Web-based Supplementary Materials for "Estimating Disease Prevalence Using Relatives of Case and Control Probands" by Javaras, Laird, Hudson, & Ripley

Web Appendix A. Proof that $\widehat{\pi}$ 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

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

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

$$\pi = \frac{\pi_{_U}}{1 - \pi_{_A} + \pi_{_U}}, \quad \text{(A.2)}$$

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

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

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

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

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

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

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

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

The estimator $\widehat{\pi}$ 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:

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

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

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

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:

$$p_A \;=\; \frac{\displaystyle\sum_{i=1}^{F}\sum_{\{j:\, j\in J_i^*\}}\delta_{ij}^A\,\mathrm{I}(Y_{ij}=1)}{\displaystyle\sum_{i=1}^{F}\sum_{\{j:\, j\in J_i^*\}}\delta_{ij}^A}$$

where $J_i^* = \{j : 1 \le j \le N_i \text{ and } r_i(j) \ne 1\}$, 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:

$$\begin{aligned}
\mathrm{Cov}\Big(p_{A}, \sum_{i=1}^{F} \sum_{\{j:\, j\in J_i^*\}} \delta_{ij}^{A}\Big) &= \mathrm{E}\Big(\sum_{i=1}^{F} \sum_{\{j:\, j\in J_i^*\}} \delta_{ij}^{A}\mathrm{I}(Y_{ij}=1)\Big) - \mathrm{E}p_{A}\cdot \mathrm{E}\Big(\sum_{i=1}^{F} \sum_{\{j:\, j\in J_i^*\}} \delta_{ij}^{A}\Big) \\
&= \sum_{i=1}^{F} \sum_{\{j:\, j\in J_i^*\}} \mathrm{E}(\delta_{ij}^{A})\mathrm{I}(Y_{ij}=1) - \mathrm{E}p_{A}\cdot \sum_{i=1}^{F} \sum_{\{j:\, j\in J_i^*\}} \mathrm{E}(\delta_{ij}^{A}) \\
&= \sum_{i=1}^{F} \sum_{\{j:\, j\in J_i^*\}} \mathrm{E}\Big(\mathrm{E}\left(\delta_{ij}^{A}|\delta_{i}^{A}\right)\Big)\mathrm{I}(Y_{ij}=1) - \mathrm{E}p_{A}\cdot \sum_{i=1}^{F} \sum_{\{j:\, j\in J_i^*\}} \mathrm{E}\Big(\mathrm{E}\left(\delta_{ij}^{A}|\delta_{i}^{A}\right)\Big).
\end{aligned}$$

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

$$\begin{aligned}
\mathrm{Cov}\Big(p_{A}, \sum_{i=1}^{F} \sum_{\{j:\, j\in J_i^*\}} \delta_{ij}^{A}\Big) &= \sum_{i=1}^{F} \sum_{\{j:\, j\in J_i^*\}} \mathrm{E}\left(s\delta_{i}^{A} + 0(1-\delta_{i}^{A})\right)\mathrm{I}(Y_{ij}=1) \\
&\quad - \mathrm{E}p_{A}\cdot \sum_{i=1}^{F} \sum_{\{j:\, j\in J_i^*\}} \mathrm{E}\left(s\delta_{i}^{A} + 0(1-\delta_{i}^{A})\right) \\
&= s\sum_{i=1}^{F} \mathrm{E}(\delta_{i}^{A}) \sum_{\{j:\, j\in J_i^*\}} \mathrm{I}(Y_{ij}=1) - \mathrm{E}p_{A}\cdot s\sum_{i=1}^{F} \mathrm{E}(\delta_{i}^{A}) \sum_{\{j:\, j\in J_i^*\}} 1 \\
&= s\sum_{i=1}^{F} \mathrm{E}(\delta_{i}^{A})(N_{i}^{A}-1) - \mathrm{E}p_{A}\cdot s\sum_{i=1}^{F} \mathrm{E}(\delta_{i}^{A})(N_{i}-1) \\
&= s\,F_{A}\,s_{A}\sum_{i=1}^{F} N_{i}^{A}(N_{i}^{A}-1) - \mathrm{E}p_{A}\cdot s\,F_{A}\,s_{A}\sum_{i=1}^{F} N_{i}^{A}(N_{i}-1)
\end{aligned}$$

where the third line follows because 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

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

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

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

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

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

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

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 $\mathrm{CV}\Big(\sum_{i=1}^{F} \sum_{\{j:\, j \in J_i^*\}} \delta_{ij}^A\Big)$ 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, $\mathrm{Cor}\Big(p_{A}, \sum_{i=1}^{F} \sum_{\{j:\, j \in J_i^*\}} \delta_{ij}^A\Big)$ 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 experiments in Section 4, where the lifetime prevalence of disease was set to 11.5% for females from all families and 5.9% for males from all families. The second is a population created in an identical fashion, except that the prevalence of disease was set to 16.5% for females and 7.9% for males from families with three or fewer members, and to 6.5% for females and 3.9% for males from families with four or more 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 $\sum_{i=1}^{F} \sum_{\{j:\, j \in J_i^*\}} \delta_{ij}^A$ for each dataset and then used the resulting values to estimate the three multiplicands in (A.8) for that population. Web Table 1 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 Web Table 1 reveals that $\text{Cor}\left(p_A, \sum_{i=1}^{F} \sum_{\{j:\, j \in J_i^*\}} \delta_{ij}^A\right)$ 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 Web Table 1 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 $\text{Cor}\left(p_A, \sum_{i=1}^{F} \sum_{\{j:\, j \in J_i^*\}} \delta_{ij}^A\right)$.

We can use the approach employed above to obtain an analogous expression for the bias of p_U as an estimator for π_U

$$\text{E}p_U - \pi_U \;\; = \;\; -\,\text{Cor}\left(p_U, \sum_{i=1}^{F} \sum_{\{j:\, j \in J_i^*\}} \delta_{ij}^U\right) \cdot \text{SD}\big(p_U\big) \cdot \text{CV}\Big(\sum_{i=1}^{F} \sum_{\{j:\, j \in J_i^*\}} \delta_{ij}^U\Big) \quad \text{(A.9)}$$

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, Web Table 2 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

Web Table 1

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 $\text{Cor}(N_i, N_i^A/N_i) \approx 0$	Population 2 $\text{Cor}(N_i, N_i^A/N_i) \approx -0.19$
$\text{Cor}\left(p_A, \sum_{i=1}^{F} \sum_{\{j:\, j \in J_i^*\}} \delta_{ij}^A\right)$	0.001459	-0.137009
$\text{SD}(p_A)$	0.033318	0.038691
$\text{CV}(\sum_{i=1}^{F} \sum_{\{j:\, j \in J_i^*\}} \delta_{ij}^A)$	0.058140	0.054310
Percentage bias of p_A †	-0.0017%	0.173%

† Percentage bias of p_A equals $100 \cdot (Ep_A - \pi_A)/\pi_A$, where $\pi_A \approx 0.16$; and where

$$Ep_A - \pi_A = \text{Cor}\left(p_A, \sum_{i=1}^{F} \sum_{\{j:\, j \in J_i^*\}} \delta_{ij}^A\right) \cdot \text{SD}\left(p_A\right) \cdot \text{CV}\left(\sum_{i=1}^{F} \sum_{\{j:\, j \in J_i^*\}} \delta_{ij}^A\right).$$

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, $\widehat{\pi}$ 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 $\text{Cov}(p_A, p_U) = 0$. Next, note that the numerators of the first two second-order terms in (A.6) can be re-written as $\left[Ep_U(1 - Ep_U)/\sum_{i^*=F_A+1}^{F_A+F_U}(n_{i^*} - 1)\right](1 - Ep_A)$ and $\left[Ep_A(1 - Ep_A)/\sum_{i^*=1}^{F_A}(n_{i^*} - 1)\right]Ep_U$, 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

Web Table 2

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 $\text{Cor}(N_i, N_i^A/N_i) \approx 0$	Population 2 $\text{Cor}(N_i, N_i^A/N_i) \approx -0.19$
$\text{Cor}\left(p_U, \sum_{i=1}^{F} \sum_{\{j:\, j \in J_i^*\}} \delta_{ij}^U\right)$	-0.008865	-0.098235
$\text{SD}(p_U)$	0.025768	0.020937
$\text{CV}(\sum_{i=1}^{F} \sum_{\{j:\, j \in J_i^*\}} \delta_{ij}^U)$	0.061170	0.060793
Percentage bias of p_U †	0.0018%	0.2281%

† Percentage bias of p_U equals $100 \cdot (Ep_U - \pi_U)/\pi_U$, where $\pi_U \approx 0.055$; and where

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

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 experiments in Section 4, where $\widehat{\pi}$ 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. ■

Web Appendix B. Proof that $\widehat{\pi}^x$ 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

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

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

$$\widehat{\pi}^x = p_{A}^{x}\widehat{\pi}+p_{U}^{x}(1-\widehat{\pi}), \qquad \text{(B.5)}$$

where

$$p_{A}^{x}=\frac{\displaystyle\sum_{i^*=1}^{F_A}\sum_{j^*=2}^{n_{i^*}}\mathrm{I}(X_{i^*j^*}=x)\mathrm{I}(Y_{i^*j^*}=1)}{\displaystyle\sum_{i^*=1}^{F_A}\sum_{j^*=2}^{n_{i^*}}\mathrm{I}(X_{i^*j^*}=x)}$$

and

$$p_{U}^{x}=\frac{\displaystyle\sum_{i^*=F_A+1}^{F_A+F_U}\sum_{i^*=2}^{n_{i^*}}\mathrm{I}(X_{i^*j^*}=x)\mathrm{I}(Y_{i^*j^*}=1)}{\displaystyle\sum_{i^*=F_A+1}^{F_A+F_U}\sum_{i^*=2}^{n_{i^*}}\mathrm{I}(X_{i^*j^*}=x)}.$$

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

around $\mathrm{E}\widehat{\pi}$, $\mathrm{E}p_{U}^{x}$ and $\mathrm{E}p_{A}^{x}$:

$$
\begin{aligned}
\widehat{\pi}^{x} \;\approx\; & \left(\mathrm{E}p_{A}^{x}\mathrm{E}\widehat{\pi} + \mathrm{E}p_{U}^{x}(1-\mathrm{E}\widehat{\pi})\right) \\
& + (\widehat{\pi} - \mathrm{E}\widehat{\pi})\left.\frac{\partial \widehat{\pi}^{x}}{\partial \widehat{\pi}}\right|_{\mathrm{E}\widehat{\pi},\mathrm{E}p_{U}^{x},\mathrm{E}p_{A}^{x}} + \left(p_{U}^{x} - \mathrm{E}p_{U}^{x}\right)\left.\frac{\partial \widehat{\pi}^{x}}{\partial p_{U}^{x}}\right|_{\mathrm{E}\widehat{\pi},\mathrm{E}p_{U}^{x},\mathrm{E}p_{A}^{x}} + \left(p_{A}^{x} - \mathrm{E}p_{A}^{x}\right)\left.\frac{\partial \widehat{\pi}^{x}}{\partial p_{A}^{x}}\right|_{\mathrm{E}\widehat{\pi},\mathrm{E}p_{U}^{x},\mathrm{E}p_{A}^{x}} \\
& + \frac{1}{2}(\widehat{\pi} - \mathrm{E}\widehat{\pi})^{2}\left.\frac{\partial^{2}\widehat{\pi}^{x}}{\partial \widehat{\pi}^{2}}\right|_{\mathrm{E}\widehat{\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}\widehat{\pi}^{x}}{\partial p_{U}^{x\,2}}\right|_{\mathrm{E}\widehat{\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}\widehat{\pi}^{x}}{\partial p_{A}^{x\,2}}\right|_{\mathrm{E}\widehat{\pi},\mathrm{E}p_{U}^{x},\mathrm{E}p_{A}^{x}} \\
& + (\widehat{\pi} - \mathrm{E}\widehat{\pi})\left(p_{A}^{x} - \mathrm{E}p_{A}^{x}\right)\left.\frac{\partial^{2}\widehat{\pi}^{x}}{\partial \widehat{\pi}\partial p_{A}^{x}}\right|_{\mathrm{E}\widehat{\pi},\mathrm{E}p_{U}^{x},\mathrm{E}p_{A}^{x}} + (\widehat{\pi} - \mathrm{E}\widehat{\pi})\left(p_{U}^{x} - \mathrm{E}p_{U}^{x}\right)\left.\frac{\partial^{2}\widehat{\pi}^{x}}{\partial \widehat{\pi}\partial p_{U}^{x}}\right|_{\mathrm{E}\widehat{\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}\widehat{\pi^{x}}}{\partial p_{U}^{x}\partial p_{A}^{x}}\right|_{\mathrm{E}\widehat{\pi},\mathrm{E}p_{U}^{x},\mathrm{E}p_{A}^{x}}.
\end{aligned}
$$

Taking the expectation of both sides of the above equation yields

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

We focus now on the leading term of the expectation of the Taylor expansion in (B.6): $\mathrm{E}p_{A}^{x}\mathrm{E}\widehat{\pi} + \mathrm{E}p_{U}^{x}(1-\mathrm{E}\widehat{\pi})$. We have already shown in Web Appendix A that, under conditions (i)-(v), $\widehat{\pi}$ 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

$$
p_{A}^{x} \;=\; \frac{\displaystyle\sum_{i=1}^{F}\sum_{\{j:\, j\in J_{i}^{*}\}}\delta_{ij}^{Ax}\mathrm{I}(X_{ij}=x)\mathrm{I}(Y_{ij}=1) + \sum_{i=1}^{F}\sum_{\{j:\, j\in J_{i}^{*}\}}\delta_{ij}^{Ax^{c}}\mathrm{I}(X_{ij}=x)\mathrm{I}(Y_{ij}=1)}{\displaystyle\sum_{i=1}^{F}\sum_{\{j:\, j\in J_{i}^{*}\}}\delta_{ij}^{Ax}\mathrm{I}(X_{ij}=x) + \sum_{i=1}^{F}\sum_{\{j:\, j\in J_{i}^{*}\}}\delta_{ij}^{Ax^{c}}\mathrm{I}(X_{ij}=x)}
$$

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

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

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 Web Appendix A:

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

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

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

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

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

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

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

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

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}:

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

We can then use the same arguments made in Web 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 $\widehat{\pi}$ 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, $\widehat{\pi}^{x}$ is a slightly downwardly biased estimator for π_{x}. However, the results of the simulation experiments in Section 4 suggest that the bias introduced by the leading term and also the higher order terms in (B.6) is very small. ■

Web Appendix C. Standard Errors and Confidence Intervals for $\widehat{\pi}$ and $\widehat{\pi}^x$

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

$$\begin{aligned} \text{se}(\widehat{\pi}) &= \pi(1-\pi) \\ &\cdot \sqrt{\frac{\pi_{_A}}{d_{_A}(1-\pi_{_A})}\left[1+\frac{2\rho_{_A}}{d_{_A}}\sum_{i^*=1}^{F_{_A}}\binom{n_{i^*}-1}{2}\right]+\frac{(1-\pi_{_U})}{d_{_U}\pi_{_U}}\left[1+\frac{2\rho_{_U}}{d_{_U}}\sum_{i^*=F_{_A}+1}^{F_{_A}+F_{_U}}\binom{n_{i^*}-1}{2}\right]}, \end{aligned} \tag{C.1}$$

where $d_{_A} = \sum_{i^*=1}^{F_{_A}}\sum_{j^*=2}^{n_{i^*}} 1$ and $d_{_U} = \sum_{i^*=F_{_A}+1}^{F_{_A}+F_{_U}}\sum_{j^*=2}^{n_{i^*}} 1$; $\rho_{_A} = \text{Cor}(Y_{i^*j^*}, Y_{i^*j^{*\prime}})$ for $i^* = 1,\ldots,F_{_A}$, $j^* > 1$, $j^{*\prime} > 1$, and $j^* \neq j^{*\prime}$; and $\rho_{_U} = \text{Cor}(Y_{i^*j^*}, Y_{i^*j^{*\prime}})$ for $i^* = F_{_A}+1,\ldots,F_{_A}+F_{_U}$, $j^* > 1$, $j^{*\prime} > 1$, and $j^* \neq j^{*\prime}$. Note that the more the disease aggregates in families, the larger $\rho_{_A}$ and $\rho_{_U}$ will be and therefore the larger the standard error for $\widehat{\pi}$ will be.

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

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

where

$$a = \left[\begin{array}{cccc} \pi & (1-\pi) & \frac{(\pi_{_A}^x-\pi_{_U}^x)\pi^2}{\pi_{_U}} & \frac{(\pi_{_A}^x-\pi_{_U}^x)(1-\pi)^2}{1-\pi_{_A}} \end{array}\right]; \tag{C.2.a}$$

and

$$\Sigma = \left[\begin{array}{cccc} \sigma_{1,1} & 0 & \sigma_{1,3} & 0 \\ 0 & \sigma_{2,2} & 0 & \sigma_{2,4} \\ \sigma_{1,3} & 0 & \sigma_{3,3} & 0 \\ 0 & \sigma_{2,4} & 0 & \sigma_{4,4} \end{array}\right] \tag{C.2.b}$$

with

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

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

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

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

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

and

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

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

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 $\widehat{\mathrm{se}}(\widehat{\pi})$ and $\widehat{\mathrm{se}}(\widehat{\pi}^{x})$, 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 (1) and (2), the parameters ρ_{A}, ρ_{U}, ρ_{A}^{x}, ρ_{U}^{x}, ρ_{A}^{x,x^c}, and ρ_{U}^{x,x^c} will need to be estimated. Estimates can be calculated from the data using Pearson correlations for all relevant pairs of relatives or instead can be based on prior information. In situations where the data do not contain sufficient information to estimate the correlations and there is no prior information about them, the value 0.30 can be used as an extremely conservative estimate (by comparison, the correlations are approximately 0.10 for a disease with heritability and prevalence similar to the Austrian case-control family study). In our simulation experiments, we calculated the correlations from the data whenever possible and otherwise set them to 0.30.

The estimated quantities $\widehat{\pi}$ and $\widehat{\mathrm{se}}(\widehat{\pi})$ could be used to form a Wald interval for π, which

would take the form CI $= [\widehat{\pi} \pm z_{\alpha/2}\widehat{se}(\widehat{\pi})]$, where $z_{\alpha/2}$ is the $\alpha/2$ quantile of the standard normal 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 $\widehat{\pi}$ and $\widehat{\text{se}}(\widehat{\pi})$, respectively, with $p_{_A}$ replaced by

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

and $p_{_U}$ replaced by

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

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

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

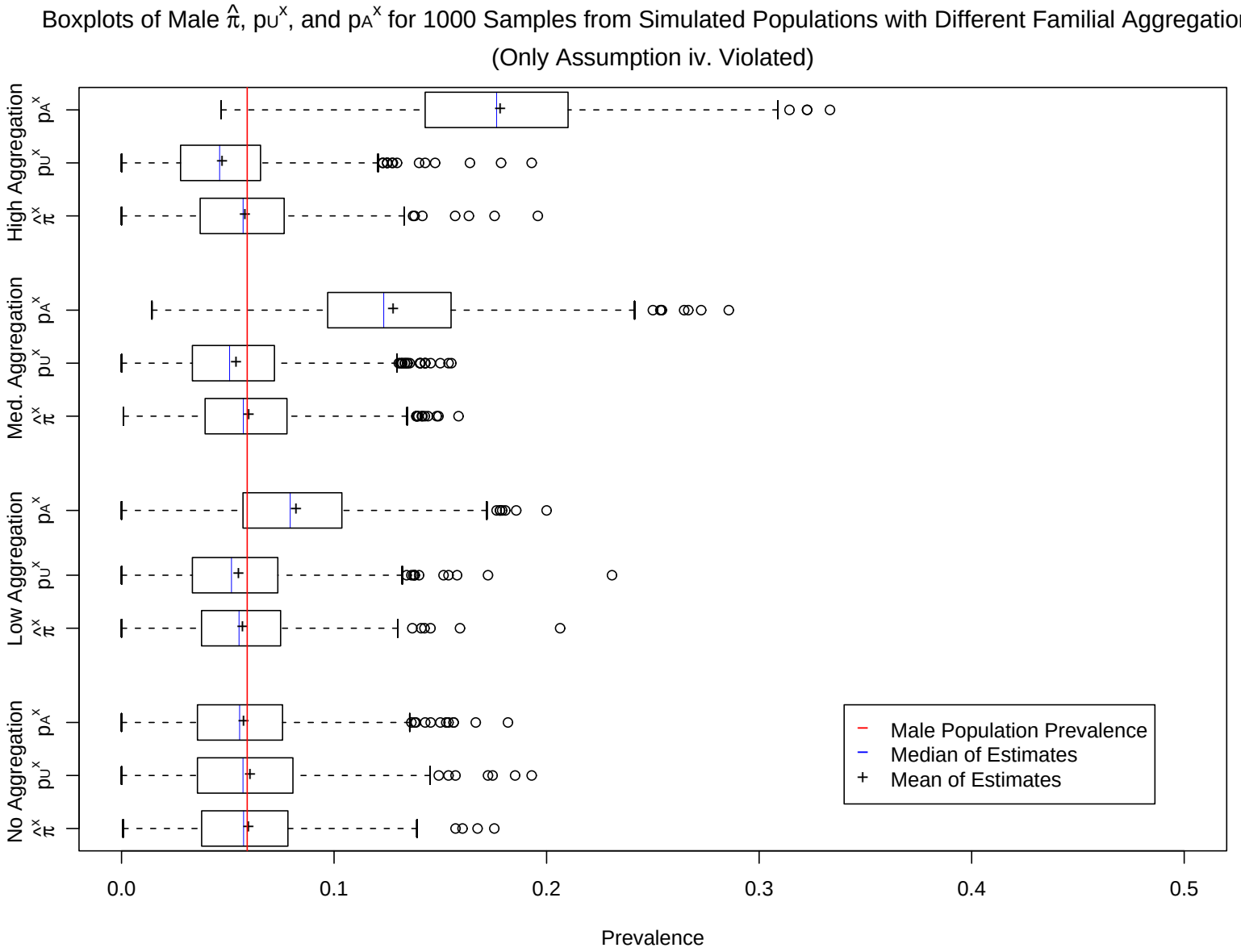
where $\widetilde{\pi}^x$ and $\widetilde{se}^x$ are calculated using the formulas for $\widehat{\pi}^x$ and $\widehat{\text{se}}(\widehat{\pi}^x)$, respectively, with $p_{_A}$ replaced by $\widetilde{p}_{_A}$, $p_{_U}$ replaced by $\widetilde{p}_{_U}$, $p^x_{_A}$ replaced by

$$\widetilde{p}^x_{_A} = \frac{d^x_{_A}p^x_{_A} + 0.5z^2_{\alpha/2}}{d^x_{_A} + z^2_{\alpha/2}}$$

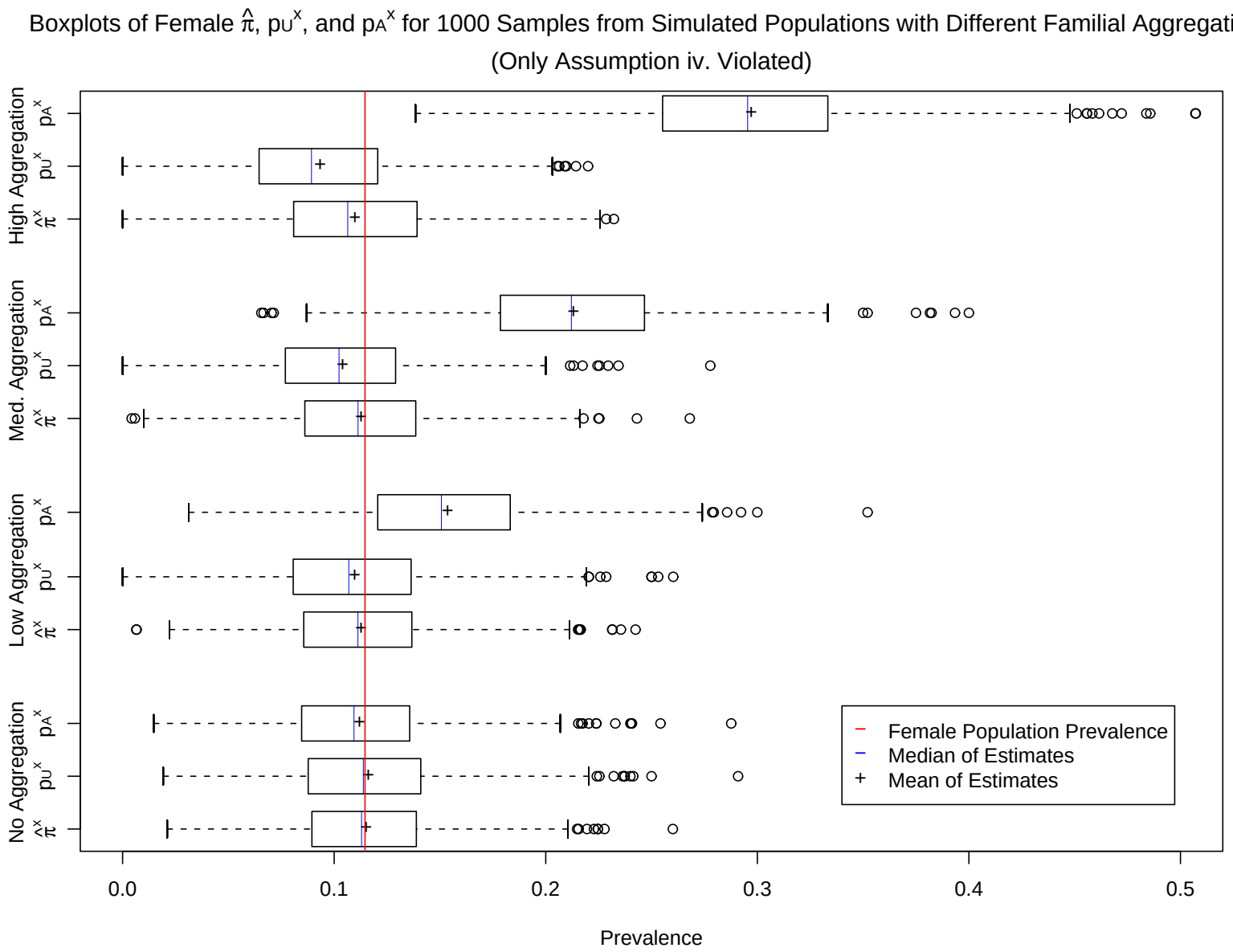
and p_U^x replaced by

$$\widetilde{p}_U^x = \frac{d_U^x p_U^x + 0.5 z_{\alpha/2}^2}{d_U^x + z_{\alpha/2}^2}.$$

Of course, if the lower (upper) bound of the confidence interval in either (C.3) or (C.4) turns out to be less (greater) than 0 (1), then it should be replaced with 0 (1).



Web Figure 1: Boxplots of male $\widehat{\pi}^x$, p_U^x, and p_A^x values calculated for 1000 samples drawn from four different populations with varying degrees of disease familiality.



Web Figure 2: Boxplots of female $\widehat{\pi}^x$, p_U^x, and p_A^x values calculated for 1000 samples drawn from four different populations with varying degrees of disease familiality.