Published in final edited form as:

Genet Epidemiol. 2010 May ; 34(4): 373–381. doi:10.1002/gepi.20493.

Estimating Gene Penetrance from Family Data

Gail Gong, Nathan Hannon, and Alice S. Whittemore

Department of Health Research and Policy Stanford University School of Medicine

Abstract

Family data are useful for estimating disease risk in carriers of specific genotypes of a given gene (penetrance). Penetrance is frequently estimated assuming that relatives' phenotypes are independent, given their genotypes for the gene of interest. This assumption is unrealistic when multiple shared risk factors contribute to disease risk. In this setting, the phenotypes of relatives are correlated even after adjustment for the genotypes of any one gene (residual correlation). Many methods have been proposed to address this problem, but their performance has not been evaluated systematically. In simulations we generated genotypes for a rare (frequency 0.35%) allele of moderate penetrance, and a common (frequency 15%) allele of low penetrance, and then generated correlated disease survival times using the Clayton-Oakes copula model. We ascertained families using both population and clinic designs. We then compared the estimates of several methods to the optimal ones obtained from the model used to generate the data. We found that penetrance estimates for common low-risk genotypes were more robust to model misspecification than those for rare, moderate-risk genotypes. For the latter, penetrance estimates obtained ignoring residual disease correlation had large biases. Also biased were estimates based only on families that segregate the risk allele. In contrast, a method for accommodating phenotype correlation by assuming the presence of genetic heterogeneity performed nearly optimally, even when the survival data were coded as binary outcomes. We conclude that penetrance estimates that accommodate residual phenotype correlation (even only approximately) outperform those that ignore it, and that coding censored survival outcomes as binary does not substantially increase the mean-square error of the estimates, provided the censoring is not extensive.

Keywords

bias; copula model; estimating equations; familial risk correlation; penetrance; genetic heterogeneity; root-mean-squared-error; survival analysis

INTRODUCTION

A high clinical priority is accurate estimation of risks in individuals whose genotypes at a locus are associated with specific chronic diseases. Genes identified through linkage studies generally have low frequencies and are associated with moderately increased risks (rare alleles of moderate penetrance). Those identified through genome-wide association studies are more common and confer smaller risks (common alleles of low penetrance). Reliable risk estimates from both types of alleles are needed for disease prevention and control. Such estimates are obtained from case-control studies of unrelated individuals, or from family studies. The latter are useful for several reasons. First, families with multiple cases of disease are enriched for causal variants, leading to increased statistical power [Risch and Teng 1999, Thornton and McPeck 2007]. Second, family-based estimates avoid the imprecision and computational difficulties that can arise in case control studies of rare variants when few of the controls carry the variant [Saunders and Begg 2003], and they are generally less sensitive to confounding by population stratification [Laird and Lange 2006].

Finally, differences between penetrance estimates from case-control and family studies could elucidate gene-gene or gene-environment interactions [Poynter et al 2007].

We shall compare methods for estimating penetrance using one of two types of study design for ascertaining families. The *population designs*, applicable to diseases monitored with population case registries, gather data from a sample of registered cases (the family probands) and their close relatives. These designs ascertain families because their probands are registered with the disease. In contrast, the *clinic designs* ascertain families containing multiple cases of disease because family members have sought clinical care or counseling. Both designs gather data on disease occurrence and, when possible, DNA from family members. Those who provide DNA are genotyped for the variant(s) of interest. Thus families include members with known phenotypes but possibly unknown genotypes.

Many methods have been proposed for using data from these population and clinic designs to estimate risks associated with a given gene. Some are likelihood-based and involve full modeling of the joint probability of observed family genotypes and phenotypes, conditioned suitably on family phenotypes to accommodate how the families are ascertained. For data from population designs, ascertainment through the affected proband is handled by modeling the conditional probability of family phenotypes Y and observed family genotypes G_{obs} , given the proband's affected phenotype Y_0 :

$$\Pr(Y, G_{obs}|Y_0) = \frac{\Pr(Y, G_{obs})}{\Pr(Y_0)} = \frac{\sum_{G_{unobs}} \Pr(G) \Pr(Y|G)}{\sum_{G_0} \Pr(G_0) \Pr(Y_0|G_0)} \quad (1)$$

Here $G = (G_{obs}, G_{unobs})$ denotes the observed and unobserved genotypes of all members with known phenotypes. The second equality in (1) is based on the assumption that the proband's marginal disease probability, given his own genotype, does not depend on the genotypes of his relatives. The penetrances are determined by parameters in the model for $\Pr(Y|G)$. Likelihoods based on the probability (1) (hereafter called *prospective likelihoods*) have been used by several investigators [e.g., Prentice and Hsu 1997, Gail et al 1999, Chatterjee et al 2006].

For data from clinic designs, penetrance estimates obtained from prospective likelihoods are generally biased, due to failure to account for family ascertainment based on multiple disease occurrence [Kraft and Thomas 2000, Wang et al 2006]. Instead, clinic data are handled by modeling the conditional probability of observed family genotypes given family phenotypes:

$$\Pr(G_{obs}|Y) = \frac{\sum_{G_{unobs}} \Pr(G) \Pr(Y|G)}{\sum_G \Pr(G) \Pr(Y|G)}, \quad (2)$$

[Risch 1984, Hodge and Elston 1994]. Likelihoods based on the probabilities (2) (hereafter called *retrospective likelihoods*) have been used for families ascertained for linkage studies [Ford et al 1998, Choi et al 2008]. And while prospective likelihoods are inappropriate for clinic data, retrospective likelihoods can be applied to population data to obtain unbiased penetrance estimates, albeit with efficiency loss due to conditioning on the phenotypes of probands' relatives. We shall report briefly on the performance of penetrance estimates obtained by analyzing population data using retrospective likelihoods.

The probability $\Pr(G)$ of family genotypes G in (1) and (2) is specified in terms of the frequency of the risk allele, using Hardy-Weinberg and Mendelian inheritance assumptions.

Specification of the family phenotype probabilities $\Pr(Y|G)$ is more problematic. Of particular importance is the extent of correlation assumed among members' phenotypes, given their genotypes at the locus of interest. Some methods assume that phenotypes are conditionally independent, given genotypes. This residual independence assumption ignores correlation due to other unmeasured disease-susceptibility genes and shared environmental factors. Its violation can produce biased estimates [Kraft and Thomas 2000, Begg 2002]. However estimation methods that attempt to fully model familial disease correlation are complex, produce estimates with large variance, and require specialized, computer-intensive software, particularly for the large families, with many unobserved genotypes, that are encountered in practice.

These problems have motivated simpler models for correlation in the joint probability $\Pr(Y|G)$. Some assume only partial disease correlation, by taking members' phenotypes to be conditionally independent, given the phenotype of an index member [Chatterjee and Wacholder 2001, Zhao et al 1998]. Others code the family phenotypes as binary rather than survival outcomes, and model the joint distribution of the families' multivariate binary outcomes. Still others restrict analysis to families known to segregate the risk allele(s) of the gene, on the grounds that nonsegregating families contribute little to penetrance estimates. Yet other methods avoid modeling $\Pr(Y|G)$ entirely and instead model the distribution of family genotype probabilities $\Pr(G_{\text{obs}}|Y)$ in (2) [Meijers-Heijboer 2002, Slager and Schaid 2001, Whittemore and Halpern 2003].

These methods need systematic head-to-head comparison. Here we use simulations to evaluate their estimates in relation to those based on the more complex copula model used to generate phenotypes with residual correlation. We consider rare alleles of moderate penetrance and common alleles of low penetrance, and use data generated with population and clinic designs to estimate the penetrance. The following section describes how we generated the family data and outlines the analytic methods we evaluated. We then present the simulation results and conclude by discussing their implications and their relationship to other findings in the literature.

METHODS

We simulated four types of family data, corresponding to the two types of risk alleles (rare moderately penetrant and common low penetrant) and the two study designs (population and clinic). Each simulation consisted of 100 replicated data sets, generated assuming a common level of residual correlation in times to disease between pairs of family members. Each data set was then analyzed by various methods to estimate disease risk by age 70 years in carriers of the risk allele (the penetrance). We averaged these penetrance estimates over the replicated data sets to obtain their bias, standard deviation (SD) and root-mean-squared-error (RMSE).

DATA GENERATION

For each of the four family data types and each of the 100 replications, we generated K families ($K = 2400$ for population data and $K = 1200$ for clinic data). We generated each family in two steps. First, we generated a family structure and then a family genotype vector G and phenotype vector Y . Second, we repeated this step until we obtained a family that met the eligibility criteria for the study design.

Family structures—Each family consisted of three generations with varying numbers of members in generations 2 and 3 (see Figure 1 for an example). Generation 1 consisted of two parents. The number of offspring in generation 2 was randomly generated according to the distribution of live births in the US in the period 1925-29. Each of these siblings had a

spouse and a number of offspring that was randomly generated according to the distribution of births in the US in the period 1955-59.

Family genotypes—For each family we generated a family genotype vector G by assigning genotypes to founders according to a risk allele frequency q and Hardy-Weinberg genotype frequencies. We then modeled genotype transmissions to nonfounders using Mendel's laws.

Family phenotypes—Given a family's genotypes G , we generated its phenotype Y as a vector of correlated (i.e., statistically dependent) and censored survival times $Y_1, \dots, Y_n$. Here $Y_i = \min(U_i, T_i)$ is the minimum of independent times U to censoring and T to disease. Although we focus on penetrance estimation using family data, our goal is estimation of disease risk for an individual randomly selected from the general population. Marginal disease models such as copula models [Louis and Shih 1995] are particularly useful for this purpose. These models specify that a person's disease probability, given his family's size, structure and genotypes, depends only on his own genotype. This assumption allows one to specify penetrance in terms of easily interpretable hazard ratios or odds ratios. Accordingly, we used a copula model to generate the family survival data for the simulations. Specifically, we generated a family's failure times $T_1; \dots, T_n$ according to the Clayton-Oakes copula model described in the online Supplement [Clayton 1978, Oakes 1989]. The gene was assumed to have a dominant effect. The marginal survival function was modeled as Weibull, with hazard $\lambda(t|c) = e^{\alpha + \beta c} (t - 20)^4$, $20 < t < \infty$. Here c is an indicator for carrier status of the risk allele, assuming the value one if the family member is a carrier, and zero otherwise. The parameter α determines the noncarrier hazard, and β governs the carrier-noncarrier hazard-ratio. This model assumes a common level of residual disease correlation for all pairs of family members.

All disease times were censored at age 70 years. In addition, random censoring times $U_1, \dots, U_n$ were generated according to a Weibull hazard of the form $\psi(u) = e^{-22.6} (u - 20)^4$, giving a 1% probability of censoring before age 70 years. The penetrance parameter of interest is $p(1)$, where $p(c)$ is the probability of uncensored disease before aged 70 years among carriers ($c = 1$) and noncarriers ($c = 0$). The distributions of disease and censoring times were the same for males and females.

Family ascertainment—For the population design, we randomly selected a family proband from generation 2 and considered the family eligible for analysis if this proband was affected. To increase the efficiency of risk estimates, we used two-stage sampling to select eligible families [Kalbfleisch and Lawless 1988, Whittemore and Halpern 1997]. Specifically, we oversampled families whose phenotypes exhibit a hereditary pattern and who thus may be more informative for penetrance. Here we classified a family as “hereditary” if it contained at least one affected member in addition to the proband. In stage 1 we screened each family's phenotype and classified it as either hereditary or nonhereditary. In stage 2 we sampled hereditary families with higher probability (80%) than nonhereditary families (20%).

For the clinic design, eligible families were those containing no more than 15 members, at least three of whom were affected, with at least one affected member in generation III. (See Figure 1.) We randomly selected one of the latter affected members as the “proband”.

Unobserved genotypes—To mimic the unobserved genotypes encountered in practice, we masked family members' genotypes with probabilities that depend on their generation in the pedigree and relation to the proband. Specifically, probands' genotypes were observed with probability one, while genotypes in generations I, II and III were observed with

probabilities 10%, 20% and 30%, respectively, for the proband's first-degree relatives, and 5%, 10% and 15%, respectively, for second-degree relatives.

Parameter Values—The parameters used to generate the data are q , α , β , θ , where q is the frequency of the risk allele, α determines the noncarrier Weibull hazard, e^β is the carrier/noncarrier hazard-ratio, and θ determines the correlation coefficient in times to disease within families. (See online Supplement <http://med.stanford.edu/epidemiology/PenCompareSup121009.pdf> for further details.) We took $\alpha = -20.9158$, which gives a noncarrier cumulative risk by age 70 years of 5%, and $\theta = 2.0$, which gives a familial disease correlation coefficient of 10%. For the rare variant simulations, we took $(q, \beta) = (0.0035, 1.153)$, which gives a carrier prevalence of 0.7% and a penetrance of 15%. For the common variant simulations, we took $(q, \beta) = (0.1500, 0.393)$, which gives a 27.8% carrier prevalence and a 7% penetrance.

DATA ANALYSIS

We examined the performance of several *phenotype-based* penetrance estimates, obtained by specifying models for the family phenotypes $\Pr(Y|G)$ of (1) and (2). For clinic data, we also examined *genotype-based* penetrance estimates obtained by methods that avoid modeling $\Pr(Y|G)$ and instead directly model $\Pr(G_{obs}|Y)$ in (2). Further details of all the methods are provided in the online Supplement (<http://med.stanford.edu/epidemiology/PenCompareSup121009.pdf>).

Phenotype-based penetrance estimates—The phenotype-based methods code the multivariate family phenotype Y as a vector whose components are either censored survival outcomes or binary outcomes. The relation between members' phenotypes and genotypes is specified by a Weibull model with the correct shape parameter for survival outcomes, and a logistic model for binary outcomes. Thus methods based on survival outcomes parameterize the penetrance using the log hazard ratio β , while those based on binary outcomes do so using the log odds-ratio β_{bin} .

The methods also vary in how they specify correlation in the family phenotype probabilities $\Pr(Y|G)$. The simplest models assume conditional independence of phenotypes given genotypes. At the next level of complexity, the genetic heterogeneity methods accommodate residual disease correlation by assuming the presence of an unmeasured disease-susceptibility gene unlinked to the gene of interest, with family phenotypes conditionally independent given genotypes for both genes. This approach is similar to that of frailty models [see Thomas 2004 for a review] and to the polygenic models of Antoniou and Easton [2003]. The marginal disease models for $\Pr(Y|G)$ assume and estimate a common pairwise correlation coefficient for the phenotypes of all pairs of relatives, based either on a Clayton-Oakes copula model for multivariate survival outcomes or on a Bahadur model [Bahadur 1961] for multivariate binary outcomes (Online Supplement <http://med.stanford.edu/epidemiology/PenCompareSup121009.pdf>). However application of marginal disease models to multivariate binary outcomes is limited by the need to constrain the familial disease correlation coefficient, to avoid negative values for $\Pr(Y|G)$ [Zhao et al 1998]. This is a particular problem for large families. To circumvent the problem when analyzing population data, Zhao et al [1998] propose estimating equations that assume only partial correlation among the family outcomes, given the family genotypes and the affected status of the proband. The simplest estimating equation arises by treating relatives' phenotypes as conditionally independent given the proband's affected phenotype. The conditional probabilities of the relatives' phenotypes are derived from the Bahadur model. A variation of this “independent doublet” approach has been used with survival outcomes [Chatterjee and Wacholder 2001].

In the classic trade-off between bias and variance, accommodating familial phenotype correlation in multivariate models for $\Pr(Y|G)$ produces penetrance estimates with less bias but with greater variance [Begg 2002, Chatterjee et al 2006]. We attempted to mitigate this problem by minimizing the number of parameters to be estimated in the likelihoods (1) and (2) of the previous section. Specifically, we assumed that all models correctly specify the family genotype probabilities $\Pr(G)$ using Hardy-Weinberg frequencies with the correct risk allele frequency q for founder genotypes and Mendelian transmission of genotypes to nonfounders. We also avoided estimating the intercept parameter in the Weibull and logistic models by imposing the constraint that all models for $\Pr(Y|G)$ specify the correct risk of disease by age 70 years in a random individual in the population from which the families were sampled. This constraint determines the intercept parameter in both the logistic and Weibull regression models as a function of the regression coefficient for carrier status.

The upper part of Table 1 summarizes the phenotype-based methods whose performance we consider. These estimates are computer-intensive, since the family phenotype probabilities $\Pr(Y|G)$ must be summed over the unobserved genotypes, which can require considerable computing time for data sets containing many large families with few typed members. The software package MENDEL [Lange et al 2001] calculates these sums efficiently for methods assuming residual phenotype independence [Ford et al 1998, Antoniou et al 2003], and it has been adapted to handle them for genetic heterogeneity methods [Antoniou and Easton 2003]. In contrast, the available software for estimating penetrance from population and clinic designs using copula models is limited.

In summary, all phenotype-based methods estimate carriers' cumulative disease risk by age 70 years as functions of an intercept and a regression coefficient for carrier status. The intercept is not estimated; instead it is determined as a function of the regression coefficient by the constraint that the marginal probability of disease by age 70 must equal that of the general population. The regression coefficient is estimated as the solution to an estimating equation formed by summing family-specific contributions. The methods vary in their choice of: 1) outcome (survival or binary); 2) strategy for handling residual phenotype correlation; and 3) extent of conditioning on phenotypes and/or genotypes used to derive a family's contribution to the estimating equation. The two-stage sampling design used to generate the population data is accommodated by weighting each family's contribution to the estimating equation by the inverse of the family's sampling probability [Kalbfleisch and Lawless 1988, Whittemore and Halpern 1997]. Only the Clayton-Oakes copula model conforms to the way the data were generated. Thus its estimates form the gold standard to which the other, computationally simpler estimates are compared.

Genotype-based penetrance estimates—Alternatives to the phenotype-based methods are those that directly model the family genotypes $\Pr(G_{obs}|Y)$ without modeling the phenotype probabilities $\Pr(Y|G)$ on the right side of (2). We considered two types of estimates obtained in this way; both treat the family phenotypes as multivariate binary outcomes. The first method, the *secondary case* method, models the probability of variant carriage for all genotyped affected family members, given their genealogical relationships to a randomly selected affected carrier called the index case [Meijers-Heijboer et al 2002]. This analysis includes only families with at least one affected carrier and at least one additional affected and genotyped member. The ratio R of risks in carriers vs. noncarriers is estimated as the solution to an estimating equation based on the carrier statuses of the affected and typed relatives of the index case. These carrier statuses are treated as independent Bernoulli variates with means that depend on the relative risk R (Online Supplement <http://med.stanford.edu/epidemiology/PenCompareSup121009.pdf>).

The second genotype-based method, the *retrospective logistic method*, obtains penetrance estimates by solving an estimating equation for the log odds-ratio β_{bin} relating genotypes to phenotypes [Slager and Schaid 2001, Whittemore and Halpern 2003]. The estimating function is the efficient score for ordinary logistic regression of phenotypes given genotypes, as applied to affected and unaffected individuals pooled across all families. Thus penetrance can be estimated using standard software for ordinary logistic regression. The lower part of Table 1 provides salient features of these two genotype-based methods.

General approach—All methods produce estimates of a risk-ratio parameter φ . For survival outcomes φ is the hazard ratio e^{β} , and for binary outcomes φ is either the odds-ratio $e^{\beta_{bin}}$ or the carrier/noncarrier relative risk R . All methods estimate penetrance as a function of the parameter estimate $\hat{\varphi}$ and correctly specified values for the risk allele frequency and the disease prevalence in the general population.

We implemented all analyses with Python code [Python 2008] using numerical routines from <www.tramy.us/numpybook.pdf> and <www.numpy.scipy.org>. Computing time for each of the 100 replications of a simulation varied across the four types of family data and the two types of variants, requiring between 0.5 and 1.4 days on a PC with 8 gigabytes memory and quad processors each running 2.66 GHz. This code is freely available by contacting us at <gail.gong@stanford.edu>.

Finally, closed-form variance estimates for all the above penetrance estimates are available from the publications describing them. However we do not evaluate the adequacy of these variance estimates, but focus instead on the bias and efficiency of the penetrance estimates themselves.

RESULTS

RARE RISK ALLELES

For the simulations involving risk alleles of low frequency, carriers comprised 0.7% of the population and the cumulative disease risks were 5% for noncarriers and 15% for carriers, giving a relative risk of 3.0. Thus the true penetrance is 15%. The histograms on the left of Figure 2 show the distributions of penetrance estimates from population data, and those on the right correspond to clinic data. Corresponding biases, SDs and RMSEs for these distributions are shown in Table 2. The first row of both the figure and table show the Clayton-Oakes estimates, to which the other estimates were compared.

Among the phenotype-based estimates, those obtained from the genetic heterogeneity method performed well compared to those from the Clayton-Oakes gold standard. In contrast, the independence estimates were biased upward for the population data, and biased downward for the clinic data, in each case with bias of about one-third the actual penetrance. When the data were analyzed using the simpler binary outcomes, the same patterns emerged: the genetic heterogeneity methods again performed well, and the independence estimates again did poorly with the same pattern of bias as noted for the survival outcomes. For the genetic heterogeneity method, the RMSEs of estimates based on binary outcomes were similar to those based on censored survival data: Table 2 shows RMSEs of 2.2% and 2.3%, respectively, for the population data, and 3.4% and 3.7%, respectively, for the clinic data. The binary-based independent doublet estimates also performed well, as evident from their histograms, their biases and their RMSEs. These results suggest that the performance of penetrance estimates is not seriously compromised by treating survival data as binary, provided the censoring is not extensive.

The genotype-based methods (secondary case and retrospective logistic) were designed for families ascertained on the basis of their phenotypes (i.e., clinic families). Table 2 and Figure 2 show that, for these families, the secondary case estimates exhibited large bias and large variance. The retrospective logistic method showed little bias but relatively large variance. We also investigated how the retrospective logistic method performed when applied to population families. The table and figure show that estimates obtained with population families exhibited more negative bias and smaller variance than those obtained with clinic families.

COMMON RISK ALLELES

For the simulations involving common risk alleles, carriers comprise 28% of the population and the cumulative risks of disease by age 70 years were 5% for noncarriers and 7% for carriers, giving a relative risk of 1.4. The histograms of Figure 3 are analogous to those of Figure 2, with graphs on the left showing the distributions of penetrance estimates from population data, and those on the right corresponding to clinic data. Biases, SDs, and RMSEs for these distributions are shown in Table 3. Again, the first row of both figure and table show the Clayton-Oakes estimates, which form the gold standard for comparison with the others. Several conclusions appear warranted.

First, comparison of Figures 2 and 3 shows that all penetrance estimates performed better for common than rare alleles; the histograms in Figure 3 show tighter clustering around the true penetrance, and the RMSE values in Table 3 are nearly an order of magnitude smaller than those in Table 2. Closer inspection shows that the estimates in Figure 3 exhibit patterns similar to those in Figure 2 for a rare risk allele. For phenotype-based methods, the genetic heterogeneity estimates again outperformed the independence methods, and although the bias is less severe, the independence methods were again biased upward for population data and downward for clinic data. The exception to the similarity between rare and common alleles is the behavior of the independent doublet method. For the common allele simulation, this method exhibited a relatively large downward bias resulting in a large RMSE. Among the genotype-based methods, the secondary case estimates again performed poorly with large bias and large variance, while the retrospective logistic estimates performed well.

OTHER FINDINGS

Several additional findings are noteworthy. First, we investigated why methods that ignore residual phenotype correlation consistently produced penetrance estimates with upward bias for population data and downward bias for clinic data. The upward bias agrees with the findings of others [Gail et al 2001, Begg 2002, Chatterjee et al 2006], but the downward bias is counterintuitive. Additional simulations indicated that the difference is due not to the study design but rather to the form of the likelihood function used to analyze the data. Specifically, we analyzed the same population data using both the prospective likelihood (1) and the retrospective likelihood (2), and found that estimates obtained from the prospective likelihood were biased upward while those obtained from the retrospective likelihood were biased downward (data not shown). The negative bias for the retrospective likelihood did not occur when we removed the constraint that the estimated disease risk in the general population must equal its true value.

Second, we found that needlessly accommodating residual phenotype correlation when the data were generated without it incurred relatively little precision loss. For example, when we generated population data using the Clayton-Oakes method with the correlation parameter set to zero, the standard deviation of the penetrance estimate increased only from 0.9% (for analysis with $\theta = 0$) to 1.2% (with θ arbitrary). This small precision loss may reflect the

parsimony we imposed by specifying externally derived values for the gene frequency and population disease risk.

Finally, we found that phenotype-based methods that were restricted to families segregating the risk allele produced estimates with large bias: penetrance estimates were more than double their true values (data not shown). Some of this bias is due to failure to accommodate the selection of families on the basis of their genotypes. However, as seen in the tables and figures, performance also was poor for the genotype-based secondary case method, which accommodates this selection. For example, its average penetrance estimate for the rare risk allele was 9.6% when the actual penetrance was 15.0%.

DISCUSSION

We used simulations to examine the performance of several methods for estimating penetrance associated with a gene of interest, when applied to families whose disease occurrences are correlated because of other shared risk factors. The simulations produced several interesting results. First, all methods performed considerably better at estimating the penetrance of a common, low-risk allele than that of a rare, moderate-risk allele: nearly all methods performed acceptably for the common allele. Exceptions were estimates based only on families that segregated the risk allele, which performed poorly in all simulations. Moreover, for the mild censoring assumed here, coding the phenotypes as binary rather than survival outcomes did not substantially increase the mean-squared error of the estimates.

We also found that the genetic heterogeneity methods, which were based on the simple assumption of a second unmeasured gene, estimated penetrance nearly as well as the optimal method used to analyze the data. This good performance is encouraging, since these methods can be implemented using the software MENDEL [Lange et al 2001, Antoniou and Easton 2003]. However when interest centers on risk-ratios (rather than absolute risks) associated with the gene of interest, genetic heterogeneity methods can be disadvantageous, because their risk-ratio parameters involve conditional risks, given one's genotype for the unobserved additional gene(s), rather than unconditional risks in the population.

We noted bias in estimates from methods that ignore genotype-based family selection, which agrees with the findings of Carayol and Bonaiti-Pellie [2004]. However, we also found large bias even for the secondary case method, which accommodates the family selection. In contrast, both Carayol and Bonaiti-Pellie [2004] and Choi et al [2008] found that accommodating this selection incurred little bias in the circumstances they considered.

We found that the clinic design provided more efficient penetrance estimates than did the population design, despite the two-stage sampling used to increase the latter's efficiency. Specifically, the variances of estimates from the two designs were comparable, even though, on average, the clinic design included only 2100 genotyped individuals per replication, compared to 6000 for the population design. (See also Gong and Whittemore [2003], for a similar conclusion.) Choi et al [2008] generated data under genetic heterogeneity and analyzed them assuming residual phenotype independence, for comparison with the correct model. They found a clinic design more efficient than a population design for estimating risk-ratios for carriers relative to noncarriers, though not for estimating penetrance. Nevertheless the results of Choi et al support the present findings, since here all analytic methods used externally derived values for the gene frequency and population risk, and so the variances of risk-ratio parameter estimates determine the variances of penetrance estimates.

The simulation results reported here apply both to risk alleles that are causally related to the disease and those that are associated only because of linkage disequilibrium with causal

variants. Even in the latter case, penetrance estimates are needed for risk assessment and prediction. The simulations also are useful because they used family data similar to those found in practice: i.e., families of highly variable size whose members provide fairly complete phenotype information but limited genotype information. In addition, the simulations provide new information on the performance of two strategies for increasing the precision of the penetrance estimates: a) using a two-stage design and analysis for the population data, and b) buttressing the analyses with externally derived values for the gene frequency and the disease prevalence in the general population.

The study also has some limitations. The data were generated assuming a simple copula model, and actual familial disease data may exhibit more complex patterns of residual disease correlation. Also, the extensive computing time needed for the simulations precluded our consideration of alternatives to a dominant effect for the variant on disease risk, and some conclusions might differ for, say a recessive mode of inheritance. And we did not investigate the effects either of attempting to estimate or of misspecifying key quantities, such as the gene frequency, the population disease risk and the form of the baseline hazard function in the survival models. In addition, we evaluated only a single gene rather than groups of several genes. Evaluating the joint effects of multiple genes is a more complex problem which needs further research. Finally, we did not evaluate the performance of the Clayton-Oakes model when applied to data generated under genetic heterogeneity. However the present results suggest that the two methods perform similarly, suggesting robustness against misspecification of the exact form of the residual correlation, as long as it is accommodated.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by NIH grant CA094069. The authors thank the reviewers of an earlier version of the manuscript for helpful suggestions.

References

- Antoniou AC, Easton DF. Polygenic inheritance of breast cancer: Implications for design of association studies. *Genet Epidemiol.* 2003; 25:190–202. [PubMed: 14557987]
- Antoniou A, Pharoah PD, Narod S, Risch HA, Eyfjord JE, Hopper JL, Loman N, Olsson H, Johannsson O, Borg A, Pasini B, Radice P, Manoukian S, Eccles DM, Tang N, Olah E, Anton-Culver H, Warner E, Lubinski J, Gronwald J, Gorski B, Tulinius H, Thorlacius S, Eerola H, Nevanlinna H, Syrjäkoski K, Kallioniemi OP, Thompson D, Evans C, Peto J, Lalloo F, Evans DG, Easton DF. Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. *Am J Hum Genet.* 2003; 72:1117–1130. [PubMed: 12677558]
- Bahadur, RR. A representation of the joint distribution of responses to n dichotomous items In *Studies in Item Analysis and Prediction*, Stanford Mathematical Studies in the Social Sciences IV. Solomon, H., editor. Stanford University Press; Stanford, CA: 1961. p. 158-68.
- Begg CB. On the use of familial aggregation in population-based case probands for calculating penetrance. *J Natl Cancer Inst.* 2002; 94:1221–1226. [PubMed: 12189225]
- Begg CB. Letter to the editor of *Biometrics*. *Biometrics.* 2007; 63:964–965. author reply 965-966. [PubMed: 17825027]
- Carayol J, Bonaiti-Pellié C. Estimating penetrance from family data using a retrospective likelihood when ascertainment depends on genotype and age of onset. *Genet Epidemiol.* 2004; 27:109–117. [PubMed: 15305327]

- Chatterjee N, Wacholder S. A marginal likelihood approach for estimating penetrance from kin-cohort designs. *Biometrics*. 2001; 57:245–252. [PubMed: 11252606]
- Chatterjee N, Zeynep Kalaylioglu Z, Shih JH, Gail MH. Case-control and case-only designs with genotype and family history data: estimating relative risk, residual familial aggregation, and cumulative risk. *Biometrics*. 2006; 62:36–48. [PubMed: 16542227]
- Chatterjee N, Kalaylioglu Z, Shih JH, Gail M. Letter to the editor of *Biometrics*. *Biometrics*. 2006; 63:965–966.
- Clayton DG. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*. 1978; 65:144–151.
- Choi YH, Kopciuk KA, Briollais L. Estimating disease risk associated with mutated genes in family-based designs. *Hum Hered*. 2008; 66:238–251. [PubMed: 18612208]
- 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 BA, Gayther SA, Zelada-Hedman M. Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. *Am J Hum Genet*. 1998; 62:676–689. [PubMed: 9497246]
- Gail MH, Pee D, Benichou J, Carroll R. Designing studies to estimate the penetrance of an identified autosomal dominant mutation: cohort, case-control, and genotyped-proband designs. *Genet Epidemiol*. 1999; 16:15–39. [PubMed: 9915565]
- Gail MH, Pee D, Benichou J, Carroll R. Effects of violations of assumptions on likelihood methods for estimating the penetrance of an autosomal dominant mutation from kin-cohort studies. *Journal of statistical planning and inference*. 2001; 96:167–177.
- Gong G, Whittemore AS. Optimal designs for estimating penetrance of rare mutations of a disease-susceptibility gene. *Genet Epidemiol*. 2003; 24:173–180. [PubMed: 12652521]
- Hodge S, Elston R. Lods, wrods, and mods: the interpretation of lod scores calculated under different models. *Genet Epidemiol*. 1994; 11:329–342. [PubMed: 7813895]
- Kalbleisch JD, Lawless JF. Likelihood analysis of multi-state models for disease incidence and mortality. *Stats in Med*. 1988; 7:149–160.
- Kraft P, Thomas DC. Bias and efficiency in family-based gene-characterization studies: conditional, prospective, retrospective, and joint likelihoods. *Am J Hum Genet*. 2000; 66:1119–1131. [PubMed: 10712222]
- Laird N, Lange C. Family-based designs in the age of large-scale gene-association studies. *Nat Rev Genet*. 2006; 7:385–394. [PubMed: 16619052]
- Lange K, Cantor R, Horvath S, Perola M, Sabatti C, Sinsheimer J, Sobel E. A complete package for the exact genetic analysis of discrete traits in pedigree and population data sets. *Am J Human Genetics*. 2001; 69:504. [PubMed: 11462172]
- Louis TA, Shih JH. Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*. 1995; 51:1384–1399. [PubMed: 8589230]
- Meijers-Heijboer H, van den Ouweland A, Klijn J, Wasielewski M, de Snoo A, Oldenburg R, Hollestelle A, Houben M, Crepin E, van Veghel-Plandsoen M, Elstrodt F, van Duijn C, Bartels C, Meijers C, Schutte M, McGuffog L, Thompson D, Easton D, Sodha N, Seal S, Barfoot R, Mangion J, Chang-Claude J, Eccles D, Eeles R, Evans DG, Houlston R, Murday V, Narod S, Peretz T, Peto J, Phelan C, Zhang HX, Szabo C, Devilee P, Goldgar D, Futreal PA, Nathanson KL, Weber B, Rahman N, Stratton MR, CHEK2-Breast Cancer Consortium. CHEK2-Breast Cancer Consortium: Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nat Genet*. 2002; 31:55–59. [PubMed: 11967536]
- Oakes D. Bivariate Survival Models Induced by Frailties. *J Am Statist Assoc*. 1989; 84:487–493.
- Poynter JN, Figueiredo JC, Conti DV, Kennedy K, Gallinger S, Siegmund KD, Casey G, Thibodeau SN, Jenkins MA, Hopper JL, Byrnes GB, Baron JA, Goode EL, Tiirikainen M, Lindor N, Grove J, Newcomb P, Jass J, Young J, Potter JD, Haile RW, Duggan DJ, Le Marchand L, Colon CFR. Variants on 9p24 and 8q24 are associated with risk of colorectal cancer: results from the Colon Cancer Family Registry. *Cancer Res*. 2007; 67:11128–11132. [PubMed: 18056436]

- Prentice RL. Correlated binary regression with covariates specific to each binary observation. *Biometrics*. 1988; 44:1033–1048. [PubMed: 3233244]
- Prentice RL, Hsu L. Regression on hazard ratios and cross ratios in multivariate failure time analysis. *Biometrika*. 1997; 84:349–363.
- Risch N. Segregation analysis incorporating linkage markers. I. Single-locus models with an application to type I diabetes. *Am J Hum Genet*. 1984; 36:363–386. [PubMed: 6608876]
- Risch N, Teng J. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. II. Individual genotyping. *Genome Res*. 1999; 9:234–241. [PubMed: 10077529]
- Saunders CL, Begg CB. Kin-cohort evaluation of relative risks of genetic variants. *Genet Epidemiol*. 2003; 24:220–229. [PubMed: 12652526]
- Slager SL, Schaid DJ. Evaluation of candidate genes in case-control studies: a statistical method to account for related subjects. *Am J Hum Genet*. 2001; 68:1457–1462. [PubMed: 11353403]
- Thomas, D. *Statistical Methods in Genetic Epidemiology*. Oxford University Press; 2004.
- Thornton T, McPeck MS. Case-control association testing with related individuals: a more powerful quasi-likelihood score test. *Am J Hum Genet*. 2007; 81:321–327. [PubMed: 17668381]
- Wang Y, Ottman R, Rabinowitz D. A method for estimating penetrance from families sampled for linkage analysis. *Biometrics*. 2006; 62:1081–1088. [PubMed: 17156282]
- Whittemore AS, Halpern J. Multi-stage sampling in genetic epidemiology *Statistics in Medicine*. 1997; 16:153–167.
- Whittemore AS, Halpern J. Logistic regression of family data from retrospective study designs. *Genet Epidemiol*. 2003; 25:177–189. [PubMed: 14557986]
- Zhao LP, Hsu L, Holte S, Chen Y, Quiaoit F, Prentice RL. Combined association and aggregation analysis of data from case-control family studies. *Biometrika*. 1998; 85:299–315.
- Zhao LP, Prentice RL. Correlated binary regression using a quadratic exponential model. *Biometrika*. 1990; 77:642–648.

Fig. 1.
Examples of three-generation families generated for the population design (proband designated with arrow) and the clinic design. Filled circles (squares) indicate affected females (males).

Fig. 2. Distribution of 100 estimates of penetrance (probability of disease occurrence by age 70 years) for a rare risk allele of frequency 0.35%, as obtained by several methods. True risk of 15% is indicated with the black arrow.

Fig. 3. Distribution of 100 estimates of penetrance (probability of disease occurrence by age 70 years) for a common risk allele of frequency 15%, as obtained by several methods. True risk of 7% is indicated with the black arrow.

Table I**Methods for Estimating Penetrance from Family Data**

Method	References	Implementation
Phenotype-based		
<i>Survival outcomes</i>		
Clayton-Oakes	Clayton 1978; Oakes 1989	Requires custom programming
Genetic heterogeneity	Antoniou and Easton 2003; Thomas 2004	Possible with MENDEL software ^a
Independence	Antoniou et al 2003; Chatterjee and Wacholder 2001	Possible with MENDEL software
<i>Binary outcomes</i>		
Genetic heterogeneity	Thomas 2004	Possible with MENDEL software
Independence	Breslow and Day 1980	Possible with MENDEL software
Independent doublet	Zhao et al 1998	Requires custom programming
Genotype-based		
<i>Binary outcomes</i>		
Secondary-case	Meijers-Heijboer et al 2002	Uses simple binomial-based likelihood
Retrospective logistic	Slager and Schaid 2001; Whittemore and Halpern 2003	Possible using logistic regression software

^aLange et al 2001

Table II
Bias, Standard Deviation (SD) and Root-Mean-Squared-Error (RMSE) ^a of Penetrance Estimates for a Rare ^b Variant

Method	Population Design			Clinic Design		
	Bias	SD	RMSE	Bias	SD	RMSE
Phenotype-based Methods (Survival Outcomes)						
Clayton-Oakes ^c	-0.2	1.5	1.5	-0.1	1.7	1.7
Genetic heterogeneity	-1.8	1.3	2.2	-2.9	1.9	3.4
Independence	5.2	2.8	5.9	-5.0	0.6	5.0
Phenotype-based Methods (Binary Outcomes)						
Genetic heterogeneity	1.0	2.1	2.3	-0.4	3.7	3.7
Independence	7.3	3.6	8.1	-5.0	0.6	5.0
Independent doublet	0.4	3.7	3.7	-	-	-
Genotype-based Methods (Binary Outcomes)						
Secondary-case	-	-	-	-5.3	9.6	11.0
Retrospective logistic	-3.8	1.5	4.0	-1.6	4.5	4.8

^a all values are multiplied by 100
^b frequency 0.35% and true penetrance 15%
^c gold standard used to generate the data

Bias, Standard Deviation (SD) and Root-Mean-Squared-Error (RMSE)^a of Penetrance Estimates for a Common ^b Variant

Table III

	Population Design			Clinic Design		
	Bias	SD	RMSE	Bias	SD	RMSE
Phenotype-based Methods (Survival Outcomes)						
Clayton-Oakes ^c	-0.0	0.2	0.2	0.0	0.2	0.2
Genetic heterogeneity	-0.1	0.2	0.2	-0.2	0.3	0.3
Independence	0.4	0.2	0.4	-0.5	0.1	0.5
Phenotype-based Methods (Binary Outcomes)						
Genetic heterogeneity	0.1	0.2	0.2	-0.0	0.4	0.4
Independence	0.5	0.2	0.5	-0.5	0.1	0.5
Independent doublet	-1.0	0.1	1.0	-	-	-
Genotype-based Methods (Binary Outcomes)						
Secondary-case	-	-	-	1.7	1.0	2.0
Retrospective logistic	-0.4	0.2	0.5	-0.2	0.3	0.4

^a all values are multiplied by 100

^b frequency 15% and true penetrance 7%

^c gold standard used to generate the data