Am. J. Hum. Genet. 63:1202–1215, 1998

Ascertainment Corrections Based on Smaller Family Units

George Ebow Bonney

Department of Biostatistics, Fox Chase Cancer Center, Philadelphia

Summary

Ascertainment concerns the manner by which families are selected for genetic analysis and how to correct for it in likelihood models. Because such families are often neither drawn at random nor selected according to well-defined rules, the problem of ascertainment correction in the genetic analysis of family data has proved durable. This paper undertakes a systematic study of ascertainment corrections in terms of smaller distinct units, which will usually be sibships, nuclear families, or small pedigrees. Three principal results are presented. The first is that ascertainment corrections in likelihood models for family data can be made in terms of smaller units, without breaking up the pedigree. The second is that the appropriate correction for single ascertainment in a unit is the reciprocal of the sum of the marginal probabilities of all the persons relevant to its ascertainment, as if affected. The third result is a generalization of the single ascertainment-correction formula to *k*-plex ascertainment, in which each unit has *k* or more affecteds. The correction is the reciprocal of the sum of the joint probabilities of all distinct sets of *k* persons in the unit, as if they were all affected. In extended families, two additional ascertainment schemes will be considered and explicit formulas will be presented. One of these schemes is "uniform-proband-status ascertainment," in which nonmembers of a given unit have the same chance as members to become probands if they are affected; the other scheme is the "inverse law of ascertainment," in which the chance that nonmembers of a unit will become probands for that unit decreases with degree of relationship. Several specific recommendations are made for further study.

Received March 13, 1998; accepted for publication August 12, 1998; electronically published October 2, 1998.

Address for correspondence and reprints: Dr. George Ebow Bonney, Department of Biostatistics, Fox Chase Cancer Center, 7701 Burholme Avenue, Philadelphia, PA 19111. E-mail: ge_bonney@fccc.edu

0002-9297/98/6304-0035$02.00

Introduction

Statistical inference presumes that the sample and the population are characterized by the same types of units, which we call "sampling units," and, further, that the units appearing in the sample can be appropriately weighted if they were nonrandomly selected from the population. Without the first, the concepts of likelihood, bias, standard error, and hypothesis-testing criteria such as χ^2 are all meaningless. Without the second, the results are often biased. In families, structure and size are often arbitrary, and so it is not obvious what the sampling units are in genetic studies utilizing families. The early workers (e.g., see Weinberg 1912; Fisher 1934; Haldane 1938; Bailey 1951; Morton 1959) concerned with segregation at a single locus developed methods for the analysis of independent nuclear families ascertained through probands, by conditioning on each family containing at least one proband. The sampling unit was clearly the sibship, if the selection was through an affected child, or the nuclear family, if selection was through an affected parent. Inference pertains to either a population of sibships each with at least one affected person or a population of nuclear families each with at least one affected person. Moreover, the π model first introduced by Fisher (1934) allowed each unit in the sample to be weighted by its selection probability, the ascertainment correction. The later workers (e.g., see Morton et al. 1971; Elston and Yelverton 1975; Stene 1977, 1978) followed this paradigm.

For extended families, or pedigrees, the picture is not so clear. Elston and Sobel (1979) proposed conditioning on the pedigree containing at least one proband (among persons who could be probands regardless of their phenotypes). This, in effect, suggests that the classic method can be extended to pedigrees, without regard to their sizes or structure, although it is not known what the actual sampling unit may be.

On the other hand, Lalouel and Morton (1981) proposed treating the sibships within pedigrees as the sampling units, in keeping with the classic approach. They recognized, however, that some sibships in pedigrees may not contain probands; they therefore based the ascertainment correction for such nuclear families on pointers: relatives (of extreme phenotypes) outside the nuclear family who may have caused the selection of

those nuclear families. In their likelihood formulation, each pedigree was broken into nuclear families, and the latter were handled as if they were independently selected from the population.

In a marked departure from the techniques reviewed so far, Cannings and Thompson (1977) showed that, under certain sequential sampling schemes, one needs to condition on the initial ascertainment event only and that it is not necessary to break up the pedigree. They noted that independently sampled pedigrees that join up remain problematic. Because of its simplicity, their method of ascertainment correction has been adapted even for families not selected according to their sequential scheme. Although the method was not presented in those terms, well-defined sampling units can be decided on and drawn sequentially, as they suggested.

Ewens and Shute (1986) and Shute and Ewens (1988) have suggested, without consideration of structure, that conditioning on parts of the data relevant to ascertainment can lead to results robust to certain ascertainment models. There is some loss in efficiency in certain important cases. Thompson (1988) showed that their procedure can be more appropriately viewed as an example of Cox's (1972, 1975) partial likelihood. Risch (1984) had proposed conditioning on affected persons when simultaneously estimating recombination fractions between marker/disease loci and parameters at the disease locus. However, there appears to be no systematic method for deciding on the data relevant for ascertainment. For example, should one condition on only the actual probands, or should one include the pointers of Lalouel and Morton as well?

The methods of ascertainment corrections, especially in pedigrees, have often generated controversy. Vieland and Hodge (1995) have recently claimed that the problem is intractable largely because of difficulties in specification of the probability distribution for the structure of a pedigree. In view of the problems that they have described, Vieland and Hodges advised that future research efforts should focus on the development of robust approximate approaches. Robustness in ascertainment corrections is taken to mean that Fisher's (1934) π model is not used but that, instead, conditioning on data relevant for ascertainment are used (Risch 1984; Ewens and Shute 1986; Shute and Ewens 1988). Elston (1995), in his invited editorial on the work of Vieland and Hodge, argued that the problem is not pedigree structure per se but, rather, the fact that independently sampled branches can join up, as Cannings and Thompson (1977) already had noted in the context of sequential sampling for which there is as yet no solution.

Another unsettling point that does not appear to have been made in the discussion of ascertainment corrections is the irony that the criticisms and the proposed robust methods are both based on the work of Fisher (1934), who considered only the simple genetic model of complete penetrance with a known mating type, so that only the segregation ratio is the unknown genetic parameter needing to be statistically inferred. Most modern approaches for segregation analysis (e.g., see Elston and Stewart 1971; Morton and MacLean 1974; Lalouel et al. 1983; Bonney 1986) and linkage analysis requiring ascertainment correction (Bonney et al. 1988) utilize models that are far more complex than that used by Fisher. It is not known for certain that the so-called robust methods apply to the more complex models and pedigree structures.

There is clearly a need to go back to basics and to derive ascertainment-correction methods that satisfy the basic paradigms of statistical inference from sample to a defined population and that can be seen as broadly applicable. This paper begins a systematic study of the problem of ascertainment corrections, thinking in terms of smaller units that make up the population and the sample, without breaking up pedigrees, and using elementary probability principles, so that the basis of the results will be clear. I show that, by thinking in terms of such units, appropriate ascertainment corrections can also be found. Moreover, by a proper generalization of Fisher's work for the defined units, more generally applicable robust ascertainment corrections can be found. The organization of the rest of the paper is as follows. First an expository review of Fisher's formulation and extensions is presented in the section on small unrelated pedigrees. This section includes some new results on single and k-plex ascertainment corrections, which are robust. This section is followed by a discussion of the second major development: the decomposition of ascertainments, which permits the extension of our results to pedigrees of arbitrary structure.

Small Independent Nuclear Families

The Classic Fisher π Model

The classic approach to correcting for ascertainment assumed independent nuclear families selected through either affected children or parents. Let Y_{js} denote the phenotype of the sth person in the jth family; let $s = 1,2,\ldots,n_j$; and let $F_j = (Y_{j1},Y_{j2},\ldots,Y_{jn_j})$ denote the phenotypes of persons in the jth family. Define the ascertainment event for family j as $A_j =$ {family unit j contains at least one proband}. A proband is an affected person through whom the family came to attention; if there are two or more probands, the classic assumption is that the family came to attention independently through each of them. The ascertainment correction for the jth family, when it is assumed that the A's are independent, is

$$C_j = \frac{\Pr(A_j|F_j)|}{\Pr(A_j)} = \frac{\Pr(A_j|F_j)|}{\sum_{F_i} \Pr(F_i) \Pr_i(A_i \mid F_i)} \quad . \tag{1}$$

The essential features of the π−model approach introduced by Fisher (1934) will be depicted with the simplest case in which the phenotype (with the subscript j dropped, for convenience) is $Y = 1$ for affected and $Y = 0$ for unaffected, sampling is through an affected child, and π, the probability that an affected person becomes a proband, is known; that is

$$\pi(Y) = \begin{cases} \pi, & \text{if } Y = 1 \\ 0, & \text{if } Y = 0 \end{cases} \quad .$$

Now suppose that the sibship size is n and that $\sum Y_j = a$, the number of affected. Then ascertainment correction (1) reduces to

$$C_j = \frac{1 - (1 - \pi)^a}{1 - (1 - \theta\pi)^n} \quad , \tag{2}$$

where θ is the segregation parameter. In this scheme, $\pi = 1$ is one of the cases of special interest, for then an affected person will certainly become a proband. This case is therefore often called "complete" ascertainment (Morton [1959] called this case "truncate"); $\pi = 1$ is used when it can be justified that complete ascertainment is likely to have been achieved. But note that the resulting ascertainment correction is the same as the correction for the probability of a affecteds in a randomly selected sibship of size n from a large population of sibships of size n of the same parental mating type and having at least one affected member. Observe that

$$\Pr(\text{a affecteds in a sibship of size } n \mid \text{complete ascertainment}) = \frac{1}{1 - (1 - \theta)^n} \binom{n}{a} \theta^a (1 - \theta)^{n-a} \quad ,$$

which is the same as the probability of obtaining a affecteds from randomly selecting from sibships of size n, each with at least one affected. Therefore, it can be supposed that, when $\pi = 1$, it is not the case that all affecteds in the population necessarily became probands but, rather, that the sample of families was randomly selected from the truncated population of families. Thus, when $\pi \neq 1$, the ascertainment-correction factor adjusts for the different probabilities of selection. The π model is therefore an approach to characterize the sampled population and to simultaneously provide a correction for nonrandom selection from the "sampled population." How well the π model will do this, can, of course, be debated in any given situation.

When π is small, Taylor-series expansions of both numerator and denominator yield, to the first order,

$$C_j \simeq \frac{a}{n\theta} \propto \frac{1}{\theta} \quad , \tag{3}$$

which is called "single ascertainment," because the probability of more than one proband per family is negligible. Sampling is proportional to the number of affecteds. I shall discuss some generalizations of Fisher's approach in the following sections.

Generalized Multiple Ascertainment

More generally, let $\pi_j(Y_{js})$ be the probability that the sth person in unit j with phenotype Y_{js} is a proband for that unit. Then the ascertainment correction for the jth family, if the ascertainment events are assumed to be independent, is

$$C_j = \frac{\Pr(A_j|F_j)|}{\Pr(A_j)} = \frac{1 - \prod_{s=1}^{n_j} [1 - \pi_j(Y_{js})]}{1 - \sum_{F_j} \Pr(F_j) \left\{ \prod_{s=1}^{n_j} [1 - \pi_j(Y_{js})] \right\}} \quad . \tag{4}$$

The summations are replaced by integrals if the Y's are continuous. This particular formula was used by Elston and Yelverton (1975) to discuss ascertainment correction for a nuclear family. An alternative generalization of formula (2), which uses the segregation parameter, θ, for a given parental mating type can be developed as follows. Let the parental mating type be Y_F, Y_M, where the subscripts "F" and "M" denote "father" and "mother," respectively; then we can generalize the segregation parameter to be $\Pr(Y_{js} = 1 \mid Y_F, Y_M)$. Hence, if we condition the analysis on mating type, then the ascertainment correction for unit j becomes

$$C_j(Y_F, Y_M) = \frac{\Pr(A_j \mid F_j)}{\Pr(A_j)} = \frac{1 - \prod_{s=1}^{n_j} [1 - \pi_j(Y_{js})]}{1 - \left\{ \prod_{s=1}^{n_j} [1 - \sum_{F_j} \Pr(Y_{js} \mid Y_F, Y_M) \pi_j(Y_{js})] \right\}} \quad . \tag{5}$$

It is easily verified that, if $\pi_j(Y_{js} = 1) = \pi$ and $\pi_j(Y_{js} \neq 1) = 0$ and $\theta = \Pr(Y_{js} = 1 \mid Y_F, Y_M)$, then formula (5) sim-

plifies to formula (2). For an analysis unconditional on parental mating type, we have

$$C_j = \frac{\Pr(A_j \mid F_j)}{\Pr(A_j)} = \left\{1 - \prod_{s=1}^{n_j}[1 - \pi_j(Y_{js})]\right\} \Big/ \left(1 - \sum_{Y_F}\sum_{Y_M} \Pr(Y_F, Y_M) \times \left\{\prod_{s=1}^{n_j}\left[1 - \sum_{Y_{js}} \Pr(Y_{js} \mid Y_F, Y_M)\pi_j(Y_{js})\right]\right\}\right). \quad (6)$$

It is a straightforward matter to write $\Pr(Y_F, Y_M)$ and $\Pr(Y_{js} \mid Y_F, Y_M)$ in terms of parameters of interest, given a model for the joint distribution of phenotypes. Other weights can be explored, but I shall not pursue them here.

Generalized Single and Multiplex Ascertainments

We now derive some new results. Suppose that the π's are small; then, by taking first-order Taylor-series approximations in ascertainment correction (4), with respect to the π's in the numerator and denominator separately, and retaining the first-order terms, we obtain

$$C_j = \left\{1 - \left[1 - \sum_s \pi_j(Y_{js}) + \sum_s \sum_{s'} \pi_j(Y_{js})\pi_j(Y_{js'}) - \ldots\right]\right\} \Big/ \left\{1 - \left[\sum_s \sum_{Y_{js}} \Pr(Y_{js})\pi_j(Y_{js}) + \sum_s \sum_{s'} \sum_{Y_{js}} \sum_{Y_{js'}} \Pr(Y_{js}, Y_{js'})\pi_j(Y_{js})\pi_j(Y_{js'}) - \ldots\right]\right\} \simeq \left[\sum_s \pi_j(Y_{js})\right] \Big/ \left[\sum_s \sum_{Y_{js}} \Pr(Y_{js})\pi_j(Y_{js})\right]. \quad (7)$$

We are now ready for our first major result.

PROPOSITION I. *If* $\pi_j(Y_{js}) = \pi$ *if* $Y_{js} = 1$, *and if it is* 0 *if* $Y_{js} = 0$, *then the correction for single ascertainment of unit* j *is*

$$C_j = \frac{\sum_{s=1}^{n_j} Y_{js}}{\sum_{s=1}^{n_j} \Pr(Y_{js} = 1)}. \quad (8)$$

Proof. The result follows directly from ascertainment correction (7).

This proposition implies that sampling is proportional to the number of affecteds ($\Sigma_s Y_{js}$), as in the classic single-ascertainment case. However, we do not divide by the probability of just the particular proband being affected, as the current practice is, but by the sum of the marginal probabilities for all the n_j persons who could be probands if they were affected. A search of the literature shows that ascertainment correction (8) is more in keeping with the generalization of Fisher's (1934) work to weighted distributions in general (Rao 1965; Patil and Rao 1978). If we condition on parental phenotypes, we have

$$C_j(Y_F, Y_M) \simeq \frac{\sum_{s=1}^{n_j} Y_{js}}{\sum_{s=1}^{n_j} \Pr(Y_{js} = 1 \mid Y_F, Y_M)};$$

so, if there are no individual differences, such as covariate effects, in the probabilities of being affected, the formula reduces to that of Fisher—that is, ascertainment correction (3).

When π is not small, the multiple-ascertainment formulas (4)–(6) can be used, but then π must be specified. An alternative is to generalize proposition I to multiplex ascertainment in which the selected units contain more than one proband. In particular, for the duplex ascertainment in which there are two or more probands per unit but the probability of a pair of affected persons becoming probands for the unit is small, I propose, by analogy with the formula (8),

$$C_j = \frac{\sum_{s=1}^{n_j} \sum_{s'>s} Y_{js} Y_{js'}}{\sum_{s=1}^{n_j} \sum_{s'>s} \Pr(Y_{js} = 1, Y_{js'} = 1)}. \quad (9)$$

The numerator does not depend on parameters of interest. The denominator is the sum of the marginal probabilities of all relevant distinct pairs on the unit being affected. If an affected sib pair is the duplex proband, so to speak, then the subscripts s and s' refer only to full sibs within the nuclear family. If an affected parent-offspring pair is considered as the duplex proband, then the relevant pairs are the distinct parent-offspring pairs, so that

$$C_j = \text{constant} / \left\{\sum_{s \neq F,M} [\Pr(Y_{jF} = 1, Y_{js} = 1) + \Pr(Y_{jM} = 1, Y_{js} = 1)]\right\}. \quad (10)$$

The generalization to a k-plex ascertainment, where $1 \leqslant k < \Sigma_{s=1}^{n_j} Y_{js}$ is

$$C_j = \frac{\text{constant}}{\sum_{s_1} \sum_{s_2 > s_1} \cdots \sum_{s_{k-1} > s_{k-2}} \Pr(Y_{js_1} = 1, \ldots, Y_{js_{k-1}} = 1)} , \quad (11)$$

in which the denominator is the sum of the probabilities of the relevant distinct sets of k persons as if affected. In practice, the case of $k > 3$ will be extremely rare.

A Note on Applications

In my development I have so far deliberately avoided the use of a particular genetic model, so that the results will be as broadly applicable as possible. It is a very simple matter to now write the phenotypic probabilities in terms of specific genetic quantities, if any. Consider, as an example, genetic analysis of a binary disease phenotype, under the assumption that there is a single locus with two alleles, *u* and *v,* in Hardy-Weinberg equilibrium. Let the population frequencies be q for *u* and $1 - q$ for *v;* then the frequencies of the three genotypes—*uu, uv,* and *vv*—are $\Pr(g = uu) = q^2$, $\Pr(g = uv) = 2q(1 - q)$, and $\Pr(g = vv) = (1 - q)^2$, respectively. For parent-offspring (or offspring-parent) pairs, the joint genotypic probabilities are

$$\begin{aligned}
\Pr(uu,uu) &= q^3 \ ; \\
\Pr(uu,uv) &= q^2(1 - q) = \Pr(uv,uu) \ ; \\
\Pr(uv,uv) &= q(1 - q) \ ; \\
\Pr(uv,vv) &= q(1 - q)^2 = \Pr(vv,uv) \ ; \\
\Pr(vv,vv) &= (1 - q)^3 \ .
\end{aligned}$$

For full-sib pairs, the joint genotypic probabilities are

$$\begin{aligned}
\Pr(uu,uu) &= \frac{1}{4} q^2(1 + q)^2 \ ; \\
\Pr(uu,uv) &= \frac{1}{2} q^2(1 - q^2) = \Pr(uv,uu) \ ; \\
\Pr(uu,vv) &= \frac{1}{4} q^2(1 - q)^2 = \Pr(vv,uu) \ ; \\
\Pr(uv,uv) &= q(1 - q)\{1 + q(1 - q)\} \ ; \\
\Pr(uv,vv) &= \frac{1}{2} q(1 - q)^2(2 - q) = \Pr(vv,uv) \ ; \\
\Pr(vv,vv) &= \frac{1}{4} (1 - q)^2(2 - q)^2 \ .
\end{aligned}$$

These and other joint genotypic probabilities among family members are calculated by standard formulas (for an excellent summary, see Elandt-Johnson 1971, chap. 7). Then, for single-ascertainment correction, formula (8) is computed as

$$\begin{aligned}
C_j &\simeq \frac{\sum_s Y_{js}}{\sum_s \Pr(Y_{js} = 1)} \\
&= \frac{\sum_s Y_{js}}{\sum_s \sum_{g_{js}} \Pr(g_{js}) \Pr(Y_{js} = 1 \mid g_{js})} ,
\end{aligned}$$

where $\Pr(Y = 1 \mid g)$'s are the penetrance functions, and the $\Pr(g)$'s are given above. Similarly, for a duplex ascertainment, ascertainment correction (13) can be calculated by

$$\begin{aligned}
C_j &= \frac{\text{constant}}{\sum_s \sum_{s' > s} \Pr(Y_{js} = 1, Y_{js'} = 1)} \\
&= \frac{\text{constant}}{\sum_s \sum_{s' > s} \Pr(g_{js}, g_{js'}) \Pr(Y_{js} = 1, Y_{js'} = 1 \mid g_{js}, g_{js'})} ,
\end{aligned}$$

where the relevant values for $\Pr(g_{js}, g_{js'})$ are read from the joint genotypic probabilities given above. Note that

$$\begin{aligned}
&\Pr(Y_{js} = 1, Y_{js'} = 1 \mid g_{js}, g_{js'}) \\
&\quad = \Pr(Y_{js} = 1 \mid g_{js}) \Pr(Y_{js'} = 1 \mid g_{js'})
\end{aligned}$$

if the postulated loci and the observed covariates completely explain the phenotypic dependence among family members. Otherwise, the appropriate form must be used. No new parameters are required by the formulation, in the specification and calculation of the ascertainment corrections.

Ascertainment Correction in Pedigrees

General Decomposition

The second basic idea of this paper is that units within a pedigree can be separately corrected for ascertainment without breaking it up. To develop the idea formally and to obtain explicit formulas, we start with four basic postulates.

POSTULATE I: *A pedigree, however complex, can be regarded as comprising distinct full sibships joined by the common biologic parents.*

If sibships are distinct, then, of course, the nuclear families are also distinct, although some of the nuclear families may partially overlap. Hence we can regard pedigrees as a sample of sibships or nuclear families drawn from a population of sibships or nuclear families. Whether purposely or not, if, after the sample is drawn, we end up with pedigrees in which every nuclear family,

for example, has one or more affected persons, then we can characterize the sampled population by that fact. Statistical inference pertains to that sampled population.
POSTULATE II: *The overall ascertainment correction for a pedigree can be derived from the ascertainment corrections of its constituent units.*

The units in a pedigree are typically not selected at random from the population, and so the likelihood for a whole pedigree should be weighted by the selection probabilities (separate ascertainment corrections) of the units that it contains. Without loss of generality, consider the family shown in the pedigree in figure 1. Let F denote the set of phenotypes and let A denote the ascertainment event. The term "ascertainment event" is used in this article (rather broadly) to denote the event (or variables) that describes the essence of the sampling of the family. In a nuclear family, it denotes the event "at least one proband," a proband being a person with the condition under study who, independently of other members of the family, caused it to come to the attention of the researchers so that the family was drawn into the study. In pedigrees, there will typically be more than one proband, and there could also be other persons in the pedigree who caused certain branches to be drawn into the pedigree. Lalouel and Morton (1981) called these persons "pointers." Clearly, probands and pointers together make up event A. The precise definition of A for a pedigree is usually unavailable. However, the probands and pointers occur in certain units, and so how those units are separately ascertained will define the overall A. The likelihood function for the data is constructed by specification of a mathematical formula for the conditional probability of the set of phenotypes F, given the ascertainment event A. Suppose, to begin with, that a nuclear family is the unit chosen, at least for the purpose of analysis. The sets F_1 and F_2 are the phenotypes measured in nuclear families 1 and 2, respectively; that is, $F = (F_1,F_2)$. Note that, although F_1 and F_2 overlap, they do not completely overlap and are therefore distinct subsets of F. Throughout this article, only distinct subsets of F are implied in the terms "sampling unit," "smaller family unit," and "unit." These terms are used here interchangeably. If two units completely overlap, then they are actually the same unit multiply ascertained, and so only one ascertainment correction is needed. I shall let Y_{js} denote the phenotype of the sth person in the jth observational unit, $s = 1,2,\ldots,n_j$, let $F_j = (Y_{j1},Y_{j2},\ldots,Y_{jn_j})$ denote the phenotypes of persons in the jth unit, and denote the joint probability of the phenotypes of all pedigree members (under the assumption that the pedigree was randomly selected from the population) as $(\Pr(F) = \Pr(F_1,F_2)$. This is the joint distribution of the phenotypes, uncorrected for ascertainment. I stress that, by this convention, I am simply recognizing the fact that F_1 and F_2 are two different random vector

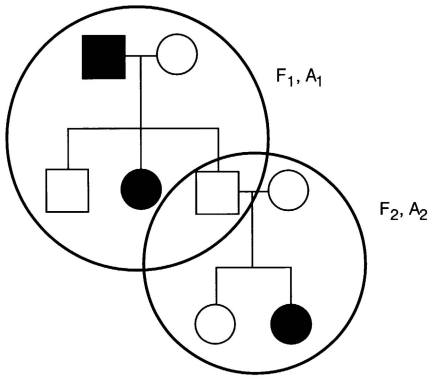

Figure 1 $F = \{F_1,F_2\}$, and $A = \{A_1,A_2\}$

variables. If the units 1 and 2, on which they are measured, completely overlap, then F_1 and F_2 are identical random variables, and so $\Pr(F_1,F_2) \equiv \Pr(F_1) \equiv \Pr(F_2)$.

I shall associate with unit i its own ascertainment event A_i, so that the ascertainment of the entire pedigree can be broken into the ascertainments of the units—that is, $A = (A_1,A_2)$. In my development of these terms, I seek to meaningfully express $\Pr(F|A)$, the joint distribution of the phenotypes, given its overall ascertainment A, in terms of the uncorrected joint probability, $\Pr(F)$, and ascertainment corrections for the units that comprise F. To proceed further, I need to formalize the description of the basic features of ascertainment procedures. So, in addition to postulates I and II above, I note the following.
POSTULATE III: *Ascertainment events of different units within a pedigree are dependent, but only through the phenotypes actually measured in the units.*

The decision to include a particular unit in the pedigree depends wholly or partly on its own phenotypes and the phenotypes of other units. Thus, the ascertainment event A_1 does not by itself provide any more information about the ascertainment event A_2 than that provided by the actual phenotypes in unit 1, and vice versa. In formal probabilistic language, we say that, given the phenotypes, the actual ascertainment events of different units are conditionally independent. A precise probability formulation of this statement is the following:

$$\Pr(A_2 \mid F_1,F_2,A_1) = \Pr(A_2 \mid F_1,F_2) \ ;$$
$$\Pr(A_1 \mid F_1,F_2,A_2) = \Pr(A_1 \mid F_1,F_2) \ . \qquad (12)$$

POSTULATE IV: *Independently selected units can join to form a bigger pedigree.*

In this case the ascertainment events are independent. Thus A_i depends only on F_i—that is,

$$\Pr(A_2 \mid F_1,F_2,A_1) = \Pr(A_2 \mid F_2) \ ;$$
$$\Pr(A_1 \mid F_1,F_2,A_2) = \Pr(A_1 \mid F_1) \ . \qquad (13)$$

Note that, if A_i depends on only F_i, then equation (12) reduces to equation (13).

If the pedigree was not formed by independently selected units joining up, then we shall say that the chosen units are dependent, or, more loosely speaking, correlated. It seems convenient to introduce a measure of the correlation in the ascertainment events (CAE), which I shall define as

$$\kappa = \log_e \left\{ \frac{\Pr(A_1,A_2)}{\Pr(A_1)\Pr(A_2)} \right\} \ . \qquad (14)$$

It is the logarithm of the ratio of the joint probability of the ascertainments to the product of the marginal probabilities. CAE takes the value zero if the ascertainments are independent; i is positive if the probability of joint selection of A_1 and A_2 exceeds the probability that they are independently selected; it is negative if the probability of joint selection of them is less than the probability that they are independently selected. Our basic result can be stated as follows.

PROPOSITION II. *Every distinct unit can be separately corrected for ascertainment, without breaking up the pedigree.* Specifically,

$$\Pr(F|A) = \Pr(F_1,F_2|A_1,A_2) = e^{-\kappa} C_1 C_2 \Pr(F) \ ,$$

where

$$C_1 = \frac{\Pr(A_1 \mid F_1,F_2)}{\Pr(A_1)} \ ,$$
$$C_2 = \frac{\Pr(A_2 \mid F_1,F_2)}{\Pr(A_2)} \ , \qquad (15)$$

and κ is defined by equation (14).

Proof. Write

$$\Pr(F_1,F_2|A_1,A_2) = \frac{\Pr(A_1,A_2,F_1,F_2)}{\Pr(A_1,A_2)}$$
$$= \frac{\Pr(A_1,A_2 \mid F_1,F_2)}{\Pr(A_1,A_2)} \Pr(F_1,F_2)$$
$$= \frac{\Pr(A_1)\Pr(A_2)}{\Pr(A_1,A_2)} \frac{\Pr(A_1 \mid F_1,F_2)}{\Pr(A_1)}$$
$$\times \frac{\Pr(A_2 \mid F_1,F_2,A_1)}{\Pr(A_2)} \Pr(F_1,F_2)$$

and apply equations (12) and (14) to obtain the result.

Some Remarks

1. The proposition implies that the likelihood of F, given A, is the unconditional likelihood of F not broken but corrected by the factors C_1 and C_2, in formula (15), corresponding to the ascertainments of the two observational units, and CAE.

2. The proposition is true regardless of the structure of the smaller family units. Thus, singletons, pairs of individuals, sibships, nuclear families, or more-extended pedigrees can be considered as the smaller family units. The proposition generalizes easily to more than two smaller family units. Note that, if there are M units within the pedigree, then

$$\Pr(F_1,F_2,\ldots,F_M|A_1,A_2,\ldots,A_M)$$
$$= \frac{\Pr(A_1,A_2,\ldots,A_M \mid F_1,F_2,\ldots,F_M)}{\Pr(A_1,A_2,\ldots,A_M)}$$
$$\times \Pr(F_1,F_2,\ldots,F_M)$$
$$= \frac{\Pr(A_1)\Pr(A_2)\ldots\Pr(A_M)}{\Pr(A_1,A_2,\ldots,A_M)} \frac{\Pr(A_1 \mid F)}{\Pr(A_1)}$$
$$\times \frac{\Pr(A_2 \mid F)}{\Pr(A_2)} \ldots \frac{\Pr(A_M \mid F)}{\Pr(A_M)} \Pr(F)$$
$$= e^{-k} C_1 C_2 \ldots C_M \Pr(F) \ ,$$

where

$$\kappa = \left\{ \frac{\Pr(A_1 A_2 \ldots, A_M)}{\Pr(A_1)\Pr(A_2)\ldots\Pr(A_M)} \right\} .$$

3. The proposition has been derived for the joint probability of the vector of phenotypes, $F = (F_1,F_2)$, given the ascertainment events, $A = (A_1,A_2)$. Clearly, F can be augmented to include marker data, measurements on covariates, proband status of each person, and other random variables—such as number of affecteds, $R = (R_1,R_2)$; sibship sizes, $S = (S_1,S_2)$; and number of probands, $T = (T_1,T_2)$—that are defined for each smaller family unit as an entity. George and Elston (1991) have provided an overview of the classic segregation-analysis model including S and T. My thesis that, in the derivation of ascertainment corrections, we can think in terms of smaller family units, still holds. If S and R are recorded but F is not, the results still hold, with S and R in F's place; but these days a lot of attention is being given to

variable age at onset, marker data, and other covariates measured on individuals, and so it is rare to have S and R recorded but to not have the actual phenotypes F of individuals in the family. Consider the decomposition

$$\begin{aligned}\Pr(F,S,R,T \mid A) &= \Pr(F \mid A) \\ &\times \Pr(S,R \mid F,A) \\ &\times \Pr(T \mid F,S,R,A) \ .\end{aligned}$$

Given F and its structure (biological relationships), S and R are determined, and therefore the middle factor is unity. If T is a function of A and/or F, or if F has been augmented to include proband status, then the third factor is, similarly, unity.

The ascertainment corrections C_1 and C_2 simplify considerably, if certain restrictions apply. In the following sections, I discuss cases that are likely to be most useful in practice.

Independently Selected Families Joining

Independent ascertainments can also occur if independently selected families are found to be related in a larger pedigree. Figure 2 illustrates this case. Two persons, *a* and *b*, are identical, and the three families are members of the larger pedigree. The theory shows that $\Pr(F) = \Pr(F_1,F_2,F_3)$ can be specified for the combined pedigree, not broken but adjusted for ascertainment by multiplication by factors for each of the three independently selected components.

But do we need to join them? If we do not join them, then the inference based on the likelihood function pertains to the "sampled population" of similar nuclear-family units. However, when they are combined, a feature of the underlying population, discovered after selection, is taken into account, and the inference is, in this sense, more pertinent to the underlying population. Moreover, parameters of the dependence of the units can be either incorporated, if known, or inferred, if unknown. This may confer some advantage. As an example, suppose that, in addition to disease status, figure 2 shows marker data. One can, of course, do linkage analysis using the nuclear families. Using the combined form now allows one to do linkage analysis while taking phase into account. Similarly, in segregation analysis, the mode of inheritance is easier to infer when the pedigrees are intact than when they are broken into nuclear families.

When families join into a larger pedigree, it is possible for some units to completely overlap. Two possibilities can be considered. One of these possibilities is to regard such units as multiply ascertained and given an appropriate correction; Fisher's (1934) π model was designed with multiple ascertainment in mind. An alternative possibility is to use the theory discussed above, which means that the C's involved are multiplied. In both cases the selection probability of individuals in the overlapping units increases. The difference is in the degree of increase. The merits of the two approaches will be left as an open research question.

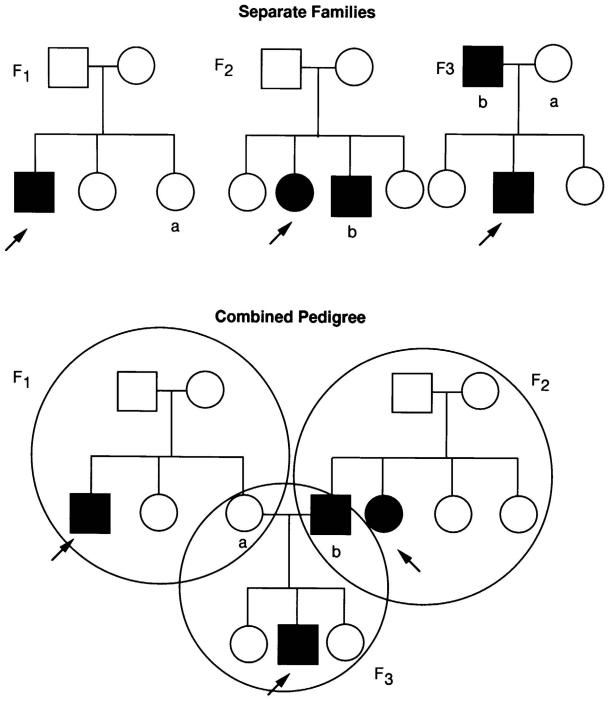

Figure 2 F_1, F_2, and F_3 are independently selected families. Persons with the same Arabic letter are found to be identical after selection, leading to the combined pedigree.

Dependent Ascertainments: Units with No Affecteds

One can speak of dependent ascertainments if the pedigree was not formed by the joining of independently ascertained units. In particular, if the jth observational unit does not contain probands, then it is in the pedigree only because other units contain probands. I stress now that the term "proband" is being used here in a broader sense, to include even the pointers described by Lalouel and Morton (1981). For a general formulation, one needs to allow for probands outside the smaller family unit. We can achieve this by defining the ascertainment event A_j to include the ascertainment events of other observational units. Let

$$A_1 = \{\text{at least one proband for unit 1 among members of unit 1 and/or unit 2}\} ,$$
$$A_2 = \{\text{at least one proband for unit 2 among members of unit 1 and/or unit 2}\} .$$

Because we have extended the definition of the ascertainment event for unit j to include the probands in other units, the information that A_1 provides about A_2 is contained in the very definition of A_2, and vice versa, so $\Pr(A_2 \mid A_1) = \theta \Pr(A_2)$ and $\Pr(A_1 \mid A_2) = \theta \Pr(A_1)$, where θ is a constant that does not depend on parameters of interest. Usually we expect $\theta = 1$. Consequently, from equation (14),

$$e^{\kappa} = \frac{\Pr(A_1,A_2)}{\Pr(A_1)\Pr(A_2)} = \frac{\Pr(A_2 \mid A_1)}{\Pr(A_2)} = \frac{\Pr(A_1 \mid A_2)}{\Pr(A_1)} = \theta .$$

Thus, the CAE is ignorable in likelihood analyses.

The ascertainment corrections for the separate units can be specified according to formula (4); that is,

$$C_1 = \frac{\Pr(A_1 \mid F_1,F_2)}{\Pr(A_1)} = \frac{1 - \prod_{s_1}[1 - \pi_1(Y_{1s_1})]\prod_{s_2}[1 - \pi_1(Y_{2s_2})]}{1 - \sum_{F_1,F_2}\Pr(F_1,F_2)\prod_{s_1}[1 - \pi_1(Y_{1s_1})]\prod_{s_2}[1 - \pi_1(Y_{2s_2})]} ,$$

$$C_2 = \frac{\Pr(A_2 \mid F_1,F_2)}{\Pr(A_2)} = \frac{1 - \prod_{s_1}[1 - \pi_2(Y_{1s_1})]\prod_{s_2}[1 - \pi_2(Y_{2s_2})]}{1 - \sum_{F_1,F_2}\Pr(F_1,F_2)\prod_{s_1}[1 - \pi_2(Y_{1s_1})]\prod_{s_2}[1 - \pi_2(Y_{2s_2})]} , \quad (16)$$

where the products with respect to s_2 are for distinct persons in smaller family unit 2 who are not also in smaller family unit 1. In the following formulas, the summations corresponding to the products over s_2 in formula (16) are also only over the corresponding distinct persons. To avoid confusion, I will denote the affected quantities by the superscript "*" (e.g., "a_2^*").

First-order Taylor-series approximations of the numerator and denominator yields an expression analogous to ascertainment correction (7),

$$C_m \simeq \frac{\sum_{u=1}^{2}\sum_{s_u}\pi_m(Y_{us_u})}{\sum_{u=1}^{2}\sum_{Y_{us_u}=0}^{1}\Pr(Y_{us_u})\pi_m(Y_{us_u})} , \; m = 1,2 . \quad (17)$$

To be explicit, substitute, in formula (17),

$$\pi_1(Y_{1s_1}) = \pi_{11} \text{ if } Y_{1s_1} = 1 ,$$
$$= 0 \text{ otherwise} ;$$
$$\pi_1(Y_{2s_2}) = \pi_{12} \text{ if } Y_{2s_2} = 1 ,$$
$$= 0 \text{ otherwise} ;$$
$$\pi_2(Y_{1s_1}) = \pi_{21} \text{ if } Y_{1s_1} = 1 ,$$
$$= 0 \text{ otherwise} ;$$
$$\pi_2(Y_{2s_2}) = \pi_{22} \text{ if } Y_{2s_2} = 1 ,$$
$$= 0 \text{ otherwise} ;$$

$$a_1 = \sum_{s_1=1}^{n_1} Y_{1s_1} ,$$
$$a_2^* = \sum_{s_2=1}^{n_2^*} Y_{2s_2} ;$$

$$\bar{p}_1 = \frac{1}{n_1}\sum_{s_1=1}^{n_1}\Pr(Y_{1s_1} = 1) ,$$
$$\bar{p}_2^* = \frac{1}{n_2^*}\sum_{s_2=1}^{n_2^*}\Pr(Y_{2s_2} = 1) .$$

We find that

$$C_1 = \frac{a_1 + a_2^*\pi_{12}/\pi_{11}}{n_1\bar{p}_1 + n_2^*\bar{p}_2^*\pi_{12}/\pi_{11}} ,$$
$$C_2 = \frac{a_1\pi_{21}/\pi_{22} + a_2^*}{n_1\bar{p}_1\pi_{21}/\pi_{22} + n_2^*\bar{p}_2^*} .$$

These formulas generalize, in an obvious manner, to the case of a pedigree containing more than two units. Thus, for a pedigree containing M units, the ascertainment corrections are

$$C_m = \frac{a_m + \sum_{u\neq m} a_u^*\pi_{mu}/\pi_{mm}}{n_m\bar{p}_m + \sum_{u\neq m} n_u^*\bar{p}_u^*\pi_{mu}/\pi_{mm}} , \; m = 1,2,\ldots,M . \quad (18)$$

In these formulas, the unknown quantity is the ratio of the π's. I will briefly discuss two possibilities.

1. *Uniform-proband-status ascertainment correction:*

Suppose that a member of a different unit, if affected with the disease, has the same chance as members of the unit under consideration to be a proband for that unit; then all the π's are equal, and ascertainment correction (18) simplifies to

$$C_1 = C_2 = \ldots = C_M = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} \Pr(Y_i = 1)} , \qquad (19)$$

where the subscript i indexes members of the whole pedigree. Therefore, the overall ascertainment correction is formula (19) raised to the power M. The equal-proband-status scenario is not unreasonable for small three-generational pedigrees.

2. *Inverse law of ascertainment corrections:* For large pedigrees, the degree of biological relationships plays a key role in the selection, so it is reasonable to construct ascertainment-probability functions on the basis of that fact. In particular, functions in which π_{mu} decreases with the biological relationships between members of smaller family units m and u can be considered. Usually first-degree and, rarely, second-degree relations are important for ascertainment. Thus, π_{mu} can be set to zero, for relationships more distant than second degree.

To be precise, let ν_{mu} be the degree of the closest biological relationship between members of the smaller family unit m and members of smaller family unit u, where $\nu_{mu} = 0$, if $m = u$. Furthermore, let $\omega > 0$ be a known or unknown number that can depend on ν_{mu}. Then, I suggest, as an illustration, the following inverse law of ascertainment probabilities, with degree of biological relationship:

$$\pi_{mu} = \frac{\pi}{(1 + \nu_{mu})^{\omega}} \text{ for } \nu_{mu} \leqslant 2, \text{and } 0 \text{ for } \nu_{mu} > 2 .$$

Then ascertainment correction (18) becomes

$$C_m = \frac{a_m + \sum_{u \neq m} \frac{a_u^*}{(1+\nu_{mu})^{\omega}}}{n_m \bar{p}_m + \sum_{u \neq m} n_u^* \bar{p}_u^* \frac{1}{(1+\nu_{mu})^{\omega}}}, \; m = 1,2,\ldots,M . \qquad (20)$$

In this formulation, ω can be chosen judiciously. Some choices that can be readily interpreted are $\omega = 1$, $\omega = 2$, and $\omega = \nu_{mu}$.

The multiplex-ascertainment-correction formula (11) can be extended in an obvious manner. For example, in the equal-proband-status scenario, the duplex ascertainment correction has the same form as formula (9), except that the summations will span the whole pedigree.

Choice of Sampling Unit and Method of Ascertainment Correction

As I have already pointed out, the basis of statistical inference is the presumption that the population and the sample are both characterized by the same units; the sample is simply a selection of units from the defined population. The type of units, how they are drawn into the sample, and how the resulting data are analyzed, are critical components of study design that need to be carefully decided before the study is begun. After the sample has been drawn, it is prudent to inspect the data to see whether the units and their characterization indeed conform to the design specifications and to make adjustments, if necessary, in the data analysis. There was no problem about the unit in the classic works that considered sibships and nuclear families. Decisions about the sample unit and its characterization, including the fact that it should contain at least one proband, were made before the units were actually drawn. There was no ambiguity about the sampled population. But this is obviously not the case for a pedigree. However, a pedigree, no matter how complex, is simply a set of distinct sibships connected by common parents, and so it is natural to characterize both the sampled population and the target population in terms of sibships. Moreover, because every full sibship is determined by the biological mother and father, one can, alternatively, consider nuclear-family units. The units will then partially overlap, but my theory allows for that. Hence I suggest that the classic units for family studies, which are sibships and nuclear families, be used for pedigrees as well, without breaking them into nuclear families. Formally, I state the following:

PROPOSITION III. *If no pedigree in the sample contains a nuclear family with no affecteds, then ascertainment corrections can be based on one of the following: selection through an affected child; selection through an affected parent; or selection through an affected parent-child pair, without breaking the pedigree into nuclear families.*

To facilitate the application of proposition III, I make seven recommendations that leave no ambiguities about either the sampling unit or the applicable model-free method of ascertainment correction.

1. If every sibship contains one (two) or more affecteds, consider the sibship as the unit and use single (duplex) ascertainment correction.

2. If every spouse pair has one or two affecteds, consider using spouse pairs as the units and use single ascertainment correction.

3. If every nuclear family has one or more affecteds, consider nuclear families as the units and use single ascertainment correction.

4. If every nuclear family has an affected parent-off-

spring pair, consider nuclear families as the units and use duplex ascertainment involving distinct parent-offspring pairs, as in equation (10).

5. If a pedigree contains a nuclear family with no affected persons, consider the nuclear family as the unit or, if more convenient, the sibship and use formula (19) or formula (20).

6. If the sample consists of several small pedigrees each with several affecteds, and if the actual ascertainment scheme is not well defined, consider each pedigree as a whole unit and use the uniform-proband-status ascertainment correction, formula (19).

7. If the pedigree was selected because it contains $\geqslant k$ probands regardless of size and structure, then use the whole pedigree as the unit and consider the k-plex ascertainment correction. Even so, if smaller units such as sibships or nuclear families are well characterized, then a correction based on that may be better, since the inference will pertain to a well-defined sampled population.

The term "proband" is not overtly used (except in recommendations 6 and 7), so that the applicable model-free method in the "Generalized Single and Multiplex Ascertainments" section (above) will be obvious.

In some applications, two or more of the above-listed recommendations may apply. The cardinal rule should be the following: stick to the sampling unit determined before the data were collected, unless, for some reason, a redefinition is required.

Is it sufficient to condition on the smaller family units containing at least one proband or, for that matter, two or more probands? There are cases of known sampling rules that may make it necessary to condition on the phenotypes of certain members of the pedigree, in addition to conditioning on some smaller family units containing at least one proband. Let us return to figure 3. Suppose, now, that the pedigree is drawn because U_1 contains at least one proband and person II-3 has a phenotype that causes ascertainment. The ascertainment event A_3 now includes the fact that person II-3 has a phenotype that causes ascertainment, and we should therefore also condition on the probability of the phenotype of II-3. The sampled population is now a proper subset of the usual one. The analyst should give some thought to what the sampled population might actually be. If all the pedigrees are of the same form as above, and if, indeed, all II-3 persons have the phenotype that causes ascertainment, then the sampled population is unambiguously characterized by the conditioning that we have indicated. But, if, for some reason, some of the pedigrees do not have unit 3, and if we wish to analyze all of them, then a less restrictive sampled population is needed. Sometimes, this can be achieved by ignoring the actual units used for the sampling and choosing instead a unit that is more convenient for the purpose of analysis. For this purpose, using the sibship or nuclear family as the smaller family unit is more in keeping with the classic work and has a natural appeal.

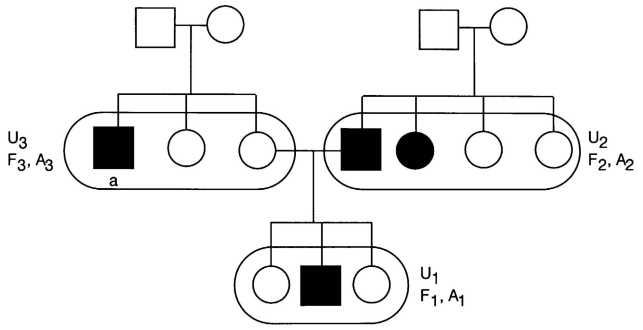

Figure 3 Extending a pedigree upward. Sibships U_1, U_2, and U_3 are the sampling units. The pedigree is selected because U_1 contains at least one proband and because individuals II-1 and II-4 have a phenotype that causes ascertainment of their sibships—U_3 and U_2, respectively.

Further Discussion of CAE

We now return to κ, the CAE defined in formula (14). There are two scenarios that make CAE ignorable in likelihood-based analysis of pedigrees with two or more units: the first is that in which independent ascertainments of the separate units join to form the pedigree; the second is that in which κ does not depend on parameters of interest. These assumptions may, in practice, be difficult to justify for all pedigrees under study. However, they can be made in the same spirit as the independence of probands in the classic formulation of ascertainment corrections. Even so, it will be helpful to have some sense about circumstances under which κ can be ignored. In family studies, we are concerned with diseases that aggregate in the selected families, and the sampling process itself is supposed to ensure this. Therefore, the units of interest are likely to be more frequent in pedigrees identified through probands than they are in pedigrees obtained by random selection. If that is the case, then we speak of aggregation of ascertainment events in the pedigrees. I shall construct a model for the joint probability, J_A, of ascertainment events A_1, A_2, …, A_M, where $A_m = 1$ if unit m is ascertained and where $A_m = 0$ otherwise, that appropriately allows clustering of these events in the pedigrees.

The disposition models of Bonney (1992, 1995, 1998), for correlated binary outcomes, can be readily adapted for this purpose. Let δ_{Am} denote the disposition of unit m to ascertainment; it can depend on characteristics or covariates of unit m. Furthermore, let δ_A be the baseline disposition to ascertainment when all covariates take a common value. If there is no aggregation of as-

certainment events in the pedigree, then we can write $\delta_A = \mu_A$ (the population rate, so to speak). Let the measure of the aggregation of the ascertained events in the pedigrees under study be defined by the ratio

$$\alpha_A = \frac{\delta_A \text{ with no aggregation of ascertained units}}{\delta_A \text{ with aggregation of ascertained units}} = \frac{\mu_A}{\delta_A} .$$

Then a joint probability model connecting these quantities is, from Bonney (1995, and and unpublished data),

$$J_A = \Pr(A_1,A_2,...,A_M) = (1-\alpha_A)\prod_{m=1}^{M}(1-A_m) + \alpha_A \prod_{m=1}^{M} \delta_{A_m}^{A_m}(1-\delta_{A_m})^{1-A_m} .$$

In this model there is positive aggregation if $\mu_A \leqslant \alpha_A < 1$, no aggregation if $\alpha_A = 1$, and negative aggregation if $\alpha_A > 1$. Under positive aggregation, the joint distribution J_A lies in the interval 0–1. Negative aggregation is also permitted, but, for J_A to be appropriately bounded, the size of negative aggregation is restricted by cluster size (M, in this case) in the same way that the negative intraclass correlation coefficient is restricted to the interval $[-1/(M-1),0]$. Moreover, α_A is related to the population rate, μ_A, of the event of interest and to their interunit baseline correlation, ρ_A, by the formula

$$\alpha_A = \left[1 + \frac{\rho_A}{\mu_A/(1-\mu_A)}\right]^{-1} .$$

Our CAE, κ, can be quickly calculated. We find that

$$\begin{aligned}\kappa &= \log_e\left[\frac{\Pr(A_1 = 1,A_2 = 1,...,A_M = 1)}{\Pr(A_1 = 1)\Pr(A_2 = 1)...\Pr(A_M = 1)}\right] \\ &= \log_e\left\{\frac{\alpha_A\delta_{A_1}\delta_{A_2}...\delta_{A_M}}{(\alpha_A\delta_{A_1})(\alpha_A\delta_{A_2})...(\alpha_A\delta_{A_M})}\right\} \\ &= -(M-1)\log\alpha_A \\ &= (M-1)\log_e\left[1 + \frac{\rho_A}{\mu_A/(1-\mu_A)}\right] .\end{aligned}$$

So it is α_A, the measure of aggregation of ascertainment events in the pedigree relative to that in the general population, that is important for κ. The actual disposition of the units to ascertainment, the δ's, do not count. It is now clear that, if α_A has no connection with the actual disease phenotypes as measured on the pedigree, then CAE can be ignored in likelihood calculations. But, as already noted, in family studies, we are concerned with diseases that aggregate in the selected families, and we often employ a selection process that guarantees this. A connection that is therefore relevant in family studies can be stated as follows.

PROPOSITION IV. *If the clustering of ascertainment of units in the pedigree, α_A, is equal to the clustering of the disease in the pedigree, then*

$$\begin{aligned}\text{CAE} &= -(M-1)\log_e\alpha_A \\ &= (M-1)\log_e\left[1 + \frac{\rho_0}{\mu_0/(1-\mu_0)}\right] ,\end{aligned}$$

where the subscript zero now refers to the corresponding quantity for the disease phenotype itself.

Note that, even if the proposition holds but ρ_0 is equal to the population odds for the disease, $\mu_0/(1-\mu_0)$, then CAE does not depend on parameters of interest. For practical purposes, we observe that, under complete ascertainment ($\pi = 1$) or a so-called population-based family study, every affected person is a proband, and so the units selected into the pedigree are only those that have the disease or are otherwise needed to fill gaps in the pedigree. Under those sampling conditions, the proposition strongly applies, and so some information can be lost by ignoring the CAE, if the phenotypic correlation is substantially different from the population odds of the disease. In applications in which π is small, the aggregation of ascertainment events in the pedigrees under study may not be as strong, and so it may be less penalizing to ignore the CAE; but this needs further study.

Concluding Remarks

Our systematic study of ascertainment corrections for pedigrees shows the following.

1. Smaller units suitably chosen, at least for the purpose of statistical analysis, can be separately corrected for ascertainment without breaking up the pedigrees and without the introduction of the pointers of Lalouel and Morton. This should encourage, in the study-design phase of a project, a consideration of smaller and well-characterized sampling units, even if large pedigrees are contemplated.

2. The appropriate correction for single ascertainment is the reciprocal of the sum of the marginal probabilities of all relevant persons in the unit as if they were affected. Extensions of the result to multiplex ascertainment are indicated. The k-plex ascertainment correction is the reciprocal of the sum of the joint probabilities of all distinct sets of k persons in the unit as if they were all affected.

3. If pedigrees join, then our formulas still apply. If, by pedigrees joining, some units completely overlap,

then those units can be regarded as multiply ascertained, or one can simply multiply the corresponding C's. Several new questions remain. Here we note just a few.

First, can we use CAE, the measure of interunit correlation of ascertainment events, as a way to decide on the optimum smaller family unit for the purpose of ascertainment correction? It seems intuitive that, if the units are chosen so that the biological phenomenon under study is adequately described for each unit, then the effect of CAE, if not zero, would be minimal, on the parameters of interest. In this regard, a nuclear family as a unit may be sufficient for segregation analysis, whereas, for linkage analysis taking into account phase, three generations may be required. Proposition II and the further discussion of the correlation of ascertainment events can be used in numerical assessments of the effects of assuming that there is independent ascertainment of the units.

Another question that can be investigated further concerns pedigrees joining. It may happen that some units completely overlap. The merits of multiple-ascertainment procedures can be studied in comparison with simple multiplication of the associated C's.

The main thesis of the work reported here is that, for the purpose of correcting for ascertainment, we can think in terms of the smaller family units that make up the pedigree, without breaking it. In this regard, a pedigree is simply a set of distinct sibships connected by common parents, and so it is natural to characterize both the sampled and the target population in terms of sibships. Alternatively, one can consider nuclear family units (two-parent sets each with the biological children). The units will then partially overlap, but the theory allows for that. Hence, the classic thinking about ascertainment corrections—that is, in terms of either selection through an affected child or through an affected parent—can be extended to pedigrees as well, without breaking them. This is the thrust of our proposition III. Moreover, with the formulas presented here for single and k-plex ascertainments, we need not use Fisher's π. Thus, advocates of the so-called robust ascertainment corrections should find this work useful. In particular, the appropriate correction for single ascertainment in our proposition I, the generalizations to k-plex ascertainment, and the six recommendations following our proposition III can be evaluated in the development of guidelines.

Furthermore, the problem of pedigree structure is not as entirely hopeless as Vieland and Hodge (1995) portray it to be, for, by thinking in terms of sibships or nuclear families, we can extend the classic approach of Bailey (1951) and Morton (1959), which includes consideration of sibship size distribution, to pedigrees as well, although the modeling should now incorporate intergenerational correlations in sibship sizes. Also, to say that inference is conditional on pedigree structure (of arbitrary size) is then not meaningless, for then we are really speaking of conditioning on sibship sizes.

However, there is still a question about whether we should always think in terms of smaller units, for the purpose of ascertainment correction. In the foregoing development for dependent ascertainments, it was noted that the equal proband status for the selected subunits is not unreasonable for small—say, three-generational—pedigrees. And so, an alternative in such cases is not to think in terms of smaller units at all but to treat the intact pedigree as one unit and to apply the formulas for k-plex ascertainment. This requires the sixth or seventh recommendation following proposition III. In conclusion, thinking in terms of smaller family units, as we have done in this article, opens the question of ascertainment corrections, for further research.

Acknowledgments

The author wishes to acknowledge helpful discussions with Dr. Florence Demenais during the initial phases of the work. This study was supported by U.S. Public Health Service research grants CA61044 and CA06927.

References

Bailey NTJ (1951) A classification of methods of ascertainment and analysis in estimating the frequencies of recessives in man. Ann Eugenics 16:223–225

Bonney GE (1986) Regressive logistic models for familial disease and other binary traits. Biometrics 42:611–625

——— (1992) Compound regressive models for family data. Hum Hered 42:28–41

——— (1995) Some new results on regressive models in family studies. In: 1995 Proceedings of the Biometrics Section. American Statistical Association, Alexandria, VA, pp 177–182

——— (1998) Regression analysis of disposition to correlated binary outcomes. Tech rep 98-01, Department of Biostatistics, Fox Chase Cancer Center, Philadelphia

Bonney GE, Lathrop GM, Lalouel J-M (1988) Combined linkage and segregation analysis using regressive models. Am J Hum Genet 43:29–37

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

Cox DR (1972) Regression models and life-tables (with discussion). J R Stat Soc Ser B 34:187–220

——— (1975) Partial likelihood. Biometrika 62:269–276

Elandt-Johnson RC (1971) Probability models and statistical methods in genetics. John Wiley & Sons, New York

Elston RC (1995) 'Twixt cup and lip: how intractable is the ascertainment problem? Am J Hum Genet 56:15–17

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

Elston, RC, Stewart, J (1971) A general model for genetic analysis of pedigree data. Hum Hered 21:523–554

Elston RC, Yelverton KC (1975) General models for segregation analysis. Am J Hum Genet 27:31–45

Ewens WJ, Shute NC (1986) A resolution of the ascertainment sampling problem. I. Theory. Theor Popul Biol 30:388–412

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

George VT, Elston RC (1991) Ascertainment: an overview of the classical segregation analysis model for independent sibships. Biometrical J 33:741–753

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

Lalouel JH, Morton NE (1981) Complex segregation analysis with pointers. Hum Hered 31:312–321

Lalouel JM, Rao DC, Morton NE, Elston RC (1983) A unified model for complex segregation analysis. Am J Hum Genet 35:816–826

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

Morton NE, MacLean CJ (1974) Analysis of family resemblances. III. Complex segregation of quantitative traits. Am J Hum Genet 26:489–503

Morton NE, Yee S, Lew R (1971) Complex segregation analysis. Am J Hum Genet 23:602–611

Patil GP, Rao CR (1978) Weighted distributions and size-biased sampling with application to wildlife populatioons and human families. Biometrics 34:179–189

Rao CR (1965) On discrete distributions arising out of methods of ascertainment. In: Patil GP (ed) Classical and contagious discrete distributions. Statistical Publishing, Calcutta, and Pergamon Press, New York, pp 320–332

Risch N (1984) Segregation analysis incorporating linkage markers. I Single-locus models with an application to type I diabetes. Am J Hum Genet 36:363–386

Shute NCE, Ewens WJ (1988) A solution of the ascertainment sampling problem. II. Generalizations and numerical results. Am J Hum Genet 43:374–386

Stene J (1977) Assumptions for different ascertainment models in human genetics. Biometrics 33:523–527

——— (1978) Choice of ascertainment model. I. Discrimination between single proband models by means of birth order data. Ann Hum Genet 42:219–229

Thompson EA (1988) Partial and conditional likelihoods in pedigree analysis. Tech rep 141, Department of Statistics, University of Washington, Seattle

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) Weitere Beiträge zur Theorie der Vererbung. IV. Über Methode und Fehlerquellen der Untersuchung auf Mendelsche Zahlen beim Menschen. Arch Rassen Gesellschaftsbiol 9:165–174