Analysis of Familial Aggregation Studies with Complex Ascertainment Schemes

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

[1]MGH Biostatistics Center, Boston, MA 02114

[2]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115

April 5, 2005

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 fmaily 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.

KEY WORDS: familial aggregation; ascertainment bias; quadratic exponential.

1. Introduction

The first step in the identification of hereditary diseases is frequently a familial aggregation study. The goal of the study is to determine whether there is an increase in the risk of disease associated with having relatives affected with disease. Familial aggregation refers to the clustering of the disease within families, which may be due to genetic factors, environmental factors, and/or infectious agents. Given the cost and complexity of finding the genes responsible for a disease, this initial step is useful as it narrows the focus for future genetic research. Some familial aggregation studies recruit subjects from a registry and the recruitment can be considered population-based. Other studies recruit subjects on the basis of their being at increased risk for the hereditary form of the disease. Alternatively, recruitment can be based on the actual presence of disease so that there are enough cases of disease to detect aggregation. Non-random sampling, such as in these examples, is particularly useful if the hereditary form of the disease is rare.

There are several methods of non-random ascertainment. Single ascertainment involves sampling individuals, called probands, on the basis of their disease status and then obtaining a detailed history of disease in their relatives. It is possible that multiple probands in a single family could be ascertained. For example, if probands are referred by a physician, two family members could be referred by the same physician. More complicated ascertainment schemes recruit probands based on certain criteria, such as their disease status, and then recruit family members based on the same or different disease criteria. For example, a study may identify affected individuals through physician referrals, and additionally require that at least two first degree relatives are also affected. We

refer to the participation criteria as the ascertainment event.

Analysis of study designs with non-random sampling must account for the ascertainment in order to avoid bias. This bias could potentially translate into a spurious finding of familial aggregation. Consider a study design in which families are sampled if they have at least two affected members. If the **naive** approach is taken and 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 (1993) provides a thorough discussion of non-random sampling, ascertainment bias and several classical approaches to adjusting for ascertainment.

In the case of single ascertainment, a simple approach is to condition the likelihood contribution of each family on the disease outcome of its proband (e.g., Betensky and Whittemore, 1996 and Hudson, Laird and Betensky, 2001). If there are multiple probands, one approach is to condition on the disease outcome of the first proband recruited to the study. We refer to this as the **first proband** approach. Tosteson, Rosner and Redline (1991) extended this and adjusted for the ascertainment of all probands in a family. They treat ascertainment status as random and condition on the ascertainment indicators of all family members and on the disease indicators of all probands. Their approach requires two strong assumptions. One is that the probability of being a proband is independent of family history. The other is that either the probability of being a proband is completely independent of disease, or the source population from which families are drawn is extremely large. Under these assumptions they show that ascertainment can be ignored, and that it is sufficient to condition the likelihood contribution of a family on

the disease statuses of all probands in the family. Alternatively, Bonney (1998) suggested that ascertainment corrections can be based on subunits of a family, such as sibships, but requires that some subunits not contain any probands.

All approaches to adjusting for ascertainment considered thus far specify the joint distribution of disease outcomes among family members, which explicity involves association parameters. Alternatively, the familial association of disease can be captured through introduction of a random effect (e.g., Howing-Duistermaat, van Houwelingen and de Winter, 2000, Commenges, Jacqmin, Letenneur and van Duijn, 1995, Stiratelli, Laird and Ware, 1984), through which familial aggregation is expressed implicitly in the variance parameters of the random effect. Random effects models also yield biased results if the ascertainment process is not properly adjusted for or if the assumed distribution of the random effects is incorrect (Epstein et al., 2002; Glidden and Liang, 2002). In a simple case-control study design, Commenges et al. (1995) proposed adjusting for ascertainment by conditioning on the marginal probability of disease for the proband. If there are multiple probands per family, Whittemore and Halpern (2003) proposed conditioning on the disease indicators of all probands and required that at least one pair of relatives be discordant with respect to disease. Neuhaus and Jewell (1990) assumed that the sampling mechanism is based on the number of affected relatives and adjusted for ascertainment by conditioning on the event that the family was sampled. The probability of ascertainment is calculated from the assumed model; it is simply the probability of a certain number of affected relatives.

In this paper, we take the former approach and express the familial association of

disease explicitly in a full multivariate model. We do this because the interpretation of the measures of association as odds ratios, as obtained from our particular models, is appealing for its simplicity and familiarity. This interpretation is not afforded by the random effects model. Also, regression modeling of the familial association is more straightforward when based on the fully specified joint model than on a random effects model. In the former, the association parameters are the canonical parameters of the model, leading naturally to regression modeling through the introduction of covariates. In the latter, regression modeling is less direct and is implemented through careful specification of the covariance structure. As in Tosteson et al. (1991), we treat ascertainment status as random and jointly model the ascertainment and disease outcomes of a family. We avoid the overly restrictive assumptions of Tosteson et al. (1991) by directly modeling the association between ascertainment and disease at the family- and individual-level. In addition, we appropriately condition on the ascertainment event that brought the family into the study.

In Section 2 we review the Tosteson et al. (1991) approach and present the multivariate model for a family's disease and ascertainment outcomes considered in this paper. In Section 3 we apply the proposed method of analysis to three commonly used study designs. In Section 4, we apply our approach to a large familial aggregation study of cancer. In Section 5 we present simulation results, and in Section 6 we conclude.

2. Joint Modeling of Disease and Ascertainment

Let y_i indicate the disease status of the ith individual in a given family (i.e., $y_i = 1$ if i has disease, 0 otherwise $i = 1, \ldots, n$), and a_i the ascertainment status (i.e., $a_i = 1$ if i

is ascertained and 0 otherwise). Several members of a single family can be ascertained. Under the assumptions of Tosteson et al. (1991) given in the Introduction, the likelihood contribution of each family is conditioned only on the disease indicators (i.e., the y_i's) of all ascertained members. Thus, for their approach, specification of the joint distribution of the disease outcomes within a family is required; that of disease and ascertainment is not required.

The assumptions made by Tosteson et al. (1991) are often not realistic. Motivated by this concern, we propose to jointly model disease and ascertainment. An advantage of this joint modeling is the straightforward adjustment for complex ascertainment events without reliance on unrealistic assumptions. Further, this joint modeling approach also allows us to introduce various types of heterogeneity through the introduction of covariates, such as ethnicity and pedigree relationship.

Any multivariate binary model can be used for the joint distribution of disease and ascertainment. Here we consider the quadratic exponential model (QEM) (Zhao and Prentice, 1990). The QEM has been used extensively in the analysis of familial aggregation (e.g., Betensky and Whittemore, 1996, Hudson et al., 2001, Hudson, Laird, Betensky, Keck and Pope, 2001, Laird and Cuenco, 2003, Rabbee and Betensky, 2004 and Matthews, Finkelstein and Betensky, 2005). It is a multivariate log-linear model with all three-way and higher-order associations set to zero. Zhao and Prentice (1990) developed the univariate model and Betensky and Whittemore (1996) extended it for two outcomes per individual. Hudson et al. (2001) derived the corresponding logistic regression equations for the multivariate case and Rabbee and Betensky (2004) 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; individuals are frequently interested in their risk of disease given their family history. Third, it models associations of outcomes within families and within individuals. Modification of these relationships is straightforward through the introduction of covariates, such as pedigree relationship.

The QEM for two binary outcomes (y_i and a_i) for a family of size n is

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

The parameters of primary interest in assessing familial aggregation are the γ_{yij}'s; they capture the increase in disease odds associated with having an affected relative. The log-odds of disease is captured by θ_{yi}. Other parameters capture the clustering and interaction terms involving ascertainment. 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. Exchangeability among family members 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 $i \neq j$. This assumption can be 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 specific to pair (i,j), such as genetic distance. To simplify our presentation, we assume exchangeability throughout

this paper.

The QEM admits a set of logistic regression equations, and thus standard statistical software may be used for estimation (Hudson et al., 2001). These regression equations are

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

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 (1986) is used to adjust the variance of the resulting parameter estimates for the correlation among relatives.

The canonical parameter, γ_y, is the log-odds ratio conditional on family history and ascertainment. However, interest is more likely to reside instead in the marginal association of disease among family members from the general population, without regard for ascertainment. Any such measure of association can be calculated from the fully specified joint probability model. One example is the unconditional pairwise odds ratio of disease (e^{δ_M}), where

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

for all $i \neq j$ and $\boldsymbol{y}_{-ij}$ is the vector of disease statuses of all family members excluding the

ith and jth individuals. The probabilities follow from (1), i.e.,

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

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.

Inference based on these transformed measures of association requires computation of the Jacobian associated with each transformation in (3). 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 $\boldsymbol{\phi}$ denote the vector of the original parameters $(\theta_y, \theta_a, \theta_{ya}, \underline{\gamma_y}, \gamma_a, \gamma_{ya})'$, $\boldsymbol{\phi}'$ 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 by $\boldsymbol{J} = \left(\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\phi}'}\right) = \left(\frac{\partial \boldsymbol{\phi}'}{\partial \boldsymbol{\phi}}\right)^{-1}$. Only the fourth element of the parameter vector is transformed. Thus, the Jacobian is an identity matrix except for the fourth row, which is the derivative of δ_M with respect to $\boldsymbol{\phi}$. Since the QEM is a member of the exponential family, the fourth row of the Jacobian is

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

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

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

where $\boldsymbol{\mathcal{I}}$ 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 confidence intervals for the various measures of association that are marginal with respect to ascertainment (e.g., δ_M).

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 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 have a *minimum* number of affected ascertained individuals, and control families 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 the event that there is at least one proband in the family (i.e., $\sum a_i \geq 1$). As an example, consider a study in which individuals are recruited from a high-risk clinic. A family is included in the study as long as at least one of its members was recruited from the clinic. The Tosteson et al. (1991) assumption of independence of ascertainment and family history of disease implies that an individual enters a high-risk clinic independent of the disease history of his/her relatives. This is suspect in the context of many diseases. For the proposed approach, under this design, a family's contribution to the likelihood is

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

The logistic regression equations in (2) can be used to obtain parameter estimates; however, the expected information matrix must account for the conditional likelihood (Appendix A).

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 sporadic disease. Case families potentially carry

the hereditary form of the disease and contain at least a *minimum* number of affected ascertained individuals. Control families with potentially sporadic disease contain at most a *maximum* number of affected ascertained individuals. One example of this study design is the recruitment of individuals from high-risk clinics, with ultimate recruitment of the family depending on the numbers of recruited individuals with and without disease. In this setting, the Tosteson et al. (1991) assumption of independence of ascertainment and family disease history is again suspect. It implies that a subject enters a high-risk clinic regardless of his/her family's disease history.

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

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

and for a control family is

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

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. The logistic regression equations in (2) can be used to obtain parameter estimates but the variances of these estimates must account for the conditional likelihoods for each type of family in (7) and (8). Derivation of the expected information matrix is similar to that given in Appendix A for the proband sampling study 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

$$\mathcal{I} = N_0 \, \mathrm{Cov}_{\boldsymbol{\phi}} \left[\boldsymbol{T} \,\middle|\, \sum y_i a_i \leq c_0, \sum a_i \geq 1 \right] + N_1 \, \mathrm{Cov}_{\boldsymbol{\phi}} \left[\boldsymbol{T} \,\middle|\, \sum y_i a_i \geq c_1 \right]. \qquad (9)$$

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. One example of this study design involves recruitment of individuals from a high-risk clinic, with the ultimate recruitment of the family depending on there being at least two affected family members. This design differs from the case-control family sampling design in that it does not require the affected family members to be among those who are ascertained. Again, in high-risk family sampling, the Tosteson et al. (1991) assumption of independence of ascertainment and family history of disease is unlikely to hold given typical ascertainment through a high-risk clinic. Appendix B shows that for this study design, under the assumptions of Tosteson et al. (1991), ascertainment drops out of the likelihood and can be ignored. 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 is given by

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

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

$$\mathcal{I} = N \, \mathrm{Cov}_{\boldsymbol{\phi}} \left(\boldsymbol{T} \,\middle|\, \sum a_i \geq 1, \sum y_i \geq c \right).$$

4. Example

We now compare the four approaches to accounting for ascertainment described in the Introduction under each of the three different sampling designs described in Section 3. Each of these approaches requires specification of the joint distribution of disease or of disease and ascertainment among family members. As described previously, we elect to use the univariate and bivariate QEM, respectively. The first method is the naive approach, which completely ignores ascertainment in the analysis. The second approach is the first proband approach, which conditions the joint likelihood of disease outcomes on the disease status of the first individual recruited to the study. The third approach is that of Tosteson et al. (1991), which conditions the univariate likelihood of disease outcomes 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 individual ascertainment.

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 population-based families from the CGN registry to obtain "pseudo-studies" that conform to these designs. We defined the ascertainment event as a cancer diagnosis

before age 65. We included sibships of size four only, as a crude form of age-matching.

The fictitious proband sampling study includes a total of 406 sibships with at least one ascertained individual. We are interested in investigating whether skin cancer clusters in these families. 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 of skin cancer (see (4)). The estimates and standard errors are listed in Table 2. All approaches, except Tosteson et al. (1991), find statistically significant familial aggregation of skin cancer. The standard error of the Tosteson et al. (1991) 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 to detect familial aggregation and to a decrease in magnitude of the estimate.

[Table 1 about here.]

[Table 2 about here.]

In the fictitious case-control family study, case families are required to have at least one affected ascertained individual (122 families) and control families must contain only unaffected ascertained individuals (284 families). The results of the proposed method are listed in Table 2 (the results of the other methods are the same as for the proband sampling design). The proposed approach finds significant aggregation of skin cancer, although the estimate is smaller than that from the proband sampling design. This is due to the fact that under this design, more cases of disease are attributed to the ascertainment scheme.

In the high-risk family study design, families were included if they contained at least one affected member and at least one ascertained individual. In total, there were 133 sibships of size four that were analyzed. The distribution of skin cancer in these sibships is given in the lower half of Table 1. Table 2 lists the results from all four analytic approaches. The naive and first proband approaches give negative, though nonsignificant, results. This anomaly arises from their improper treatment of ascertainment; they do not account for the required affected family member. Because most families (82%) in the study have only one affected member, the sampling design induces a negative disease association without proper adjustment. The Tosteson et al. (1991) approach yields a small and nonsignificant log-odds ratio. As with the case-control family sampling design, the standard error of the Tosteson et al. (1991) estimate is large relative to magnitude of the estimate. The estimate of the pairwise log-odds ratio of skin cancer based on the proposed approach is statistically significant, agreeing with the results from the other two studies. We note that this can be viewed as evidence of either an environmental or genetic cause, 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 approach yields estimates that are larger in magnitude and have smaller standard errors than the Tosteson et al. (1991) approach. This suggests that despite the fact that it involves estimation of more parameters than the Tosteson et al. (1991) 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.

5. Simulation Studies

We conducted several simulation studies to compare the proposed approach with the naive, first proband and Tosteson et al. (1991) approaches for the three study designs considered in this paper. We considered two different parameter configurations for each study design. The first configuration contains a moderate association between family history of disease and ascertainment, and the second contains a strong association. The Tosteson et al. (1991) assumption of independence is violated under each configuration. We report simulation results for families of size four; results for families of size three are similar. For each study design, 300 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 (4). Results are listed in Table 3.

[Table 3 here.]

For the proband sampling design, the naive, first proband, and Tosteson et al. (1991) 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 estimates are 0.19, when they should be 0.91. In the case of strong disease-ascertainment association, the estimates are 0.40–0.48 when the true value is 1.25. 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 150 case families from (7) and 150 control families from (8). The results are similar to those from the proband design, but even more extreme. In particular, the naive and first proband approaches yield a *negative* familial association. The Tosteson et al. (1991) estimate is positive, though quite biased and nonsignificant.

Results are similar for the high-risk family sampling study design. Under strong familial association, the the Tosteson et al. (1991) estimate is negative (-0.60); 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 tere are some discrepancies. The discrepancies 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 approach 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. In particular,

we set $P(a_i = 1|y_i = 1, \sum y_j \leq 2) = 0.4$, $P(a_i = 1|y_i = 0, \sum y_j \leq 2) = 0.1$, $P(a_i = 1|y_i = 1, \sum y_j > 2) = 0.8$, and $P(a_i = 1|y_i = 0, \sum y_j > 2) = 0.5$.

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 150 or 100 families. Results are given in Table 4. For all three study designs, and for sample sizes of 150 and 100 families, the power of the proposed approach far exceeds that of the other approaches. Interestingly, the power of the Tosteson et al. (1991) approach is exceedingly low, likely due to the violation of its assumptions by the probability model from which we simulated. In addition, it decresaes as the complexity of the ascertainment event increases.

[Table 4 here]

6. Discussion

The simulation studies performed in this paper confirm that if ascertainment is related to disease, then ascertainment must be fully adjusted for in any analysis in order to avoid bias. Partial adjustment, as afforded by the Tosteson et al. (1991) 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 (that is, having an affected relative decreases the risk of disease). In other simulations (not reported here), the proposed approach is comparable in performance to the Tosteson et al. (1991) approach when the Tosteson et al. (1991) assumptions are

valid. Since we condition on less information, in some cases the proposed approach is even more precise. In addition, the proposed approach appears to perform well under one example 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 the Tosteson et al. (1991) assumptions.

We assumed the QEM for the joint model of disease and ascertainment that we used in our analyses. This model has the drawback of being *irreproducible*; that is, if the model holds for a family of size n, then it necessarily does not hold for families of a different size. Cox and Wermuth (1994) and Betensky and Whittemore (1996) identified circumstances under which approximate reproducibility holds, and Matthews et al. (2005) proposed a method of analysis to allow for varying family sizes in the univariate QEM. This method could be applied to the proposed joint modeling approach for disease and ascertainment, as well. Alternatively, any joint model for disease and ascertainment could be used as the basis for the proposed approach.

Acknowledgements

This research was supported in part by the Cancer Genetics Network (CGN) under NCI contract U01 CA78284-04 and NIH grants R01 CA 74302 and R01 CA 75971. 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.

Appendix A

Calculation of the expected information matrix for the proband sampling study design

Calculation of the expected information matrix follows that of an exponential family, except that the proposed likelihood is conditional. Letting k index families, and assuming that the joint distribution of a family of size n is given by (1), the log-likelihood of all N families in the observed data is

$$\begin{aligned}\ell_N &= \sum_{k=1}^{N}\Bigg[\theta_y\sum_i y_{ki} + \theta_a\sum_i a_{ki} + \theta_{ya}\sum_i y_{ki}a_{ki} \\ &\qquad + \gamma_y\sum_{i<j} y_{ki}y_{kj} + \gamma_a\sum_{i<j} a_{ki}a_{kj} + \gamma_{ya}\sum_{i\neq j} y_{ki}a_{kj}\Bigg] \\ &\quad -N\log\Bigg[\sum_{\boldsymbol{Y},\boldsymbol{A}^{(1)}}\exp\Bigg(\theta_y\sum_i y_i + \theta_a\sum_i a_i + \theta_{ya}\sum_i y_i a_i \\ &\qquad +\gamma_y\sum_{i<j} y_i y_j + \gamma_a\sum_{i<j} a_i a_j + \gamma_{ya}\sum_{i\neq j} y_i a_j\Bigg)\Bigg],\end{aligned}$$

where $\boldsymbol{Y}$ denotes all possible values of $\boldsymbol{y}$, and $\boldsymbol{A}^{(1)}$ denotes all possible values of $\boldsymbol{a}$ where $\sum a_i \geq 1$. The score equations, obtained by differentiating the above log-likelihood are

$$\frac{\partial \ell_N}{\partial \boldsymbol{\phi}} = \sum_{k=1}^{N}\boldsymbol{T}_k - N\,\mathrm{E}_{\boldsymbol{\phi}}\left[\boldsymbol{T}\,\middle|\,\sum a_i \geq 1\right].$$

Further differentiation is performed to obtain the expected information matrix, $\mathcal{I}$, which is

$$\mathcal{I} = N\,\mathrm{Cov}_{\boldsymbol{\phi}}\left[\boldsymbol{T}\,\middle|\,\sum a_i \geq 1\right]. \tag{10}$$

Appendix B

Extension of the Tosteson et al. (1991) approach to a high-risk family sampling study design

To adjust the Tosteson et al. (1991) 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

$$\mathrm{P}\left(y_{r+1},\ldots,y_n \,\middle|\, y_1,\ldots,y_r,a_1,\ldots,a_n,\sum y_i \geq c\right), \tag{11}$$

where r is the number of probands in the family.

The likelihood in (11) can be shown to equal

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

The Tosteson et al. (1991) assumption 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

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

Thus, the likelihood becomes

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

since the terms involving only ascertained individuals cancel.

The second set of Tosteson et al. (1991) assumptions are: (i) a large source population (i.e., $\tau_1, \tau_2 \to 0$), or (ii) independence between ascertainment and disease within an individual (i.e., $\tau_1 = \tau_2$). If either holds,

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

Thus, the assumptions of Tosteson et al. (1991) imply that the ascertainment indicators in (11) can be ignored when computing the likelihood contribution of a family under this study design.

References

Betensky, R. and Whittemore, A. (1996). An analysis of correlated multivariate binary data: Application to familial cancers of the ovary and breast. *Appl Statist* **45**, 411–429.

Bonney, G. (1998). Ascertainment corrections based on smaller family units. *American Journal of Human Genetics* **63**, 1202–1215.

Commenges, D., Jacqmin, H., Letenneur, L. and van Duijn, C. (1995). Score test for

familial aggregation in proband studies: Application to Alzheimer's disease. *Biometrics* **51**, 542–551.

Cox, D. and Wermuth, N. (1994). A note on the quadratic exponential binary distribution. *Biometrika* **81**, 403–408.

Epstein, M., Lin, X. and Boehnke, M. (2002). Ascertainment-adjusted parameter estimates revisited. *American Journal of Human Genetics* **70**, 886–895.

Glidden, D. and Liang, K.-Y. (2002). Ascertainment adjustment in complex diseases. *Genetic Epidemiology* **23**, 201–208.

Howing-Duistermaat, J., van Houwelingen, H. and de Winter, J. (2000). Estimation of individual genetic effects from binary observations on relatives applied to a family history of respiratory illnesses and chronic lung disease of newborns. *Biometrics* **56**, 808–814.

Hudson, J., Laird, N. and Betensky, R. (2001). Multivariate logistic regression for familial aggregation of two disorders: I. Development of models and methods. *American Journal of Epidemiology* **53**, 500–505.

Hudson, J., Laird, N., Betensky, R., Keck, P. J. and Pope, H. J. (2001). Multivariate logistic regression for familial aggregation of two disorders: II. Analysis of studies of eating and mood disorders. *American Journal of Epidemiology* **53**, 506–514.

Laird, N. and Cuenco, K. (2003). Regression methods for assessing familial aggregation of disease. *Statistics in Medicine* **22**, 1447–1455.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Matthews, A., Finkelstein, D. and Betensky, R. (2005). Analysis of familial aggregation

in the presence of varying family sizes. To appear, *Journal of the Royal Statistical Society, Series C.*

Neuhaus, J. and Jewell, N. (1990). The effect of retrospective sampling on binary regression models for clusteres data. *Biometrics* **46**, 977–990.

Rabbee, N. and Betensky, R. (2004). Power calculations for familial aggregation studies. *Genetic Epidemiology* **26**, 316–327.

Stiratelli, R., Laird, N. and Ware, J. (1984). Random-effects models for serial observations with binary response. *Biometrics* **40**, 961–971.

Thompson, E. (1993). Sampling and ascertainment in genetic epidemiology; a tutorial review. Unpublished manuscript.

Tosteson, T., Rosner, B. and Redline, S. (1991). Logistic regression for clustered binary data in proband studies with application to familial aggregation of sleep disorders. *Biometrics* **47**, 1257–1265.

Whittemore, A. and Halpern, J. (2003). Logistic regression of family data from retrospective study designs. *Genetics epidemiology* **25**, 177–189.

Zhao, L. and Prentice, R. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika* **77**, 642–648.

Table 1: Distribution of Skin Cancer and Ascertainment in CGN. Only families of size 4 are considered.

		# Affected Non-Probands					# Affected Probands					
Study Design	# Probands*	0	1	2	3	4	0	1	2	3	4	Total # Families
I., II. Proband[1] and Case-control family sampling[2]	1	247	13	0	0	0	204	56	0	0	0	260
	2	118	3	1	0	0	69	41	12	0	0	122
	3	21	1	0	0	0	10	9	2	1	0	22
	4	2	0	0	0	0	1	0	0	0	1	2
III. High-risk family sampling[3]	1	52	13	0	0	0	9	56	0	0	0	65
	2	51	3	1	0	0	2	41	12	0	0	55
	3	11	1	0	0	0	0	9	2	1	0	12
	4	1	0	0	0	0	0	0	0	0	1	1

* Probands have a cancer diagnosis before age 65.
[1] Families have at least one proband.
[2] Case families have at least *one* proband with skin cancer; control families have *no* probands with skin cancer.
[3] Families have at least one proband, and at least one member with skin cancer.

Table 2: Analysis of Skin Cancer in the CGN where Probands have a Cancer Diagnosis Before Age 65

Study Design	# Families	Approach	$\delta_M^{\dagger}$ (se)
I. Proband sampling[1]	406	Naive	**1.91** (0.08)
		First Proband	**2.00** (0.13)
		Tosteson	1.64 (0.97)
		Proposed	**2.67** (0.33)
II. Case-control family sampling[2]	406	Proposed	**2.12** (0.35)
III. High-risk family sampling[3]	133	Naive	-1.35 (1.34)
		First Proband	-1.50 (0.82)
		Tosteson	0.09 (3.12)
		Proposed	**1.68** (0.75)

† δ_M is the pairwise log-odds ratio of skin cancer.

[1] All families have at least one proband.

[2] Case families have *one* proband with skin cancer; control families have *no* probands with skin cancer.

[3] All families have at least one proband and at least one member with skin cancer.

Table 3: Comparison of Ascertainment Adjustment Approaches with Respect to Estimation of the Pairwise Log Odds Ratio (δ_M). Each simulated dataset consists of 300 families of size 4.

	Estimate (se†)			
Study Design	Naive	FP	Tosteson	Proposed
I. Proband sampling[1]	0.19 (0.27; 0.24)	0.19 (0.27; 0.25)	0.19 (0.30; 0.28)	0.88 (0.26; 0.26)
I. Proband sampling[2]	0.46 (0.28; 0.21)	0.48 (0.29; 0.19)	0.40 (0.35; 0.25)	1.22 (0.26; 0.32)
II. Case-control family sampling[3]	-0.51 (0.18; 0.26)	-0.44 (0.20; 0.29)	0.18 (0.28; 0.37)	0.88 (0.26; 0.27)
II. Case-control family sampling[4]	-0.30 (0.17; 0.22)	-0.17 (0.19; 0.24)	0.33 (0.23; 0.30)	1.09 (0.23; 0.29)
III. High-risk family sampling[5]	-0.60 (0.09; 0.34)	-0.63 (0.10; 0.37)	-0.60 (0.09; 0.34)	1.24 (0.21; 0.17)
III. High-risk family sampling[6]	-0.61 (0.10; 0.34)	-0.75 (0.10; 0.49)	-0.60 (0.10; 0.34)	1.61 (0.21; 0.24)

† First value is Monte Carlo standard error; second is the square root of the average analytic variance.

[1] $\theta_y = -2.5, \theta_a = -1.0, \theta_{ya} = 0.2, \gamma_y = 0.2, \gamma_a = 0.1, \gamma_{ya} = 0.1, \delta_M = 0.91$

[2] $\theta_y = -3.5, \theta_a = -1.0, \theta_{ya} = 0.2, \gamma_y = 0.2, \gamma_a = 0.1, \gamma_{ya} = 0.7, \delta_M = 1.25$

[3] $\theta_y = -3.0, \theta_a = -1.0, \theta_{ya} = 0.2, \gamma_y = 0.2, \gamma_a = 0.1, \gamma_{ya} = 0.25, \delta_M = 0.95$

[4] $\theta_y = -3.5, \theta_a = -2.0, \theta_{ya} = 0.2, \gamma_y = 0.2, \gamma_a = 0.1, \gamma_{ya} = 0.75, \delta_M = 1.18$

[5] $\theta_y = -2.0, \theta_a = -1.0, \theta_{ya} = 0.4, \gamma_y = 0.5, \gamma_a = 0.1, \gamma_{ya} = 0.1, \delta_M = 1.27$

[6] $\theta_y = -3.0, \theta_a = -2.0, \theta_{ya} = 0.4, \gamma_y = 0.5, \gamma_a = 0.1, \gamma_{ya} = 0.75, \delta_M = 1.63$

Table 4: Comparison of Power of $\delta_M^{\dagger}$ under Model Misspecification

Design	N	Method	Power(%)
I. Proband sampling[1]	150	Naive	33.8
		First Proband	35.2
		Tosteson	16.2
		Proposed	83.2
	100	Naive	27.6
		First Proband	27.5
		Tosteson	12.4
		Proposed	69.7
II. Case-control family sampling[2]	75 cases 75 controls	Naive	20.8
		First Proband	21.6
		Tosteson	1.4
		Proposed	94.2
	50 cases 50 controls	Naive	15.6
		First Proband	16.2
		Tosteson	3.0
		Proposed	65.2
III. High-risk family sampling[3]	150	Naive	0.0
		First Proband	0.0
		Tosteson	0.0
		Proposed	95.0
	100	Naive	0.0
		First Proband	0.0
		Tosteson	0.0
		Proposed	84.0

† δ_M is the pariwise log-odds ratio of disease.

†† Monte Carlo estimate of the standard error.

[1] $\delta_M = 0.33$.

[2] $\delta_M = 0.27$.

[3] $\delta_M = 0.74$.