Collection of Biostatistics Research Archive
COBRA Preprint Series

Year 2007 *Paper* 31

Estimating the Prevalence of Disease Using Relatives of Case and Control Probands

Kristin N. Javaras* Nan M. Laird†
James I. Hudson‡ Brian D. Ripley**

*University of Wisconsin - Madison, javaras@wisc.edu
†Harvard School of Public Health - Biostatistics
‡McLean Hospital / Harvard Medical School
**Oxford University - Statistics

Estimating the Prevalence of Disease Using Relatives of Case and Control Probands

Kristin N. Javaras, Nan M. Laird, James I. Hudson, and Brian D. Ripley

Abstract

We introduce a method for estimating the prevalence of disease using data from a case-control family study performed to investigate the aggregation of disease in families. The families are sampled via case and control probands, and the resulting data consist of information on disease status and covariates for the probands and their relatives. We introduce estimators for overall prevalence and for covariate stratum-specific prevalence (e.g., sex-specific prevalence) that yield approximately unbiased estimates of their population counterparts. We also introduce corresponding confidence intervals that have good coverage properties even for small prevalences. The estimators and intervals address the over-representation of diseased individuals in case-control family data by using only the relatives (of the probands) and by taking into account whether each relative was selected via a case or a control proband. Finally, we describe a simulation experiment in which the estimators and intervals were applied to case-control family datasets sampled from a fictional population that resembled the catchment area for an Austrian family study of major depressive disorder. The resulting estimates varied closely and symmetrically around their population counterparts, and the resulting intervals had good coverage properties.

1. INTRODUCTION

The gold standard approach to estimating prevalence involves first obtaining a cross-sectional (or prevalence) sample from the population of interest, then assessing whether the disease is present in the sampled individuals, and finally calculating the proportion of sampled individuals with the disease, sometimes with individuals weighted to reflect the probability that they were sampled and responded. Often, researchers do not have access to an existing cross-sectional sample that is relevant to both the population and the disease of interest, and the cost of collecting one would be prohibitive. However, if they do have access to a case-control family sample from the population of interest that was originally collected to investigate familial aggregation of the disease, it can be used to obtain valid estimates of prevalence, as we show below.

Case-control family studies are conducted to investigate the extent to which a disease aggregates (with itself) within families, or co-aggregates with other diseases within families (Hudson et al., 2001). In these studies, researchers select case probands who are affected by the disease and control probands who are not, and then select relatives from among the case and control probands' family members (e.g., first-degree relatives). The resulting data consist of information on disease status and covariates for the case and control probands and their relatives. When the data is used to investigate familial aggregation, the most basic analysis entails comparing the proportion of affected relatives for case probands to the proportion of affected relatives for control probands. Here, we refer to an example that is a case-control family study of major depressive disorder (MDD) conducted at Innsbruck University Clinics in Innsbruck, Austria (Hudson et al., 2003). In the study, 64 adults with MDD (case probands) were selected from the psychiatric unit, and 58 adults without MDD (control probands) were selected from the surgical and ophthalmology units. Three hundred and thirty of the probands' adult first-degree relatives (parents, siblings, children) consented to participate in the study. Table 1 presents the numbers of relatives with and without MDD, by proband disease status and relative sex.

The probands provide no information on prevalence because the proportion of affected (or

case) probands is fixed by design. The relatives, on the other hand, do provide information on prevalence, but the simple proportion of affected relatives is a biased estimate of prevalence if the disease aggregates in families because, in that case, the relatives' probability of selection depends on their disease status, albeit indirectly (through the probands' disease statuses). However, by using only the relatives and conditioning on the disease status of the probands through which the relatives were selected, we can obtain valid estimates and confidence intervals for overall and stratum-specific (e.g., sex-specific) prevalence, provided that certain commonly-made assumptions about sampling and the population structure hold. Our method yields estimates that are biased only slightly downwards for their population counterparts. Further, they are less seriously biased than estimates from other commonly-used methods of estimating prevalence from case-control family data, such as the proband or propositus method (Kendler and Eaton, 1988; Strömgren, 1948). Our method performs very well when applied to datasets sampled from a fictional population: the resulting estimates vary closely and symmetrically around their population counterparts, with only a very small downwards bias, and the resulting intervals have good coverage properties.

The paper is organized as follows. Section 2 introduces our estimators for overall prevalence and stratum-specific prevalence, as well as the assumptions on which they rely. Section 3 presents the results of the simulation experiment, and Section 4 is a discussion of the advantages and limitations of the method. Appendices A and B contain proofs that the overall and stratum-specific estimators, respectively, are approximately unbiased for their population counterparts. Finally, Appendix C introduces standard errors and confidence intervals for overall and stratum-specific prevalence.

2. ESTIMATION

Before presenting estimators for overall and stratum-specific prevalence, it is necessary to introduce some notation, as well as several assumptions. These assumptions are commonly, if

implicitly, made when analyzing data from case-control family studies; here, they are used to guarantee that the proposed estimators will be approximately unbiased. The assumptions describe a simplified model for the underlying population and for the ascertainment of case-control families from it. Although not a perfect representation of reality, this simplified model is an adequate approximation to reality when the size of the population is sufficiently large (relative to the sizes of the families that comprise the population and relative to the number of probands ascertained in the study). Further, the results of the simulation experiment in Section 3 suggest that our method is robust to violations of the assumptions underlying the simplified model.

We will assume that the population of interest is finite (but very large) and that it can be partitioned into F mutually exclusive and exhaustive families of siblings. These families are indexed by i. Family i has N_i members, who are indexed by ij, where $j = 1, \ldots, N_i$. For individual ij, we use Y_{ij} to denote disease status, with 1 corresponding to presence of the disease and 0 corresponding to absence of the disease. The population prevalence, π, is defined as $f(Y_{ij} = 1)$, where individual ij is randomly selected from the population. Similarly, the stratum-specific prevalence, π^x, is defined as $f(Y_{ij} = 1|X_{ij} = x)$, where X_{ij} is a categorical variable whose levels define covariate strata of interest (e.g., males and females); x is a particular value of X_{ij} (e.g., the female stratum); and individual ij is randomly selected from the population in stratum x. Note that X_{ij} may result from coarsening the values of a continuous variable (e.g., age) or from crossing the levels of multiple categorical variables (e.g., sex and race).

Families are ascertained for the case-control family study via F_A unrelated probands with the disease and F_U unrelated probands without the disease. Once families have been ascertained, they are renumbered, as are their members. The re-numbered families are now indexed by $\underset{\sim}{i}$, where, for the sake of convenience, the values $\underset{\sim}{i} = 1, \ldots, F_A$ refer to families ascertained via case probands, the values $\underset{\sim}{i} = F_A + 1, \ldots, F_A + F_U$ refer to families ascertained via control probands, and the values $\underset{\sim}{i} = F_A + F_U, \ldots, F$ refer to unascertained families. For ascertained

family $\underset{\sim}{i}$, disease status and covariate information is obtained for only $n_{\underset{\sim}{i}} - 1$ of the $N_{\underset{\sim}{i}} - 1$ remaining (i.e., non-proband) family members. The re-numbered members of ascertained family $\underset{\sim}{i}$ are now indexed by $\underset{\sim}{ij}$, where $\underset{\sim}{j} = 1$ refers to the proband, $\underset{\sim}{j} = 2, \ldots, n_{\underset{\sim}{i}}$ refer to the sampled relatives, and $\underset{\sim}{j} = n_{\underset{\sim}{i}} + 1, \ldots, n_{\underset{\sim}{i}} + N_{\underset{\sim}{i}}$ refer to the unsampled relatives. The original index j, which refers to an individual as a member of a family in the population, has a 1:1 mapping to the index $\underset{\sim}{j}$, which refers to the individual as a member of his or her family once it has been ascertained. We use $r_i(j)$ to refer to the renumbered index for the j^{th} member of the i^{th} family in the population once his or her family has been ascertained.

Below, we show how data from a case-control family study can be used to obtain estimates of overall prevalence and stratum-specific prevalence. Several more assumptions must hold for the proposed estimators to yield approximately unbiased estimates:

(i.) *Availability of Relatives*: Each member of the population of interest has at least one living relative.

(ii.) *Family Size and Disease Status are Uncorrelated*: $\text{Cor}(N_i, \frac{N_i^A}{N_i}) = 0$, where $N_i^A = \sum_{j=1}^{N_i} \text{I}(Y_{ij} = 1)$, the number of affected members in family i.

(iii.) *Proband Sampling*: The case probands are randomly sampled from the affected members of the population, and the control probands are randomly sampled from the unaffected members of the population.

(iv.) *Single Ascertainment*: The number of case (control) probands is sufficiently small relative to the number of affected (unaffected) members of the population to guarantee that no family will be selected via more than one proband.

(v.) *Relative Sampling*: Given that family i has been ascertained, the probability that individual $\underset{\sim}{ij}$ ($\underset{\sim}{j} \neq 1$) is included in the study is a constant (referred to as s) and, thus, does not depend on $Y_{\underset{\sim}{ij}}$ (his or her disease status), $X_{\underset{\sim}{ij}}$ (his or her covariates), $Y_{\underset{\sim}{i}(-\underset{\sim}{j})}$ (the disease statuses for the other members of the family), $X_{\underset{\sim}{i}(-\underset{\sim}{j})}$ (the covariates for the other

members of the family), or on $N_{\underset{\sim}{i}}$ (the family's size).

(vi.) *Disease Status is Independent of Other Family Members' Covariates*: For individual ij, Y_{ij} (his or her disease status) is independent of $X_{i(-j)}$ (the covariates for the other members of the family), conditional on X_{ij} (the individual's covariate)

If Assumptions (i.)-(v.) hold, then the following estimator is approximately unbiased at the first-order for the overall prevalence of disease in the population (see Appendix A for a proof):

$$\hat{\pi} = \frac{p_{U}}{1 - p_{A} + p_{U}}, \tag{2.1}$$

where p_{A} is the proportion of case probands' relatives who are affected,

$$p_{A} = \frac{\sum_{\underset{\sim}{i}=1}^{F_{A}} \sum_{\underset{\sim}{j}=2}^{n_{\underset{\sim}{i}}} \mathrm{I}(Y_{\underset{\sim}{i}\underset{\sim}{j}} = 1)}{\sum_{\underset{\sim}{i}=1}^{F_{A}} \sum_{\underset{\sim}{j}=2}^{n_{\underset{\sim}{i}}} 1};$$

and p_{U} is the proportion of control probands' relatives who are affected,

$$p_{U} = \frac{\sum_{\underset{\sim}{i}=F_{A}+1}^{F_{A}+F_{U}} \sum_{\underset{\sim}{j}=2}^{n_{\underset{\sim}{i}}} \mathrm{I}(Y_{\underset{\sim}{i}\underset{\sim}{j}} = 1)}{\sum_{\underset{\sim}{i}=F_{A}+1}^{F_{A}+F_{U}} \sum_{\underset{\sim}{i}=2}^{n_{\underset{\sim}{i}}} 1}.$$

Further, we can show that the slight bias introduced by the second-order terms is downward when $F_{A} \approx F_{U}$ (the number of case probands is approximately equal to the number of control probands) and when $E(1 - p_{U}) > E(p_{A})$ (the expected proportion of control probands' relatives who are unaffected is greater than the expected proportion of case probands' relatives who are affected).

Note that the estimator in (2.1) adjusts p_U, an estimate of prevalence based on relatives of control probands only, by the factor $\frac{1}{1-p_A+p_U}$. Since $E(p_A) > E(p_U)$ for diseases that aggregate in families, this adjustment will usually have the effect of moving the prevalence estimate upwards from p_U. Thus, using p_U to estimate overall prevalence—an approach that is referred to as the proband or propositus method and has been widely used in genetic-epidemiologic studies of psychiatric disorders (Kendler and Eaton, 1988; Strömgren, 1948)—results in greater downward bias than using the estimator in (2.1), except when the disease does not aggregate in families. Similar arguments reveal that p_A overestimates prevalence except when the disease does not aggregate in families.

Next, if Assumptions (i.)-(vi.) hold, then the following estimator is biased only slightly at the first-order for the prevalence of disease in stratum x (see Appendix B for a proof):

$$\hat{\pi}^x = p_A^x \hat{\pi} + p_U^x (1 - \hat{\pi}), \tag{2.2}$$

where p_A^x is the proportion of case probands' relatives who have covariate value x and are affected

$$p_A^x = \frac{\sum_{\underset{\sim}{i}=1}^{F_A} \sum_{j=2}^{n_{\underset{\sim}{i}}} \mathrm{I}(X_{\underset{\sim}{ij}} = x)\mathrm{I}(Y_{\underset{\sim}{ij}} = 1)}{\sum_{\underset{\sim}{i}=1}^{F_A} \sum_{\underset{\sim}{j}=2}^{n_{\underset{\sim}{i}}} \mathrm{I}(X_{\underset{\sim}{ij}} = x)};$$

and p_U^x is the proportion of control probands' relatives who have covariate value x and are affected

$$p_U^x = \frac{\sum_{\underset{\sim}{i}=F_A+1}^{F_A+F_U} \sum_{j=2}^{n_{\underset{\sim}{i}}} \mathrm{I}(X_{\underset{\sim}{ij}} = x)\mathrm{I}(Y_{\underset{\sim}{ij}} = 1)}{\sum_{\underset{\sim}{i}=F_A+1}^{F_A+F_U} \sum_{\underset{\sim}{j}=2}^{n_{\underset{\sim}{i}}} \mathrm{I}(X_{\underset{\sim}{ij}} = x)}.$$

Further, we can show that the slight first-order bias is downwards when, again, $F_A \approx F_U$ and $E(1 - p_U) > E(p_A)$. Note that, as above, an examination of Equation (2.2) reveals that using only the relatives of control probands to estimate stratum-specific prevalence results in more serious underestimation than using the estimator in (2.2), except when the disease does not aggregate in families.

In Appendix C, we provide approximate standard errors and confidence intervals for $\hat{\pi}$ and $\hat{\pi}^x$. The standard errors and confidence intervals are appropriate for dependent observations since disease status will be positively correlated within families when the disease aggregates in families. The confidence intervals are based on the same concept as the Agresti-Coull (1998) interval, which modifies the standard Wald interval for binomial proportions so that it will attain actual coverage levels near the nominal coverage level even for small proportions. The modification, which has strong roots in the work of Wilson (1927), involves replacing the maximum likelihood estimate of the proportion used to calculate the center and standard error of the Wald interval with an estimate that is smoothed towards the uniform probability distribution by adding a small number (e.g., two) of successes and the same number of failures to the observed data. Because the Agresti-Coull interval appears to perform well for small independent samples (1998) and, more relevantly for our data, medium-sized dependent samples (Miao and Gastwirth, 2004), we use a similar approach to form confidence intervals: the intervals' center and spread are calculated using $\widetilde{p}_A$, $\widetilde{p}_U$, $\widetilde{p}_A^x$, and $\widetilde{p}_U^x$, which smooth p_A, p_U, p_A^x, and p_U^x, respectively, towards the uniform distribution by adding two failures and two successes for every 100 observations.

To illustrate the use of our method, we apply it to the data from the Austrian case-control family study. Equations (2.1) and (C.3) yield an estimate of 8.8% and a 95% confidence interval of $[5.9\%, 15\%]$, respectively, for the overall lifetime prevalence of MDD in the Tyrol region. Equations (2.2) and (C.4) yield an estimate of 6.0% and a 95% confidence interval of $[2.3\%, 13\%]$ for male lifetime prevalence, and 11.3% and $[6.4\%, 20.0\%]$ for female lifetime

prevalence. Note that the overall, male, and female prevalence estimates are larger than the affected proportions of all relatives, male relatives, and female relatives of control probands (7.9%, 5.5%, and 10.1%, respectively), but considerably smaller than the affected proportions of all relatives, male relatives, and female relatives of case probands (18.5%, 11.0%, and 23.8%, respectively).

3. Results of the Simulation Experiment

We conducted a simulation experiment to investigate how well the estimators from Section 2 and the confidence intervals from Appendix C perform in practice. The experiment was designed to mimic the Austrian case-control family study of MDD, which is at the smaller end of case-control family studies.

We created a fictional population with approximately $500,000$ individuals, which corresponds to the number of people between 18 and 70 years old reported to be living in the Tyrol region of Austria, the catchment area for the Austrian study, in 2003 (Statistik Austria, 2003). To create a population of this size, we generated data for approximately $125,000$ families, which involved three steps: (a) generating family sizes based on the distribution of family sizes in the Austrian data; (b) generating the sexes of and relationships between (e.g., siblings, parents, etc.) family members based on the percentage of females between 18 and 70 years in the Tyrolean population in 2003 (=50.5%) and the distribution of family relationships and sex in the Austrian data, and; (c) generating lifetime disease statuses for the family members conditional on their sexes and relationships, based on parameter estimates from the Austrian data.

To generate the disease statuses in step (d), we used the ACE (A = additive genetic effects, C = common or shared family environment, and E = unique environment) model for case-control family data (Javaras et al., 2007). In this model, a subject is affected if his or her 'liability to the disease' exceeds a threshold that corresponds to disease prevalence for the relevant covariate

stratum. The liabilities for subjects from family i are modeled by an N_i-variate normal distribution with mean vector set to zero and correlations that are a function of a^2 (the percentage of variation in liability due to A) and c^2 (the percentage of variation in liability due to C). In our experiments, we set the ACE model's parameters to values based on analysis of the actual MDD data (Javaras et al., 2007, Section 6): we set a^2 to 0.45, c^2 to 0, lifetime disease prevalence among males to 6.0%, and lifetime disease prevalence among females to 11.3%. Note that the male and female prevalences, along with the proportion of females, determine the overall lifetime prevalence of disease ($= 8.8\%$) for the fictional population.

Next, we sampled $1,000$ small case-control family datasets from the fictional population. Each dataset was formed by selecting $F_A = 64$ case probands and $F_U = 58$ control probands, and then including all of the probands' family members ($s = 1$). (F_A, F_U, and s were set equal to their values in the Austrian study.) For each sampled dataset, Equation (2.1) was used to estimate overall prevalence, and Equation (2.2) was used to estimate the male and female prevalences. In addition, we used Equation (C.3) to form two-sided and lower and upper one-sided 95% confidence intervals for the overall prevalence, and we used Equation (C.4) to form the same confidence intervals for the male and female prevalences.

In the $1,000$ case-control family datasets sampled, the number of included individuals (relatives plus probands) ranged between 449 and 541. Even for this relatively small study size, the population was not sufficiently large to ensure single ascertainment: in 122 of the $1,000$ datasets, at least one family was doubly ascertained. In these instances, the first family member to be selected as a proband was retained as the sole proband for his or her family. Figure 1 presents boxplots of the resulting prevalence estimates for the $1,000$ datasets. The plots reveal that the prevalence estimates vary symmetrically and closely around the population prevalences, which are indicated by "X"s. The downward bias in the estimates is extremely small (especially relative to the length of the confidence intervals): the means of the $1,000$ estimates are within -0.0008 (-0.9%), -0.0006 (-0.5%), and -0.0011 (-1.9%) of the overall, male, and female

prevalences, respectively. The two-sided 95% confidence intervals for overall, male, and female prevalence have mean lengths 0.105, 0.146, and 0.114, respectively. Although these intervals are fairly wide, especially for such small estimates, this is to be expected due to the positive correlation of MDD status within families. The actual coverage probabilities of the intervals for overall, male, and female prevalence are 94.9%, 90.5%, and 98.8% for the two-sided 95% confidence intervals, 96.2%, 94.4%, and 97.2% for the one-sided lower 95% confidence intervals, and 94.7%, 93.1%, and 96.8% for the one-sided upper 95% confidence intervals. Note that the two-sided intervals, which will be used most often in practice, attain actual coverage levels very close to the nominal level of 95%. Note also that the actual coverage levels are generally a little too high for the upper one-sided intervals and generally a little too low for the lower one-sided intervals. This phenomenon reflects the fact that the intervals are calculated using estimated proportions that are smoothed away from zero.

The simulation experiment suggests that the prevalence estimators in (2.1) and (2.2) are approximately unbiased and reasonably efficient, even when the population size is relatively small and the assumption of single ascertainment does not hold. As would be expected, our estimators and confidence intervals perform even better in additional similar experiments (not described here) that involve a larger fictional population with over 2 million individuals.

4. Discussion

We have introduced a method of forming estimates and confidence intervals for overall and stratum-specific prevalence based on case-control family data.

It is clear from the simulation experiment (Section 3) and proofs (Appendices A and B) that the proposed estimators and intervals yield valid information about the prevalence of disease. The ability to glean valid information about disease prevalence from case-control family data is useful to medical researchers when no population-based data (from a cross-sectional sample)

are available for the population of interest. Knowledge of prevalence augments epidemiological understanding of the disease and also informs resource allocation. In addition, knowledge of prevalence makes it possible to estimate other parameters of epidemiological interest. For instance, data from a case-control sample can be weighted to create data representative of the population by using weights equal to the inverse sampling probabilities for the cases and controls, the calculation of which requires knowledge of prevalence. The weighted data that result can be used to obtain approximately unbiased estimates of population parameters, such as the exposure-disease risk difference and the exposure-disease risk ratio, that cannot be obtained from case-control studies unless the sampling fractions of cases and controls is known. (In contrast, the exposure-disease odds ratio can, of course, be obtained from case-control data without weighting them.)

Several limitations should be noted. For one, when the disease of interest aggregates in families, disease status will be positively correlated for individuals within the same family, which will have the effect of inflating the errors and intervals for $\hat{\pi}$ and $\hat{\pi}^x$. Thus, in this case, the prevalence estimators in Equations (2.1) and (2.2) will be less precise than corresponding estimators based on the same number of unrelated individuals from a cross-sectional sample. Further, the estimates and intervals would probably not perform well for very small proportions unless the sample size were very large, but this would also be true for estimates and intervals calculated from cross-sectional samples.

Second, the prevalence estimators may no longer be unbiased if one or more of the assumptions enumerated in Section 2 are violated. For example, if smaller families have a greater proportion of affected individuals, a violation of Assumption (ii.), then prevalence may be underestimated (Kendler and Eaton, 1988). This scenario is plausible for early-onset diseases that impair individuals' ability to have children or for diseases that result in early death. However, results from the simulation experiment suggest that violations of Assumption (ii.) introduce only a very small amount of bias, as shown in the right-hand columns of Tables 2 and 3 in

the appendices. As another example, if probands are selected based not only on their disease status but also on disease characteristics such as severity, a violation of Assumption (iii.), then the resulting prevalence estimates may be biased. This is a potential problem for the Austrian study because the case probands were sampled from a psychiatric clinic rather than from the community. In contrast, if probands are selected based not only on disease status but also on measured covariates such as sex or age, another violation of Assumption (iii.), estimates of stratum-specific prevalence can still be obtained by applying Equation (2.2) only to the relatives of those probands who belong to the stratum of interest. As for violations of Assumption (iv.), the simulation experiment in Section 3 suggests that our method is robust to at least moderate departures from single ascertainment. Next, if the affected relatives of the probands are less likely to participate in the study, a violation of Assumption (v.), then prevalence will be underestimated. Finally, if the disease of interest is extremely common or if it is somewhat common and aggregates extensively in families, then it may not be true that $E(1 - p_{_U}) > E(p_{_A})$. It is easy to see why this inequality will not hold if the disease in question is extremely common (prevalence over 50%), since in that case $E(p_{_A})$ will be large and $E(1 - p_{_U})$ will be small even if the disease does not aggregate in families. Another case where the inequality will not hold is when the disease aggregates in families to such an extent that $Ep_{_A}$ is large and when the disease is common enough so that $E(1 - p_{_U})$ is not large. However, for most diseases (including MDD), the inequality will hold. Further, since the assumption that $E(1 - p_{_U}) > E(p_{_A})$ is required only to ensure that the bias in $\hat{\pi}$ is downwards, our method will still be approximately unbiased even when this assumption is violated.

In general, though, our method appears to be reasonably robust to the violation of most assumptions. The most crucial assumption is likely to be the one about relative sampling, which assumes that individuals with the disease are no more or less likely to be included in the sample than individuals without the disease. This assumption would apply equally to cross-sectional samples. The second-most crucial assumption is likely to be the assumption that family size and disease status are uncorrelated in the population of interest. If these two crucial assumptions

hold, then our method of estimating disease prevalence from case-control family data is a useful tool, especially for diseases and populations where no cross-sectional samples are available.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Barbara Mangweth-Matzek (Department of Psychiatry, Innsbruck Medical University) for access to the data from the case-control family study of depression.

FUNDING

National Institute of Health Training Program in Psychiatric Epidemiology and Biostatistics (5 T32 MN17119-22 to K.J.).

APPENDIX A

Proof that $\hat{\pi}$ is approximately unbiased at the first-order for π

The overall population prevalence is defined as $\pi \equiv f(Y_{ij} = 1)$, where individual ij is randomly selected from the population of interest. Assumption (i.) about the availability of relatives allows us to expand $f(Y_{ij} = 1)$ as follows

$$\begin{aligned} \pi &\equiv f(Y_{ij} = 1) \\ &= f(Y_{ij} = 1|Y_{ij'} = 1)f(Y_{ij'} = 1) + f(Y_{ij} = 1|Y_{ij'} = 0)f(Y_{ij'} = 0), \end{aligned} \tag{A.1}$$

where individual ij' is randomly selected from among Y_{ij}'s relatives with disease status $Y_{ij'}$. (In the remainder of this proof, we will assume that $j' \neq j$.) We can rewrite the above equation as

$$\pi = f(Y_{ij} = 1|Y_{ij'} = 1)\pi + f(Y_{ij} = 1|Y_{ij'} = 0)(1 - \pi),$$

which can be rearranged to give

$$\pi = \frac{\pi_{U}}{1 - \pi_{A} + \pi_{U}}, \tag{A.2}$$

where $\pi_{U} \equiv f(Y_{ij} = 1|Y_{ij'} = 0)$ and $\pi_{A} \equiv f(Y_{ij} = 1|Y_{ij'} = 1)$. The parameters π_{A} and π_{U} can be defined in terms of the finite population:

$$\begin{aligned} \pi_{A} &\equiv f(Y_{ij} = 1|Y_{ij'} = 1) \\ &= \frac{\sum_{i=1}^{F}\sum_{j=1}^{N_i}\sum_{j'\neq j} \mathrm{I}(Y_{ij} = 1)\,\mathrm{I}(Y_{ij'} = 1)}{\sum_{i=1}^{F}\sum_{j=1}^{N_i}\sum_{j'\neq j} \mathrm{I}(Y_{ij'} = 1)} \\ &= \frac{\sum_{i=1}^{F}\sum_{j=1}^{N_i} \mathrm{I}(Y_{ij} = 1)(N_i^A - 1)}{\sum_{i=1}^{F}\sum_{j=1}^{N_i} \left(\mathrm{I}(Y_{ij} = 0)N_i^A + \mathrm{I}(Y_{ij} = 1)(N_i^A - 1)\right)} \\ &= \frac{\sum_{i=1}^{F}(N_i^A - 1)N_i^A}{\sum_{i=1}^{F}(N_i - 1)N_i^A}; \end{aligned} \tag{A.3}$$

where $N_i^A = \sum_{j=1}^{N_i} \mathrm{I}(Y_{ij} = 1)$, the number of affected members in family i, and

$$\begin{aligned}
\pi_{_U} &\equiv f(Y_{ij} = 1 | Y_{ij'} = 0) \\
&= \frac{\sum_{i=1}^{F} \sum_{j=1}^{N_i} \sum_{j' \neq j} \mathrm{I}(Y_{ij} = 1)\, \mathrm{I}(Y_{ij'} = 0)}{\sum_{i=1}^{F} \sum_{j=1}^{N_i} \sum_{j' \neq j} \mathrm{I}(Y_{ij'} = 0)} \\
&= \frac{\sum_{i=1}^{F} \sum_{j=1}^{N_i} \mathrm{I}(Y_{ij} = 1) N_i^U}{\sum_{i=1}^{F} \sum_{j=1}^{N_i} \left(\mathrm{I}(Y_{ij} = 0)(N_i^U - 1) + \mathrm{I}(Y_{ij} = 1) N_i^U \right)} \\
&= \frac{\sum_{i=1}^{F} N_i^A N_i^U}{\sum_{i=1}^{F} (N_i - 1) N_i^U}, \qquad \text{(A.4)}
\end{aligned}$$

where $N_i^U = \sum_{j=1}^{N_i} \mathrm{I}(Y_{ij} = 0)$, the number of unaffected members in family i.

Now, recall that

$$\hat{\pi} = \frac{p_{_U}}{1 - p_{_A} + p_{_U}}, \qquad \text{(A.5)}$$

where $p_{_A}$ is the proportion of case probands' relatives who are affected,

$$p_{_A} = \frac{\sum_{\underset{\sim}{i}=1}^{F_A} \sum_{\underset{\sim}{j}=2}^{n_{\underset{\sim}{i}}} \mathrm{I}(Y_{\underset{\sim}{i}\underset{\sim}{j}} = 1)}{\sum_{\underset{\sim}{i}=1}^{F_A} \sum_{\underset{\sim}{j}=2}^{n_{\underset{\sim}{i}}} 1};$$

and $p_{_U}$ is the proportion of control probands' relatives who are affected,

$$p_{_U} = \frac{\sum_{\underset{\sim}{i}=F_A+1}^{F_A+F_U} \sum_{\underset{\sim}{j}=2}^{n_{\underset{\sim}{i}}} \mathrm{I}(Y_{\underset{\sim}{i}\underset{\sim}{j}} = 1)}{\sum_{\underset{\sim}{i}=F_A+1}^{F_A+F_U} \sum_{\underset{\sim}{j}=2}^{n_{\underset{\sim}{i}}} 1}.$$

The estimator $\hat{\pi}$ in (A.5) can be approximated by a second-order Taylor expansion around

$\mathrm{E}p_U$ and $\mathrm{E}p_A$, the expected values of p_U and p_A, respectively:

$$\begin{aligned}\hat{\pi} \approx\ & \frac{\mathrm{E}p_U}{1-\mathrm{E}p_A+\mathrm{E}p_U} + (p_U-\mathrm{E}p_U)\left.\frac{\partial\hat{\pi}}{\partial p_U}\right|_{\mathrm{E}p_U,\mathrm{E}p_A} + (p_A-\mathrm{E}p_A)\left.\frac{\partial\hat{\pi}}{\partial p_A}\right|_{\mathrm{E}p_U,\mathrm{E}p_A} \\ & + \frac{1}{2}(p_U-\mathrm{E}p_U)^2\left.\frac{\partial^2\hat{\pi}}{\partial {p_U}^2}\right|_{\mathrm{E}p_U,\mathrm{E}p_A} + \frac{1}{2}(p_A-\mathrm{E}p_A)^2\left.\frac{\partial^2\hat{\pi}}{\partial {p_A}^2}\right|_{\mathrm{E}p_U,\mathrm{E}p_A} \\ & + (p_U-\mathrm{E}p_U)(p_A-\mathrm{E}p_A)\left.\frac{\partial^2\hat{\pi}}{\partial p_U \partial p_A}\right|_{\mathrm{E}p_U,\mathrm{E}p_A}\end{aligned}$$

Inserting expressions for the derivatives and then taking the expectation of both sides of the above equation yields

$$\begin{aligned}E\hat{\pi} \approx\ & \frac{\mathrm{E}p_U}{1-\mathrm{E}p_A+\mathrm{E}p_U} - \frac{\mathrm{Var}(p_U)(1-\mathrm{E}p_A)}{(1-\mathrm{E}p_A+\mathrm{E}p_U)^3} + \frac{\mathrm{Var}(p_A)\mathrm{E}p_U}{(1-\mathrm{E}p_A+\mathrm{E}p_U)^3} \\ & + \frac{\mathrm{Cov}(p_U,p_A)(1-\mathrm{E}p_A-\mathrm{E}p_U)}{(1-\mathrm{E}p_A+\mathrm{E}p_U)^3}\end{aligned} \tag{A.6}$$

In order to determine the bias in the leading term on the right-hand side of (A.6), we must derive expressions for the bias of p_A and p_U as estimators for π_A and π_U, respectively. Beginning with the former, we introduce indicators in order to rewrite p_A as a sum over every member of every family in the population, except for one affected member of each family who is arbitrarily designated as the (case) proband:

$$p_A = \frac{\sum_{i=1}^{F}\sum_{\{j:\, j\in \underset{\sim}{J}_i\}} \delta_{ij}^A\, \mathrm{I}(Y_{ij}=1)}{\sum_{i=1}^{F}\sum_{\{j:\, j\in \underset{\sim}{J}_i\}} \delta_{ij}^A}$$

where $\underset{\sim}{J}_i = \{j : 1 \le j \le N_i \text{ and } r_i(j) \ne 1\}$, a set containing the indices of the non-proband members of family i (after ascertainment); and δ_{ij}^A equals 1 if family member ij is sampled as part of a case-ascertained family and equals 0 otherwise. The indicator δ_{ij}^A will depend on δ_i^A, the case-ascertainment indicator for family i, which equals 1 if family i is ascertained via an

affected proband and 0 otherwise (with the constraint that $\sum_{i=1}^{F} \delta_i^A = F_A$.) If $\delta_i^A = 0$, then $\delta_{ij}^A = 0$ by definition, but if $\delta_i^A = 1$, then δ_{ij}^A can equal 0 or 1.

To obtain an expression for the bias of p_A as an estimator for π_A, we employ the strategy of Hartley and Ross (1954) for determining the bias of a ratio estimator. This strategy begins by expanding the covariance between p_A and its denominator:

$$\begin{aligned}
\mathrm{Cov}\Big(p_A, \sum_{i=1}^{F} \sum_{\{j:\, j\in \underset{\sim}{J}_i\}} \delta_{ij}^A\Big) &= \mathrm{E}\Big(\sum_{i=1}^{F} \sum_{\{j:\, j\in \underset{\sim}{J}_i\}} \delta_{ij}^A \mathrm{I}(Y_{ij}=1)\Big) - \mathrm{E}p_A \cdot \mathrm{E}\Big(\sum_{i=1}^{F} \sum_{\{j:\, j\in \underset{\sim}{J}_i\}} \delta_{ij}^A\Big) \\
&= \sum_{i=1}^{F} \sum_{\{j:\, j\in \underset{\sim}{J}_i\}} \mathrm{E}(\delta_{ij}^A)\mathrm{I}(Y_{ij}=1) - \mathrm{E}p_A \cdot \sum_{i=1}^{F} \sum_{\{j:\, j\in \underset{\sim}{J}_i\}} \mathrm{E}(\delta_{ij}^A) \\
&= \sum_{i=1}^{F} \sum_{\{j:\, j\in \underset{\sim}{J}_i\}} \mathrm{E}\Big(\mathrm{E}\left(\delta_{ij}^A|\delta_i^A\right)\Big)\mathrm{I}(Y_{ij}=1) - \mathrm{E}p_A \cdot \sum_{i=1}^{F} \sum_{\{j:\, j\in \underset{\sim}{J}_i\}} \mathrm{E}\Big(\mathrm{E}\left(\delta_{ij}^A|\delta_i^A\right)\Big).
\end{aligned}$$

Under Assumption (v.), the probability that relative ij is sampled is a constant referred to as s. Using this fact to replace $\mathrm{E}\left(\delta_{ij}^A|\delta_i^A\right)$ in the last line of the above equation yields

$$\begin{aligned}
\mathrm{Cov}\Big(p_A, \sum_{i=1}^{F} \sum_{\{j:\, j\in \underset{\sim}{J}_i\}} \delta_{ij}^A\Big) &= \sum_{i=1}^{F} \sum_{\{j:\, j\in \underset{\sim}{J}_i\}} \mathrm{E}\left(s\delta_i^A + 0(1-\delta_i^A)\right)\mathrm{I}(Y_{ij}=1) \\
&\quad - \mathrm{E}p_A \cdot \sum_{i=1}^{F} \sum_{\{j:\, j\in \underset{\sim}{J}_i\}} \mathrm{E}\left(s\delta_i^A + 0(1-\delta_i^A)\right) \\
&= s\sum_{i=1}^{F} \mathrm{E}(\delta_i^A) \sum_{\{j:\, j\in \underset{\sim}{J}_i\}} \mathrm{I}(Y_{ij}=1) - \mathrm{E}p_A \cdot s\sum_{i=1}^{F} \mathrm{E}(\delta_i^A) \sum_{\{j:\, j\in \underset{\sim}{J}_i\}} 1 \\
&= s\sum_{i=1}^{F} \mathrm{E}(\delta_i^A)(N_i^A - 1) - \mathrm{E}p_A \cdot s\sum_{i=1}^{F} \mathrm{E}(\delta_i^A)(N_i - 1) \\
&= s\,F_A\,s_A \sum_{i=1}^{F} N_i^A(N_i^A - 1) - \mathrm{E}p_A \cdot s\,F_A\,s_A \sum_{i=1}^{F} N_i^A(N_i - 1)
\end{aligned}$$

where the third line follows because $\underset{\sim}{J}_i$ does not include one affected member of family i who is designated as the case proband; and where the fourth line follows because, under Assumption

(iii.) about proband selection and Assumption (iv.) about single ascertainment, $\mathrm{E}(\delta_i^A)$ can be rewritten as $N_i^A\, F_{_A}\, s_{_A}$, where $s_{_A} = 1/\sum_{i=1}^{F} N_i^A$, the sampling fraction for case probands. The last line above can be rearranged to give

$$\mathrm{E}p_{_A} - \frac{\sum_{i=1}^{F} N_i^A(N_i^A - 1)}{\sum_{i=1}^{F} N_i^A(N_i - 1)} = -\frac{\mathrm{Cov}\Big(p_{_A}, \sum_{i=1}^{F} \sum_{\{j:\, j \in \underset{\sim}{J}_i\}} \delta_{ij}^A\Big)}{s\, F_{_A}\, s_{_A} \sum_{i=1}^{F} N_i^A(N_i - 1)}.$$

Since the second term on the lefthand side of the above equation is just $\pi_{_A}$ as written in (A.3), the bias of $p_{_A}$ can be written as

$$\mathrm{E}p_{_A} - \pi_{_A} = -\frac{\mathrm{Cor}\Big(p_{_A}, \sum_{i=1}^{F} \sum_{\{j:\, j \in \underset{\sim}{J}_i\}} \delta_{ij}^A\Big) \cdot \mathrm{SD}\big(p_{_A}\big) \cdot \mathrm{SD}\Big(\sum_{i=1}^{F} \sum_{\{j:\, j \in \underset{\sim}{J}_i\}} \delta_{ij}^A\Big)}{s\, F_{_A}\, s_{_A} \sum_{i=1}^{F} N_i^A(N_i - 1)}. \qquad \text{(A.7)}$$

Since the denominator on the righthand side of the above equation equals the expectation of $\sum_{i=1}^{F} \sum_{\{j:\, j \in \underset{\sim}{J}_i\}} \delta_{ij}^A$, we can rewrite (A.7) as

$$\mathrm{E}p_{_A} - \pi_{_A} = -\,\mathrm{Cor}\Big(p_{_A}, \sum_{i=1}^{F} \sum_{\{j:\, j \in \underset{\sim}{J}_i\}} \delta_{ij}^A\Big) \cdot \mathrm{SD}\big(p_{_A}\big) \cdot \mathrm{CV}\Big(\sum_{i=1}^{F} \sum_{\{j:\, j \in \underset{\sim}{J}_i\}} \delta_{ij}^A\Big), \qquad \text{(A.8)}$$

where $\mathrm{CV}(\sum_{i=1}^{F} \sum_{\{j:\, j \in \underset{\sim}{J}_i\}} \delta_{ij}^A)$, the coefficient of variation, is defined as the ratio of the standard deviation of $\sum_{i=1}^{F} \sum_{\{j:\, j \in \underset{\sim}{J}_i\}} \delta_{ij}^A$ to the mean of $\sum_{i=1}^{F} \sum_{\{j:\, j \in \underset{\sim}{J}_i\}} \delta_{ij}^A$.

We examine the magnitude of each of the three multiplicands on the righthand side of (A.8). First, $\mathrm{SD}(p_{_A})$ must be less than 0.5 because $p_{_A}$ is a proportion. Second, it is difficult (if not impossible) to construct a population where $\mathrm{CV}\Big(\sum_{i=1}^{F} \sum_{\{j:\, j \in \underset{\sim}{J}_i\}} \delta_{ij}^A\Big)$ is larger than 2, as would be expected for a quantity that is effectively the sum of binary variables (albeit non-identical,

non-independent ones). Third, under Assumption (ii.), which states that family size and disease status are uncorrelated, $\text{Cor}\left(p_{A}, \sum_{i=1}^{F} \sum_{\{j:\, j \in \underset{\sim}{J}_i\}} \delta_{ij}^{A}\right)$ is negligible. This is true because Assumption (ii.) ensures that the average size of the case-ascertained families included in the study will have a negligible correlation with the proportion of the relatives in those families who have the disease. Putting these three facts together, we see that the bias of p_{A} is negligible when Assumption (ii.) holds.

To illustrate the importance of Assumption (ii.) in guaranteeing that p_{A} is approximately unbiased for π_{A}, we examine the bias of p_{A} in two fictional populations. The first is the same population used in the simulation experiment in Section 3. The second is a population created in an identical fashion, except that the prevalence of disease was set lower for males and females from families with more than three members and set higher for males and females from families with fewer than three members. Note that the overall prevalence of disease is equal in both populations, but in the second population, a disproportionately large number of the diseased individuals belong to small families. For each population, we sampled $1,000$ datasets, each consisting of 64 case probands ($F_{A} = 64$) and all of their relatives ($s = 1$). We calculated p_{A} and $\sum_{i=1}^{F} \sum_{\{j:\, j \in \underset{\sim}{J}_i\}} \delta_{ij}^{A}$ for each dataset and then used the resulting values to estimate the three multiplicands in (A.8) for that population. Table 2 presents the values of the three multiplicands in the two populations, as well as the bias of p_{A} in percentage terms. The middle column of Table 2 reveals that $\text{Cor}\left(p_{A}, \sum_{i=1}^{F} \sum_{\{j:\, j \in \underset{\sim}{J}_i\}} \delta_{ij}^{A}\right)$ and the percentage bias of p_{A} are negligible in the first population, where Assumption (ii.) holds. Comparing the middle column to the right-most column in Table 2 reveals that the percentage bias of p_{A} is approximately 100 times larger for the second population, where Assumption (ii.) does not hold. This increase in bias is due to the larger value of $\text{Cor}\left(p_{A}, \sum_{i=1}^{F} \sum_{\{j:\, j \in \underset{\sim}{J}_i\}} \delta_{ij}^{A}\right)$.

We can use the approach employed above to obtain an analogous expression for the bias of

p_U as an estimator for π_U

$$\mathrm{E}p_U - \pi_U \;=\; -\,\mathrm{Cor}\Big(p_U, \sum_{i=1}^{F} \sum_{\{j:\, j \in \underset{\sim}{J}_i\}} \delta_{ij}^U\Big) \cdot \mathrm{SD}\big(p_U\big) \cdot \mathrm{CV}\big(\sum_{i=1}^{F} \sum_{\{j:\, j \in \underset{\sim}{J}_i\}} \delta_{ij}^U\big) \quad \text{(A.9)}$$

where δ_{ij}^U equals 1 if relative ij is sampled as part of a control-ascertained family and equals 0 otherwise. Since the same arguments made for the multiplicands in (A.8) also apply to the multiplicands in (A.9), we see that the bias of p_U will be negligible when Assumption (ii.) holds. For both populations described above, Table 3 presents values of the three multiplicands in (A.9) and the percentage bias of p_U, calculated from $1,000$ datasets containing $F_U = 58$ control probands and all of their relatives ($s = 1$). The bias of p_U is negligible for the first population, but is again approximately 100 times larger for the second population because the correlation between p_U and its denominator is larger. Thus, the bias of p_U is negligible when Assumption (ii.) holds.

The fact that p_A and p_U have negligible bias under the assumptions enumerated in Section 2 implies that the bias in the leading term of (A.6) is negligible. Thus, to a first-degree approximation, $\hat{\pi}$ is an unbiased estimator for π.

We now turn to the bias introduced by the second-order terms in (A.6). First, note that the final second-order term in (A.6) introduces no bias because $\mathrm{Cov}(p_A, p_U) = 0$. Next, note that the numerators of the first two second-order terms in (A.6) can be re-written as $\Big[Ep_U(1 - Ep_U)/\sum_{\underset{\sim}{i}=F_A+1}^{F_A+F_U} (n_{\underset{\sim}{i}} - 1)\Big]\,(1 - Ep_A)$ and $\Big[Ep_A(1 - Ep_A)/\sum_{\underset{\sim}{i}=1}^{F_A} (n_{\underset{\sim}{i}} - 1)\Big]\,Ep_U$, respectively. We can ignore the summations in the denominators of these terms because they are approximately equal under assumption (ii.) that disease status is uncorrelated with family size and under the assumption that $F_A \approx F_U$. Now, the only difference between the two terms is that $(1 - Ep_A)Ep_U$ is multiplied by $(1 - Ep_U)$ in the first term and by Ep_A in the second term. Thus, if we assume that $E\,(1 - p_U) > Ep_A$, then the first second-order term, which has a negative sign in front of it, is larger in magnitude than the second second-order term, which

has a positive sign in front of it. As a result, the bias introduced through the second-order terms in (A.6) will be non-positive if $E(1-p_U) > Ep_A$ and $F_A \approx F_U$. However, the results of the simulation experiment in Section 3, where $\hat{\pi}$ underestimates π by only a very small amount, suggest that the bias introduced through the second- (and higher-) order terms is very small in practice. ■

APPENDIX B

Proof that $\hat{\pi}^x$ is only slightly biased at the first-order for π^x

We define $\pi^x \equiv f(Y_{ij} = 1|X_{ij} = x)$, where individual ij is randomly selected from among the members of the population with $X_{ij} = x$. Assumption (i.) allows us to expand $f(Y_{ij} = 1|X_{ij} = x)$ as

$$\begin{aligned} \pi^x &= f(Y_{ij} = 1|Y_{ij'} = 1, X_{ij} = x) f(Y_{ij'} = 1|X_{ij} = x) \\ &\quad + f(Y_{ij} = 1|Y_{ij'} = 0, X_{ij} = x) f(Y_{ij'} = 0|X_{ij} = x), \end{aligned} \tag{B.1}$$

where individual ij' is randomly selected from among Y_{ij}'s relatives with disease status $Y_{ij'}$. (In the remainder of this proof, we will assume that $j' \neq j$.) Under Assumption (vii.) about the independence of probands' disease status and relatives' covariates,

$$\pi^x = f(Y_{ij} = 1|Y_{ij'} = 1, X_{ij} = x) f(Y_{ij'} = 1) + f(Y_{ij} = 1|Y_{ij'} = 0, X_{ij} = x) f(Y_{ij'} = 0),$$

follows from (B.1). We can rewrite the preceding equation as

$$\pi^x = \pi^x_{A} \pi + \pi^x_{U} (1 - \pi), \tag{B.2}$$

where $\pi^x_{A} \equiv f(Y_{ij} = 1|Y_{ij'} = 1, X_{ij} = x)$ and $\pi^x_{U} \equiv f(Y_{ij} = 1|Y_{ij'} = 0, X_{ij} = x)$. The parameters π^x_{A} and π^x_{U} can be defined in terms of the finite population:

$$\begin{aligned} \pi^x_{A} &= \frac{\sum_{i=1}^{F} \sum_{j=1}^{N_i} \mathrm{I}(Y_{ij} = 1) \mathrm{I}(X_{ij} = x) \sum_{j' \neq j} \mathrm{I}(Y_{ij'} = 1)}{\sum_{i=1}^{F} \sum_{j=1}^{N_i} \mathrm{I}(X_{ij} = x) \sum_{j' \neq j} \mathrm{I}(Y_{ij'} = 1)} \\ &= \frac{\sum_{i=1}^{F} \sum_{j=1}^{N_i} \mathrm{I}(Y_{ij} = 1) \mathrm{I}(X_{ij} = x)(N_i^A - 1)}{\sum_{i=1}^{F} \sum_{j=1}^{N_i} \left(\mathrm{I}(X_{ij} = x) \mathrm{I}(Y_{ij} = 1)(N_i^A - 1) + \mathrm{I}(X_{ij} = x) \mathrm{I}(Y_{ij} = 0) N_i^A \right)} \\ &= \frac{\sum_{i=1}^{F} N_i^{Ax} (N_i^A - 1)}{\sum_{i=1}^{F} \left(N_i^x N_i^A - N_i^{Ax} \right)}, \end{aligned} \tag{B.3}$$

and

$$\begin{aligned}\pi_U^x &= \frac{\sum_{i=1}^{F}\sum_{j=1}^{N_i}\mathrm{I}(Y_{ij}=1)\mathrm{I}(X_{ij}=x)\sum_{j'\neq j}\mathrm{I}(Y_{ij'}=0)}{\sum_{i=1}^{F}\sum_{j=1}^{N_i}\mathrm{I}(X_{ij}=x)\sum_{j'\neq j}\mathrm{I}(Y_{ij'}=0)} \\ &= \frac{\sum_{i=1}^{F}\sum_{j=1}^{N_i}\mathrm{I}(Y_{ij}=1)\mathrm{I}(X_{ij}=x)N_i^U}{\sum_{i=1}^{F}\sum_{j=1}^{N_i}\big(\mathrm{I}(X_{ij}=x)\mathrm{I}(Y_{ij}=1)N_i^U+\mathrm{I}(X_{ij}=x)\mathrm{I}(Y_{ij}=0)(N_i^U-1)\big)} \\ &= \frac{\sum_{i=1}^{F}N_i^{Ax}N_i^U}{\sum_{i=1}^{F}\big(N_i^xN_i^U-N_i^{Ux}\big)} \qquad \text{(B.4)}\end{aligned}$$

where $N_i^A = \sum_{j=1}^{N_i}\mathrm{I}(Y_{ij}=1)$; $N_i^U = \sum_{j=1}^{N_i}\mathrm{I}(Y_{ij}=0)$; $N_i^x = \sum_{j=1}^{N_i}\mathrm{I}(X_{ij}=x)$; $N_i^{Ax} = \sum_{j=1}^{N_i}\mathrm{I}(Y_{ij}=1)\mathrm{I}(X_{ij}=x)$; and $N_i^{Ux} = \sum_{j=1}^{N_i}\mathrm{I}(Y_{ij}=0)\mathrm{I}(X_{ij}=x)$.

Now, recall that

$$\hat{\pi}^x = p_A^x\hat{\pi} + p_U^x(1-\hat{\pi}), \qquad \text{(B.5)}$$

where

$$p_A^x = \frac{\sum_{\underset{\sim}{i}=1}^{F_A}\sum_{\underset{\sim}{j}=2}^{n_{\underset{\sim}{i}}}\mathrm{I}(X_{\underset{\sim}{ij}}=x)\mathrm{I}(Y_{\underset{\sim}{ij}}=1)}{\sum_{\underset{\sim}{i}=1}^{F_A}\sum_{\underset{\sim}{j}=2}^{n_{\underset{\sim}{i}}}\mathrm{I}(X_{\underset{\sim}{ij}}=x)}$$

and

$$p_U^x = \frac{\sum_{\underset{\sim}{i}=F_A+1}^{F_A+F_U}\sum_{\underset{\sim}{i}=2}^{n_{\underset{\sim}{i}}}\mathrm{I}(X_{\underset{\sim}{ij}}=x)\mathrm{I}(Y_{\underset{\sim}{ij}}=1)}{\sum_{\underset{\sim}{i}=F_A+1}^{F_A+F_U}\sum_{\underset{\sim}{i}=2}^{n_{\underset{\sim}{i}}}\mathrm{I}(X_{\underset{\sim}{ij}}=x)}.$$

The estimator $\hat{\pi}^x$ in (B.5) can be approximated by a second-order Taylor expansion around

$\mathrm{E}\hat{\pi}$, $\mathrm{E}p_U^x$ and $\mathrm{E}p_A^x$:

$$\begin{aligned}
\hat{\pi}^x \approx\ & \left(\mathrm{E}p_A^x \mathrm{E}\hat{\pi} + \mathrm{E}p_U^x(1-\mathrm{E}\hat{\pi})\right) \\
& + (\hat{\pi} - \mathrm{E}\hat{\pi})\left.\frac{\partial \hat{\pi}^x}{\partial \hat{\pi}}\right|_{\mathrm{E}\hat{\pi},\mathrm{E}p_U^x,\mathrm{E}p_A^x} + \left(p_U^x - \mathrm{E}p_U^x\right)\left.\frac{\partial \hat{\pi}^x}{\partial p_U^x}\right|_{\mathrm{E}\hat{\pi},\mathrm{E}p_U^x,\mathrm{E}p_A^x} + \left(p_A^x - \mathrm{E}p_A^x\right)\left.\frac{\partial \hat{\pi}^x}{\partial p_A^x}\right|_{\mathrm{E}\hat{\pi},\mathrm{E}p_U^x,\mathrm{E}p_A^x} \\
& + \frac{1}{2}(\hat{\pi} - \mathrm{E}\hat{\pi})^2\left.\frac{\partial^2 \hat{\pi}^x}{\partial \hat{\pi}^2}\right|_{\mathrm{E}\hat{\pi},\mathrm{E}p_U^x,\mathrm{E}p_A^x} + \frac{1}{2}\left(p_U^x - \mathrm{E}p_U^x\right)^2\left.\frac{\partial^2 \hat{\pi}^x}{\partial {p_U^x}^2}\right|_{\mathrm{E}\hat{\pi},\mathrm{E}p_U^x,\mathrm{E}p_A^x} \\
& + \frac{1}{2}\left(p_A^x - \mathrm{E}p_A^x\right)^2\left.\frac{\partial^2 \hat{\pi}^x}{\partial {p_A^x}^2}\right|_{\mathrm{E}\hat{\pi},\mathrm{E}p_U^x,\mathrm{E}p_A^x} \\
& + (\hat{\pi} - \mathrm{E}\hat{\pi})\left(p_A^x - \mathrm{E}p_A^x\right)\left.\frac{\partial^2 \hat{\pi}^x}{\partial \hat{\pi} \partial p_A^x}\right|_{\mathrm{E}\hat{\pi},\mathrm{E}p_U^x,\mathrm{E}p_A^x} + (\hat{\pi} - \mathrm{E}\hat{\pi})\left(p_U^x - \mathrm{E}p_U^x\right)\left.\frac{\partial^2 \hat{\pi}^x}{\partial \hat{\pi} \partial p_U^x}\right|_{\mathrm{E}\hat{\pi},\mathrm{E}p_U^x,\mathrm{E}p_A^x} \\
& + \left(p_U^x - \mathrm{E}p_U^x\right)\left(p_A^x - \mathrm{E}p_A^x\right)\left.\frac{\partial^2 \hat{\pi^x}}{\partial p_U^x \partial p_A^x}\right|_{\mathrm{E}\hat{\pi},\mathrm{E}p_U^x,\mathrm{E}p_A^x}.
\end{aligned}$$

Taking the expectation of both sides of the above equation yields

$$\begin{aligned}
\mathrm{E}\hat{\pi}^x \approx\ & \left(\mathrm{E}p_A^x \mathrm{E}\hat{\pi} + \mathrm{E}p_U^x(1-\mathrm{E}\hat{\pi})\right) \\
& + \frac{1}{2}\mathrm{Var}(\hat{\pi})\left.\frac{\partial^2 \hat{\pi}^x}{\partial \hat{\pi}^2}\right|_{\mathrm{E}\hat{\pi},\mathrm{E}p_U^x,\mathrm{E}p_A^x} + \frac{1}{2}\mathrm{Var}(p_U^x)\left.\frac{\partial^2 \hat{\pi}^x}{\partial {p_U^x}^2}\right|_{\mathrm{E}\hat{\pi},\mathrm{E}p_U^x,\mathrm{E}p_A^x} + \frac{1}{2}\mathrm{Var}(p_A^x)\left.\frac{\partial^2 \hat{\pi}^x}{\partial {p_A^x}^2}\right|_{\mathrm{E}\hat{\pi},\mathrm{E}p_U^x,\mathrm{E}p_A^x} \\
& + \mathrm{Cov}(\hat{\pi}, p_A^x)\left.\frac{\partial^2 \hat{\pi}^x}{\partial \hat{\pi} \partial p_A^x}\right|_{\mathrm{E}\hat{\pi},\mathrm{E}p_U^x,\mathrm{E}p_A^x} + \mathrm{Cov}(\hat{\pi}, p_U^x)\left.\frac{\partial^2 \hat{\pi}^x}{\partial \hat{\pi} \partial p_U^x}\right|_{\mathrm{E}\hat{\pi},\mathrm{E}p_U^x,\mathrm{E}p_A^x} \\
& + \mathrm{Cov}(p_U^x, p_A^x)\left.\frac{\partial^2 \hat{\pi^x}}{\partial p_U^x \partial p_A^x}\right|_{\mathrm{E}\hat{\pi},\mathrm{E}p_U^x,\mathrm{E}p_A^x}.
\end{aligned} \tag{B.6}$$

We focus now on the leading term of the expectation of the Taylor expansion in (B.6): $\mathrm{E}p_A^x \mathrm{E}\hat{\pi} + \mathrm{E}p_U^x(1-\mathrm{E}\hat{\pi})$. We have already shown in Appendix A that, under conditions (i.)-(v.), $\hat{\pi}$ has a very small negative bias as an estimator for π. To derive the bias in p_A^x and p_U^x, we introduce indicators in order to rewrite them as

$$p_A^x = \frac{\sum_{i=1}^{F}\sum_{\{j:\, j\in J_i\}} \delta_{ij}^{Ax}\mathrm{I}(X_{ij}=x)\mathrm{I}(Y_{ij}=1) + \sum_{i=1}^{F}\sum_{\{j:\, j\in J_i\}} \delta_{ij}^{Ax^c}\mathrm{I}(X_{ij}=x)\mathrm{I}(Y_{ij}=1)}{\sum_{i=1}^{F}\sum_{\{j:\, j\in J_i\}} \delta_{ij}^{Ax}\mathrm{I}(X_{ij}=x) + \sum_{i=1}^{F}\sum_{\{j:\, j\in J_i\}} \delta_{ij}^{Ax^c}\mathrm{I}(X_{ij}=x)}$$

where δ_{ij}^{Ax} (or $\delta_{ij}^{Ax^c}$) equals 1 if family member ij is sampled as part of a family ascertained through a case proband with covariate value x (or covariate value in the complement of x) and equals 0 otherwise; and

$$p_{U}^{x} = \frac{\sum_{i=1}^{F}\sum_{\{j:\, j\in\underset{\sim}{J}_i\}}\delta_{ij}^{Ux}\mathrm{I}(X_{ij}=x)\mathrm{I}(Y_{ij}=1)+\sum_{i=1}^{F}\sum_{\{j:\, j\in\underset{\sim}{J}_i\}}\delta_{ij}^{Ux^c}\mathrm{I}(X_{ij}=x)\mathrm{I}(Y_{ij}=1)}{\sum_{i=1}^{F}\sum_{\{j:\, j\in\underset{\sim}{J}_i\}}\delta_{ij}^{Ux}\mathrm{I}(X_{ij}=x)+\sum_{i=1}^{F}\sum_{\{j:\, j\in\underset{\sim}{J}_i\}}\delta_{ij}^{Ux^c}\mathrm{I}(X_{ij}=x)}$$

where δ_{ij}^{Ux} (or $\delta_{ij}^{Ux^c}$) equals 1 if family member ij is sampled as part of a family ascertained through a control proband with covariate value x (or covariate value in the complement of x) and equals 0 otherwise. For the sake of brevity, we will refer to the denominators of p_{A}^{x} and p_{U}^{x} as d_{A}^{x} and d_{U}^{x}, respectively.

To derive the bias of p_{A}^{x} as an estimator for π_{A}^{x}, we use the same Hartley-Ross (1954) approach employed for p_{A} in Appendix A:

$$\begin{aligned}\mathrm{Cov}\big(p_{A}^{x}, d_{A}^{x}\big) &= \mathrm{E}\Big(\sum_{i=1}^{F}\sum_{\{j:\, j\in\underset{\sim}{J}_i\}}\delta_{ij}^{Ax}\mathrm{I}(X_{ij}=x)\mathrm{I}(Y_{ij}=1)+\sum_{i=1}^{F}\sum_{\{j:\, j\in\underset{\sim}{J}_i\}}\delta_{ij}^{Ax^c}\mathrm{I}(X_{ij}=x)\mathrm{I}(Y_{ij}=1)\Big)\\ &\quad - \mathrm{E}p_{A}^{x}\cdot\mathrm{E}\Big(\sum_{i=1}^{F}\sum_{\{j:\, j\in\underset{\sim}{J}_i\}}\delta_{ij}^{Ax}\mathrm{I}(X_{ij}=x)+\sum_{i=1}^{F}\sum_{\{j:\, j\in\underset{\sim}{J}_i\}}\delta_{ij}^{Ax^c}\mathrm{I}(X_{ij}=x)\Big)\end{aligned}$$

$$
\begin{aligned}
= \quad & \Big[\sum_{i=1}^{F} \sum_{\{j:\, j \in \underset{\sim}{J}_i\}} \mathrm{E}\Big(\mathrm{E}\left(\delta_{ij}^{Ax} | \delta_i^{Ax}\right)\Big) \mathrm{I}(X_{ij} = x) \mathrm{I}(Y_{ij} = 1) \\
& + \sum_{i=1}^{F} \sum_{\{j:\, j \in \underset{\sim}{J}_i\}} \mathrm{E}\Big(\mathrm{E}\left(\delta_{ij}^{Ax^c} | \delta_i^{Ax^c}\right)\Big) \mathrm{I}(X_{ij} = x) \mathrm{I}(Y_{ij} = 1)\Big] \\
& - \mathrm{E}p_{A}^{x} \cdot \Big[\sum_{i=1}^{F} \sum_{\{j:\, j \in \underset{\sim}{J}_i\}} \mathrm{E}\Big(\mathrm{E}\left(\delta_{ij}^{Ax} | \delta_i^{Ax}\right)\Big) \mathrm{I}(X_{ij} = x) + \sum_{i=1}^{F} \sum_{\{j:\, j \in \underset{\sim}{J}_i\}} \mathrm{E}\Big(\mathrm{E}\left(\delta_{ij}^{Ax^c} | \delta_i^{Ax^c}\right)\Big) \mathrm{I}(X_{ij} = x)\Big]
\end{aligned}
$$

where δ_i^{Ax} equals 1 if family i is ascertained via an affected proband with covariate value x and 0 otherwise, and where $\delta_i^{Ax^c}$ equals 1 if family i is ascertained via an affected proband with covariate value in the complement of x and 0 otherwise. Invoking Assumption (v.), the preceding line reduces to

$$
\begin{aligned}
\mathrm{Cov}(p_{A}^{x}, d_{A}^{x}) \quad = \quad & \Big[s \sum_{i=1}^{F} \sum_{\{j:\, j \in \underset{\sim}{J}_i\}} \mathrm{E}(\delta_i^{Ax}) \mathrm{I}(X_{ij} = x) \mathrm{I}(Y_{ij} = 1) + s \sum_{i=1}^{F} \sum_{\{j:\, j \in \underset{\sim}{J}_i\}} \mathrm{E}(\delta_i^{Ax^c}) \mathrm{I}(X_{ij} = x) \mathrm{I}(Y_{ij} = 1)\Big] \\
& - \mathrm{E}p_{A}^{x} \cdot \Big[s \sum_{i=1}^{F} \sum_{\{j:\, j \in \underset{\sim}{J}_i\}} \mathrm{E}(\delta_i^{Ax}) \mathrm{I}(X_{ij} = x) + s \sum_{i=1}^{F} \sum_{\{j:\, j \in \underset{\sim}{J}_i\}} \mathrm{E}(\delta_i^{Ax^c}) \mathrm{I}(X_{ij} = x)\Big] \\
= \quad & \Big[s \sum_{i=1}^{F} \mathrm{E}(\delta_i^{Ax})(N_i^{Ax} - 1) + s \sum_{i=1}^{F} \mathrm{E}(\delta_i^{Ax^c}) N_i^{Ax}\Big] \\
& - \mathrm{E}p_{A}^{x} \cdot \Big[s \sum_{i=1}^{F} \mathrm{E}(\delta_i^{Ax})(N_i^{x} - 1) + s \sum_{i=1}^{F} \mathrm{E}(\delta_i^{Ax^c}) N_i^{x}\Big] \\
= \quad & \Big[s\, F_{A}\, s_{A} \sum_{i=1}^{F} N_i^{Ax}(N_i^{Ax} - 1) + s\, F_{A}\, s_{A} \sum_{i=1}^{F} N_i^{Ax^c} N_i^{Ax}\Big] \\
& - \mathrm{E}p_{A}^{x} \cdot \Big[s\, F_{A}\, s_{A} \sum_{i=1}^{F} N_i^{Ax}(N_i^{x} - 1) + s\, F_{A}\, s_{A} \sum_{i=1}^{F} N_i^{Ax^c} N_i^{x}\Big]
\end{aligned}
$$

where $N_i^{Ax^c} = \sum_{j=1}^{N_i} \mathrm{I}(X_{ij} \neq x) \mathrm{I}(Y_{ij} = 1)$. Note that the last expression above follows from the second-to-last expression above under Assumption (iii.) about proband selection and As-

sumption (iv.) about single ascertainment. The last expression above can be rewritten as

$$\text{Cov}(p_{A}^{x}, d_{A}) \;=\; s\,F_{A}\,s_{A}\sum_{i=1}^{F} N_i^{Ax}(N_i^{A}-1) - \text{E}p_{A}^{x}\cdot s\,F_{A}\,s_{A}\sum_{i=1}^{F}\left(N_i^{A}N_i^{x}-N_i^{Ax}\right),$$

which, when rearranged and combined with (B.3), gives

$$\begin{aligned}\text{E}p_{A}^{x}-\pi_{A}^{x} &= -\frac{\text{Cov}(p_{A}^{x}, d_{A}^{x})}{s\,F_{A}\,s_{A}\sum_{i=1}^{F}\left(N_i^{A}N_i^{x}-N_i^{Ax}\right)} \\ &= -\,\text{Cor}(p_{A}^{x}, d_{A}^{x})\cdot\text{SD}(p_{A}^{x})\cdot\text{CV}(d_{A}^{x}). \qquad \text{(B.7)}\end{aligned}$$

We can use the Hartley-Ross (1954) approach to obtain an analogous expression for the bias of p_{U}^{x} as an estimator for π_{U}^{x}:

$$\begin{aligned}\text{E}p_{U}^{x}-\pi_{U}^{x} &= -\frac{\text{Cov}(p_{U}^{x}, d_{U}^{x})}{s\,F_{U}\,s_{U}\sum_{i=1}^{F}\left(N_i^{U}N_i^{x}-N_i^{Ux}\right)} \\ &= -\,\text{Cor}(p_{U}^{x}, d_{U}^{x})\cdot\text{SD}(p_{U}^{x})\cdot\text{CV}(d_{U}^{x}). \qquad \text{(B.8)}\end{aligned}$$

We can then use the same arguments made in Appendix A to establish that the right-hand sides of (B.7) and (B.8) will be negligible when Assumption (ii.) holds. Thus, under Assumption (ii.), p_{A}^{x} and p_{U}^{x} are approximately unbiased estimators for π_{A}^{x} and π_{U}^{x}, respectively. Using our previous finding that $\hat{\pi}$ slightly underestimate π, along with the fact that p_{A}^{x} will typically exceed p_{U}^{x} for diseases that aggregate in families, we see that the leading term in (B.6) underestimates π_x slightly. Thus, to a first-degree approximation, $\hat{\pi}^{x}$ is a slightly downwardly biased estimator for π_x. However, the results of the simulation experiment in Section 3 suggest that the bias introduced by the leading term and also the higher order terms in (B.6) is very small. ■

APPENDIX C

Standard Errors and Confidence Intervals for $\hat{\pi}$ and $\hat{\pi}^x$

The delta method can be used to obtain approximate standard errors for $\hat{\pi}$ and $\hat{\pi}^x$. The approximate standard error for $\hat{\pi}$ is

$$\begin{aligned} \text{se}(\hat{\pi}) &= \pi(1-\pi) \\ &\cdot \sqrt{\frac{\pi_A}{d_A(1-\pi_A)}\left[1+\frac{2\rho_A}{d_A}\sum_{\underset{\sim}{i}=1}^{F_A}\binom{n_{\underset{\sim}{i}}-1}{2}\right]+\frac{(1-\pi_U)}{d_U\pi_U}\left[1+\frac{2\rho_U}{d_U}\sum_{\underset{\sim}{i}=F_A+1}^{F_A+F_U}\binom{n_{\underset{\sim}{i}}-1}{2}\right]}, \end{aligned} \tag{C.1}$$

where $d_A=\sum_{\underset{\sim}{i}=1}^{F_A}\sum_{\underset{\sim}{j}=2}^{n_{\underset{\sim}{i}}}1$ and $d_U=\sum_{\underset{\sim}{i}=F_A+1}^{F_A+F_U}\sum_{\underset{\sim}{j}=2}^{n_{\underset{\sim}{i}}}1$; $\rho_A=\text{Cor}(Y_{\underset{\sim}{i}\underset{\sim}{j}},Y_{\underset{\sim}{i}\underset{\sim}{j}'})$ for $\underset{\sim}{i}=1,\ldots,F_A$, $\underset{\sim}{j}>1$, $\underset{\sim}{j}'>1$, and $\underset{\sim}{j}\neq\underset{\sim}{j}'$; and $\rho_U=\text{Cor}(Y_{\underset{\sim}{i}\underset{\sim}{j}},Y_{\underset{\sim}{i}\underset{\sim}{j}'})$ for $\underset{\sim}{i}=F_A+1,\ldots,F_A+F_U$, $\underset{\sim}{j}>1$, $\underset{\sim}{j}'>1$, and $\underset{\sim}{j}\neq\underset{\sim}{j}'$. Note that the more the disease aggregates in families, the larger ρ_A and ρ_U will be and therefore the larger the standard error for $\hat{\pi}$ will be.

The approximate standard error for $\hat{\pi}^x$ is

$$\text{se}(\hat{\pi}^x)=\sqrt{a\Sigma a^T}, \tag{C.2}$$

where

$$a=\begin{bmatrix} \pi & (1-\pi) & \frac{(\pi_A^x-\pi_U^x)\pi^2}{\pi_U} & \frac{(\pi_A^x-\pi_U^x)(1-\pi)^2}{1-\pi_A} \end{bmatrix}; \tag{C.2.a}$$

and

$$\Sigma=\begin{bmatrix} \sigma_{1,1} & 0 & \sigma_{1,3} & 0 \\ 0 & \sigma_{2,2} & 0 & \sigma_{2,4} \\ \sigma_{1,3} & 0 & \sigma_{3,3} & 0 \\ 0 & \sigma_{2,4} & 0 & \sigma_{4,4} \end{bmatrix} \tag{C.2.b}$$

with

$$\sigma_{1,1}=\frac{\pi_A^x(1-\pi_A^x)}{d_A^x}\left[1+\frac{2\rho_A^x}{d_A^x}\sum_{\underset{\sim}{i}=1}^{F_A}\binom{n_{\underset{\sim}{i}}^x}{2}\right],$$

$$\sigma_{2,2} = \frac{\pi_{U}^{x}(1-\pi_{U}^{x})}{d_{U}^{x}}\left[1+\frac{2\rho_{U}^{x}}{d_{U}^{x}}\sum_{\underset{\sim}{i}=F_{A}+1}^{F_{A}+F_{U}}\binom{n_{\underset{\sim}{i}}^{x}}{2}\right],$$

$$\sigma_{3,3} = \frac{\pi_{A}(1-\pi_{A})}{d_{A}}\left[1+\frac{2\rho_{A}}{d_{A}}\sum_{\underset{\sim}{i}=1}^{F_{A}}\binom{n_{\underset{\sim}{i}}-1}{2}\right],$$

$$\sigma_{4,4} = \frac{\pi_{U}(1-\pi_{U})}{d_{U}}\left[1+\frac{2\rho_{U}}{d_{U}}\sum_{\underset{\sim}{i}=F_{A}+1}^{F_{A}+F_{U}}\binom{n_{\underset{\sim}{i}}-1}{2}\right],$$

$$\sigma_{1,3} = \frac{1}{d_{A}^{x}d_{A}}\left[\sigma_{1,1}+\rho_{A}^{x,x^{c}}\sum_{\underset{\sim}{i}=1}^{F_{A}}n_{\underset{\sim}{i}}^{x}n_{\underset{\sim}{i}}^{x^{c}}\right],$$

and

$$\sigma_{2,4} = \frac{1}{d_{U}^{x}d_{U}}\left[\sigma_{2,2}+\rho_{U}^{x,x^{c}}\sum_{\underset{\sim}{i}=F_{A}+1}^{F_{A}+F_{U}}n_{\underset{\sim}{i}}^{x}n_{\underset{\sim}{i}}^{x^{c}}\right].$$

Further, $d_{A}^{x}=\sum_{\underset{\sim}{i}=1}^{F_{A}}\sum_{\underset{\sim}{j}=2}^{n_{\underset{\sim}{i}}}\mathrm{I}(X_{\underset{\sim}{i}\underset{\sim}{j}}=x)$ and $d_{U}^{x}=\sum_{\underset{\sim}{i}=F_{A}+1}^{F_{A}+F_{U}}\sum_{\underset{\sim}{j}=2}^{n_{\underset{\sim}{i}}}\mathrm{I}(X_{\underset{\sim}{i}\underset{\sim}{j}}=x)$; $n_{\underset{\sim}{i}}^{x}=\sum_{\underset{\sim}{j}=2}^{n_{\underset{\sim}{i}}}\mathrm{I}(X_{\underset{\sim}{i}\underset{\sim}{j}}=x)$ and $n_{\underset{\sim}{i}}^{x^{c}}=\sum_{\underset{\sim}{j}=2}^{n_{\underset{\sim}{i}}}\mathrm{I}(X_{\underset{\sim}{i}\underset{\sim}{j}}\neq x)$; $\rho_{A}^{x}=\mathrm{Cor}(Y_{\underset{\sim}{i}\underset{\sim}{j}},Y_{\underset{\sim}{i}\underset{\sim}{j}'})$ for $\underset{\sim}{i}=1,\ldots,F_{A}$, $\underset{\sim}{j}>1$, $\underset{\sim}{j}'>1$, $\underset{\sim}{j}\neq\underset{\sim}{j}'$, $X_{\underset{\sim}{i}\underset{\sim}{j}}=X_{\underset{\sim}{i}\underset{\sim}{j}'}=x$; $\rho_{U}^{x}=\mathrm{Cor}(Y_{\underset{\sim}{i}\underset{\sim}{j}},Y_{\underset{\sim}{i}\underset{\sim}{j}'})$ for $\underset{\sim}{i}=F_{A}+1,\ldots,F_{A}+F_{U}$, $\underset{\sim}{j}>1$, $\underset{\sim}{j}'>1$, $\underset{\sim}{j}\neq\underset{\sim}{j}'$, $X_{\underset{\sim}{i}\underset{\sim}{j}}=X_{\underset{\sim}{i}\underset{\sim}{j}'}=x$; $\rho_{A}^{x,x^{c}}=\mathrm{Cor}(Y_{\underset{\sim}{i}\underset{\sim}{j}},Y_{\underset{\sim}{i}\underset{\sim}{j}'})$ for $\underset{\sim}{i}=1,\ldots,F_{A}$, $\underset{\sim}{j}>1$, $\underset{\sim}{j}'>1$, $\underset{\sim}{j}\neq\underset{\sim}{j}'$, $X_{\underset{\sim}{i}\underset{\sim}{j}}=x$, $X_{\underset{\sim}{i}\underset{\sim}{j}'}\neq x$; and $\rho_{U}^{x,x^{c}}=\mathrm{Cor}(Y_{\underset{\sim}{i}\underset{\sim}{j}},Y_{\underset{\sim}{i}\underset{\sim}{j}'})$ for $\underset{\sim}{i}=F_{A}+1,\ldots,F_{A}+F_{U}$, $\underset{\sim}{j},\underset{\sim}{j}'>1$, $\underset{\sim}{j}\neq\underset{\sim}{j}'$, $X_{\underset{\sim}{i}\underset{\sim}{j}}=x$, $X_{\underset{\sim}{i}\underset{\sim}{j}'}\neq x$.

In practice, the population parameters in Equations (C.1) and (C.2) are replaced with estimates, which yields estimated standard errors that we refer to as $\hat{\mathrm{se}}(\hat{\pi})$ and $\hat{\mathrm{se}}(\hat{\pi}^{x})$, respectively. Although estimating the parameters π, π_{A}, π_{U}, π_{A}^{x}, and π_{U}^{x} will not require additional calculation because they appear in Equations (2.1) and (2.2), the parameters ρ_{A}, ρ_{U}, ρ_{A}^{x}, ρ_{U}^{x}, $\rho_{A}^{x,x^{c}}$, and $\rho_{U}^{x,x^{c}}$ will need to be calculated from the data using Pearson correlations.

The estimated quantities $\hat{\pi}$ and $\hat{\mathrm{se}}(\hat{\pi})$ could be used to form a Wald interval for π, which would take the form $\mathrm{CI}=[\hat{\pi}\pm z_{\alpha/2}\hat{se}(\hat{\pi})]$, where $z_{\alpha/2}$ is the $\alpha/2$ quantile of the standard nor-

mal distribution. However, for small population prevalences ($\pi < 0.2$), Wald intervals do not achieve their nominal coverage level because of the frequent occurrence of $[0, 0]$ intervals for samples with no diseased relatives. Various adjusted confidence intervals with improved coverage probabilities have been proposed for population proportions, including the Agresti-Coull (1998) interval, which has its roots in the work of Wilson (1927). Simply put, the Agresti-Coull interval improves the Wald interval's coverage properties by smoothing the proportion estimates and the estimated standard errors away from zero. Miao and Gastwirth (2004) have performed simulations to examine the performance of an Agresti-Coull-type interval (and various other intervals) for proportions estimated from moderately-sized samples containing dependent clusters. Since the Agresti-Coull-type interval appears to perform well in the simulations and, further, is easy to compute, we adopt confidence intervals based on the same concept.

For the overall prevalence, the $100 \cdot (1 - \alpha)\%$ interval takes the form:

$$\text{CI} = \widetilde{\pi} \pm z_{\alpha/2}\sqrt{\widetilde{se}} \tag{C.3}$$

where $\widetilde{\pi}$ and $\widetilde{se}$ are calculated using the formulas for $\hat{\pi}$ and $\hat{\text{se}}(\hat{\pi})$, respectively, with $p_{_A}$ replaced by

$$\widetilde{p}_{_A} = \frac{d_{_A} p_{_A} + (0.5 z^2_{\alpha/2})/100}{d_{_A} + (z^2_{\alpha/2})/100}$$

and $p_{_U}$ replaced by

$$\widetilde{p}_{_U} = \frac{d_{_U} p_{_U} + (0.5 z^2_{\alpha/2})/100}{d_{_U} + (z^2_{\alpha/2})/100}.$$

For the stratum-specific prevalence, π^x, the $100 \cdot (1 - \alpha)\%$ interval takes the form:

$$\text{CI} = \widetilde{\pi}^x \pm z_{\alpha/2}\sqrt{\widetilde{se}^x} \tag{C.4}$$

where $\widetilde{\pi}^x$ and $\widetilde{se}^x$ are calculated using the formulas for $\hat{\pi}^x$ and $\hat{\text{se}}(\hat{\pi}^x)$, respectively, with $p_{_A}$

replaced by $\widetilde{p}_{A}$, p_{U} replaced by $\widetilde{p}_{U}$, p_{A}^{x} replaced by

$$\widetilde{p}_{A}^{x} = \frac{d_{A}^{x} p_{A}^{x} + (0.5 z_{\alpha/2}^{2})/100}{d_{A}^{x} + (z_{\alpha/2}^{2})/100}$$

and p_{U}^{x} replaced by

$$\widetilde{p}_{U}^{x} = \frac{d_{U}^{x} p_{U}^{x} + (0.5 z_{\alpha/2}^{2})/100}{d_{U}^{x} + (z_{\alpha/2}^{2})/100}.$$

REFERENCES

AGRESTI, A. & COULL, B. A. (1998). Approximate is better than 'exact' for interval estimation of binomial proportions. *The American Statistician* **52**, 119–126.

FIRST, M. B., SPITZER, R. L., GIBBONS, M. & WILLIAMS, J. B. W. (1994). *Structured Clinical Interview for Axis I DSM-IV Disorders - Patient Edition (SCID-I/P, version 2.0).* New York: Biometrics Research Department, New York State Psychiatric Institute.

HARTLEY, H. O. & ROSS, A. (1954). Unbiased ratio estimators. *Nature* **174**, 270–271.

HUDSON, J. I., LAIRD, N. M. & BETENSKY, R. A. (2001). Multivariate logistic regression for familial aggregation of two disorders: I. Development of models and methods. *American Journal of Epidemiology* **153**, 501–505.

HUDSON, J. I., MANGWETH, B., POPE JR., H. G., DE COL, C., HAUSMANN, A., GUTWENIGER, S., LAIRD, N. M., BIEBL, W. & TSUANG, M. T. (2003). Family study of affective spectrum disorder. *Archives of General Psychiatry* **60**, 170–177.

JAVARAS, K. N., LAIRD, N. M. & HUDSON, J. I. (2007). Structural equation models for familial aggregation of binary outcomes in relatives of case-control-sampled probands .

KENDLER, K. S. & EATON, W. W. (1988). The proband method in psychiatric epidemiology: A bias associated with differences in family size. *Acta Psychiatrica Scandinavica* **77**, 511–514.

MIAO, W. & GASTWIRTH, J. L. (2004). The effect of dependence on confidence intervals for a population proportion. *The American Statistician* **58**, 124–130.

STATISTIK AUSTRIA (2003). Population stastistics: Tables for population 2003. http://www.statistik.at/englisch/results/population.

STRÖMGREN, E. (1948). Social surveys. *Journal of Mental Science* **94**, 266–276.

WILSON, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* **22**, 209–212.

WITTCHEN, H. U., ZAUDIG, M., SCHRAMM, E., SPENGLER, P., MOMBOUR, W., KLUG, J. & HORN, R. (1996). *Das Strukturierte Klinische Interviewnach DSM-IV.* Beltz: Weinheim.

Table 1: Number of Relatives With (Without) Major Depressive Disorder*

Proband Disease Status	Sex of Relatives	
	Male	Female
Case	8 (65)	25 (80)
Control	4 (69)	8 (71)

* MDD was diagnosed by interviewing probands and their relatives using the German translation (Wittchen et al. 1996) of the Structured Clinical Interview for DSM-IV (First et al. 1994).

Table 2: Components of the Bias of p_{A} When Assumption (ii.) Does and Does Not Hold

Term	Value (for $F_{A} = 64$ and $s = 1$)	
	Population 1 $\text{Cor}(N_i, \frac{N_i^A}{N_i}) \approx 0$	Population 2 $\text{Cor}(N_i, \frac{N_i^A}{N_i}) \approx -0.19$
$\text{Cor}\left(p_{A}, \sum_{i=1}^{F} \sum_{\{j:\, j \in \mathcal{L}_i\}} \delta_{ij}^{A}\right)$	0.001459	-0.137009
$\text{SD}(p_{A})$	0.033318	0.038691
$\text{CV}(\sum_{i=1}^{F} \sum_{\{j:\, j \in \mathcal{L}_i\}} \delta_{ij}^{A})$	0.058140	0.054310
Percentage bias of p_{A} †	-0.0017%	0.173%

† Percentage bias of p_{A} equals $100 \cdot \left(\frac{Ep_{A} - \pi_{A}}{\pi_{A}}\right)$, where $\pi_{A} \approx 0.16$; and where

$$Ep_{A} - \pi_{A} = \text{Cor}\left(p_{A}, \sum_{i=1}^{F} \sum_{\{j:\, j \in \mathcal{L}_i\}} \delta_{ij}^{A}\right) \cdot \text{SD}(p_{A}) \cdot \text{CV}\left(\sum_{i=1}^{F} \sum_{\{j:\, j \in \mathcal{L}_i\}} \delta_{ij}^{A}\right).$$

Table 3: Components of the Bias of p_U When Assumption (ii.) Does and Does Not Hold

Term	Value (for $F_U = 58$ and $s = 1$)	
	Population 1 $\text{Cor}(N_i, \frac{N_i^A}{N_i}) \approx 0$	Population 2 $\text{Cor}(N_i, \frac{N_i^A}{N_i}) \approx -0.19$
$\text{Cor}\left(p_U, \sum_{i=1}^{F} \sum_{\{j:\, j \in \mathcal{L}_i\}} \delta_{ij}^U\right)$	-0.008865	-0.098235
$\text{SD}(p_U)$	0.025768	0.020937
$\text{CV}(\sum_{i=1}^{F} \sum_{\{j:\, j \in \mathcal{L}_i\}} \delta_{ij}^U)$	0.061170	0.060793
Percentage bias of p_U †	0.0018%	0.2281%

† Percentage bias of p_U equals $100 \cdot \left(\frac{Ep_U - \pi_U}{\pi_U}\right)$, where $\pi_U \approx 0.055$; and where

$$Ep_U - \pi_U = \text{Cor}\left(p_U, \sum_{i=1}^{F} \sum_{\{j:\, j \in \mathcal{L}_i\}} \delta_{ij}^U\right) \cdot \text{SD}(p_U) \cdot \text{CV}\left(\sum_{i=1}^{F} \sum_{\{j:\, j \in \mathcal{L}_i\}} \delta_{ij}^U\right).$$

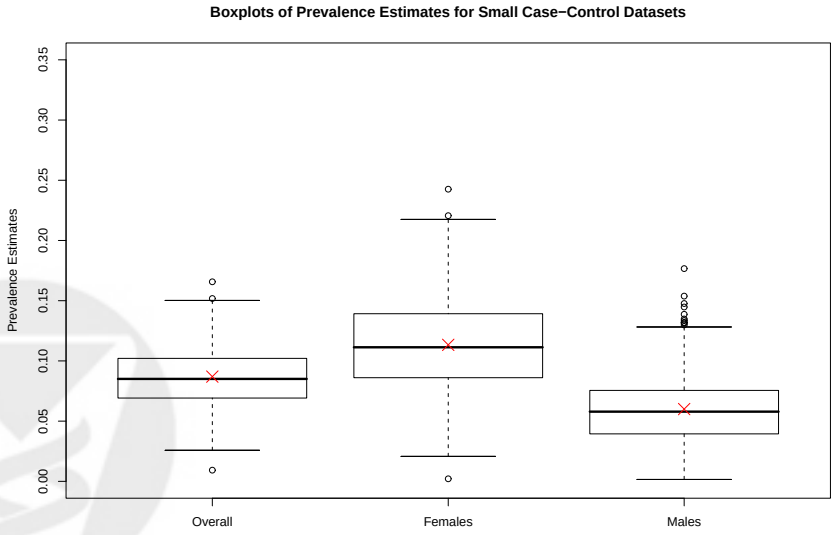

Figure 1. Boxplots of overall, male, and female prevalence estimates from $1,000$ small case-control family datasets with 64 case probands and 58 control probands. Xs indicate the corresponding population prevalences.