Ascertainment-Adjusted Parameter Estimates Revisited

Michael P. Epstein, Xihong Lin, and Michael Boehnke

Department of Biostatistics, University of Michigan, Ann Arbor

Ascertainment-adjusted parameter estimates from a genetic analysis are typically assumed to reflect the parameter values in the original population from which the ascertained data were collected. Burton et al. (2000) recently showed that, given unmodeled parameter heterogeneity, the standard ascertainment adjustment leads to biased parameter estimates of the population-based values. This finding has important implications in complex genetic studies, because of the potential existence of unmodeled genetic parameter heterogeneity. The authors further stated the important point that, given unmodeled heterogeneity, the ascertainment-adjusted parameter estimates reflect the true parameter values in the ascertained subpopulation. They illustrated these statements with two examples. By revisiting these examples, we demonstrate that if the ascertainment scheme and the nature of the data can be correctly modeled, then an ascertainment-adjusted analysis returns population-based parameter estimates. We further demonstrate that if the ascertainment scheme and data cannot be modeled properly, then the resulting ascertainment-adjusted analysis produces parameter estimates that generally do not reflect the true values in either the original population or the ascertained subpopulation.

Introduction

Adjusting for nonrandom sampling, or ascertainment, has been an important topic in the genetics literature for many years (e.g., Weinberg 1912; Apert 1914; Fisher 1934; Haldane 1938; Morton 1959; Cannings and Thompson 1977; Elston and Sobel 1979; Ewens and Shute 1986*a,* 1986*b;* Vieland and Hodge 1995; de Andrade and Amos 2000). Ascertainment issues arise often in genetic studies because of the frequent use of nonrandom sampling, particularly when the trait of interest is rare. For a family-based genetic study of a rare disease, a common ascertainment sampling procedure is to collect families with at least one or at least two affected members. Ascertainment usually results in oversampling subjects from the affected subset of the original population and undersampling subjects from the complementary set. Failure to account for this ascertainment effect may lead to biased estimates of the parameters of interest.

After proper adjustment for ascertainment has been made, it is generally assumed that the resulting analysis will yield parameter estimates that reflect the values of the parameters in the original population from which the ascertained data were collected. Recently, Burton et al. (2000) stated that, in the presence of unmodeled parameter heterogeneity, a standard ascertainment-adjusted analysis returns parameter estimates that are biased with respect to the population-based values. This finding has important implications in genetic studies because of the probable existence of unmodeled parameter heterogeneity in a complex genetic trait. The authors' finding implies that it can be difficult, if not impossible, to interpret the results of an ascertainment-adjusted genetic analysis with respect to the original population. This raises the question of whether it is futile even to attempt an ascertainment-adjusted analysis in a genetic study.

Burton et al. (2000) went on to state the important point that, given unmodeled heterogeneity, ascertainment-adjusted parameter estimates reflect the true parameter values in the ascertained subpopulation. We interpret this statement to mean that, in the presence of unmodeled heterogeneity, ascertainment-adjusted parameter estimates converge to the true parameter values in the ascertained subpopulation. Burton and colleagues illustrated their statements with two examples.

In the present article, we make two points regarding ascertainment-adjusted analyses in the presence of latent parameter heterogeneity. First, we demonstrate that the proper construction of the ascertainment-adjusted likelihood (which properly models both the ascertainment mechanism and the true nature of the data) yields population-based parameter estimates. Second, we demonstrate that if one is unable to properly construct the correct ascertainment-adjusted likelihood (as Burton et al. [2000] pointed out, this can occur), then resulting parameter estimates need not reflect the true values in either the original population or the ascertained sub-

Received October 1, 2001; accepted for publication January 7, 2002; electronically published March 5, 2002.

Address for correspondence and reprints: Dr. Michael Boehnke, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029. E-mail: boehnke@umich.edu

0002-9297/2002/7004-0007$15.00

population. We support our points by revisiting the two examples of Burton et al. (2000). For each example, we describe the authors' ascertainment-adjusted methods. We then describe ascertainment-adjustment procedures that yield parameter estimates that (when identifiable) reflect the true parameter values in the original population. Finally, we show that using the standard ascertainment-adjusted analyses in the two examples produce parameter estimates that do not reflect the true parameter values in the ascertained subpopulation.

Material and Methods

Assumptions and Definitions

Suppose our original population consists of a set of n independent sibships. Let n_{ASC} denote the total number of sibships ascertained from the original population and let J_i denote the number of siblings in ascertained sibship i. Let D_{ij} represent an indicator variable for the presence or absence of the disease in the jth sibling in the ith sibship, where $D_{ij} = 1$ if the disease is present and $D_{ij} = 0$ otherwise.

General Form of the Ascertainment-Adjusted Likelihood

In general, one constructs the standard ascertainment-adjusted likelihood by dividing the unconditional likelihood by the probability of the ascertainment event. We let ASC_i denote the ascertainment event for sibship i. For example, ASC_i could represent ascertainment based on the presence of at least one affected sibling, such that

$$ASC_i = \left\{ \sum_{j=1}^{J_i} D_{ij} \geqslant 1 \right\} \ .$$

The ascertainment-adjusted likelihood then takes the form

$$\begin{aligned} L(D \mid ASC) &= \prod_{i=1}^{n_{ASC}} L(D_{i1},D_{i2},\ldots,D_{iJ_i} \mid ASC_i) \\ &= \prod_{i=1}^{n_{ASC}} \frac{L(D_{i1},D_{i2},\ldots,D_{iJ_i})}{L(ASC_i)} \end{aligned} \qquad (1)$$

Example 1: Estimating Disease Prevalence

In their first example, Burton et al. (2000) were interested in estimating disease prevalence under the assumption of a population of n sibships, each of size J. They distributed the sibships into one of K discrete strata, each with a different disease prevalence p_k ($k = 1,\ldots,K$). The affection status of each sibling depended only on the sibship's stratum-specific disease prevalence. Burton and colleagues collected an ascertained subpopulation by ascertaining all n_{ASC} sibships that included at least one affected sibling. Let $N^{(k)}$ and $N^{(k)}_{ASC}$ denote the number of sibships from stratum k in the original population and ascertained subpopulation, respectively. By definition,

$$n = \sum_{k=1}^{K} N^{(k)}$$

and

$$n_{ASC} = \sum_{k=1}^{K} N^{(k)}_{ASC} \ .$$

Burton et al. (2000) estimated the overall disease prevalence p as the average of the prevalence of each stratum weighted by its stratum size, which is asymptotically equivalent to being weighted by the probability of stratum membership. We denote the overall disease prevalence p in the original population by p_P and that in the ascertained subpopulation by p_A. By definition, p_P is estimated by

$$\hat{p}_P = \frac{\sum_{k=1}^{K} p_k N^{(k)}}{n} \ ,$$

whereas p_A is estimated by

$$\hat{p}_A = \frac{\sum_{k=1}^{K} p_k N^{(k)}_{ASC}}{n_{ASC}} \ .$$

Burton et al. (2000) assumed that stratum membership was unobservable and estimated p by combining the ascertained subpopulation of each of the K strata into one overall subpopulation; they then analyzed the resulting sample, using the classical approaches for a homogeneous sample. Because of prevalence heterogeneity across strata, sibships in the higher-risk strata were more likely to be ascertained than were sibships in the lower-risk strata. This leads to differences in the distribution of the values of the overall prevalence between the ascertained subpopulation (p_A) and the original population (p_P).

Burton et al. (2000) assumed that, for a given sibship, D_{i1},D_{i2},...,D_{iJ} were independent Bernoulli random variables with disease probability p. They then constructed the ascertainment-adjusted likelihood across the n_{ASC} ascertained sibships as

$$\prod_{i=1}^{n_{ASC}} \frac{L(D_{i1},D_{i2},\ldots,D_{iJ})}{L\left(\sum_{j=1}^{J} D_{ij} \geqslant 1\right)} = \prod_{i=1}^{n_{ASC}} \frac{p^{\sum_{j=1}^{J} D_{ij}}(1-p)^{J-\sum_{j=1}^{J} D_{ij}}}{1-(1-p)^J}$$

$$= \frac{\prod_{j=1}^{J} [p^j(1-p)^{J-j}]^{n_j}}{[1-(1-p)^J]^{n_{ASC}}} \ , \qquad (2)$$

where n_j represents the number of (ascertained) sibships with j affected members ($j = 1,\ldots,J$) and $n_{ASC} = \sum_{j=1}^{J} n_j$.

The authors' motivation for considering the likelihood (2) is that one would have difficulty constructing the correct likelihood because of the inherent inability to resolve all the latent stratification in the analysis. They acknowledged that likelihood (2) was incorrect because it did not properly account for the prevalence heterogeneity due to the effect of unobserved strata. We note that, in fact, the main reason for likelihood (2) to fail is that it assumes that the disease statuses of all subjects in the ascertained subpopulation are independent. However, under the data-generating mechanism assumed by the authors, $D_{i1},D_{i2},\ldots,D_{iJ}$ are independent only when conditioned on their sibship's stratum membership and therefore are marginally dependent. The likelihood (2) does not account for the marginal dependence of these observations in the pooled subpopulation.

We now illustrate our first point: that an analysis based on the correct likelihood (which properly models the ascertainment criterion and the dependent nature of the data) leads to population-based estimates. Later, we demonstrate our second point: that if the data cannot be modeled properly, then ascertainment-adjusted parameter estimates do not reflect the true values in either the ascertained subpopulation or the original population. It actually is not difficult mathematically to replace the incorrect likelihood (2) with one that correctly accounts for the dependence among the disease status indicators D_{ij}, under the sampling frame assumed by the authors. To allow for the dependence, we must account for the stratum membership of the various sibships within the likelihood. Let π_k be the proportion of the population that is in stratum k. Initially, we assume that π_k is known for all k. Conditional on sibship i being in stratum k, $D_{i1},D_{i2},\ldots,D_{iJ}$ are independent and each follows a Bernoulli distribution with disease probability p_k. The unconditional likelihood for sibship i then has the form

$$L(D_{i1},D_{i2},\ldots,D_{iJ})$$

$$= \textstyle\sum_{k=1}^{K} \pi_k L(D_{i1},D_{i2},\ldots,D_{iJ} \mid \text{stratum}_k)$$

$$= \textstyle\sum_{k=1}^{K} \pi_k [p_k^{\sum_{j=1}^{J} D_{ij}}(1-p_k)^{J-\sum_{j=1}^{J} D_{ij}}] \ .$$

The ascertainment-adjusted likelihood across all n_{ASC} ascertained sibships is then

$$\prod_{i=1}^{n_{ASC}} \frac{L(D_{i1},D_{i2},\ldots,D_{iJ})}{L\left(\sum_{j=1}^{J} D_{ij} \geqslant 1\right)} = \frac{\prod_{j=1}^{J}\left[\sum_{k=1}^{K} \pi_k p_k^j (1-p_k)^{J-j}\right]^{n_j}}{\left[1-\sum_{k=1}^{K} \pi_k (1-p_k)^J\right]^{n_{ASC}}} \ . \qquad (3)$$

Using the ascertainment-adjusted likelihood (3), we can, in principle, obtain estimates $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_K$ of the stratum-specific prevalences $p_1, p_2, \ldots, p_K$ and estimate the overall disease prevalence by

$$\hat{p} = \sum_{k=1}^{K} \pi_k \hat{p}_k \ .$$

However, we show in the Appendix that the estimates of $p_1, p_2, \ldots, p_K$ are only identifiable when the sibship size J is strictly greater than the number of strata K.

A second issue for our ascertainment-adjusted likelihood (3) is that we are assuming both the number of strata K and the probabilities of stratum membership $\pi_1,\pi_2,\ldots,\pi_K$ are known. However, as stated by Burton et al. (2000), these are typically unknown in genetic analyses. In such cases, we might apply latent-class analysis methods and mixture models (Roeder et al. 1999) to the data to obtain valid estimates of the overall disease prevalence $\hat{p}$. If marker genotype data are available for individuals within the original population, we could also estimate K and $\pi_1,\pi_2,\ldots,\pi_K$, using the methods suggested by Pritchard et al. (2000), and then estimate $\hat{p}$ by use of the likelihood (3).

We use an example to contrast the results of the ascertainment-adjusted likelihood (2) with the ascertainment-adjusted likelihood (3). Burton et al. (2000) originally examined a simulated data set of $n = 8{,}000$ sibships, each of size $J = 3$, that were distributed into one of $K = 4$ strata, each with its own stratum-specific disease prevalence. Within stratum k, the authors simulated the disease status of a sibling using a Bernoulli random variable with disease probability p_k. After simulating the disease phenotypes within the sibships ($n = 8{,}000$), the authors ascertained all n_{ASC} sibships with one or more affected siblings.

Burton et al. (2000) estimated the overall disease prevalence p in the ascertained subpopulation using two different analyses. Using the likelihood (2), they estimated p by use of Gibbs sampling procedures (Gelfand and Smith 1990). They also estimated p by use of the method-of-moments Li-Mantel (1968) estimator (see Appendix). Like the ascertainment-adjusted likelihood (2), the validity of the Li-Mantel estimator requires that D_{i1}, D_{i2}, and D_{i3} be independent and be identically distributed as Bernoulli random variables with disease probability p. Application of the Li-Mantel method in this example fails because of the dependence among the D_{ij}. It should be noted that if the D_{ij} are marginally independent, the Gibbs sampling method and Li-Mantel method used by Burton et al. (2000) would yield consistent estimates of the population-based disease prevalence p, even when the population is composed of latent subpopulations with heterogeneous disease prevalences. In the Appendix, we show that this statement holds for the Li-Mantel method.

Burton et al. (2000) found that estimates of disease prevalence p, based on both Gibbs sampling and the Li-Mantel estimator, more closely resembled the prevalence in the ascertained subpopulation than that in the original population. They then asserted that overall prevalence estimates using these two methods reflect the overall disease prevalence in the ascertained subpopulation. We interpret this to mean both estimates asymptotically converge to the true prevalence in the ascertained subpopulation. However, we show in the Appendix that the Li-Mantel estimator does not converge to the true prevalence in the ascertained subpopulation. To verify our theoretical findings, we use the data in the example of Burton et al. (2000) and apply equations (B1) and (B2) in the Appendix. The theoretical overall prevalence is 0.132 in the original population and 0.223 in the ascertained subpopulation. Using equation (B3) in the Appendix, we calculate that the asymptotic theoretical value of the Li-Mantel estimator is 0.238. These values are in nearly perfect agreement with those reported by the authors. It should be noted that the difference between 0.238 and 0.223 is intrinsic and is not due to sampling error in finite samples. Thus, the Li-Mantel estimate that ignores the strata does not reflect the true value in either the original population or the ascertained subpopulation, which validates our second point.

We could not apply our ascertainment-adjusted likelihood (3) to the ascertained data set of Burton et al. (2000), since there are $K = 4$ strata and the sibship size is $J = 3$, which makes $p_1, p_2 \ldots, p_K$ unidentifiable. To assure identifiable prevalence estimates, we modified the example to assume only $K = 2$ disease strata. We simulated a population of $n = 10{,}000$ sibships each of size $J = 3$. Stratum 1 contained 8,000 sibships of size 3 and had a simulated disease prevalence p_1 of 0.10. Stratum 2 contained the remaining 2,000 sibships of size 3 and had a simulated disease prevalence p_2 of 0.40. The population characteristics are shown in table 1. The overall population prevalence is then $p_P = (0.10)(8{,}000/10{,}000) + (0.40)\,(2{,}000/10{,}000) = 0.16$.

To help in interpretation, we simulated the number of sibships with zero, one, two, and three affected siblings within each stratum to be the numbers expected. We then ascertained all $n_{ASC} = 3{,}736$ sibships with at least one affected sibling. The characteristics of the ascertained subpopulation are shown in table 2. The prevalence in the ascertained subpopulation is $p_A = (0.10)(2{,}168/3{,}736) + (0.40)\,(1{,}568/3{,}736) = 0.226$.

Table 1

Original Population Characteristics

Stratum	No. of Sibships	No. of Siblings	No. of Affected Siblings	Disease Prevalence
1	8,000	24,000	2,400	.10
2	2,000	6,000	2,400	.40
Total	10,000	30,000	4,800	.16

From table 2, the numbers of sibships with one affected sibling (n_1), two affected siblings (n_2), and three affected siblings (n_3) across both strata are 2,808, 792, and 136, respectively. Using these ascertained counts and knowing $\pi_1 = 4/5$ and $\pi_2 = 1/5$, we applied our ascertainment-adjusted likelihood (3). Using a Fisher-scoring estimation procedure, we obtained stratum-specific prevalence estimates of $\hat{p}_1 = 0.10$ (SE $= 0.020$) and $\hat{p}_2 = 0.40$ (SE $= 0.008$), consistent with the values of p_1 and p_2 in the original population and not that in the ascertained subpopulation. We then estimated the overall prevalence as $\hat{p} = 0.16$ (SE $= 0.017$), which also reflects the overall disease prevalence in the original population. This validates our first point.

We then compared our results with those obtained by means of the classical procedures used by Burton et al. (2000), which did not use any information about the dependent nature of the data and were therefore biased. We applied a Fisher-scoring procedure using the likelihood (2) and obtained a biased prevalence estimate of 0.241 (SE $= 0.004$). Likewise, when we applied the authors' Li-Mantel estimator, we obtained a biased prevalence estimate of 0.237 (SE $= 0.006$). Using (B3) in the Appendix, we found that the asymptotic theoretical value of the Li-Mantel estimator is 0.237. These estimates do not reflect the overall disease prevalence in either the ascertained subpopulation ($p_A = 0.226$) or the original population ($p_P = 0.16$). These results are consistent with our second point.

The results from this example support our two main points. We can consistently estimate the overall disease prevalence in the original population from the disease statuses of the siblings in the ascertained subpopulation if we can correctly model the dependent structure of the data in the ascertainment-adjusted likelihood. If not, the resulting estimates need not reflect the true parameter values in either the original population or the ascertained subpopulation. Not surprisingly, incorrect specification of the likelihood, as in equation (2), can lead to biased estimates of the disease prevalence. If a non–likelihood-based approach, such as the method-of-moments Li-Mantel estimator, is used, then it is important to make sure the assumptions regarding the nature of the data (such as independent observations) are valid.

Table 2

Ascertainment Subpopulation Characteristics

	No. of Sibships with			Total No. of Ascertained		
Stratum	1 Affected Sibling	2 Affected Siblings	3 Affected Siblings	Sibships	Siblings	Affected Siblings
1	1,944	216	8	2,168	6,504	2,400
2	864	576	128	1,568	4,704	2,400
Total	2,808	792	136	3,736	11,208	4,800

Example 2: Estimating Parameters in a Logistic Variance-Component Model

In their second example, Burton et al. (2000) investigated the effect of ascertainment on parameter estimates in a logistic variance components model. They simulated the disease-status indicator D_{ij} as a Bernoulli random variable with mean μ_{ij}, using a logistic variance-components model where $\eta_{ij} = \ln[\mu_{ij}/(1-\mu_{ij})]$ and $\eta_{ij} = \alpha + \beta_B z_{ij,B} + \beta_N z_{ij,N} + C_i$ (Breslow and Clayton 1993). In this model, α represents the overall intercept, β_B is the regression coefficient for a binary covariate z_B, β_N is the regression coefficient for a normally distributed covariate z_N, and C_i is a random effect shared by all members of the *i*th sibship. Fixed covariates were centered about their means, to have expected values of zero. The random effect C_i was assumed to follow a normal distribution with mean of zero and variance σ_C^2. After simulating sibships under the logistic variance-components model, the authors ascertained all sibships with at least one affected member from the original population, to form their ascertained subpopulation.

In the example, we focus on illustrating our first point: that an ascertainment-adjusted analysis based on a properly constructed ascertainment-adjusted likelihood returns population-based parameter estimates. To demonstrate this, we first examined the ascertainment-adjusted likelihood that Burton et al. (2000) used for analysis. After viewing the computer code that Burton et al. (2000) used, we determined that the authors constructed their ascertainment-adjusted likelihood by dividing the likelihood of the data by the probability of ascertainment conditional on the random effects. They then integrated the conditional ascertainment-adjusted likelihood over the random effects C_i. Specifically, their ascertainment-adjusted likelihood had the form

$$\prod_{i=1}^{n_{ASC}} \int \frac{\left[\prod_{j=1}^{J_i} L(D_{ij} \mid C_i)\right]}{\left[L\left(\sum_{j=1}^{J_i} D_{ij} \geqslant 1 \mid C_i\right)\right]} f(C_i)dC_i$$

$$= \prod_{i=1}^{n_{ASC}} \int \frac{\left[\prod_{j=1}^{J_i} L(D_{ij} \mid C_i)\right]}{\left[1 - \prod_{j=1}^{J_i} L(D_{ij} = 0 \mid C_i)\right]} f(C_i)dC_i \ , \qquad (4)$$

where

$$L(D_{ij} \mid C_i)$$

$$= \left(\frac{e^{\alpha+\beta_B z_{ij,B}+\beta_N z_{ij,N}+C_i}}{1+e^{\alpha+\beta_B z_{ij,B}+\beta_N z_{ij,N}+C_i}}\right)^{D_{ij}} \left(\frac{1}{1+e^{\alpha+\beta_B z_{ij,B}+\beta_N z_{ij,N}+C_i}}\right)^{1-D_{ij}}$$

and where $f(C_i)$ denotes the probability-density function of the normally distributed random variable C_i.

However, using the usual ascertainment-adjusted likelihood (1), we obtained the following ascertainment-adjusted likelihood for this example:

$$\prod_{i=1}^{n_{ASC}} \frac{L(D_{i1},D_{i2},\ldots,D_{iJ_i})}{L\left(\sum_{j=1}^{J_i} D_{ij} \geqslant 1\right)}$$

$$= \prod_{i=1}^{n_{ASC}} \frac{\int \left[\prod_{j=1}^{J_i} L(D_{ij} \mid C_i)\right] f(C_i)dC_i}{\int \left[1 - \prod_{j=1}^{J_i} L(D_{ij} = 0 \mid C_i)\right] f(C_i)dC_i} \ . \qquad (5)$$

The correct ascertainment-adjusted likelihood (5) is different from (4). The likelihood (5) requires integrating over the distribution of the random effects C_i in the numerator and denominator separately before taking their ratio. In contrast, the likelihood (4) is misspecified and conditions on both the ascertainment and the random effects first, followed by integration over the distribution of the random effects. Results based on the likelihood (5) are consistent with the suggestion by the authors that a likelihood-based model can be constructed that returns population-based parameter estimates (see below).

Burton et al. (2000) applied the ascertainment-adjusted likelihood (4) to analyze a simulated data set. The authors set $\alpha = -5$, $\beta_B = -0.4$, $\beta_N = 0.3$, and $\sigma_C^2 = 4.5$ in their logistic variance-components model. They simulated sibships with five members and then ascertained samples of 1,000 sibships, each with at least one affected member. The authors correctly noted that this ascertainment criterion selects sibships in which values of C_i are primarily in the upper tail of the normal distribution, so that the features of the random effects C_i in the ascertained subpopulation are different from those in the original population. They also noted that, although the random effects are still approximately normally distributed in the ascertained subpopulation, the empirical mean and variance of the C_i were 2.76 and 2.42, respectively, in contrast to 0 and 4.5 in the original population. This affects the values of the grand mean (α) and the variance parameter (σ_C^2) in the ascertained subpopulation. In the subpopulation, the grand mean (α) is $E[\eta_{ij}] = E[\alpha + \beta_B z_{ij,B} + \beta_N z_{ij,N} + C_i] = -5 + 2.76 = -2.24$, whereas the variance parameter σ_C^2 is 2.42. So, although the true parameter values of (α, σ_C^2) were (-5, 4.5) in the original population, the authors expected (α, σ_C^2) to be closer to (-2.24, 2.42) in the ascertained subpopulation.

Burton et al. (2000) performed their ascertainment-adjusted analysis by applying the likelihood (4), using Gibbs sampling procedures (Gelfand and Smith 1990; Zeger and Karim 1991) in the software package WinBUGS (Spiegelhalter et al. 2000). The results of their analysis yielded parameter estimates of $\hat{\alpha} = -2.15$ (SE = 0.11) and $\hat{\sigma}_C^2 = 1.98$ (SE = 0.32) as reported in an erratum by Burton et al. (2000). These estimates were closer to the expected values (-2.24, 2.42) in the ascertained subpopulation than those in the original population (-5, 4.5). From these results, the authors argued that the ascertainment-adjusted parameter estimates reflected the values of the parameters in the ascertained subpopulation rather than those in the original population. We suggest instead that this conclusion results from the use of a misspecified likelihood and does not represent the true nature of the problem.

To study whether we can recover mean values of (α, σ_C^2) in the original population by use of the ascertainment-adjusted likelihood (5), we simulated 100 data sets of 1,000 ascertained sibships, each of size 5, using the same logistic variance-components model and same ascertainment criterion as Burton et al. (2000). We analyzed the ascertained subpopulation by maximizing the likelihood (5), which we evaluated using adaptive Gaussian quadrature (Pinheiro and Bates 1995). To ensure a high degree of accuracy, we used 20 quadrature points in our analyses. We implemented these estimation procedures using the SAS version 8 procedure PROC NLMIXED (SAS Institute). Our SAS code is available upon request.

Our analyses yielded mean estimates of α and σ_C^2 of -4.77 (SD = 0.74) and 4.21 (SD = 1.01), respectively, over the 100 simulated data sets. These results are consistent with the generating values of -5.0 and 4.5 in the original population and are inconsistent with those of -2.24 and 2.42 in the ascertained subpopulation. Appealing to asymptotics, we repeated the simulations with 100 data sets of 10,000 ascertained sibships of size five. Analyses yielded even better mean estimates of α and σ_C^2 of -4.95 (SD = 0.24) and 4.43 (SD = 0.33), respectively. Our results for this example support our first point that, for a well-specified model, ascertainment-adjusted parameter estimates reflect the true values of the parameters in the original population when the correct ascertainment-adjusted likelihood is used.

Discussion

Given a well-defined ascertainment scheme, it has long been assumed that ascertainment correction leads to parameter estimates that reflect parameter values in the population. Burton et al. (2000) recently demonstrated that, given unmodeled heterogeneity, the usual ascertainment adjustment leads to parameter estimates that do not reflect those in the original population. This conclusion is certainly true and is a useful warning to avoid performing genetic analyses uncritically.

Burton et al. (2000) go on to state the important finding that, given unmodeled heterogeneity, ascertainment-adjusted parameter estimates reflect parameter values in the ascertained subpopulation, and they support their claim with two examples. We demonstrate instead that: (1) if the genetic mechanism and ascertainment scheme can be appropriately modeled, the genetic analysis should yield estimates consistent with the parameter values in the original population; and (2) if not, the estimates using the conventional method cannot be expected to reflect the parameters in either the original population or the ascertained subpopulation.

To support our argument, we revisited the two examples of Burton et al. (2000) and showed that, for these examples, properly-specified analyses yield ascertainment-adjusted parameter estimates that reflect parameter values in the original population. As we have shown, the key to recovering estimates that reflect parameter values in the original population is correct specification of the ascertainment-adjusted likelihood in the analysis. Incorrect specification of the ascertainment-adjusted likelihood owing to, for example, use of the conventional method, unknown model features, nonidentifiability of the correct model, or uncertain ascertainment scheme, can be expected to lead to parameter estimates that do not reflect the true values in either the original population or the ascertained subpopulation. Similar conclusions likely hold for non–likelihood-based ascertainment-adjusted estimation procedures. We showed this clearly in example 1, where we demonstrated that the conventional Li-Mantel method in this context failed to consistently estimate the true prevalence value in either the original population or the ascertained subpopulation.

Although we did not prove that the ascertainment-correction equation (1) works in general to obtain population-based parameter estimates, it is reasonable to assume that it does in cases for which the correct ascertainment-adjusted likelihood can be derived. We feel it is important to emphasize that proper construction of the ascertainment-adjusted likelihood (1) is necessary in order for the ascertainment-adjusted analysis to return valid population-based estimates. As Burton et al. (2000) pointed out, circumstances exist in the analysis of complex traits in which one will be unable to correctly model the true nature of the data by use of (1), owing, perhaps, to the inability to resolve all the hidden data-influencing strata. In such cases, the resulting ascertainment-adjusted parameter estimates cannot be expected to reflect the true values of the parameters in either the original population or the ascertained subpopulation. To avoid this unpleasant predicament in genetic studies, we should seek, when possible, to apply current statistical methods, such as those described by Pritchard et al. (2000), and to develop new approaches, such as mixture models, to identify hidden strata.

Acknowledgments

We thank Drs. Robert Elston and Jane Olson for their helpful comments. We thank Dr. Paul Burton for generously providing us his WinBUGS computer code. This work was supported by National Institutes of Health grants T32 HG00040 (to M.P.E.), R29 CA76404 (to X.L.), and R01 HG00376 (to M.B.).

Appendix A

Identifiability of $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_K$ by Use of the Ascertainment-Adjusted Likelihood (3)

In this Appendix, we briefly describe why estimates of stratum-specific prevalences p_1, $p_2, \ldots, p_K$, by use of the likelihood (3) are identifiable only when sibship size J is strictly greater than the number of strata K. To show this holds, define the function

$$R_j(p_1, \ldots, p_K) = \frac{\sum_{k=1}^{K} \pi_k p_k^j (1 - p_k)^{J-j}}{1 - \sum_{k=1}^{K} \pi_k (1 - p_k)^J}$$

for $j = 1, \ldots, J - 1$. We can rewrite the ascertainment-adjusted likelihood (3) as

$$\left(\prod_{j=1}^{J-1} [R_j(p_1, p_2, \ldots, p_K)]^{n_j}\right) \left[1 - \sum_{j=1}^{J-1} R_j(p_1, p_2, \ldots, p_K)\right]^{n_{ASC} - \sum_{j=1}^{J-1} n_j} .$$

We can easily obtain maximum-likelihood estimates of $\hat{R}_j(p_1, p_2, \ldots, p_K)$ ($j = 1, \ldots, J - 1$) from equation (3), and, from these estimates, determine maximum likelihood estimates of $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_K$. However, if $K > J - 1$, then $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_K$ are clearly nonidentifiable. Therefore, we will only obtain identifiable estimates of $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_K$ when the sibship size J is strictly greater than the number of strata K ($J \geqslant K + 1$).

Appendix B

The Li-Mantel (1968) Estimator of Disease Prevalence Assuming Complete Ascertainment

Assume we have a population consisting of n sibships, each of size three. Let n_j denote the number of sibships in the population with j affected siblings ($j = 0, .., 3$) such that $n = n_0 + n_1 + n_2 + n_3$. As before, let D_{ij} denote the affection status of the jth sibling in the ith sibship. Also, let $a = n_1 + 2n_2 + 3n_3$ denote the total number of affected siblings in the population.

To estimate the overall disease prevalence p, we collect all sibships from the population that have at least one affected sibling, to form the ascertained subpopulation. As defined earlier, we let $n_{ASC} = n_1 + n_2 + n_3$ denote the total number of sibships in the ascertained subpopulation. Also, let $m_{ASC} = 3n_{ASC}$ denote the total number of siblings in the ascertained subpopulation and define a_{ASC} as the number of affected siblings in the ascertained subpopulation. Under our complete ascertainment model, $a_{ASC} = a$. The Li-Mantel (1968) estimator of p then takes the form $\hat{p}_{LM} = (a_{ASC} - n_1)/(m_{ASC} - n_1)$. If the values of p are the same for all subjects within the population, then $\hat{p}_{LM}$ is a consistent, but not unbiased, estimator of p that solves the estimating equation $a_{ASC} - n_1 = p\,(m_{ASC} - n_1)$ (Li and Mantel 1968; Burton et al. 2000).

Li-Mantel Estimator Assuming Multiple Strata and Marginal Dependence of Siblings within a Sibship

Now, assume that the disease prevalence varies across strata within the original population. To be consistent with the first example of Burton et al. (2000), assume that the original population contains $K = 4$ strata with prevalences p_1, p_2, p_3, and p_4. Assume that the disease statuses of siblings are independent only when conditioned on stratum membership (so the disease statuses of siblings are marginally dependent). Let π_k denote the proportion of the original population found in stratum k. Also, let $N^{(k)}$ and $N^{(k)}_{ASC}$ denote the number of sibships from stratum k in the original population and ascertained subpopulation, respectively. By definition,

$$n = \sum_{k=1}^{4} N^{(k)}$$

and

$$n_{ASC} = \sum_{k=1}^{4} N^{(k)}_{ASC} \ .$$

Therefore, the overall disease prevalences in the original population and the ascertained subpopulation, which we denote as p_P and p_A, respectively, converge in probability to the following forms:

$$p_P = \frac{\sum_{k=1}^{4} p_k N^{(k)}}{n} \rightarrow \sum_{k=1}^{4} \pi_k p_k \tag{B1}$$

and

$$p_A = \frac{\sum_{k=1}^{4} p_k N^{(k)}_{ASC}}{n_{ASC}} \rightarrow \frac{\sum_{k=1}^{4} \pi_k p_k [1-(1-p_k)]^3}{\sum_{k=1}^{4} \pi_k [1-(1-p_k)]^3} \ . \tag{B2}$$

Suppose we fail to detect strata and only observe the pooled ascertained sibship counts (n_1, n_2, and n_3). Burton et al. (2000) stated that the Li-Mantel (1968) estimator $\hat{p}_{LM}$ should reflect the disease prevalence in the ascertained subpopulation p_A, but not that in the original population p_P. We show that the Li-Mantel estimate $\hat{p}_{LM}$ does not consistently estimate p_P or p_A. To show this, we evaluate the marginal expectations $E[a_{ASC}]$, $E[m_{ASC}]$, and $E[n_1]$ by conditioning on all possible strata. We obtain the following expected values:

$$E[a_{ASC}] = \sum_{i=1}^{n}\sum_{j=1}^{3} E[D_{ij}] = \sum_{i=1}^{n}\sum_{j=1}^{3}\sum_{k=1}^{4} \pi_k E[D_{ij} \mid \text{stratum } k] = 3n\sum_{k=1}^{4} \pi_k p_k = 3np_P$$

$$E[m_{ASC}] = 3E[n_{ASC}] = 3n\sum_{k=1}^{4} \pi_k (1-(1-p_k)^3)$$

$$E[n_1] = n\sum_{k=1}^{4} 3\pi_k p_k (1-p_k)^2 = 3n\sum_{k=1}^{4} \pi_k p_k (1-p_k)^2 \ .$$

Using these expected values, we have

$$\hat{p}_{LM} \rightarrow \frac{E[a_{ASC} - n_1]}{E[m_{ASC} - n_1]} = \frac{3n\sum_{k=1}^{4}\pi_k p_k - 3n\sum_{k=1}^{4}\pi_k p_k(1-p_k)^2}{3n\sum_{k=1}^{4}\pi_k(1-(1-p_k)^3) - 3n\sum_{k=1}^{4}\pi_k p_k(1-p_k)^2} = \frac{\sum_{k=1}^{4}\pi_k p_k(2p_k - p_k^2)}{\sum_{k=1}^{4}\pi_k(2p_k - p_k^2)} \ . \tag{B3}$$

By comparison of (B3) with the theoretical forms of p_P and p_A in (B1) and (B2), it is clear that, when the disease statuses are marginally dependent and we fail to account for strata, the Li-Mantel estimate fails to consistently estimate the overall disease prevalence in either the original population (p_P) or the ascertained subpopulation (p_A). Olson and Cordell (2000) demonstrated a similar result in the analysis of sibling recurrence risk.

Li-Mantel Estimator Assuming Multiple Strata and Marginal Independence of Siblings in a Sibship

Now, let us assume that the disease statuses of siblings are marginally independent. We show in such a case that the Li-Mantel estimator will consistently estimate the population prevalence p_P, even when the population contains strata with heterogeneous disease prevalences. As before, assume that the original population contains $K = 4$ strata with prevalences p_1, p_2, p_3, and p_4. Let π_k denote the proportion of the original population found in stratum k. It can easily be shown that the population disease prevalence converges in probability to

$$p_P = \sum_{k=1}^{4}\pi_k p_k \ .$$

Assuming marginal independence of siblings in a sibship, the expected values $E[a_{ASC}]$, $E[m_{ASC}]$, and $E[n_1]$ are evaluated as

$$E[a_{ASC}] = 3np_P$$

$$E[m_{ASC}] = 3E[n_{ASC}] = 3n[1-(1-p_P)^3]$$

$$E[n_1] = n[3p_P(1-p_P)^2] \ .$$

Using these expected values, we have

$$\hat{p}_{LM} \rightarrow \frac{E[a_{ASC} - n_1]}{E[m_{ASC} - n_1]} = \frac{3np_P - 3np_P(1-p_P)^2}{3n[1-(1-p_P)^3] - 3np_P(1-p_P)^2} = p_P \ .$$

This shows that, in the presence of hidden stratification, the Li-Mantel estimator consistently estimates the population prevalence when the disease statuses of siblings in a sibship are marginally independent. This might occur when disease statuses of siblings are determined entirely by environmental factors that have no tendency to be excessively shared by siblings.

References

Apert E (1914) The laws of Naudin-Mendel. J Hered 5:492–497

Breslow NE, Clayton DG (1993) Approximate inference in generalized linear mixed models. J Am Stat Assoc 88:9–25

Burton PR, Palmer LJ, Jacobs K, Keen KJ, Olson JM, Elston RC (2000) Ascertainment adjustment: where does it take us? Am J Hum Genet 67:1505–1514 (erratum: 69:672 [2001])

Cannings C, Thompson EA (1977) Ascertainment in the sequential sampling of pedigrees. Clin Genet 12:208–212

de Andrade M, Amos CI (2000) Ascertainment issues in variance components models. Genet Epidemiol 19:333–344

Elston RC, Sobel E (1979) Sampling considerations in the gathering and analysis of pedigree data. Am J Hum Genet 31: 62–69

Ewens WJ, Shute NC (1986*a*) The limits of ascertainment. Ann Hum Genet 50: 399–402

——— (1986*b*) A resolution of the ascertainment sampling problem. I. Theory. Theor Popul Biol 30:388–412

Fisher RA (1934) The effects of methods of ascertainment upon the estimation of frequencies. Ann Eugen 6:13–25

Gelfand AE, Smith AFM (1990) Sampling based approaches to calculating marginal densities. J Am Stat Assoc 85:398–409

Haldane JBS (1938) The estimation of the frequencies of recessive conditions in man. Ann Eugen 8:255–262

Li CC, Mantel N (1968) A simple method of estimating the segregation ratio under complete ascertainment. Am J Hum Genet 20:61–81

Morton NE (1959) Genetic tests under incomplete ascertainment. Am J Hum Genet 11:1–16

Olson JM, Cordell HJ (2000) Ascertainment bias in the estimation of sibling genetic risk parameters. Genet Epidemiol 18:217–235

Pinheiro JC, Bates DM (1995) Approximations to the loglikelihood function in the nonlinear mixed-effects model. J Comput Graph Statist 4:12–35

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Roeder K, Lynch KG, Nagin DS (1999) Modeling uncertainty in latent class membership: a case study in criminology. J Am Stat Assoc 94:766–776

Spiegelhalter D, Thomas A, Best N (2000) WinBUGS version 1.3 user manual. MRC Biostatistics Unit, Cambridge, UK

Vieland VJ, Hodge SE (1995) Inherent intractability of the ascertainment problem for pedigree data: a general likelihood framework. Am J Hum Genet 56:33–43

Weinberg W (1912) Methode und Fehlerquellen der Untersuchung auf Mendelschen Zahlen beim Menschen. Arch Rass Ges Biol 9:165–174

Zeger SL, Karim MR (1991) Generalized linear models with random effects: a Gibbs sampling approach. J Am Stat Assoc 86:79–86