Published in final edited form as:
Stat Med. 2008 October 30; 27(24): 5076–5092. doi:10.1002/sim.3327.

Analysis of familial aggregation studies with complex ascertainment schemes

Abigail G. Matthews[1,2,3,*], **Dianne M. Finkelstein**[1,2], and **Rebecca A. Betensky**[1,2]

[1] MGH Biostatistics Center, Boston, MA 02114 U.S.A.

[2] Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115 U.S.A.

[3] Laboratory of Statistical Genetics, Rockefeller University, New York, NY 10065 U.S.A.

SUMMARY

Familial aggregation studies are a common first step in the identification of genetic determinants of disease. If aggregation is found, more refined genetic studies may be undertaken. Complex ascertainment schemes are frequently employed to ensure that the sample contains a sufficient number of families with multiple affected members, as required to detect aggregation. For example, an eligibility criterion for a family might be that both the mother and daughter have disease. Adjustments must be made for ascertainment to avoid bias. We propose adjusting for complex ascertainment schemes through a joint model for the outcomes of disease and ascertainment. This approach improves upon previous simplifying assumptions regarding the ascertainment process.

1. INTRODUCTION

The first step in the identification of hereditary diseases is frequently a familial aggregation study. Such a study seeks to determine whether having relatives with disease increases one's risk of that disease. Familial aggregation refers to this clustering of disease within families. This clustering may be due to genetic and/or environmental factors, or even infectious agents. Given the cost and complexity of finding the disease-causing genes, this initial step is useful as it narrows the focus for future genetic research.

Ascertainment in familial aggregation studies generally falls into three categories. Some studies are population-based, however, this design is inefficient if disease or its hereditary form is rare. Case-control sampling has also been employed, but again, if most cases of disease are sporadic there may be insufficient power to detect agregation within families. As a remedy, other studies recruit subjects on the basis of their family's history of disease.

The general approach of familial aggregation studies is to sample individual(s), called *proband(s)*, and obtain their detailed family history of disease. Families may contain multiple probands, for example, if probands are recruited through physician referral and several family members are attended to by the same physician. We refer to the sampling of probands based on family- and individual-specific criteria as *complex ascertainment*. For example, a study may identify affected individuals through physician referral, but additionally require that at least two first degree relatives have disease. We refer to these participation criteria as the *ascertainment event*. We only consider study designs for which

[*] Correspondence to: Abigail G. Matthews, Ott Laboratory, Box 192, 1230 York Avenue, New York, NY 10065. E-mail: amatthews@rockefeller.edu

the proband statuses of all family members are known, which is generally true of registry data.

Analysis of study designs with non-random sampling must account for the ascertainment scheme in order to avoid bias. This bias could potentially translate into spurious findings of familial aggregation. Consider a study design in which families are sampled if they have at least two affected members. If ascertainment is completely ignored, even in the absence of true familial aggregation, there will appear to be a familial association solely due to the study design. Thompson [1] provided a thorough discussion of non-random sampling, ascertainment bias and several classical approaches to adjusting for ascertainment.

In the case where a single proband is sampled, the simplest approach to ascertainment adjustment is to condition the likelihood contribution of each family on the disease outcome of its proband (e.g., Betensky and Whittemore [2] and Hudson *et al.* [3]). If there are multiple probands, a simplistic approach is to condition on the disease outcome of the first proband recruited to the study, as in Matthews *et al.* [4]. We refer to this as the **first proband** approach.

Tosteson *et al.* [5] extended the latter approach by adjusting for the ascertainment of all the probands in a family. They treated ascertainment status (that is, an indicator of proband status) as random, so that each individual contributes two binary outcomes to the likelihood. They conditioned on the disease outcomes of all probands as well as the ascertainment indicators of the entire family. Two strong assumptions imply that it is sufficient to condition the likelihood contribution of a family on the disease statuses of all probands and to ignore the ascertainment indicators. Thus, under these assumptions, specification of a model for only disease is required. The first assumption is that the probability of being a proband is independent of family history of disease and the second is that either (*i*) the probability of being a proband is additionally independent of disease, or (*ii*) the source population from which families are drawn is extremely large. Tosteson *et al.* [5] propose this method for these specific situations in which the ascertainment ratio is small and the odds of being selected as a proband are solely a function of one's own disease status. In the more complex ascertainment schemes we consider, these assumptions may be inappropriate and unrealistic. However, it is a valuable tool and we will refer to the application of such an approach to complex ascertainment, albeit inappropriately, as the **individual-based** approach. Another approach, proposed by Bonney [6], bases the ascertainment correction on subunits of a family, such as sibships, but requires that some subunits not contain any probands.

The above approaches explicitly model the familial association of disease. Alternatively, the association can be captured through introduction of a random effect (e.g., Houwing-Duistermaat *et al*. [7]; Commenges *et al*. [8]; and Stiratelli *et al*. [9]), through which familial aggregation is expressed implicitly in the variance parameters of the random effect. For a simple case-control design, Commenges *et al*. [8] conditioned on the disease outcome of the proband and included a random effect for each family. For multiple probands per family in the context of a random effects model, Whittemore and Halpern [10] conditioned on the disease indicators of all probands, and required that one pair of relatives be discordant for disease. If the familial association is very strong or disease is very common, this requirement may not be appropriate. Neuhaus and Jewell [11] assumed that the sampling mechanism is based on the number of affected relatives and conditioned the random effects likelihood contribution from a given family on the event that the family contains that number of affected relatives.

In this paper, we directly model the familial association via a full multivariate model. Appropriate choices of models can provide simple and familiar measures of association, such as odds ratios, while random effects models do not. In addition, covariate-specific associations are more straightforward in the multivariate framework through regression modeling. Random effects models incorporate covariates through careful specification of the covariance structure. As in Tosteson *et al*. [5], we treat ascertainment status as random, and jointly model the ascertainment and disease outcomes of a family but assume a different multivariate distribution. We relax the restrictive assumptions of the individual-based approach by directly modeling the association between ascertainment and disease at the family- and individual-level, and we further appropriately condition on the ascertainment event. This proposed approach is referred to as **family-based** because it allows for the odds of being ascertained to depend on one's family history of disease. Since we consider complex ascertainment schemes that involve conditioning on both disease and ascertainment indicators, a full joint model facilitates analysis. Use of a univariate distribution of disease and another for ascertainment conditional on disease would not permit simple conditioning on the ascertainment event. Therefore, we utilize a multivariate model for the joint distribution of disease and ascertainment within families.

In Section 2 we present the multivariate model for a family's disease and ascertainment outcomes considered in this paper. In Section 3 we present the proposed method of analysis as applied to three commonly used study designs with complex ascertainment. In Section 4, we apply our approach to a large familial aggregation study of cancer, and in Section 5 we present simulation results. We conclude in Section 6.

2. JOINT MODELING OF DISEASE AND ASCERTAINMENT

2.1. Multivariate model for disease and ascertainment

Any multivariate binary model can be used for the joint distribution of disease and ascertainment. Here we consider the quadratic exponential model (QEM) [12]. The QEM has been used extensively in the analysis of familial aggregation (e.g., Betensky and Whittemore [2], Hudson *et al*. [3], Hudson *et al*. [13], Laird and Cuenco [14], Rabbee and Betensky [15], Matthews *et al*. [16] and Matthews *et al*. [4]). It is a multivariate log-linear model with all three-way and higher-order associations set to zero. Zhao and Prentice [12] proposed use of this model for univariate outcomes for each family member and Betensky and Whittemore [2] extended it for two outcomes per individual. Hudson *et al*. [3] derived the corresponding logistic regression equations for the multivariate case and Rabbee and Betensky [15] derived sample size calculations. The QEM has several attractive features. First, it is easily implemented using standard statistical software. Second, the parameters have interpretations as conditional odds and odds ratios. This is of particular interest in the context of familial diseases in which the risk of disease given the family history is of primary interest. Third, it enables modeling of associations of outcomes within families and within individuals. Modification of these relationships is straightforward through the introduction of covariates, such as pedigree relationship.

Let y_i indicate the disease status of the ith individual in a given family (i.e., $y_i = 1$ if i has disease and 0 otherwise), and a_i the ascertainment status (i.e., $a_i = 1$ if i is ascertained and 0 otherwise) for $i = 1, \ldots, n$. Several members of a single family can be ascertained. The QEM for two binary outcomes (y_i and a_i) for a family of size n is

$$P\left(y_1,\ldots,y_n,a_1,\ldots,a_n\right) \propto \exp\left\{\sum_{i=1}^{n}\theta_{yi}\,y_i+\sum_{i=1}^{n}\theta_{ai}\,a_i+\sum_{i=1}^{n}\theta_{yai}\,y_i\,a_i+\sum_{i<j}\gamma_{yij}\,y_i\,y_j+\sum_{i<j}\gamma_{aij}\,a_i\,a_j+\sum_{i\neq j}\gamma_{yaij}\,y_i\,a_j\right\}. \tag{1}$$

The parameters of primary interest in assessing familial aggregation are the γ_{yij}'s; they capture the increase in disease log-odds ratios associated with having an affected relative. The log-odds of disease is captured by θ_{yi}, and ascertainment by θ_{ai}. The association between disease and ascertainment within families is captured by γ_{yaij}, while θ_{yai} captures this association at the individual-level. It is important to note that these parameter interpretations are conditional on all other outcomes. For example, γ_{yij} measures familial aggregation of disease conditional on the disease and ascertainment outcomes of all other individuals.

The general QEM in (1) can be simplified by assuming that relatives are exchangeable, that is, they are identical with respect to the disease and ascertainment processes. This assumption implies that $\theta_{yi} = \theta_y$, $\theta_{ai} = \theta_a$, $\theta_{yai} = \theta_{ya}$, $\gamma_{yij} = \gamma_y$, $\gamma_{aij} = \gamma_a$ and $\gamma_{yaij} = \gamma_{ya}$ for all i and j. To simplify our presentation, we assume exchangeability throughout this paper. However, exchangeability can be easily relaxed through the introduction of covariates, for example, $\gamma_{yij} = \gamma_{y,0} + \gamma_{y,1}z_{ij}$, where z_{ij} is a pair-level covariate, such as genetic distance. Consider a family that contains full and half siblings and let z_{ij} be an indicator of full sibling status. Then, $\gamma_{y,0}$ captures the disease association between siblings who share only one parent and $\gamma_{y,1}$ is the increase in the association due to having two parents in common.

The QEM implies a set of logistic regression equations, and thus standard statistical software may be used for estimation [3]. These regression equations are

$$\begin{aligned} \text{logit}\ [\mathrm{P}\ (y_i{=}1\,|\,\boldsymbol{y}_{-i},\boldsymbol{a})] &= \theta_y+\theta_{ya}\,a_i+\gamma_y\sum_{i\neq j}y_j+\gamma_{ya}\sum_{i\neq j}a_j \\ \text{logit}\ [\mathrm{P}\ (a_i{=}1\,|\,\boldsymbol{a}_{-i},\boldsymbol{y})] &= \theta_a+\theta_{ya}\,y_i+\gamma_a\sum_{i\neq j}a_j+\gamma_{ya}\sum_{i\neq j}y_j, \end{aligned} \tag{2}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)'$ and $\boldsymbol{y}_{-i} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)'$. The vectors $\boldsymbol{a}$ and $\boldsymbol{a}_{-i}$ are defined similarly. The robust variance estimator of Liang and Zeger [17] is used to adjust the variance of the resulting parameter estimates for the correlation among relatives.

2.2. Marginal measure of disease association

The canonical parameter, γ_y, is the log-odds ratio conditional on family history and ascertainment. Interest may lie instead in the log-odds ratio of disease conditional solely on family history of disease, especially for genetic counseling. Alternatively, interest may reside in the marginal association of disease among family members, without regard for family history of disease or ascertainment. Both of these marginalized measures of association can be calculated from the fully specified joint probability model. One example is the unconditional pairwise odds ratio of disease, $e^{\delta}{}_M$, where

$$e^{\delta_M} = \frac{\mathrm{P}\left(y_i{=}1, y_j{=}1\right)\times\mathrm{P}\left(y_i{=}0, y_j{=}0\right)}{\mathrm{P}\left(y_i{=}1, y_j{=}0\right)\times\mathrm{P}\left(y_i{=}0, y_j{=}1\right)} \tag{3}$$

for all $i \neq j$. Letting $\boldsymbol{y}_{-ij}$ denote the vector of disease statuses of all family members excluding the ith and jth individuals, the probability (3) follows from (1), such that

$$\delta_M = \log\left[\sum\sum_{\boldsymbol{A}\boldsymbol{Y}_{-ij}}\mathrm{P}\left(y_i{=}0, y_j{=}0, \boldsymbol{y}_{-ij}, \boldsymbol{a}\right)\right]+\log\left[\sum\sum_{\boldsymbol{A}\boldsymbol{Y}_{-ij}}\mathrm{P}\left(y_i{=}1, y_j{=}1, \boldsymbol{y}_{-ij}, \boldsymbol{a}\right)\right]-2\log\left[\sum\sum_{\boldsymbol{A}\boldsymbol{Y}_{-ij}}\mathrm{P}\left(y_i{=}1, y_j{=}0, \boldsymbol{y}_{-ij}, \boldsymbol{a}\right)\right], \tag{4}$$

where $\boldsymbol{A}$ denotes all possible values of the vector $\boldsymbol{a}$, and $\boldsymbol{Y}_{-ij}$ denotes all possible values of the vector $\boldsymbol{y}_{-ij}$. Note that the last term is multiplied by 2 due to the assumption of exchangeability. For a family of size two, δ_M is given by

$$\delta_M \quad = \quad \log\left[\sum_{a_1,a_2} \mathrm{P}(y_1=0, y_2=0, a_1, a_2)\right] + \log\left[\sum_{a_1,a_2} \mathrm{P}(y_1=1, y_2=1, a_1, a_2)\right] - 2\log\left[\sum_{a_1,a_2} \mathrm{P}(y_1=1, y_2=0, a_1, a_2)\right].$$

These probabilities are then calculated using the distribution in (1) and the corresponding parameter estimates.

Inference based on transformed measures of association requires computation of the appropriate Jacobian. Advantageously, the QEM is a member of the exponential family of distributions. Consider the transformation from γ_y in (1) to δ_M in (3). Let ϕ denote the vector of the original parameters ($\theta_y, \theta_a, \theta_{ya}, \underline{\gamma_y}, \gamma_a, \gamma_{ya})'$, ϕ' denote the transformed vector of parameters ($\theta_y, \theta_a, \theta_{ya}, \underline{\delta_M}$, $\gamma_a, \gamma_{ya})'$ and $\boldsymbol{T}$ denote the vector of sufficient statistics,

$$\boldsymbol{T} = \left(\sum_i y_i, \sum_i a_i, \sum_i y_i a_i, \sum_{i<j} y_i y_j, \sum_{i<j} a_i a_j, \sum_{i\neq j} y_i a_j\right)'. \tag{5}$$

The Jacobian of this transformation is given $\boldsymbol{J} = \left(\frac{\partial \phi}{\partial \phi'}\right) = \left(\frac{\partial \phi'}{\partial \phi}\right)^{-1}$. Only the fourth element of the parameter vector is transformed. Thus, the Jacobian is an identity matrix except for the fourth row, which is $\frac{\partial \delta_M}{\partial \phi}$. Since the QEM is a member of the exponential family, the fourth row is

$$\left[\mathrm{E}_\phi\left(\boldsymbol{T} \mid y_i=1, y_j=1\right) + \mathrm{E}_\phi\left(\boldsymbol{T} \mid y_i=0, y_j=0\right) - 2\,\mathrm{E}_\phi\left(\boldsymbol{T} \mid y_i=1, y_j=0\right)\right]^{-1}.$$

Jacobians for other transformations are similar in form. The variance of $\hat{\phi}'$ is given by

$$\mathrm{Cov}_\phi\left[\widehat{\phi'}\right] = \left(\boldsymbol{J}\,\mathscr{I}\,\boldsymbol{J}'\right)^{-1} \tag{6}$$

where $\mathscr{I}$ is the expected information matrix of the original parameters ($\hat{\phi}$). Testing for familial aggregation of disease, independent of ascertainment, is then performed by using (4) and (6) to construct a confidence interval for δ_M .

2.3. Varying family sizes

The quadratic exponential model in (1) can only be assumed to hold for one family size, that is, it is *irreproducible*. This can be seen by the fact that the logistic regression equations in (2) are a function of the number of relatives with disease and the number probands. In the rare disease case or under approximate independence, Betensky and Whittemore [2] and Cox and Wermuth [18] showed that the QEM is reproducible.

Matthews *et al*. [16] consider several approaches to implementing the QEM in the presence of varying family sizes. One approach, termed the “missing data” approach, assumes the QEM to hold for a large family size, then treats smaller families as having “missing” relatives [19]. A likelihood-based approach is employed to account for this missing data in which the likelihood contribution of a small family is the sum of the full data likelihood in (1) over all possible values of the missing data. For example, assume the QEM for a family of size three, then the likelihood contribution for a family of size two is proportional to

$$\sum_{\mathbf{Y}_3,\mathbf{A}_3} \mathrm{P}\left(y_1, y_2, y_3, a_1, a_2, a_3\right),$$

where $\mathbf{Y}_3$ and $\mathbf{A}_3$ denote all possible values of y_3 and a_3, respectively. This summation precludes use of the logistic regression equations in (2), however, the maximum likelihood estimates can be obtained through direct minimization of the negative log-likelihood. This missing data approach is inappropriate if there are pair- or individual-specific covariates, or the range of family sizes is large. A major drawback of this approach is in interpretation, as smaller families do not actually contain any missing data.

A second, “marginalization” approach proposed by Matthews *et al*. [16] takes the opposite tactic and assumes the QEM to hold for a small family size, say *n*. The QEM is also assumed for each subset of size *n* from larger families. Each subset then contributes to the likelihood for estimation and generalized estimating equations (GEEs) [17] are used to adjust for the correlation among subsets arising from the same family. Only those subsets that satisfy the ascertainment criteria are used for estimation. For example, in the proband sampling paradigm each subset must contain at least one proband. The use of GEEs requires specification of the score equations for each family or subset.

A third approach, termed the “hybrid approach,” is to combine the first two approaches. This involves assuming the QEM to hold for a moderate family size and using the missing data approach for the smaller families and the marginalization approach for the larger families.

3. STUDY DESIGNS

Likelihood-based analysis of family studies must condition on the ascertainment event that brought the family into the study. For example, if a family is required to have at least two affected members in order to participate in the study, each family's contribution to the likelihood must condition on the event that there are at least two affected members and at least one proband. *Any* joint model of disease and ascertainment within families and within individuals facilitates this analysis. In our analyses, we elect to use the QEM, specified in (1).

We consider three commonly used study designs used for family studies of disease. The family's ascertainment event in the first of these designs is simply the ascertainment of at least one family member; we refer to this study design as **proband sampling**. There are two different familial ascertainment events utilized in the second study design. Case families are required to have a *minimum* number of affected ascertained individuals, and control families are required to have a *maximum* number of affected ascertained individuals. We refer to this study design as **case-control family sampling**. The third study design requires a minimum number of affected relatives in a family and at least one ascertained individual. We refer to this last study design as **high-risk family sampling**.

3.1. Proband sampling

Proband sampling involves recruiting individuals and then obtaining their family history of disease. As there is the possibility of multiple probands per family, the conditioning event is that there is at least one proband in the family (i.e., $\Sigma\, a_i \geq 1$). As an example, consider the Framingham Offspring Study [20], in which children of diabetic parents from the original Framingham Heart Study were recruited. These probands then provided a family history of diabetes. The assumption of the individual-based approach regarding the independence of ascertainment and family history of disease requires that the offspring are sampled independently of the parental disease statuses. This is clearly violated in this study suggesting a family-based approach would be more appropriate. For the proposed approach, under this design, a family's contribution to the likelihood is

$$\mathrm{P}\left(y_1, \ldots, y_n, a_1, \ldots, a_n \,\middle|\, \Sigma a_i \geq 1\right).$$

The logistic regression equations in (2) can be used to obtain parameter estimates; however, the expected information matrix must account for the conditional likelihood. Simple differentiation of the score equations, gives an expected information matrix of

$$\mathscr{I} = N \,\mathrm{Cov}_{\phi}\left[\boldsymbol{T} \middle| \Sigma a_i \geq 1\right]$$

where $\boldsymbol{T}$ is the vector of sufficient statistics in (5).

3.2. Case-control family sampling

The case-control design aims to sample two types of families: one with the hereditary form of the disease, and the other with no or sporadic disease. Case families potentially carry the hereditary form of the disease and are required to contain a *minimum* number of affected ascertained individuals. Control families with potentially sporadic disease are required to contain at most a *maximum* number of affected ascertained individuals. As an example, consider the study of alcoholism as described in Hill *et al*. [21]. “High-risk” (case) families required at least two affected brothers, while “low-risk” (control) families had both non-alcoholic probands and their non-alcoholic first-degree relatives. In this setting, the assumption of independence of ascertainment and family disease history from an individual-based approach is obviously inappropriate.

Consider case-control family sampling in which case families have at least c_1 affected ascertained individuals, and control families have c_0 or fewer affected ascertained individuals ($0 \leq c_0 < c_1 \leq n$). The likelihood contribution for a case family is

$$\mathrm{P}\left(y_1, \ldots, y_n, a_1, \ldots, a_n \,\middle|\, \Sigma y_i a_i \geq c_1\right), \tag{7}$$

and for a control family is

$$\mathrm{P}\left(y_1, \ldots, y_n, a_1, \ldots, a_n \,\middle|\, \Sigma y_i a_i \leq c_0, \Sigma a_i \geq 1\right). \tag{8}$$

Note that the contribution from control families must condition explicitly on the presence of at least one ascertained individual; this is implicit in the conditioning event for case families. Parameter estimation and derivation of the expected information matrix follow that of the proband sampling design. Letting N_0 denote the number of control families, and N_1 denote the number of case families, the expected information matrix is given by

$$\mathscr{I} = N_0 \operatorname{Cov}_\phi \left[\boldsymbol{T} \middle| \Sigma y_i a_i \le c_0, \Sigma a_i \ge 1 \right] + N_1 \operatorname{Cov}_\phi \left[\boldsymbol{T} \middle| \Sigma y_i a_i \ge c_1 \right].$$

3.3. High-risk family sampling

To increase the power to detect familial aggregation, the high-risk family design samples families with multiple affected members. To accomplish this, a family is required to have at least a certain number of affected members *and* at least one ascertained individual. This study design is advantageous in the case of a rare disease or if the risk of disease is small for those with the hereditary form of the disease. This design differs from the case-control family sampling design in that it does not require the affected family members to be probands. An example is a study of alcoholism that sampled families with three or more affected first degree relatives [22]. Alcoholism is a fairly common disease, but high-risk family sampling was used because of the existence of many sporadic cases. Again, the individual-based approach's assumption of independence of ascertainment and family history of disease is clearly violated. The appendix adapts the individual-based approach to this study design and shows that ascertainment can still be ignored under the original assumptions. However, when these assumptions do not hold, as they likely do not in most disease contexts, we condition the full joint distribution for the family on the appropriate ascertainment events. Letting c denote the required number of affected family members ($c < n$), the family's contribution to the likelihood for the proposed approach is given by

$$\mathrm{P}\left(y_1, \ldots, y_n, a_1, \ldots, a_n \middle| \Sigma a_i \ge 1, \Sigma y_i \ge c \right).$$

Derivation of the information matrix follows that of the first study design. For a set of N families, it is given by

$$\mathscr{I} = N \operatorname{Cov}_\phi \left(\boldsymbol{T} \middle| \Sigma a_i \ge 1, \Sigma y_i \ge c \right).$$

.

4. EXAMPLE

We now compare four approaches to accounting for ascertainment for each of the three sampling designs described in Section 3. The first approach completely ignores ascertainment and is referred to as the naive approach. The second is the first proband approach, which conditions the likelihood of a family's disease outcomes on the disease status of the first individual recruited to the study regardless of the actual ascertainment employed. The third approach is the individual-based, and conditions a family's likelihood on the disease outcomes of all ascertained individuals. The fourth approach is the one proposed here based on specification of the joint distribution of disease and ascertainment, and conditioning on the precise ascertainment event. These approaches require specification of the joint distribution of disease or the joint distribution of disease and ascertainment among family members. We use the QEM for these joint distributions, though other choices are possible as well.

To study the different analytic approaches as applied to the three study designs, we sampled from a study of 18,028 individuals recruited by the National Cancer Institute-sponsored

Cancer Genetics Network (CGN). Specifically, we applied the three sampling designs to the 11,028 population-based families from the CGN registry to obtain "pseudo-studies" that conform to these designs. Every individual with a cancer diagnosis before age 65 was "recruited" into the psuedo-study. We aimed to test the hypothesis that skin cancer aggregates in families. Initially, we only analyzed sibships of size three, as a crude form of age-matching; other sizes will be considered later. The data are summarized in the upper half of Table 1.

To compare the four approaches for analysis of these data, we computed δ_M, the pairwise log-odds ratio (4) of skin cancer. The estimates and standard errors are listed in Table 2. Both the first proband and proposed approaches found statistically significant familial aggregation of skin cancer. The first proband approach accounts for the ascertainment quite well because 83% of sibships have only one proband. The standard error of the individual-based approach is large due to the fact that it conditions on more information than the others. This example illustrates that improper adjustment for ascertainment can lead to a decrease in power.

In the case-control family study, case families are required to have at least one affected ascertained individual (334 families) and control families must contain only unaffected ascertained individuals (1056 families). Only the results of the proposed approach are listed in Table 2 under this design heading (the results of the other approaches are the same as for the proband sampling design). There is significant evidence of aggregation of skin cancer, although the estimate is smaller than that from the proband sampling design. This occurs because more of the observed cases of disease are attributable to the ascertainment scheme.

In the high-risk family study design, families were included if they contained at least one skin cancer and at least one ascertained individual. In total, 355 sibships of size three were analyzed. The distribution of skin cancer in these sibships is given in the lower half of Table 1. Results from all four analytic approaches are given in Table 2. The first proband approach yields significant negative familial aggregation. Because most families (89%) in the pseudo-study have only one affected member, the sampling design induces a negative disease association without proper adjustment. The individual-based approach does not converge. The estimate of the pairwise log-odds ratio in (4) of skin cancer based on the proposed family-based approach is 1.53 and statistically significant, agreeing with the results from the other two analyses. We note that this can be viewed as evidence of either environmental and/or genetic causes, or an interaction between the two. Sorting this out will require further study.

These comparisons highlight the necessity of adjusting for complex ascertainment in the analysis of familial aggregation studies. In all three study designs, the proposed family-based approach yields estimates that are larger in magnitude and have smaller standard errors than the individual-based approach. This suggests that despite the fact that it involves estimation of more parameters than the individual-based approach, the proposed approach is more powerful since it conditions on less information and does not require unrealistic assumptions of independence. Finally, not surprisingly, we observe that the more complex the ascertainment event, the smaller the degree of familial association detected.

Lastly, we illustrate application of the proposed family-based approach in the presence of varying family sizes using the marginalization approach introduced in Section 2.3. In this analysis sibships of sizes three, four and five from the proband sampling pseudo-study are considered. There are 1390, 1079 and 726 sibships of each size, respectively. The bivariate QEM was assumed to hold for sibships of size three. Including additional family sizes also yields significant aggregation of skin cancer, but the magnitude of the association is smaller.

The estimated pairwise log-odds ratio of skin cancer is 0.81 (95% CI: 0.54 - 1.08). The estimate using only sibships of size three is greater (1.04) because 78% of the larger sibships have no affecteds at all. When multiple family sizes are analyzed the standard error increases likely due to heterogeneity in the familial association over different sibship sizes. These measures are directly comparable because both analyses assume the QEM to hold for families of size three.

5. SIMULATION STUDIES

We conducted several simulation studies to compare the proposed family-based approach for analyses with the naive, first proband and individual-based approaches for the three study designs considered in this paper. We considered two different parameter configurations for each study design. The first contains a moderate association between family history of disease and ascertainment, and the second contains a strong association. The individual-based approach assumption of independence is violated under each configuration. We report simulation results for families of size four. For each study design, 500 families were generated from the corresponding conditional likelihood based on the bivariate QEM, and each simulation consists of 500 iterations. We focus our comparisons on the pairwise log-odds ratio parameter, δ_M, in (4). Results are listed in Table 3.

For the proband sampling design, the naive, first proband, and individual-based approaches all exhibit substantial bias in their estimates of δ_M. In the case of a moderate association between disease and ascertainment (parameter configuration 1), the naive and first proband estimates are 0.11 and 0.12, when they should be 0.26. In the case of strong disease-ascertainment association, the estimates are 0.38–0.48 when the true value is 0.58. It is apparent that even in this simple design, it is essential to fully account for ascertainment when assessing familial aggregation.

For the case-control family study design, case families contain at least one affected proband, and control families have no probands with disease. We generated 250 case families from (7) and 250 control families from (8). The results are similar to those from the proband sampling design, but even more extreme. In particular, the naive approach yields a *negative* familial association. The estimate from the individual-based approach is positive, though biased.

Results are more extreme for the high-risk family sampling study design. The individual-based approach estimate is negative (−0.93); as in the skin cancer example, this is induced through not properly adjusting for the design requirement of an affected family member.

The Monte Carlo and analytic standard errors of the estimates are listed throughout the table. These are generally close, though there are some discrepancies. The differences are due to the fact that δ_M is a transformation of the canonical parameters; any instability in those estimates is magnified through the analytic calculations for δ_M.

Lastly, we evaluated the performance of the proposed family-based approach, with respect to bias, efficiency, power and type I error rates for one- and two-sided tests, in comparison to the other three approaches when the model is misspecified. We generated disease indicators for each family from the univariate QEM. We then generated ascertainment indicators from Bernoulli distributions with probability of ascertainment being dependent on disease and the number of affected family members. We assumed ascertainment to be independent among relatives conditional on the disease indicators of all family members. To calculate the type I error rate, parameters and probabilities were chosen such that δ_M was zero.

For each study design we assessed the power of the different approaches to detect familial aggregation as measured by the pairwise log-odds ratio, δ_M. We used 500 simulated datasets consisting of either 500 or 350 families. Results are given in Table 4. The minimum bias and maximum power are achieved by the proposed family-based approach. For the high-risk family sampling design, the other approaches appear to have higher power (two-sided tests) but they are detecting a negative familial association. The type I error rate for the naive and first proband approaches are slightly inflated for the first two study designs. The proposed approach has a slightly inflated one-sided type I error for these sample sizes, however, for larger samples the rate decreases to 5%. Interestingly, the power of the individual-based approach is exceedingly low, likely due to the violation of its assumptions by the probability model from which we simulated. In addition, the power decreases as the complexity of the ascertainment event increases.

6. DISCUSSION

The simulation studies performed in this paper confirm that if ascertainment is related to disease, the analyses must fully adjust for ascertainment in order to avoid bias. Partial adjustment, as afforded by the individual-based approach, is insufficient in many realistic scenarios of genetic epidemiologic studies. In fact, as seen in both the example and the simulations, in the case of a large positive association between ascertainment and disease, an unadjusted approach may indicate *negative* disease aggregation when it is truly positive. In other simulations (not reported here), the proposed family-based approach is comparable in performance to the individual-based approach when the latter's assumptions are valid. Since we condition on less information, in some cases the family-based approach is even more powerful. In addition, the proposed approach appears to perform well under one example of model misspecification. The proposed approach is not well-suited for datasets in which there are only a few families with multiple probands or only a few families with multiple affected members. It is well-suited, however, for studies in which the mode of ascertainment violates the individual-based assumptions, as in two studies described in Hill *et al.* [22], in which probands had to be seeking treatment for alcoholism and have at least one affected brother or sister.

In addition to allowing for straightforward adjustment for complex family-based ascertainment events, the QEM has the advantage of producing estimates of disease risk as functions of family history, which is quite useful in genetic counseling situations. Its parameters are easily estimated using standard logistic regression software, and are easily transformed to marginalize over ascertainment, as required for practical use. The QEM has the drawback of being irreproducible, though we have illustrated its implementation for varying family sizes using a marginalization approach. In disease scenarios in which the irreproducibility is problematic (e.g., due to a wide range of family sizes), an alternative multivariate model for disease and ascertainment among family members can be used, as long as it allows for an association at the individual- and family-levels. The QEM is also somewhat cumbersome if it is desired to incorporate external estimates of disease prevalence into the estimation procedure; this would involve a difficult constraint on a complicated transformation of the natural parameters. This can also be remedied through the use of an alternative model in which the marginal disease prevalences are natural parameters.

Acknowledgments

This research was supported in part by the Cancer Genetics Network (CGN) under NCI contract U01 CA78284-04, NIH grants R01 CA074302 and R01 CA075971, and NIMH grants 5T32MH020053-05 and MH44292. The authors wish to thank the Cancer Genetics Network Investigators who allowed us to use their data in our example:

CGN Participating Centers Principal Investigators: Claudine Isaacs, M.D., Georgetown University Lombardi Cancer Center, Washington D.C.; Geraldine Minaeu, Ph.D., University of Utah, Salt Lake City, UT; and Joellen Schildkraut, Ph.D., Duke University Medical Center, Durham, NC.

NCI CGN Program Directors: Carol H. Kasten-Sportes, M.D., and Susan G. Nayfield, M.D., M.Sc.

Contract/grant sponsor: National Cancer Institute; contract/grant number: U01 CA78284-04

Contract/grant sponsor: National Institutes of Health; contract/grant number: R01 CA074302, R01 CA 075971

Contract/grant sponsor: National Institute of Mental Health; contract/grant number: 5T32MH020053-05, MH44292

APPENDIX: EXTENSION OF THE INDIVIDUAL-BASED APPROACH TO A HIGH-RISK FAMILY SAMPLING STUDY DESIGN

To adjust the individual-based approach to the high-risk family study design (Section 3.3), we condition on three quantities: the ascertainment indicators of all family members, the disease indicators of all ascertained individuals, and the presence of at least *c* affected members. Thus the likelihood is

$$\mathrm{P}\left(y_{r+1}, \ldots, y_n \middle| y_1, \ldots, y_r, a_1, \ldots, a_n, \Sigma y_i \geq c\right), \tag{9}$$

where *r* is the number of probands in the family.

The likelihood in (9) can be shown to equal

$$\frac{\mathrm{P}(y_1,\ldots,y_n,a_1,\ldots,a_n \mid \Sigma y_i \geq c)}{\mathrm{P}(y_1,\ldots,y_r,a_1,\ldots,a_n \mid \Sigma y_i \geq c)} = \frac{\mathrm{P}(a_1,\ldots,a_n \mid y_1,\ldots,y_n,\Sigma y_i \geq c) \times \mathrm{P}(y_1,\ldots,y_n \mid \Sigma y_i \geq c)}{\underbrace{\sum_{Y_{r+1}} \cdots \sum_{Y_n}}_{\text{where } \sum_{i=1}^{n} y_i \geq c} \mathrm{P}(a_1,\ldots,a_n \mid y_1,\ldots,y_n,\Sigma y_i \geq c) \times \mathrm{P}(y_1,\ldots,y_n \mid \Sigma y_i \geq c)}.$$

The assumption of the individual-based approach that an individual's ascertainment status is only dependent on their disease status (and not that of family members) implies that the distribution of ascertainment given disease is binomially distributed. Letting $\tau_1 = \mathrm{P}(a = 1 \mid y = 1)$ and $\tau_2 = \mathrm{P}(a = 1 \mid y = 0)$, it follows that

$$\mathrm{P}\left(a_1, \ldots, a_n \mid y_1, \ldots, y_n\right) = \prod_{i=1}^{r} \tau_1^{y_i} \tau_2^{1-y_i} \times \prod_{j=r+1}^{n} \left(1 - \tau_1^{y_j} \tau_2^{1-y_j}\right).$$

Thus, the likelihood becomes

$$\frac{\mathrm{P}\left(y_1, \ldots, y_n \mid \Sigma y_i \geq c\right)}{\underbrace{\sum_{Y_{r+1}} \cdots \sum_{Y_n}}_{\text{where } \sum_{i=1}^{n} y_i \geq c} \prod_{j=r+1}^{n} \left(1 - \tau_1^{y_j} \tau_2^{1-y_j}\right) \times \mathrm{P}\left(y_1, \ldots, y_n \mid \Sigma y_i \geq c\right)}$$

since the terms involving only ascertained individuals cancel.

The second set of assumptions of the individual-based approach are: (*i*) a large source population (i.e., τ_1, $\tau_2 \to 0$), or (*ii*) independence between ascertainment and disease within an individual (i.e., $\tau_1 = \tau_2$). If either holds,

$$\begin{aligned} \mathrm{P} & \left(y_{r+1}, \ldots, y_n \middle| y_1, \ldots, y_r, a_1, \ldots, a_n, \Sigma y_i \geq c\right) \\ & = \mathrm{P}\left(y_{r+1}, \ldots, y_n \middle| y_1, \ldots, y_r, \Sigma y_i \geq c\right). \end{aligned}$$

Thus, the assumptions of the individual-based approach imply that the ascertainment indicators in (9) can be ignored when computing the likelihood contribution of a family under this study design.

REFERENCES

1. Thompson E. Sampling and ascertainment in genetic epidemiology; a tutorial review. 1993 Unpublished manuscript.
2. Betensky R, Whittemore A. An analysis of correlated multivariate binary data: Application to familial cancers of the ovary and breast. Applied Statistics. 1996; 45:411–429.
3. Hudson JI, Laird NM, Betensky RA. Multivariate logistic regression for familial aggregation of two disorders: I. Development of models and methods. American Journal of Epidemiology. 2001; 153:500–505. [PubMed: 11226971]
4. Matthews AG, Betensky RA, Anton-Culver H, Bowen D, Griffin C, Isaacs C, Kasten C, Mineau G, Nayfield S, Schildkraut J, Strong L, Weber B, Finkelstein DM. Analysis of co-aggregation of cancer based on registry data. Community Genetics. 2006; 9:87–92. [PubMed: 16612058]
5. Tosteson T, Rosner B, Redline S. Logistic regression for clustered binary data in proband studies with application to familial aggregation of sleep disorders. Biometrics. 1991; 47:1257–1265. [PubMed: 1786318]
6. Bonney G. Ascertainment corrections based on smaller family units. American Journal of Human Genetics. 1998; 63:1202–1215. [PubMed: 9758615]
7. Houwing-Duistermaat J, van Houwelingen H, de Winter J. Estimation of individual genetic effects from binary observations on relatives applied to a family history of respiratory illesses and chronic lung disease of newborns. Biometrics. 2000; 56:808–814. [PubMed: 10985220]
8. Commenges D, Jacqmin H, Letenneur L, van Duijn C. Score test for familial aggregation in proband studies: Application to Alzheimer's disease. Biometrics. 1995; 51:542–551. [PubMed: 7662843]
9. Stiratelli R, Laird N, Ware J. Random-effects models for serial observations with binary response. Biometrics. 1984; 40:961–971. [PubMed: 6534418]
10. Whittemore A, Halpern J. Logistic regression of family data from retrospective study designs. Genetics Epidemiology. 2003; 25:177–189.
11. Neuhaus J, Jewell N. The effect of retrospective sampling on binary regression models for clustered data. Biometrics. 1990; 46:977–990. [PubMed: 2085642]
12. Zhao L, Prentice R. Correlated binary regression using a quadratic exponential model. Biometrika. 1990; 77:642–648.
13. Hudson JI, Laird NM, Betensky RA, Keck PE Jr, Pope HJ Jr. Multivariate logistic regression for familial aggregation of two disorders: II. Analysis of studies of eating and mood disorders. American Journal of Epidemiology. 2001; 153:506–514. [PubMed: 11226983]
14. Laird N, Cuenco K. Regression methods for assessing familial aggregation of disease. Statistics in Medicine. 2003; 22:1447–1455. [PubMed: 12704608]
15. Rabbee N, Betensky R. Power calculations for familial aggregation studies. Genetic Epidemiology. 2004; 26:316–327. [PubMed: 15095391]
16. Matthews AG, Finkelstein DM, Betensky RA. Analysis of familial aggregation in the presence of varying family sizes. Applied Statistics. 2005; 54:847–862.

17. Liang K, Zeger S. Longitudinal data analysis using generalized linear models. Biometrika. 1986; 73:13–22.

18. Cox D, Wermuth N. A note on the quadratic exponential binary distribution. Biometrika. 1994; 81:403–408.

19. Fitzmaurice G, Laird N. A likelihood-based method for analysing longitudinal binary responses. Biometrika. 1993; 80:141–151.

20. Meigs J, Cupples L, Wilson P. Parental transmission of type 2 diabetes: The Framingham Offspring Study. Diabetes. 2000; 49:2201–2207. [PubMed: 11118026]

21. Hill S, Yuan H, Locke J. Path analysis of P300 amplitude of individuals from families at high and low risk for developing alcoholism. Biological Psychiatry. 1999; 45:346–359. [PubMed: 10023513]

22. Hill S, Zezza N, Wipprecht G, Xu J, Neiswanger K. Linkage studies of D2 and D4 receptor genes and alcoholism. American Journal of Medical Genetics (Neuropsychiatric Genetics). 1999; 88:676–685. [PubMed: 10581489]