data have been described previously. Rosenbaum (1987) described sampling weights modeled as a function of data descriptive of class/stratum features; Reilly, Gelman, and Katz (2001) used a similar approach applied to a time series of weekly political polling data; and Lloyd and Jones (2000) corrected for sampling bias in nonparametric density estimation using two samples from the population, both biased by the same function.

Assuming (1), ascertainment correction may also be done by conditioning on phenotype. In the extreme, the retrospective (RET) model defined by $\Pr(\mathbf{T} | \mathbf{F}, \boldsymbol{\theta}, \boldsymbol{\pi})$ (Kraft and Thomas 2000) conditions on the full family history. The correction may also be achieved, with less loss of efficiency, by conditioning only on the variables involved in ascertainment, that is, by using $\Pr(\mathbf{F}_2, \mathbf{T} | \mathbf{F}_1, \boldsymbol{\theta}, \boldsymbol{\pi})$ (Ewens and Shute 1986; Hodge 1988; Pfeiffer, Gail, and Pee 2001). We refer to this as the conditional-on-ascertainment variables (CAV) model. The RET and CAV models are said to be “ascertainment assumption-free” for not explicitly modeling the ascertainment event.

In the case of population-based (i.e., random) sampling, the ACJ model simplifies to the unconditional joint model $\Pr(\mathbf{F}, \mathbf{T} | \boldsymbol{\theta}, \boldsymbol{\pi})$, because $\Pr(A | \mathbf{F}_1 = f)$ does not depend on f . Under weak ascertainment rules [those where $\Pr(A | \mathbf{F}_1 = f)$ is almost constant], ACJ will be nearly as efficient

as a population-based analysis, more efficient than CAV, and far more efficient than RET. On the other extreme, if ascertainment is very strong and sampling is from only one stratum, then ACJ and CAV correspond. Intermediate to these extremes, ACJ is more efficient than CAV.

The foregoing models assume the correct choice of F_1 and of the underlying genetic model. Unaccounted-for population substructure resulting from genetic and environmental modifiers of risk (Begg 2002) and allelic heterogeneity in risk are examples of failures to correctly specify the latter. When the genetic model and ascertainment are correctly specified, the analysis should lead to consistent estimators of the population quantities. But these estimates may be quite sensitive to model misspecification at the level of familial random effects (Epstein, Lin, and Boehnke 2002; Glidden and Liang 2002). Although estimating common parameters in presence of allelic heterogeneity may result in estimates that are useful in certain circumstances, their interpretability and applicability will be diminished and the degree to which they measure a common risk will be questionable. We investigate implications of failure to account for a specific form of allelic heterogeneity in the next section.

3. SIMULATION

In this section we compare efficiency of the retrospective, CAV and ACJ approaches in context of a simulation experiment. We investigate situations where the genetic model and sampling variable F_1 are correctly specified (case 1); where the genetic model is correctly specified but the sampling variable is misspecified (case 2); and where the genetic model is misspecified but the sampling variable is correctly specified (case 3). In all cases we assume a genetic model in which two independent autosomal dominant disease genes predispose individuals in a population to a common age- and gender-independent disease, that the probability of disease among noncarriers is $\theta_0 = .20$, and that the allele frequencies of the two genes are $\pi_1 = .10$ and $\pi_2 = .05$.

For each experimental condition studied, we simulated 250 datasets of 300 individuals. For each individual, we simulated the individual's test results for both genes and family data—specifically, disease incidence among the sampled individual, her parents, and her grandparents. Within each family we simulated segregation of the disease locus alleles down through the family. The genetic tests were assumed to have sensitivity .9 and specificity 1.0. We considered population-based sampling (i.e., a simple random sample of individuals) and five high-risk sampling schemes, where F_1 is defined as the number of disease-affected family members. Table 1 tabulates ascertainment probabilities, $\Pr(A|F_1 = f)$, associated with each scenario. Columns represent scenarios, and rows represent values of F_1 . S1 corresponds to population-based sampling, and subsequent scenarios correspond to progressively stronger ascertainment rules. In the extreme, S6 corresponds to sampling only families with four or more affected members.

In case 1 we assume that there is no allelic heterogeneity in penetrance—the probability of disease for carriers of mutation—of genes 1 and 2, and set $\theta_1 = \theta_2 = .70$. Further, we assume that ascertainment is carried out according to the rules set out in Table 1, that those weights are used in ACJ, and that F_1 is taken to be the number of affected family members when fitting CAV. In case 2 we assume the same, except that in generating the data, we modify ascertainment so that affected individuals with an affected mother and at least one affected maternal grandparent or an affected father and at least one affected paternal grandparent are ascertained according to the probabilities in Table 1, whereas the remaining individuals are ascertained at one-quarter that rate. In case 3 we assume the existence of allelic heterogeneity when generating the data, but ignore this and fit ACJ and CAV as in case 1. Here, when generating the data, $\theta_2 = .70$ and $\theta_1 = .70 + \rho$, where ρ is a uniform $[-.05, .05]$ [case 3(a)] or uniform $[-.25, .25]$ [case 3(b)] family-specific random effect.

In fitting the models, the five parameters ($\theta_0, \theta_1, \theta_2, \pi_1, \pi_2$) are assumed unknown. The sensitivity and specificity of the genetic tests are assumed known and are explicitly adjusted for in the sampling model $\Pr(\mathbf{F}, \mathbf{T} | \boldsymbol{\theta}, \boldsymbol{\pi})$. The ACJ model is fit once when its ascertainment weights are assumed known and once when they are estimated. In the latter case, we simulate independent population-based datasets of sample size 1,000 and estimate the weights by the ratio of frequencies in the high-risk set to a beta-binomial fit to the population-based data. We use the CSIRO/Rothamsted implementation of the Nelder-Mead simplex method (O'Neill 1971) with quadratic surface fitting for locating the posterior mode of $(\theta_0, \theta_1, \theta_2, \pi_1, \pi_2)$ under a uniform prior or, equivalently, its maximum likelihood estimator (MLE). The same algorithm is used to calculate the MLE of the beta-binomial parameters when estimating ACJ weights. Bias and relative efficiency estimates calculated under cases 3(a) and 3(b) for θ_1 assume no heterogeneity and a common population value $\theta_1 = .70$, mean allelic penetrance, as truth.

Table 2 presents simulation-based estimates of bias relative to true parameter values using CAV, ACJ, and RET models by scenario and case averaged over simulation replicates. For example, in case 1 all three models evidently yield overestimates of prevalence and underestimates of penetrance and phenocopy rate. The effect is least pronounced with ACJ (bias $<2.5\%$ in relative magnitude when the weights are known and $<3.6\%$ otherwise) and CAV ($<6.9\%$ in relative magnitude) and most pronounced under RET ($<59\%$ in relative magnitude), a fact likely due to loss of information through conditioning. The apparent systematic biases may reflect incompletely realized asymptotics. Figure 1 presents boxplots of relative bias in estimating θ_1 by its MLE across simulation realizations for each scenario, model, and case. If they exist in case 1, then biases are subtle. In case 2, RET is evidently unbiased, whereas ACJ appears to systematically overestimate θ_1 and CAV appears to systematically underestimate θ_1 . All three models are fairly robust to allelic heterogeneity in θ_1 of the form simulated in cases 3(a) and 3(b). In all cases and scenarios, variability across simulations is smallest for ACJ and largest for RET, reflecting the increasing amount of information in $\mathbf{F}$ for θ_1 conditioned on.

Table 3 tabulates relative efficiency of the MLE under ACJ using known weights to CAV, under ACJ using known weights to RET, and under ACJ using known weights to ACJ using estimated weights by case, scenario, and parameter. When F_1 and the genetic model are correctly specified (case 1), the gain in efficiency using ACJ over CAV or RET can be substantial. These gains persist in magnitude when there is unaccounted for population heterogeneity in penetrance similar to that simulated in cases 3(a) and 3(b). Finally, under some situations where F_1 is misspecified (e.g., case 2), ACJ still may yield more precise (but biased) estimates than CAV and RET. Interestingly, in most cases ACJ is more efficient when the weights are estimated than when known. This effect has been demonstrated in other contexts (Robins, Rotnitzky, and Zhao 1995).

4. DATA

In the following section, we estimate breast and ovarian cancer penetrance associated with the breast cancer susceptibility genes BRCA1 and BRCA2 given data from two sources: a population-based case-control study and a group of six cancer genetics counseling (high-risk) centers. The high-risk dataset comprises detailed family histories and genetic test results for 301 individuals tested for both BRCA1 and BRCA2 mutations using full sequencing. Most probands in the high-risk sample are female; the seven males were removed for purposes of this analysis. Each individual was the first in his or her family to receive genetic testing; hence in all cases the decision to test was made solely on the basis of family history. For the purpose of this analysis, results of genetic tests were simplified to presence or absence of a deleterious mutation. Although there is evidence to the contrary (Risch et al. 2001), we assume that breast

and ovarian cancer penetrance does not depend on the region of the gene in which the mutation occurs. A more detailed description of the high-risk sample was given by Berry et al. (2002).

Population-based sampling probabilities are estimated using data from the Cancer and Steroid Hormone Study (CASH), a case-control study of breast, ovarian, and endometrial cancers (Lee et al. 1987). Excluding endometrial cancer cases, the CASH dataset comprises 4,730 breast cancer cases, 575 ovarian cancer cases, and 4,186 age- and region-matched controls. All cases were first diagnoses of cancer, and no control had been previously diagnosed with any form of cancer. Subjects were recruited from the 1997 National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) program (National Cancer Institute 2005). Women age 20–54 diagnosed with breast, ovarian, or endometrial cancer between December 1, 1980 and December 31, 1982 and residing in one of the eight SEER reporting areas were eligible for the study. Of the 9,491 nonendometrial cases, 264 identified themselves as Jewish. We assumed all to be of Ashkenazi Jewish (AJ) descent. Family history variables include occurrence and age of onset of breast, ovarian, and endometrial cancer for female first- and second-degree relatives, except age at diagnosis for aunts.

We weight case and control data to reflect the early 1980s U.S. population of women age 20–54 using weights estimated from SEER data. Weights were estimates of the probability of surviving breast and ovarian cancer-free to the individual's age for controls, as the probability of developing breast cancer between age 20 and 54 and surviving ovarian cancer-free to the individual's age for breast cancer cases, and as the probability of developing ovarian cancer between age 20 and 54 and surviving breast cancer-free to the individual's age for ovarian cancer cases.

5. ANALYSIS

In what follows we use the ACJ likelihood with estimated weights to estimate age-specific risk of breast and ovarian cancer among BRCA1 and BRCA2 mutation carriers using the datasets just described. In this analysis we use a parametric penetrance model and a simple stratification variable.

5.1 Model for $\Pr(\mathbf{F}, \mathbf{T} | \boldsymbol{\theta}, \boldsymbol{\pi})$

We use a modified version of the BRCAPRO (Parmigiani, Berry, and Aguilar 1998) software to calculate $\Pr(\mathbf{F}, \mathbf{T} | \boldsymbol{\theta}, \boldsymbol{\pi})$. BRCAPRO implements a modified segregation model to calculate the probability that a proband is a carrier of a BRCA1 or BRCA2 mutation given the individual's family history of breast and ovarian cancer among first- and second-degree relatives. We set the sensitivity of full sequencing for mutations at BRCA1 and BRCA2 to .90 and its specificity to 1.0.

Ford and Easton (1995) estimated population prevalence of deleterious BRCA1 and BRCA2 mutations. Because mutation prevalence is notably higher among individuals of AJ descent (Oddoux et al. 1996; Roa, Boyd, Volcik, and Richards 1996), we allow for different BRCA1 and BRCA2 mutation prevalences in the AJ and non-AJ populations. We place independent uniform priors on these, the components of $\boldsymbol{\pi}$. Penetrance of breast and ovarian cancer for noncarriers and penetrance of male breast cancer is fixed at the values assumed by BRCAPRO and described by Parmigiani et al. (1998) and Parmigiani and Berry (2005). The four remaining penetrance curves—one for each gene (BRCA1, BRCA2)-cancer (breast, ovarian) combination—are parameterized as Weibull cumulative distribution functions with asymptote and offset parameters. In particular, penetrance of cancer c ($c = 1$ for breast cancer and $c = 2$ for ovarian cancer) for individuals with a deleterious mutation at BRCA_j ($j = 1$ or $j = 2$) at age a is denoted by

$$\rho_{jc}(a) = \begin{cases} \alpha_{jc} \left(1 - \exp \left(- \left(\frac{a - \omega_{jc}}{\lambda_{jc}} \right)^{\gamma_{jc}} \right) \right) & \text{if } a > \omega_{jc} \\ 0 & \text{otherwise,} \end{cases} \quad (4)$$

where the asymptote α_{jc} represents lifetime risk of disease, the offset parameter ω_{jc} represents the minimum age at diagnosis, γ_{jc} is the curve's shape parameter, and λ_{jc} is the curve's scale parameter. Hence θ is 16-dimensional and comprises the asymptote, offset, shape, and scale parameters for each of the 4 curves. We constructed the prior on θ by fitting penetrance curves, using the Weibull parameterization given earlier, to penetrance data published by Ford et al. (1998) and Struwing et al. (1997), following Iversen, Parmigiani, Berry, and Schildkraut (2000). The resulting prior comprises four independent Student t distributions, one for each block of four parameters that specify a penetrance curve.

5.2 Model for $\Pr(A|\theta, \pi)$

In the following analysis, we use the number of breast and ovarian cancers among female first- and second-degree relatives of the proband as F_1 [eq. (3)]. Although the proband's affected status, ages at diagnosis, the types of cancers, evidence of a maternal or paternal lineage, and the size of the pedigree all come into play along with other, less formal criteria, the number of diseased individuals is probably the single most important aspect of family history used in ascertainment and probably the variable with the most influence on estimating lifetime penetrance.

We model the distribution of $F_1 + 1$ in the population-based data using a geometric distribution. A plot of sample-weighted (see Sec. 4 for details on estimation of sample weights) empirical counts appears to be linear on the log scale. We estimate the failure probability as the exponential of the slope coefficient in the regression of log sample-weighted counts on the predictor variable F_1 . We model $\Pr(F_1|A)$, the distribution of F_1 in the high-risk data, using a Poisson distribution. The ratio of the fitted high-risk to population-based distributions increases monotonically at a slower-than-exponential rate. The effect of this is that values of (θ, π) that are consistent with larger numbers of breast and ovarian cancers among first- and second-degree relatives will down-weight the likelihood relative to those that are not.

The expectation of the ratio of high-risk to population-based distributions of F_1 with respect to $\Pr(F_1|\theta, \pi)$ in (3) is replaced with a Monte Carlo average; the exact expression is computationally intractable. We estimate the expectation by repeatedly drawing samples from $\Pr(F|\theta, \pi)$, calculating F_1 , and tabulating its resulting distribution. To increase the efficiency of this approach, we use importance sampling to oversample families informative for the right tail of the distribution. After 1,000 iterations, the expression $\Pr(A|\theta, \pi)$ is estimated by the sum

$$\sum_{i=1}^{1,000} \frac{\hat{\Pr}(F_1 = f_i|A)}{\hat{\Pr}(F_1 = f_i)} w_i, \quad (5)$$

where i indexes samples f_i of F_1 and w_i represents the importance sampling weight associated with the i th sample. Using this approach, we introduce an additional source (on top of that introduced by the sampling scheme that we use) of Monte Carlo error into the estimation procedure. We have made no attempt to estimate the magnitude of this source of error.

5.3 Analysis and Results

We use a Markov chain Monte Carlo (MCMC) algorithm to draw a sample from the posterior distribution defined by (2). In this algorithm the components of θ associated with each penetrance curve are each blocked with π , and each of the four resulting eight-dimensional blocks is updated in turn. Hence at each iteration the components of θ are updated once, and the components of π are updated four times. An independence step (Tierney 1994) is used for

each block update. In particular, for block j_c of θ , we generate a proposal point (θ'_{jc}, π') from an independence proposal distribution constructed from the output of a preliminary long random-walk Metropolis chain (Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller 1953; Gilks, Richardson, and Spiegelhalter 1996). In each proposal distribution, $(\log(-\log(\omega_{jc})), \gamma_{jc}, \lambda_{jc}, \log(-\log(\alpha_{jc})), \log(-\log(\pi)))$ is Student t distributed with 20 degrees of freedom. In calculating the acceptance ratio, we calculate $\Pr(A|\theta, \pi)$ using (5) under the current and proposed sampled values; at each iteration, these two calculations are carried out with the same set of pseudorandom numbers to remove the effect of sampling error in estimating $\Pr(A|\theta, \pi)$ on the acceptance of proposed samples. Acceptance rates ranged from 17.4% to 25.4%.

The algorithm was run for 67,000 iterations. Plots of sampled values appear stationary, and the Raftery and Lewis (1996) diagnostic indicates that the sample is sufficient to estimate the 2.5th percentile of a marginal posterior to within 1% with probability .99. Although these observations are consistent with convergence of the algorithm and indicate that the sample is informative for the kind of posterior summaries of interest to us, they are no guarantee that the sample is a representative sample from the posterior distribution. We assume that it is and base inference on the sample.

Table 4 provides marginal posterior summaries of prevalence and penetrance parameters, including estimates of the posterior mean, standard deviation (SD), and selected posterior quantiles, as well as means of the assumed prior distributions. Posterior means of the lifetime penetrance parameters α_{11} (BRCA1 penetrance of breast cancer), α_{21} (BRCA2 penetrance of breast cancer), α_{12} (BRCA1 penetrance of ovarian cancer), α_{22} (BRCA2 penetrance of ovarian cancer) are .41 (95% posterior interval, .38–.45), .52 (.42–.61), .18 (.16–.21), and .06 (.02–.11). These are consistent with estimates of penetrance at age 70 derived by Antoniou et al. (2002), who estimated these quantities to be .35, .50, .26, and .091 using a retrospective model; by Hopper, Southey, and Dite (1999), who estimated penetrance of breast cancer at age 70 to be .40 (95% interval, .27–.53) for BRCA1 or BRCA2 carriers in the Australian population; and by Satagopan et al. (2001), who estimated breast cancer penetrance at age 80 to be .59 (.40–.93) among AJ carriers of a BRCA1 founder mutation and .38 (.20–.68) among carriers of a BRCA2 founder mutation. Our breast cancer estimates are slightly higher than those obtained by Fodor et al. (1998), whose combined population-based estimate of breast cancer penetrance among AJ founder mutation carriers is .36, but lower than those of Struewing et al. (1997), who estimated among AJ founder mutation carriers a BRCA1/2 penetrance of breast cancer of about .60 at age 80, a BRCA2 (6174delT) penetrance of ovarian cancer of .10 at age 80, and a BRCA1 (185delAG and 5382insC) penetrance of ovarian cancer of about .20 at age 80. Our estimates are significantly lower than those of Ford et al. (1998), who estimated a BRCA1 breast cancer penetrance at age 70 of .71 and a BRCA2 breast cancer penetrance at age 70 of .84. These estimates are likely to be biased upward, because they are derived from families linked to BRCA1 and hence ascertained a given genotype as well as a phenotype (Siegmund, Gaurderman, and Thomas 1999).

Our estimates of prevalence are largely consistent with those that appear in the literature. Estimates of prevalence among non-AJ individuals (Ford and Easton 1995) are on the order of .0012 (for BRCA1) and .00044 (for BRCA2). We estimate these values to be .0071 (95% interval, .0048–.010) and .0014 (.00070–.0023). Among AJ individuals, estimates derived from Oddoux et al. (1996) and Roa et al. (1996) put these values at about .012 and .014. We estimate these values to be .015 (.011–.021) and .0022 (.0012–.0036). Two of four interval estimates bracket the published estimates; one reason for observed discrepancies may be due to reporting error on ethnicity. Hopper et al. (1999) found, combining BRCA1 and BRCA2, that estimates of penetrance derived from a Mendelian model were insensitive to (combined BRCA1/2) allele frequencies ranging between .010 and .001.

Figure 2 plots summaries of the posterior distribution on the penetrance of breast cancer [Figs. 2(a) and 2(b)] and ovarian cancer [Figs. 2(c) and 2(d)] among carriers of deleterious mutations at BRCA1 [Figs. 2(a) and 2(c)] or BRCA2 [Figs. 2(b) and 2(d)]. Twenty five sample curves each from the four posterior distribution are depicted by narrow gray-dashed lines, whereas posterior means are depicted by wide solid lines and the associated priors are depicted by heavy dotted lines. Notice that in all cases, posterior estimates of penetrance are everywhere lower than corresponding prior estimates and, in particular, that lifetime penetrance is lower and median age at diagnosis among those who develop cancer is later under the posterior than under the prior. This suggests that estimates derived under a noninformative prior may be consistent with even lower penetrance. Figure 2 also reports quartiles of the implied age at diagnosis distribution under the prior and posterior mean penetrance models.

6. DISCUSSION

Uncertainty in cancer penetrance among carriers of cancer susceptibility genes complicates clinical management of patients. Counseling unaffected mutation-positive patients who are not selected for a family history—a problem likely to be faced with greater frequency as the availability of laboratory assays for carrier status become more widely available—is particularly difficult. Our findings—that lifetime risk of disease is lower and age at onset of disease is later, in this population—could have significant clinical implications. In particular, if these findings are verified, then individuals in this population may choose to delay prophylactic actions in expectation of a later age at diagnosis if cancer does develop.

Bias is a critical issue in gene characterization. Ascertainment and study design may have large, unintended effects on parameter estimates. Ascertainment of high-risk individuals as a function of their risk has been termed “size-biased” and results in more extreme members of the population being sampled disproportionate to their representation in it. The method that we describe in this article fills a gap in the literature on penetrance estimation by addressing the problem of size-biased sampling in gene characterization studies (Begg 2002).

Our analysis demonstrates that the ascertainment-corrected likelihood that we describe is a practical approach to characterization of BRCA1 and BRCA2 phenotypes using high-risk family data. In addition, it corroborates recent studies suggesting that penetrance of breast and ovarian cancer among BRCA1 and BRCA2 mutation carriers may be lower than estimated earlier. Furthermore, our simulation experiment demonstrates that the ascertainment-corrected likelihood is an efficient alternative to phenotype-conditional specifications while remaining sensitive to misspecification of F_1 .

A larger dataset of the kind being constructed by the Cancer Genetics Network would allow for several important modifications to the current analysis. The first of these is inclusion of family-specific random effects to account for familial aggregation of disease due to other genetic and environmental factors (Li and Thompson 1997; Gauderman and Thomas 1994). In this formulation of the model, high-risk families may represent a skewed version of the population-based random-effects distribution, much as the high-risk data is itself representative of a skewed version of the population-based distribution over pedigrees. There is evidence, however, that careful ascertainment-adjusted analysis can account for this and deliver consistent population-based estimates (Epstein et al. 2002; Glidden and Liang 2002). Second, a more comprehensive stratification variable could be used. Finally, the assumed genetic model should be modified to allow for measurement error in self-reported incidence of cancer and age at diagnosis of cancer among relatives (Love, Evans, and Josten 1985; Schildkraut and Thompson 1988).

The approach that we have outlined is rather general. Given the appropriate set of population-based calibration data, similar models could be used in a variety of gene-characterization and other problems as long as systematic differences, such as population substructure, are properly accounted for. In context of observational studies, the primary challenge in using this approach is the availability of problem-specific calibration data, a problem that would be reduced if population-based registries such as SEER would collect family history data. In context of a planned study with a defined, possibly size-biased, ascertainment rule, the ascertainment-corrected likelihood that we describe requires no population-based data and is more efficient than a likelihood conditional on the variable on which the ascertainment rule is based.

Acknowledgements

The authors thank Giovanni Parmigiani for numerous conversations on the topic of penetrance estimation and the problem of ascertainment bias and the reviewers for their patience and for their numerous valuable suggestions. They also thank Ellaine Hiller, Judy Garber (Dana Farber Cancer Center), Beth N. Peshkin, Caryn Lerman (Georgetown University Medical Center), Patrice Watson, Henry Lynch (Creighton University), Susan Hilsenbeck (Breast Center at Baylor College of Medicine), Wendy S. Rubinstein (University of Pittsburgh Medical Center), and Kevin Hughes (Lahey Clinic) for their contribution of family data. The authors also acknowledge the contributors to the Cancer and Steroid Hormone Study. Study design and coordination at the Division of Reproductive Health, Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control: Principal Investigator, George L. Rubin; Project Director, Phyllis A. Wingo; and Project Associates, Nancy C. Lee, Michele G. Mandel, and Herbert B. Peterson. Principal investigators at Data Collection Centers: Atlanta, Raymond Greenberg; Connecticut, J. Wister Meigs, and W. Douglas Thompson; Detroit, G. Marie Swanson; Iowa, Elaine Smith; New Mexico, Charles Key, and Dorothy Pathak; San Francisco, Donald Austin; Seattle, David Thomas; Utah, Joseph Lyon, Dee West. Principal Investigators for Pathology Review: Fred Goorstein, Robert McDivitt, and Stanely J. Robboy. Project Consultants: Lonnie Burnett, Robert Hoover, Peter M. Layde, Howard W. Ory, James J. Schlesselman, David Schottenfeld, Bruce Stadel, Linda A. Webster, and Colin White. Pathology Consultants: Walter Bauer, William Christopherson, Deborah Gersell, Robert Kurman, Allen Paris, and Frank Vellios. This work was funded by the National Cancer Institute through the Specialized Program of Research Excellence (SPORE) in Breast Cancer at Duke University, P50 CA68438, through the Cancer Genetics Network, U24 CA78157, and based on work supported by the National Science Foundation under grant No. 0112340.

REFERENCES

- Antoniou A, Pharoah P, McMullan G, Day N, Stratton M, Peto J, Ponder B, Easton D. A Comprehensive Model for Familial Breast Cancer Incorporating BRCA1, BRCA2 and Other Genes. *British Journal of Cancer* 2002;86:76–83. [PubMed: 11857015]
- Begg CB. On the Use of Familial Aggregation in Population-Based Case Probands for Calculating Penetrance. *Journal of the National Cancer Institute* 2002;94:1221–1226. [PubMed: 12189225]
- Berry DA, Iversen ES Jr, Gudbjartsson DF, Hiller E, Garber J, Peshkin BN, Lerman C, Watson P, Lynch H, Hilsenbeck S, Rubinstein WS, Hughes K, Parmigiani G. BRCAPRO Validation, Sensitivity of Genetic Testing of BRCA1/BRCA2, and Prevalence of Other Breast Cancer Susceptibility Genes. *Journal of Clinical Oncology* 2002;20:2701–2712. [PubMed: 12039933]
- Elston RC, Stewart J. A General Model for the Genetic Analysis of Pedigree Data. *Human Heredity* 1971;21:523–542. [PubMed: 5149961]
- Epstein M, Lin X, Boehnke M. Ascertainment-Adjusted Parameter Estimates Revisited. *American Journal of Human Genetics* 2002;70:886–895. [PubMed: 11880949]
- Ewens WJ, Shute NCE. A Resolution of the Ascertainment Sampling Problem, I: Theory. *Theoretical Population Biology* 1986;30:388–412. [PubMed: 3810507]
- Fodor FH, Weston A, Bleiweiss IJ, McCurdy LD, Walsh MM, Tartter PI, Brower ST, Eng CM. Frequency and Carrier Risk Associated With Common BRCA1 and BRCA2 Mutations in Ashkenazi Jewish Breast Cancer Patients. *American Journal of Human Genetics* 1998;63:45–51. [PubMed: 9634504]
- Ford D, Easton DF. The Genetics of Breast and Ovarian Cancer. *British Journal of Cancer* 1995;72:805–812. [PubMed: 7547224]
- Ford D, Easton DF, Stratton M, Narod S, Goldgar D, Devilee P, Bishop DT, Weber B, Lenoir G, Chang-Claude J, Sobol H, Teare MD, Struwing J, Arason A, Scherneck S, Peto J, Rebbeck TR, Tonin P, Neuhausen S, Barkardottir R, Eyfjord J, Lynch H, Ponder BAJ, Gayther SA, Birch JM, Lindblom A, Stoppa-Lyonnet D, Bignon Y, Borg A, Hamann U, Haites N, Scott RJ, Maugard CM, Vasen H, Seitz

- S, Cannon-Albright LA, Schofield A, Zelada-Hedman M. the Breast Cancer Linkage Consortium. Genetic Heterogeneity and Penetrance Analysis of the BRCA1 and BRCA2 Genes in Breast Cancer Families. *American Journal of Human Genetics* 1998;62:676–689. [PubMed: 9497246]
- Gauderman WJ, Thomas DC. Censored Survival Models for Genetic Epidemiology: A Gibbs Sampling Approach. *Genetic Epidemiology* 1994;11:171–188. [PubMed: 8013897]
- Gilks, WR.; Richardson, S.; Spiegelhalter, DJ. Introducing Markov Chain Monte Carlo. In: Gilks, WR.; Richardson, S.; Spiegelhalter, DJ., editors. *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall; 1996. p. 1-19.
- Glidden D, Liang K. Ascertainment Adjustment in Complex Diseases. *Genetic Epidemiology* 2002;23:201–208. [PubMed: 12384971]
- Hodge S. Conditioning on Subsets of the Data: Application to Ascertainment and Other Genetic Problems. *American Journal of Human Genetics* 1988;43:364–373. [PubMed: 3177379]
- Hopper J, Southey M, Dite GS. Population-Based Estimate of the Average Age-Specific Cumulative Risk of Breast Cancer for a Defined Set of Protein-Truncating Mutations in BRCA1 and BRCA2. Australian Breast Cancer Family Study. *Cancer Epidemiology, Biomarkers & Prevention* 1999;8:741–747.
- Iversen ES Jr, Parmigiani G, Berry D, Schildkraut J. Genetic Susceptibility and Survival: Application to Breast Cancer. *Journal of the American Statistical Association* 2000;95:28–42.
- Kraft P, Thomas DC. Bias and Efficiency in Family-Based Gene-Characterization Studies: Conditional, Prospective, Retrospective, and Joint Likelihoods. *American Journal of Human Genetics* 2000;66:1119–1131. [PubMed: 10712222]
- Lee NC, Wingo PA, Gwinn ML, Rubin GL, Kendrick JS, Webster LA, Ory HW. The Reduction of Risk of Ovarian Cancer Associated With Oral Contraceptive Use. *New England Journal of Medicine* 1987;316:650–655. [PubMed: 3821795]
- Li H, Thompson E. Semiparametric Estimation of Major Gene and Family-Specific Random Effects for Age of Onset. *Biometrics* 1997;53:282–293. [PubMed: 9147596]
- Lloyd CJ, Jones MC. Nonparametric Density Estimation From Biased Data With Unknown Biasing Function. *Journal of the American Statistical Association* 2000;95:865–876.
- Love R, Evans A, Josten D. The Accuracy of Patient Reports of Family History of Cancer. *Journal of Chronic Disease* 1985;38:289–293.
- Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equations of State Calculated by Fast Computing Machines. *Journal of Chemical Physics* 1953;21:1087–1092.
- National Cancer Institute. Surveillance, Epidemiology, and End Results (SEER) Program. 2005. available at <http://seer.cancer.gov>
- Oddoux C, Struewing JP, Clayton CM, Neuhausen S, Brody LC, Kaback M, Haas B, Norton L, Borgen P, Jhanwar S, Goldgar D, Ostrer H, Offit K. The Carrier Frequency of the BRCA2 6174delT Mutation Among Ashkenazi Jewish Individuals Is Approximately 1%. *Nature Genetics* 1996;14:188–190. [PubMed: 8841192]
- O'Neill R. Algorithm as 47 Function Minimization Using a Simplex Procedure. *Applied Statistics* 20:338–345.
- Parmigiani, G.; Berry, DA. BRCAPRO homepage. 2005. available at <http://astor.som.jhmi.edu/BayesMendel/brcapro.html>
- Parmigiani G, Berry DA, Aguilar O. Determining Carrier Probabilities for Breast Cancer Susceptibility Genes BRCA1 and BRCA2. *American Journal of Human Genetics* 1998;62:145–158. [PubMed: 9443863]
- Pfeiffer R, Gail M, Pee D. Inference for Covariates That Accounts for Ascertainment and Random Genetic Effects in Family Studies. *Biometrika* 2001;88:933–948.
- Raftery, AE.; Lewis, SM. Implementing MCMC. In: Gilks, WR.; Richardson, S.; Spiegelhalter, DJ., editors. *Markov Chain Monte Carlo Practice*. London: Chapman & Hall; 1996. p. 115-127.
- Reilly C, Gelman A, Katz J. Poststratification Without Population-Level Information on the Postratifying Variable, With Application to Political Polling. *Journal of the American Statistical Association* 2001;96:1–11.
- Risch HA, McLaughlin JR, Cole DEC, Rosen B, Bradley L, Kwan E, Jack E, Vesprini DJ, Kuperstein G, Abrahamson JLA, Fan I, Wong B, Narod SA. Prevalence and Penetrance of Germline BRCA1

- and BRCA2 Mutations in a Population Series of 649 Women With Ovarian Cancer. *American Journal of Human Genetics* 2001;68:700–710. [PubMed: 11179017]
- Roa BB, Boyd AA, Volcik K, Richards CS. Ashkenazi Jewish Population Frequencies for Common Mutations in BRCA1 and BRCA2. *Nature Genetics* 1996;14:185–187. [PubMed: 8841191]
- Robins J, Rotnitzky A, Zhao L. Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. *Journal of the American Statistical Association* 1995;90:106–121.
- Rosenbaum PR. Model-Based Direct Adjustment. *Journal of the American Statistical Association* 1987;82:387–394.
- Satagopan J, Offit K, Foulkes W, Robson M, Wacholder S, Eng C, Karp S, Begg C. The Lifetime Risks of Breast Cancer in Ashkenazi Jewish Carriers of BRCA1 and BRCA2 Mutations. *Cancer Epidemiology, Biomarkers & Prevention* 2001;10:467–473.
- Schildkraut JM, Thompson WD. Familial Ovarian Cancer: A Population-Based Case-Control Study. *American Journal of Epidemiology* 1988;128:456–466. [PubMed: 3414654]
- Siegmund KD, Gaurderman WJ, Thomas DC. Gene Characterization Using High-Risk Families: A Sensitivity Analysis of the Mod Score Approach. *American Journal of Human Genetics* 1999;65:A398.[abstract 2251]
- Struwing JP, Hartge P, Wacholder S, Baker SM, Berlin M, McAdams M, Timmerman MM, Brody LC, Tucker MA. The Risk of Cancer Associated With Specific Mutations of BRCA1 and BRCA2 Among Ashkenazi Jews. *New England Journal of Medicine* 1997;336:1401–1408. [PubMed: 9145676]
- Thomas DC. Design of Gene Characterization Studies: An Overview. *Journal of the National Cancer Institute Monographs* 1999;26:17–23. [PubMed: 10854481]
- Tierney L. Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics* 1994;22:1701–1728.

Figure 1.

Boxplots of Relative Bias When Estimating θ_1 by Its MLE Across Simulation Realizations by Scenario, Model, and Case. Each panel represents a case with sampling scenarios indexing the x-axis and relative bias plotted on the y-axis. Boxplots for CAV (shaded dark gray), ACJ using known weights (shaded black), RET (unshaded), and ACJ using estimated weights (shaded light gray) are arrayed left to right above the associated scenario labels.

Figure 2.

Posterior and Prior Estimates of the Full Penetrance Curves for Breast Cancer [(a) and (b)] and Ovarian Cancer [(c) and (d)] Among Carriers of Deleterious Mutations at BRCA1 [(a) and (c)] or BRCA2 [(b) and (d)]. Samples from the four posterior distributions on the curves are depicted using narrow gray-dashed lines, posterior means are depicted using wide solid lines, and the associated priors are depicted as heavy dotted lines.

Table 1
Table of Ascertainment Probabilities $\Pr(A|F)=f$ Under the Six Ascertainment Scenarios Used in the Simulation Experiment

No. affected members (<i>f</i>)	Scenario					
	S1	S2	S3	S4	S5	S6
0	1	1/4	1/16	1/64	1/256	0
1	1	1/4	1/16	1/64	1/256	0
2	1	1/2	1/4	1/8	1/16	0
3	1	1/2	1/4	1/8	1/16	0
4	1	1	1	1	1	1
5	1	1	1	1	1	1
6	1	1	1	1	1	1
7	1	1	1	1	1	1

NOTE: Columns represent scenarios, and rows represent values *f* of *F*. S1 corresponds to population-based sampling, and subsequent scenarios correspond to progressively stronger ascertainment rules. In the extreme, S6 corresponds to sampling only families with four or more affected members.

Table 2

Relative Bias (ratio of estimated bias to true value) for Estimating Model Parameters Using CAV, ACJ, and RET Under Six Sampling Schemes (S1,...,S6) When Model Specification Is Correct (case 1), When F1 is Misspecified Under CAV and ACJ (case 2), and When Family-Specific Heterogeneity in θ_1 Is Ignored [cases 3(a) and 3(b)]

Scenario	CAV						ACJ, weights known						RET						ACJ, weights estimated					
	π_1	π_2	θ_0	θ_1	θ_2		π_1	π_2	θ_0	θ_1	θ_2		π_1	π_2	θ_0	θ_1	θ_2		π_1	π_2	θ_0	θ_1	θ_2	
Case 1: F_1 and genetic model specified correctly																								
S1	.048	.048	.006	.019	.023	0	.006	.001	.003	.007	.007	.007	.482	.588	.202	.039	.035	.035	.003	.002	.001	.003	.008	.008
S2	.069	.062	.008	.011	.022	.004	.010	.001	.002	.008	.008	.008	.294	.346	.088	.050	.058	.058	.004	.002	.004	.003	.008	.008
S3	.050	.042	.013	.019	.006	.010	.017	.011	.005	.008	.008	.198	.230	.048	.057	.039	.039	.039	.017	.008	.011	.004	.008	.008
S4	.038	.032	.015	.022	.016	.014	.023	.003	.005	.001	.001	.091	.128	.017	.070	.064	.064	.064	.018	.010	.006	.004	.002	.002
S5	.030	.024	.026	.030	.026	.014	.025	.005	.005	.003	.003	.150	.172	.016	.057	.053	.053	.053	.019	.008	.004	.005	.003	.003
S6	.026	.024	.032	.033	.026	.008	.011	.009	.002	.002	.002	.084	.112	.030	.068	.061	.061	.061	.036	.032	.003	.001	.007	.007
Case 2: F_1 specified incorrectly																								
S1	.826	.814	.708	.244	.260	.374	.353	.061	.176	.175	.175	.207	.228	.060	.044	.037	.037	.037	.046	.038	.202	.129	.129	.129
S2	.866	.828	.733	.259	.312	.330	.278	.066	.195	.190	.190	.192	.226	.047	.039	.033	.033	.033	.125	.088	.230	.156	.147	.147
S3	.900	.838	.760	.282	.393	.246	.196	.082	.214	.194	.194	.139	.146	.016	.052	.052	.052	.202	.154	.256	.182	.151	.151	.151
S4	.913	.832	.783	.321	.451	.160	.094	.122	.221	.202	.202	.171	.188	.021	.028	.018	.018	.018	.261	.196	.277	.192	.157	.157
S5	.922	.812	.791	.314	.493	.070	.001	.160	.228	.194	.194	.159	.180	.016	.036	.036	.036	.304	.228	.290	.200	.146	.146	.146
S6	.940	.790	.812	.333	.569	.106	.213	.235	.231	.164	.164	.155	.218	.034	.014	.021	.021	.372	.260	.301	.210	.125	.125	.125
Case 3(a): Unaccounted-for genetic heterogeneity																								
S1	.075	.102	.047	.006	.005	.024	.016	.004	.004	.008	.008	.522	.696	.266	.037	.031	.031	.031	.009	.018	.006	.002	.011	.011
S2	.018	.046	.002	.015	.018	.008	.005	.003	.003	.003	.003	.279	.338	.089	.042	.035	.035	.035	.009	.004	.002	.003	.003	.003
S3	.032	.044	.014	.028	.039	.022	.015	.005	.001	.012	.012	.178	.208	.020	.062	.071	.071	.071	.007	.014	.002	.002	.014	.014
S4	.005	.014	.043	.036	.044	.024	.010	.002	.001	.008	.008	.052	.100	.037	.082	.093	.093	.093	.008	.022	.005	.002	.007	.007
S5	.010	.004	.045	.049	.056	.024	.017	.007	.001	.010	.010	.073	.088	.015	.070	.072	.072	.072	.010	.016	.003	.002	.011	.011
S6	.026	.022	.068	.051	.057	.020	.020	.008	.001	.007	.007	.022	.028	.080	.094	.095	.095	.095	.033	.034	.001	.004	.002	.002
Case 3(b): Unaccounted-for genetic heterogeneity																								
S1	.063	.078	.032	.007	.003	.020	.012	.006	.004	.010	.010	.500	.632	.269	.033	.035	.035	.035	.005	.014	.005	.007	.012	.012
S2	.007	.028	.004	.001	.022	.008	.014	.001	.020	.004	.004	.231	.286	.093	.037	.058	.058	.058	.015	.006	.004	.021	.004	.004
S3	.047	.076	.003	.024	.025	.020	.005	.005	.037	.012	.012	.155	.180	.036	.028	.071	.071	.071	.009	.016	.007	.036	.013	.013
S4	.019	.034	.026	.026	.035	.028	.024	.006	.053	.009	.009	.116	.168	.016	.020	.073	.073	.073	.031	.022	.009	.051	.011	.011

Scenario	CAV			ACJ, weights known			RET			ACJ, weights estimated						
	π_1	π_2	θ_0	θ_1	θ_2	π_1	π_2	θ_0	θ_1	θ_2	π_1	π_2	θ_0	θ_1	θ_2	
S5	-.035	.044	.026	.036	.039	-.030	.052	-.001	.063	.013	-.003	.086	.035	.019	.090	-.015
S6	-.066	.034	.047	.031	.057	-.034	.066	.004	.074	.016	-.028	.112	.017	.012	.063	-.017

Table 3

Estimated Relative Efficiency of ACJ Using Known Weights to CAV, ACJ Using Known Weights to RET, and ACJ Using Known Weights to ACJ Using Estimated Weights for Estimating Model Parameters, by Case and Scenario

Scenario	ACJ to CAV					ACJ to RET					ACJ to ACJ (estimated weights)				
	π_1	π_2	θ_0	θ_1	θ_2	π_1	π_2	θ_0	θ_1	θ_2	π_1	π_2	θ_0	θ_1	θ_2
Case 1: F_1 and genetic model specified correctly															
S1	13.04	6.51	11.55	4.59	2.39	132.49	77.43	110.57	22.83	9.16	.87	.99	.83	.90	.99
S2	13.91	7.45	13.24	5.75	2.67	73.80	49.24	78.19	23.57	8.68	.85	.99	.87	.99	.97
S3	10.68	6.58	8.78	4.03	2.16	63.34	42.56	70.46	18.04	7.84	.91	.97	.90	.97	.98
S4	8.14	5.35	8.91	3.86	2.35	36.40	32.96	50.84	14.55	7.52	.89	.98	.90	.96	.99
S5	7.09	5.51	7.21	4.72	2.81	33.46	25.32	38.12	13.04	7.83	.88	.94	.79	.94	.98
S6	5.64	3.70	6.88	4.69	2.35	20.73	15.14	27.93	13.49	6.02	.88	.93	.85	.96	.95
Case 2: F_1 specified incorrectly															
S1	4.87	3.86	46.00	3.73	3.83	4.80	5.05	26.81	2.20	1.79	.09	.13	3.98	.61	.68
S2	8.34	5.05	45.13	2.93	3.93	7.49	5.76	22.61	1.63	1.60	.26	.19	4.66	.68	.72
S3	14.97	7.88	39.10	2.92	5.21	11.41	7.31	19.53	1.58	1.82	.91	.47	4.65	.76	.73
S4	35.86	12.96	27.08	3.45	6.18	24.53	12.36	9.67	1.17	1.51	3.33	1.07	3.53	.79	.73
S5	59.29	20.70	19.64	3.25	7.22	40.44	18.50	7.24	1.23	1.81	7.10	2.20	2.73	.81	.73
S6	14.68	15.58	11.00	3.39	10.56	8.04	15.85	3.34	1.07	1.79	2.42	2.20	1.55	.85	.77
Case 3(a): Unaccounted-for genetic heterogeneity															
S1	13.30	8.99	16.06	4.27	2.12	124.97	87.77	140.55	27.37	9.90	.92	.98	.90	.96	.97
S2	10.79	7.59	9.98	3.74	2.36	79.00	51.72	83.22	20.95	8.44	.98	.94	.83	.91	1.00
S3	11.15	7.94	10.76	6.23	2.85	47.63	33.45	42.42	19.35	7.51	.88	.95	.81	.90	1.01
S4	10.42	6.59	8.73	6.32	3.04	37.77	25.53	36.65	19.65	9.11	.95	.89	.87	.97	.99
S5	10.28	5.91	9.10	8.64	4.32	36.64	21.21	30.14	15.73	6.89	.99	.95	.79	.95	.96
S6	6.70	4.75	8.47	6.49	3.21	22.67	15.31	28.84	17.56	7.84	.91	1.01	1.00	.84	.93
Case 3(b): Unaccounted-for genetic heterogeneity															
S1	12.22	8.07	13.55	4.81	2.19	106.87	72.23	121.61	23.34	8.51	.92	.99	.92	.98	.95
S2	12.30	6.16	8.58	3.41	2.30	84.30	51.38	69.24	17.18	9.43	.98	.91	.86	1.00	.99
S3	11.09	7.22	8.38	2.79	1.96	47.35	27.35	44.35	10.72	6.04	.89	.99	.87	.94	.97
S4	9.73	5.17	6.95	2.75	2.56	48.44	25.03	37.86	8.67	7.37	1.02	.92	.83	.94	1.02
S5	7.64	3.58	6.39	2.25	2.71	32.60	15.58	29.37	6.42	6.87	1.00	.86	.82	.96	1.01
S6	8.25	3.76	7.95	2.71	2.87	20.83	9.77	21.92	4.69	4.71	.94	.90	.99	.92	.99

Table 4
Marginal Posterior Summaries for the Model's 16 Penetrance and 4 Prevalence Parameters Grouped by Penetrance Curve (defined by Gene+ Site; Gene+ Ethnicity for prevalence) and Parameter

Gene	Cancer site	Parameter	Prior mean	Posterior mean	Posterior sd	Posterior percentiles				
						2.5th	25th	50th	75th	97.5th
ω_{11}	BRCA1	Offset	10.4	11.52	.37	10.58	11.34	11.60	11.79	11.96
γ_{11}	BRCA1	Shape	3.0	2.91	.11	2.70	2.84	2.91	2.98	3.13
λ_{11}	BRCA1	Scale	33.0	38.69	.97	36.81	38.03	38.67	39.33	40.64
α_{11}	BRCA1	Penetrance	.70	.41	.020	.38	.40	.41	.43	.45
ω_{21}	BRCA2	Offset	10.0	8.17	2.17	3.68	6.68	8.38	9.88	11.68
γ_{21}	BRCA2	Shape	3.5	3.41	.35	2.74	3.17	3.40	3.64	4.14
λ_{21}	BRCA2	Scale	40.0	47.34	2.81	41.68	45.50	47.37	49.22	52.85
α_{21}	BRCA2	Penetrance	.77	.52	.050	.42	.48	.52	.55	.61
ω_{12}	BRCA1	Offset	14.0	14.46	.93	12.29	13.91	14.61	15.17	15.81
γ_{12}	BRCA1	Shape	5.0	3.94	.22	3.52	3.80	3.95	4.09	4.38
λ_{12}	BRCA1	Scale	35.0	44.07	1.23	41.68	43.23	44.05	44.88	46.51
α_{12}	BRCA1	Penetrance	.30	.18	.015	.16	.17	.18	.19	.21
ω_{22}	BRCA2	Offset	14.0	10.59	2.89	4.64	8.56	10.81	12.81	15.41
γ_{22}	BRCA2	Shape	5.0	6.93	1.70	3.83	5.76	6.86	8.01	10.46
λ_{22}	BRCA2	Scale	50.0	52.67	6.22	42.02	48.37	52.09	56.51	66.51
α_{22}	BRCA2	Penetrance	.30	.062	.021	.024	.047	.061	.076	.11
π_{10}	BRCA1	Prevalence	.50	.0071	.0013	.0048	.0062	.0070	.0079	.010
π_{20}	BRCA2	Prevalence	.50	.0014	.00042	.00070	.0011	.0013	.0016	.0023
π_{11}	BRCA1	Prevalence	.50	.015	.0025	.011	.014	.015	.017	.021
π_{21}	BRCA2	Prevalence	.50	.0022	.00063	.0012	.0018	.0021	.0026	.0036

NOTE: The parameterization is given by Expression 4. Estimates of each parameter's posterior mean, standard deviation (S.D.), and 2.5th, 25th, 50th, 75th, and 97.5th percentiles are given. Prior means are also provided.