7-7-2006

ON THE POTENTIAL FOR ILL-LOGIC WITH LOGICALLY DEFINED OUTCOMES

Xianbin Li
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, xli@jhsph.edu

Brian S. Caffo
The Johns Hopkins Bloomberg School of Public Health

Daniel O. Scharfstein
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

On the Potential for Ill-logic with Logically Defined Outcomes

Xianbin Li[†,‡], Brian Caffo[†] and Daniel Scharfstein[†]

† Department of Biostatistics

‡ Department of Population and Family Health Sciences

Johns Hopkins Bloomberg School of Public Health

Baltimore, MD 21205

Email: xli@jhsph.edu, bcaffo@jhsph.edu, dscharf@jhsph.edu

July 6, 2006

Abstract

Logically defined outcomes are commonly used in medical diagnoses and epidemiological research. When missing values in the original outcomes exist, the method of handling the missingness can have unintended consequences, even if the original outcomes are missing completely at random. Complicating the issue is that the default behavior of standard statistical packages yields different results. In this paper, we consider two binary original outcomes, which are missing completely at random. For estimating the prevalence of a logically defined "or" outcome, we discuss the properties of four estimators: complete case estimator, all-available case estimator, maximum likelihood estimator (MLE), and moment-based estimator. With the exception of the all-available case estimator, the estimators are consistent.

A simulation study is conducted to evaluate the finite sample performance of the four estimators and an analysis of hypertension data from the Sleep Heart Health Study is presented.

Keywords: All-Available Case Estimator, Complete-Case Estimator, Hypertension, Maximum Likelihood Estimator, Missing Data, Moment-Based Estimator

1 Introduction

Logically defined outcomes arise frequently in biomedical practice and research. For example, in epidemiologic studies, a common definition of hypertension requires systolic blood pressure over 140 mmHg, diastolic blood pressure over 90 mmHg or use of antihypertensive medications (see Nieto et al., 2000; Peppard et al., 2000; Banks et al., 2006, for example). Estimation of the prevalence of the logically defined outcome is straightforward, if there is no missing information in the original outcomes (in this example, the diagnosis criteria) or the missingness in the outcomes is completely concordant. However, when there is missing information in one or more of the original outcomes, the estimation of the prevalence may be complex. The most straightforward approach to addressing the missing data is to discard all logical outcomes where any of the original outcomes have missing values, referred to as the complete-case analysis. However, such an approach may discard known logical outcomes. For example, if a subject has missing blood pressure measurements, but is known to be taking anti-hypertensive medication, then he or she is hypertensive, as per the operational definition; hence their logical outcome is known, despite some of the original outcomes being missing. In this manuscript we investigate the utility and problems arising from using logical outcomes where some of the original outcomes are missing.

1.1 Mathematical Formulation

For ease of exposition, we only consider two binary (yes/no) outcomes, labeled $Y^{(1)}$ and $Y^{(2)}$. We define the associated observed data 0/1 indicators as $R^{(1)}$ and $R^{(2)}$, where $R^{(j)}$ equals one if $Y^{(j)}$ is observed. The logically defined outcome Y is 1 if $Y^{(1)} = 1$ or $Y^{(2)} = 1$;

otherwise it is 0. Mathematically,

$$Y = Y^{(1)}(1 - Y^{(2)}) + (1 - Y^{(1)})Y^{(2)} + Y^{(1)}Y^{(2)}.$$

Let π_{jk} be the probability that $Y^{(1)} = j$ and $Y^{(2)} = k$ and γ_{lm} indicate the probability of $R^{(1)} = l$ and $R^{(2)} = m$ (see Table 1), where $\sum_{j=0}^{1}\sum_{k=0}^{1}\pi_{jk} = 1$ and $\sum_{l=0}^{1}\sum_{m=0}^{1}\gamma_{lm} = 1$. We assume throughout that the original outcomes, $Y^{(1)}$ and $Y^{(2)}$, are independent of their observed data indicators, $R^{(1)}$ and $R^{(2)}$; that is, the missingness is completely at random (Rubin, 1976). However, the outcomes can be dependent, as well as the observed data indicators.

Of scientific interest is estimation of

$$\mu = P[Y = 1] = P[Y^{(1)} = 1 \text{ or } Y^{(2)} = 1] = \pi_{11} + \pi_{01} + \pi_{10} = \pi_{1+} + \pi_{+1} - \pi_{11} = 1 - \pi_{00}. \quad (1)$$

In what follows we discuss the impact of the choice of R, the observed data indicator for Y, on estimation of μ. A complete case analysis sets $R = R^{*}$ where

$$R^{*} = R^{(1)} \times R^{(2)}. \quad (2)$$

This approach discards all of the available information from when $Y^{(1)} = 1$ and $R^{(2)} = 0$, where Y is known to be 1 despite the missing $Y^{(2)}$ value, as well as all of the cases where $Y^{(2)} = 1$ and $R^{(1)} = 0$. The observed data indicator that uses all of the known values of Y sets $R = R^{\dagger}$, where

$$R^{\dagger} = R^{(1)}R^{(2)} + R^{(2)}(1 - R^{(1)})Y^{(2)} + R^{(1)}(1 - R^{(2)})Y^{(1)}. \quad (3)$$

This is the so-called all-available case analysis. While such an approach "seems" better, because it does not discard known outcomes, it must be noted that the observed data indicator now depends on the outcome, Y, so that this approach induces informative missingness, *even if the original data are missing completely at random*.

1.2 Implementation in Statistical Packages

It should be mentioned that the default behavior of some of the popular statistical program languages is not consistent. For example, the programs SAS and R (not to be confused with our observed data indicator, R) deal with missing values differently in standard usage. By default, SAS uses a complete case definition, as in (2), while the program R uses a all-available case definition like (3). Of course, either program can be made to exhibit the opposite behavior with appropriate care. Howeever, the main point is that many users are probably unaware of which of the two schemes their program implements.

1.3 Illustration

To illustrate the distinction between the complete case and all-available case estimators, suppose that we observe the 16 patterns with the frequencies given in Table 2. The estimated prevalence using only the complete cases (2) is

$$\frac{n_1 + n_5 + n_9 + n_{11}}{n_1 + n_5 + n_9 + n_{11} + n_{13}};$$

whereas, using all-available cases (3), it is

$$\frac{n_1 + n_2 + n_3 + n_5 + n_6 + n_9 + n_{11}}{n_1 + n_2 + n_3 + n_5 + n_6 + n_9 + n_{11} + n_{13}}.$$

Consider an extreme scenario in which n_1, n_5, n_9 and n_{11} are equal 0 while n_2, n_3 and n_6 are large. Then, the complete case estimate is 0 while the all-available case estimate is close to 1. Such data would occur if there was a high degree of negative correlation in the individual observed data indicators, implying largely discordant missingness in the two responses.

1.4 Outline

The paper is organized as follows. In Section 2, we introduce four estimators of μ: complete case, all-available case, moment-based, maximum-likelihood. We also derive their asymptotic properties. In Section 3, we present a simulation study to evaluate the performance of these estimators in finite samples. Section 4 is devoted to an analysis of hypertension data from the Sleep Heart Health Study. The final section is devoted to a discussion.

2 Four Estimators and their Asymptotic Properties

We assume that we have n independent and identically distributed copies of $O = (R^{(1)}, R^{(2)}, R^{(1)} \cdot Y^{(1)}, R^{(2)} \cdot Y^{(2)})$. We reserve the subscript i to indicate individuals when necessary. We focus on the following four estimators of μ:

$$\begin{aligned}
\text{Complete case} \quad & \hat{\mu}_c && = \frac{\sum_{i=1}^{n} Y_i R_i^*}{\sum_{i=1}^{n} R_i^*} \\
\text{All-available case} \quad & \hat{\mu}_a && = \frac{\sum_{i=1}^{n} Y_i R_i^\dagger}{\sum_{i=1}^{n} R_i^\dagger} \\
\text{Moment based} \quad & \hat{\mu}_m && = \frac{\sum_{i=1}^{n} R_i^{(1)} Y_i^{(1)}}{\sum_{i=1}^{n} R_i^{(1)}} + \frac{\sum_{i=1}^{n} R_i^{(2)} Y_i^{(2)}}{\sum_{i=1}^{n} R_i^{(2)}} - \frac{\sum_{i=1}^{n} R_i^{(1)} R_i^{(2)} Y_i^{(1)} Y_i^{(2)}}{\sum_{i=1}^{n} R_i^{(1)} R_i^{(2)}} \\
\text{Maximum likelihood} \quad & \hat{\mu}_{ML} &&
\end{aligned}$$

The first two estimators are simple averages of the observed values of Y; $\hat{\mu}_c$ uses only the instances where both of the original outcomes are observed while $\hat{\mu}_a$ uses all of the available logical outcomes. While the complete case estimator is consistent, the all-available case estimator converges in probability (see Appendix B) to

$$\frac{E[R^\dagger Y]}{E[R^\dagger]} = \frac{\gamma_{11}\mu + \gamma_{10}\pi_{1+} + \gamma_{01}\pi_{+1}}{\gamma_{11} + \gamma_{10}\pi_{1+} + \gamma_{01}\pi_{+1}} = \mu + \pi_{00}\frac{\gamma_{10}\pi_{1+} + \gamma_{01}\pi_{+1}}{\gamma_{11} + \gamma_{10}\pi_{1+} + \gamma_{01}\pi_{+1}},$$

The second term indicates non-negative bias; it is zero if and only if $\gamma_{11} = 1$ (i.e., no missing data) or $\mu = 1$ (i.e., the probability that both $Y^{(1)}$ and $Y^{(2)}$ are both zero is zero, $\pi_{00} = 0$). Notice that the bias converges to one when γ_{11} and π_{00} converge to one. As ratio estimators, these estimators are asymptotically normal with an asymptotic variance of the form:

$$\frac{\mathrm{Var}[RY]}{\{\mathrm{E}[R]\}^2} - \frac{2\mathrm{E}[RY]\mathrm{Cov}[R, RY]}{\{\mathrm{E}[R]\}^3} + \frac{\{\mathrm{E}[RY]\}^2\mathrm{Var}[R]}{\{\mathrm{E}[R]\}^4}, \tag{4}$$

where $R = R^*$ for the complete-case estimator and $R = R^\dagger$ for the all-available case estimator. In the complete case setting, this expression simplifies to $\frac{\pi_{00}(1-\pi_{00})}{\gamma_{11}}$ (see Appendix A), which is the Bernoulli variance divided by the probability of observing a complete case. The corresponding form for $\hat{\mu}_a$ is more complicated, and is provided in Appendix B.

The moment-based estimator $\hat{\mu}_m$ is a direct estimator based on the fact that $\mu = \pi_{1+} + \pi_{+1} - \pi_{11}$. Because the first two terms depend only on the individual original outcomes, this estimate makes use of more information than the complete case estimator. It is consistent as the the first, second, and third terms converge in probability to π_{1+}, π_{+1}, and π_{11}, respectively. The estimator is asymptotically normal with an asymptotic variance given in Appendix C.

The final estimator is based on maximum likelihood. Since $(R^{(1)}, R^{(2)})$ is ancillary (Basu, 1977) for $\boldsymbol{\pi} = (\pi_{01}, \pi_{10}, \pi_{11})'$, the maximum likelihood estimator for $\boldsymbol{\pi}$ can be found by maximizing the conditional likelihood for the observed data given $(R^{(1)}, R^{(2)})$. The conditional likelihood contribution for an random individual with observed data O is

$$\begin{aligned} L(\boldsymbol{\pi}; O) &= \left[\pi_{11}^{Y^{(1)}Y^{(2)}} \pi_{10}^{Y^{(1)}(1-Y^{(2)})} \pi_{01}^{(1-Y^{(1)})Y^{(2)}} (1-\pi_{01}-\pi_{10}-\pi_{11})^{(1-Y^{(1)})(1-Y^{(2)})}\right]^{R^{(1)}R^{(2)}} \times \\ & \left[(\pi_{10}+\pi_{11})^{Y^{(1)}} (1-\pi_{10}-\pi_{11})^{(1-Y^{(1)})}\right]^{R^{(1)}(1-R^{(2)})} \times \\ & \left[(\pi_{01}+\pi_{11})^{Y^{(2)}} (1-\pi_{01}-\pi_{11})^{(1-Y^{(2)})}\right]^{(1-R^{(1)})R^{(2)}} \end{aligned}$$

The overall conditional likelihood is $\prod_{i=1}^{n} L(\boldsymbol{\pi}; O_i)$. The first, second, and third terms of the conditional likelihood function, $L(\boldsymbol{\pi}; O)$, are the contributions from observations where: both $Y^{(1)}$ and $Y^{(2)}$ are observed, $Y^{(1)}$ is available and $Y^{(2)}$ is missing, and $Y^{(1)}$ is missing and $Y^{(2)}$ is available, respectively. To obtain the maximum likelihood estimators of $\boldsymbol{\pi}$, one can maximize the likelihood numerically, using a quasi-Newton algorithm. It is useful to re-parameterize $\boldsymbol{\pi}$ in terms of $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)'$, where $\beta_1 = \log\{\pi_{10}/(1-\pi_{01}-\pi_{10}-\pi_{11})\}$, $\beta_2 = \log\{\pi_{01}/(1-\pi_{01}-\pi_{10}-\pi_{11})\}$, and $\beta_3 = \log\{\pi_{11}/(1-\pi_{01}-\pi_{10}-\pi_{11})\}$, to eliminate boundary constraints. Assuming that the solution lies within the interior of a compact set, the maximum likelihood estimate of $\boldsymbol{\pi}$, $\hat{\boldsymbol{\pi}} = (\hat{\pi}_{01}, \hat{\pi}_{10}, \hat{\pi}_{11})'$, will be consistent and asymptotically normal with asymptotic variance equal to the inverse of the Fisher information matrix (see Appendix E). By the invariance property, the maximum likelihood estimator of μ is $\hat{\mu}_{ML} = \hat{\pi}_{01} + \hat{\pi}_{10} + \hat{\pi}_{11}$. This estimator will be consistent and asymptotically normal with asymptotic variance found using the delta method (see Appendix E).

The contour plots in Figure 1 assume $\pi_{01} = \pi_{10}, \pi_{00} = \pi_{11}$ and $\gamma_{01} = \gamma_{10}, \gamma_{00} = \gamma_{11}$. The first row of Figure 1 shows the contours of the asymptotic variance of $\hat{\mu}_c$ and $\hat{\mu}_m$ relative to that of $\hat{\mu}_{ML}$. As expected, the MLE performs uniformly better than the complete-case and moment-based estimators. The contour in the second row of Figure 1 shows the asymptotic variance of $\hat{\mu}_m$ relative to that of $\hat{\mu}_c$. When all π's are equal to 0.25, the complete-cases estimator has the same variance as the moment-based estimator. The complete-case has higher variance than the moment-based estimator when $\pi_{01} = \pi_{10}$ is small and $\gamma_{01} = \gamma_{10}$ are near to 0.25. On the other hand, the complete-case estimator has lower variance than the moment-based estimator when $\pi_{01} = \pi_{10}$ is close to 0.5.

To further explore the asymptotic efficiency of the complete-case estimator relative to the moment-based estimator, in Appendix D, we present a general formula for the difference between the asymptotic variance of the moment-based estimator and the complete-case estimator. We prove a proposition that shows that, when there is some discordant missingness and the proportion of complete cases is small, the choice between the complete case and moment based estimators relies on whether it is better to estimate π_{00} or π_{11} with the complete cases. Specifically, the moment-based (complete-case) estimator is dramatically more efficient if π_{11} (π_{00}) is further from 0.5 than π_{00} (π_{11}).

3 Simulation Study

A simulation study was performed to evaluate the finite-sample biases and variances of the four estimates. Because our focus is on epidemiologic settings, large sample sizes of $n = 300, 400, 600, 800$, and $1,000$ were used; in each case $1,000$ Monte Carlo simulations were performed.

Estimated bias and mean squared error (MSE) results for the case when $n = 300$ are shown in Table 3 (simulations with $n = 400, 600, 800$ or 1000 yielded similar results). The first, second, third, and fourth panels (each consisting of four rows) show the results for scenarios of increasingly discordant outcomes and increased prevalence. Within each panel, the rows are ordered according to increasingly discordant missingness. By design, the diagonals in the two-by-two tables in Table 1, are set to be equal (i.e., $\pi_{01} = \pi_{10}, \pi_{00} =$

$\pi_{11}, \gamma_{01} = \gamma_{10}, \gamma_{00} = \gamma_{11}$). As expected, Table 3 shows that $\hat{\mu}_c$, $\hat{\mu}_m$, and $\hat{\mu}_{ML}$ have very little bias, while $\hat{\mu}_a$ can be quite biased. In addition, the maximum likelihood estimator has the smallest mean squared error. The moment-based estimator has smaller mean-squared error than the complete-case estimator when $\pi_{11} = \pi_{00}$ is 0.3 or 0.4. When π_{00} is 0.25 or 0.10, the mean squared error of the complete-case estimator is smaller than the moment-based estimator. In each panel, as the degree of discordant missingness increases, the mean squared error increases.

4 Hypertension

We use hypertension data from the Sleep Heart Health Study (see Quan et al., 1997) as an illustration. The Sleep Heart Health Study is a multi-center cohort study with participants recruited from the Atherosclerosis Risk in Communities Study, the Cardiovascular Health Study, the Framingham Heart Study, the Strong Heart Study, and the Tucson Health and Environment Study. Here (as in Peppard et al., 2000) hypertension in a subject is defined as the presence of high systolic or diastolic blood pressure measurements or if the subject is taking anti-hypertensive medications. Technically, the logical outcome is then the "or" operator applied to three variables. However, because they were recorded at the same time the missingness between the two blood pressure measurements was completely concordant. Therefore, we combine these into one measurement, "High BP". Table 4 shows the counts for blood pressure and medication status.

The estimated prevalences of hypertension are $\hat{\mu}_c = 0.548, \hat{\mu}_a = 0.549, \hat{\mu}_m = 0.549, \hat{\mu}_{ML} = 0.548$ with standard errors all near 0.007. Because the number of observations with missing outcomes is negligibly small, there is no difference between four different estimates. In order to illustrate our proposed estimates, we artificially induce missingness completely at random in this data set; a process that was replicated $1,000$ times. Table 5 shows the mean estimates and standard deviations of the estimates from three different missingness scenarios. The first three lines contain the results from data simulated with 20, 40, 60% completely random missingness in $Y^{(1)}$ and $Y^{(2)}$, respectively, above and beyond the existing missingness in the original data set. The estimates $\hat{\mu}_c$, $\hat{\mu}_m$, and $\hat{\mu}_{ML}$ yield nearly identical average estimates, with the ML estimate having the smallest standard deviation. As the proportion of missingness increases, the standard deviations increase as well as the bias in $\hat{\mu}_a$. The final four lines show results from other interesting combinations of the four components of γ. Again the average of the complete case, moment-based and maximum likelihood estimators are very close with $\hat{\mu}_{ML}$ having the smallest standard deviation. The estimate $\hat{\mu}_a$ can have very severe bias, especially when the missingness is very discordant.

Finally, we compare the normalized profile likelihood functions between the complete case conditional likelihood and the full conditional likelihood from one simulation with 60 percent random missingness in each outcome. The MLEs from the full conditional likelihood function are: $\hat{\pi}_{01} = 0.128$, $\hat{\pi}_{10} = 0.233$, and $\hat{\pi}_{11} = 0.176$. Therefore, the maximum likelihood estimate is $\hat{\mu}_{ML} = 0.537$. The profile likelihood functions were obtained by performing a grid search over 1000 × 1000 targeted values of π_{01} and π_{10} for each fixed value of π_{00}. Fig 2 shows the normalized profile likelihood functions from the two data sets and associated 1/8 and 1/16 reference lines for the estimated prevalence of hypertension. The benefit of considering these likelihoods is the ability to visualize the additional evidence contained in the discordant missing cases.

5 Discussion

Logically defined outcomes are commonly used in medical diagnosis and epidemiological research. Without missing values in the original outcomes, the estimation of the prevalence of the logically defined outcomes is straightforward. However, when there are missing values in some of the original outcomes, the method of handling the missingness can have unintended consequences, even if the original outcomes are missing completely at random. We believe that this potential problem is largely unknown. Complicating the issue is that the default behavior of standard statistical packages yields different results.

In this manuscript, we considered two binary outcomes, which were assumed to be missing completely at random, and discussed four estimators of the prevalence of a logically defined “or” outcome. We derived the asymptotic properties of our estimators. The maximum likelihood estimator was shown to be the optimal choice, though it requires the use of numerical optimization techniques. Regardless, we would recommend its general use in these problems. We would hesitate to ever recommend the all-available case estimator, though it is probably the most commonly used in practice. This is especially true when the missing data patterns are particularly discordant. In the event where the missingness is largely concordant, all of the estimators are nearly identical.

In this manuscript we reduced the missing data problem to the simplest setting. For future work more complicated logical structures involving more than two original outcomes should be considered. Such structures arise frequently is medical research, such as in more general definitions of hypertension. In addition, regression models for logical outcomes that address the missing data issue, is also a potentially fruitful area for future research.

References

Banks, J., Marmot, M., Oldfield, Z., and Smith, J. P. (2006). Disease and Disadvantage in the United States and in England. *Journal of the American Medical Association*, 295(17):2037–2045.

Basu, D. (1977). On the elimination of nuisance parameters. *Journal of the American Statistical Association*, 72:355–366.

Nieto, J., MD, Young, T., Lind, B., Shahar, E., Samet, J., Redline, S., D'Agostino, R., Newman, A., Lebowitz, M., and Pickering, J. (2000). Association of sleep-disordered breathing, sleep apnea, and hypertension in a large community-based study. *Journal of the American Medical Association*, 283:1829–1836.

Peppard, P., Young, T., Palta, M., and Skatrud, J. (2000). Prospective study of the association between sleep-disordered breathing and hypertension. *New England Journal of Medicine*, 342(19):1378–1384.

Quan, S., Howard, B., Iber, C., Kiley, J., Nieto, F., O'Connor, G., Rapoport, D., Redline, S., Robbins, J., Samet, J., and Wahl, P. (1997). The Sleep Heart Health Study: design, rationale, and methods. *Sleep.*, 20(12):1077–1085.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63:581–592.

Appendix

Let $\boldsymbol{\xi} = (\boldsymbol{\pi}', \boldsymbol{\gamma}')$, where $\boldsymbol{\gamma} = (\gamma_{00}, \gamma_{01}, \gamma_{10}, \gamma_{11})'$.

The asymptotic variance for $\hat{\mu}_c$ and $\hat{\mu}_a$ is given in formula (4). To utilize this formula, it is sufficent to write expressions for $\mathrm{E}[R]$ and $\mathrm{E}[RY]$. This is because $\mathrm{Var}[R] = \mathrm{E}[R](1 - \mathrm{E}[R])$, $\mathrm{Var}[RY] = \mathrm{E}[RY](1 - \mathrm{E}[RY])$, and $\mathrm{Cov}[RY, R] = \mathrm{E}[RY](1 - \mathrm{E}[R])$.

A Asymptotic Variance of $\hat{\mu}_c$

$$\begin{aligned}
\mathrm{E}[R^*] &= \mathrm{E}[R^{(1)}R^{(2)}] = \gamma_{11} \\
\mathrm{E}[R^*Y] &= \mathrm{E}[R^*]\mathrm{E}[Y] = (1 - \pi_{00})\gamma_{11}
\end{aligned}$$

Therefore the asymptotic variance of $\hat{\mu}_c$ in Eq. (4) can be simplified to

$$\frac{\pi_{00}(1 - \pi_{00})}{\gamma_{11}}. \tag{5}$$

B Asymptotic Variance of $\hat{\mu}_a$

$$\begin{aligned}
\mathrm{E}[R^\dagger] &= \mathrm{E}[R^{(1)}R^{(2)} + R^{(1)}(1 - R^{(2)})Y^{(1)} + (1 - R^{(1)})R^{(2)}Y^{(2)}] \\
&= \gamma_{11} + \gamma_{10}\pi_{1+} + \gamma_{01}\pi_{+1} \\
\mathrm{E}[R^\dagger Y] &= \mathrm{E}[R^{(1)}R^{(2)}(Y^{(1)} + Y^{(2)} - Y^{(1)}Y^{(2)})] + \\
&\quad \mathrm{E}[R^{(1)}(1 - R^{(2)})Y^{(1)}] + \mathrm{E}[(1 - R^{(1)})R^{(2)}Y^{(2)}] \\
&= \gamma_{11}(\pi_{1+} + \pi_{+1} - \pi_{11}) + \gamma_{10}(\pi_{11} + \pi_{10}) + \gamma_{01}(\pi_{11} + \pi_{01})
\end{aligned}$$

The asymptotic variance of $\hat{\mu}_a$ in Eq. (4) can be easily computed, but is too long to display here.

C Asymptotic Variance of $\hat{\mu}_m$

Let

$$\boldsymbol{Z} = (R^{(1)}Y^{(1)}, R^{(1)}, R^{(2)}Y^{(2)}, R^{(2)}, R^{(1)}R^{(2)}Y^{(1)}Y^{(2)}, R^{(1)}R^{(2)})'$$

and

$$\begin{aligned}\boldsymbol{\mu_Z} &= (\mathrm{E}[R^{(1)}Y^{(1)}], \mathrm{E}[R^{(1)}], \mathrm{E}[R^{(2)}Y^{(2)}], \mathrm{E}[R^{(2)}], \mathrm{E}[R^{(1)}R^{(2)}Y^{(1)}Y^{(2)}], \mathrm{E}[R^{(1)}R^{(2)}])' \\ &= (\gamma_{1+}\pi_{1+}, \gamma_{1+}, \gamma_{+1}\pi_{+1}, \gamma_{+1}, \gamma_{11}\pi_{11}, \gamma_{11})\end{aligned}$$

By the multivariate central limit theorem, we know

$$\sqrt{n}\left(\overline{\boldsymbol{Z}} - \boldsymbol{\mu_Z}\right) \xrightarrow{D} MVN_6(0, \Sigma(\boldsymbol{\xi}))$$

where $\overline{\boldsymbol{Z}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i$, $\Sigma(\boldsymbol{\xi})$ is a 6×6 matrix whose row i, column j components are denoted by $\Sigma_{ij}(\boldsymbol{\xi})$, and

$$\begin{aligned}
\Sigma_{11}(\boldsymbol{\xi}) &= \mathrm{Var}[R^{(1)}Y^{(1)}] = \gamma_{1+}\pi_{1+}(1-\gamma_{1+}\pi_{1+}) \\
\Sigma_{22}(\boldsymbol{\xi}) &= \mathrm{Var}[R^{(1)}] = \gamma_{1+}(1-\gamma_{1+}) \\
\Sigma_{33}(\boldsymbol{\xi}) &= \mathrm{Var}[R^{(2)}Y^{(2)}] = \gamma_{+1}\pi_{+1}(1-\gamma_{+1}\pi_{+1}) \\
\Sigma_{44}(\boldsymbol{\xi}) &= \mathrm{Var}[R^{(2)}] = \gamma_{+1}(1-\gamma_{+1}) \\
\Sigma_{55}(\boldsymbol{\xi}) &= \mathrm{Var}[R^{(1)}R^{(2)}Y^{(1)}Y^{(2)}] = \gamma_{11}\pi_{11}(1-\gamma_{11}\pi_{11}) \\
\Sigma_{66}(\boldsymbol{\xi}) &= \mathrm{Var}[R^{(1)}R^{(2)}] = \gamma_{11}(1-\gamma_{11}) \\
\Sigma_{12}(\boldsymbol{\xi}) &= \Sigma_{21}(\boldsymbol{\xi}) = \mathrm{Cov}[R^{(1)}Y^{(1)}, R^{(1)}] = \gamma_{1+}\pi_{1+}(1-\gamma_{1+}) \\
\Sigma_{13}(\boldsymbol{\xi}) &= \Sigma_{31}(\boldsymbol{\xi}) = \mathrm{Cov}[R^{(1)}Y^{(1)}, R^{(2)}Y^{(2)}] = \gamma_{11}\pi_{11} - \gamma_{1+}\gamma_{+1}\pi_{1+}\pi_{+1} \\
\Sigma_{14}(\boldsymbol{\xi}) &= \Sigma_{41}(\boldsymbol{\xi}) = \mathrm{Cov}[R^{(1)}Y^{(1)}, R^{(2)}] = \gamma_{11}\pi_{1+} - \gamma_{1+}\gamma_{+1}\pi_{1+} \\
\Sigma_{15}(\boldsymbol{\xi}) &= \Sigma_{51}(\boldsymbol{\xi}) = \mathrm{Cov}[R^{(1)}Y^{(1)}, R^{(1)}R^{(2)}Y^{(1)}Y^{(2)}] = \pi_{11}\gamma_{11}(1-\gamma_{1+}\pi_{1+}) \\
\Sigma_{16}(\boldsymbol{\xi}) &= \Sigma_{61}(\boldsymbol{\xi}) = \mathrm{Cov}[R^{(1)}Y^{(1)}, R^{(1)}R^{(2)}] = \gamma_{11}\pi_{1+} - \gamma_{1+}\pi_{1+}\gamma_{11} \\
\Sigma_{23}(\boldsymbol{\xi}) &= \Sigma_{32}(\boldsymbol{\xi}) = \mathrm{Cov}[R^{(1)}, R^{(2)}Y^{(2)}] = \gamma_{11}\pi_{+1} - \gamma_{1+}\gamma_{+1}\pi_{+1} \\
\Sigma_{24}(\boldsymbol{\xi}) &= \Sigma_{42}(\boldsymbol{\xi}) = \mathrm{Cov}[R^{(1)}, R^{(2)}] = \gamma_{11} - \gamma_{1+}\gamma_{+1} \\
\Sigma_{25}(\boldsymbol{\xi}) &= \Sigma_{52}(\boldsymbol{\xi}) = \mathrm{Cov}[R^{(1)}, R^{(1)}R^{(2)}Y^{(1)}Y^{(2)}] = \gamma_{11}\pi_{11}(1-\gamma_{1+}) \\
\Sigma_{26}(\boldsymbol{\xi}) &= \Sigma_{62}(\boldsymbol{\xi}) = \mathrm{Cov}[R^{(1)}, R^{(1)}R^{(2)}] = \gamma_{11} - \gamma_{1+}\gamma_{11} \\
\Sigma_{34}(\boldsymbol{\xi}) &= \Sigma_{43}(\boldsymbol{\xi}) = \mathrm{Cov}[R^{(2)}Y^{(2)}, R^{(2)}] = \gamma_{+1}\pi_{+1} - \gamma_{+1}^{2}\pi_{+1} \\
\Sigma_{35}(\boldsymbol{\xi}) &= \Sigma_{53}(\boldsymbol{\xi}) = \mathrm{Cov}[R^{(2)}Y^{(2)}, R^{(1)}R^{(2)}Y^{(1)}Y^{(2)}] = \gamma_{11}\pi_{11}(1-\gamma_{+1}\pi_{+1}) \\
\Sigma_{36}(\boldsymbol{\xi}) &= \Sigma_{63}(\boldsymbol{\xi}) = \mathrm{Cov}[R^{(2)}Y^{(2)}, R^{(1)}R^{(2)}] = \gamma_{11}\pi_{+1}(1-\gamma_{+1}) \\
\Sigma_{45}(\boldsymbol{\xi}) &= \Sigma_{54}(\boldsymbol{\xi}) = \mathrm{Cov}[R^{(2)}, R^{(1)}R^{(2)}Y^{(1)}Y^{(2)}] = \gamma_{11}\pi_{11}(1-\gamma_{+1}) \\
\Sigma_{46}(\boldsymbol{\xi}) &= \Sigma_{64}(\boldsymbol{\xi}) = \mathrm{Cov}[R^{(2)}, R^{(1)}R^{(2)}] = \gamma_{11} - \gamma_{+1}\gamma_{11} \\
\Sigma_{56}(\boldsymbol{\xi}) &= \Sigma_{65}(\boldsymbol{\xi}) = \mathrm{Cov}[R^{(1)}R^{(2)}Y^{(1)}Y^{(2)}, R^{(1)}R^{(2)}] = \gamma_{11}\pi_{11} - \gamma_{11}^{2}\pi_{11}
\end{aligned}$$

Now, define $f(\boldsymbol{x}) = \frac{x_1}{x_2} + \frac{x_3}{x_4} - \frac{x_5}{x_6}$, where $\boldsymbol{x} = (x_1, \ldots, x_6)'$. By the multivariate delta method, we know that

$$\sqrt{n}\left(f\left(\overline{\boldsymbol{Z}}\right) - f\left(\boldsymbol{\mu_Z}\right)\right) = \sqrt{n}(\hat{\mu}_m - \mu) \xrightarrow{D} N\left(0, \bigtriangledown f\left(\boldsymbol{\mu_Z}\right)' \Sigma(\boldsymbol{\xi}) \bigtriangledown f\left(\boldsymbol{\mu_Z}\right)\right)$$

where

$$\bigtriangledown f\left(\boldsymbol{x}\right) = \left(\frac{1}{x_2}, -\frac{x_1}{x_2^2}, \frac{1}{x_4}, -\frac{x_3}{x_4^2}, -\frac{1}{x_6}, \frac{x_5}{x_6^2}\right)'$$

The asymptotic variance can be simplified to:

$$\frac{\pi_{1+}(1-\pi_{1+})}{\gamma_{1+}} + \frac{\pi_{+1}(1-\pi_{+1})}{\gamma_{+1}} + \frac{\pi_{11}(1-\pi_{11})}{\gamma_{11}} + 2\frac{\gamma_{11}(\pi_{11}-\pi_{1+}\pi_{+1})}{\gamma_{1+}\gamma_{+1}} - 2\frac{\pi_{11}(1-\pi_{1+})}{\gamma_{1+}} - 2\frac{\pi_{11}(1-\pi_{+1})}{\gamma_{+1}}$$

D Comparison of the Asymptotic Efficiency of $\hat{\mu}_m$ vs. $\hat{\mu}_c$

The difference between the asymptotic variance of $\hat{\mu}_m$ and $\hat{\mu}_c$ is

$$\begin{aligned}\Delta(\boldsymbol{\xi}) \;=\; & \frac{\pi_{1+}(1-\pi_{1+})}{\gamma_{1+}} + \frac{\pi_{+1}(1-\pi_{+1})}{\gamma_{+1}} + \frac{\pi_{11}(1-\pi_{11}) - \pi_{00}(1-\pi_{00})}{\gamma_{11}} + \\ & 2\frac{\gamma_{11}(\pi_{11}-\pi_{1+}\pi_{+1})}{\gamma_{1+}\gamma_{+1}} - 2\frac{\pi_{11}(1-\pi_{1+})}{\gamma_{1+}} - 2\frac{\pi_{11}(1-\pi_{+1})}{\gamma_{+1}}\end{aligned}$$

Proposition: Assume that $\gamma_{10} > 0$ and $\gamma_{01} > 0$. Suppose that $\pi_{11}(1-\pi_{11}) > (<) \pi_{00}(1-\pi_{00})$ and $\gamma_{11} \to 0$, then $\Delta(\boldsymbol{\xi})$ converges to $+(-)\infty$.

Proof: This lemma follows since, when γ_{10} and γ_{01} are strictly positive, all terms except the third term on the right hand side of the above equation are finite and the third term converes to $+(-)\infty$ when $\pi_{11}(1-\pi_{11}) > (<) \pi_{00}(1-\pi_{00})$.

E Asymptotic Variance of $\hat{\mu}_{ML}$

Let $\ell(\boldsymbol{\pi}; O) = \log L(\boldsymbol{\pi}; O)$. The Fisher information matrix (i.e., minus the expected value of the second derivative of $\ell(\boldsymbol{\pi}; O)$ with respect to $\boldsymbol{\pi}$), $I(\boldsymbol{\pi})$ is a 3×3 matrix with ith row,

jth column denoted by $I_{ij}(\boldsymbol{\pi})$, where

$$\begin{aligned}
I_{11}(\boldsymbol{\pi}) &= \left(\frac{\gamma_{11}}{1-\pi_{01}-\pi_{10}-\pi_{11}}+\frac{\gamma_{01}}{1-\pi_{01}-\pi_{11}}+\frac{\gamma_{01}}{\pi_{01}+\pi_{11}}+\frac{\gamma_{11}}{\pi_{01}}\right)\\
I_{12}(\boldsymbol{\pi}) &= I_{21}(\boldsymbol{\pi}) = \frac{\gamma_{11}}{1-\pi_{01}-\pi_{10}-\pi_{11}}\\
I_{13}(\boldsymbol{\pi}) &= I_{31}(\boldsymbol{\pi}) = \left(\frac{\gamma_{11}}{1-\pi_{01}-\pi_{10}-\pi_{11}}+\frac{\gamma_{01}}{1-\pi_{01}-\pi_{11}}+\frac{\gamma_{01}}{\pi_{01}+\pi_{11}}\right)\\
I_{22}(\boldsymbol{\pi}) &= \left(\frac{\gamma_{10}}{1-\pi_{10}-\pi_{11}}+\frac{\gamma_{10}}{\pi_{10}+\pi_{11}}+\frac{\gamma_{11}}{1-\pi_{01}-\pi_{10}-\pi_{11}}+\frac{\gamma_{11}}{\pi_{10}}\right)\\
I_{23}(\boldsymbol{\pi}) &= I_{32}(\boldsymbol{\pi}) = \left(\frac{\gamma_{10}}{1-\pi_{10}-\pi_{11}}+\frac{\gamma_{10}}{\pi_{10}+\pi_{11}}+\frac{\gamma_{11}}{1-\pi_{01}-\pi_{10}-\pi_{11}}\right)\\
I_{33}(\boldsymbol{\pi}) &= \left(\frac{\gamma_{10}}{1-\pi_{10}-\pi_{11}}+\frac{\gamma_{10}}{\pi_{10}+\pi_{11}}+\frac{\gamma_{01}}{1-\pi_{01}-\pi_{11}}+\frac{\gamma_{11}}{1-\pi_{01}-\pi_{10}-\pi_{11}}+\frac{\gamma_{01}}{\pi_{01}+\pi_{11}}+\frac{\gamma_{11}}{\pi_{11}}\right).
\end{aligned}$$

By the theory of maximum likelihood, we know that $\sqrt{n}(\widehat{\boldsymbol{\pi}}-\boldsymbol{\pi}) \xrightarrow{D} MVN_3(0, I(\boldsymbol{\xi})^{-1})$. Let $g(\boldsymbol{y}) = y_1 + y_2 + y_3$, where $\boldsymbol{y} = (y_1, y_2, y_3)'$. By the multivariate delta method, we know that

$$\sqrt{n}\left(g\left(\widehat{\boldsymbol{\pi}}\right) - g\left(\boldsymbol{\pi}\right)\right) = \sqrt{n}(\hat{\mu}_{ML} - \mu) \xrightarrow{D} N\left(0, \bigtriangledown g\left(\boldsymbol{\pi}\right)' I(\boldsymbol{\pi}) \bigtriangledown g\left(\boldsymbol{\pi}\right)\right)$$

where $\bigtriangledown g\left(\boldsymbol{y}\right) = (1,1,1)'$.

Figure 1: Asymptotic variances of the complete case and moment-based estimators relative to maximum likelihood estimator of the prevalence of the logically defined outcome and asymptotic variances of the complete case relative to moment-based estimator. By design $\pi_{01} = \pi_{10}, \pi_{00} = \pi_{11}, \gamma_{01} = \gamma_{10}, \gamma_{00} = \gamma_{11}$.

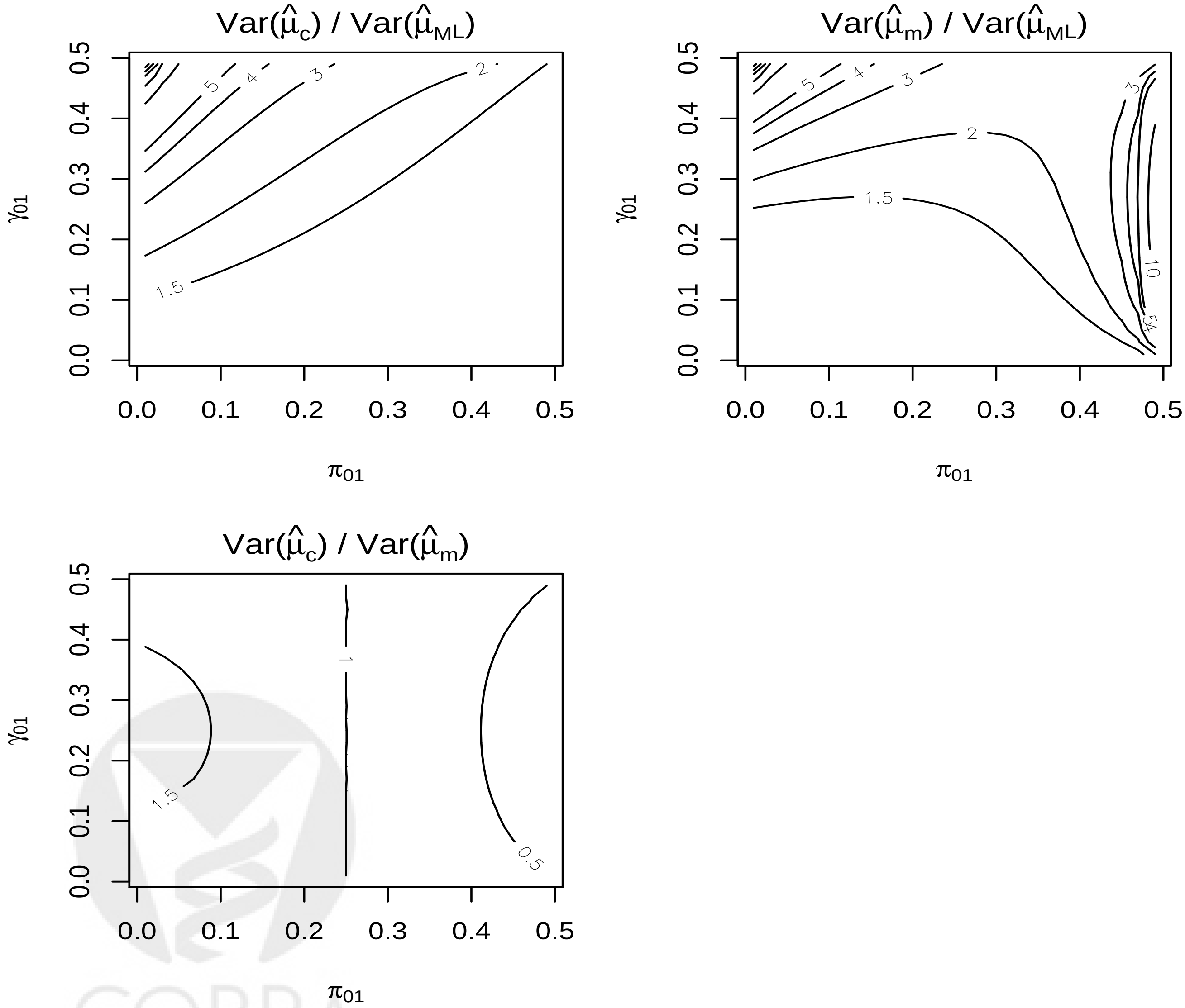
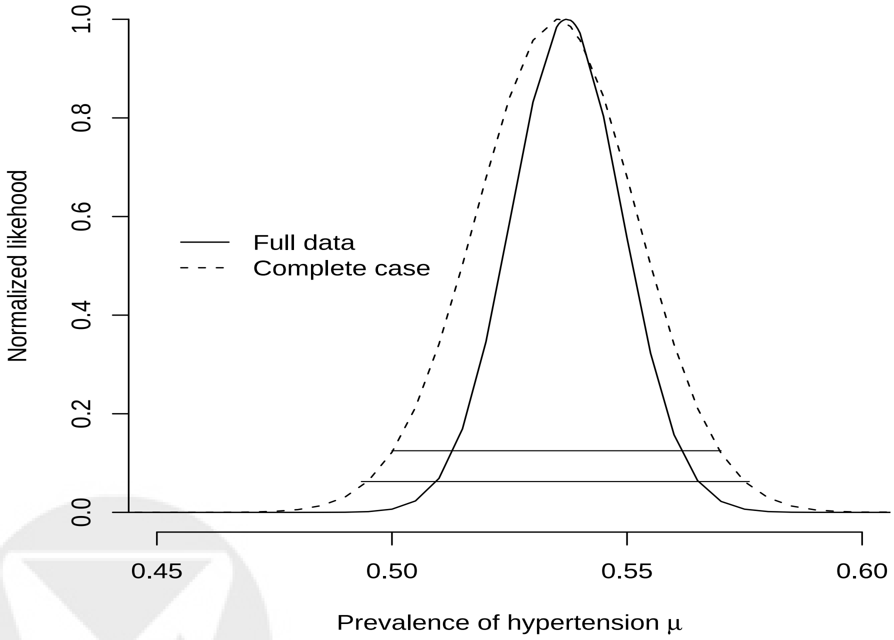


Figure 2: Normalized profile likelihood functions from complete-case conditional likelihood function and full conditional likelihood function from data with 60% completely at random missingness in original outcomes hypertension and anti-hypertensive medication from Sleep Heart Health Study



	Outcome				Observed Indicator		
	$Y^{(2)}$				$R^{(2)}$		
$Y^{(1)}$	0	1		$R^{(1)}$	0	1	
0	π_{00}	π_{01}	π_{0+}	0	γ_{00}	γ_{01}	γ_{0+}
1	π_{10}	π_{11}	π_{1+}	1	γ_{10}	γ_{11}	γ_{1+}
	π_{+0}	π_{+1}			γ_{+0}	γ_{+1}	

Table 1: Outcome probabilities and data availability probabilities

Table 2: Possible binary original outcomes $Y^{(1)}, Y^{(2)}$, logical outcomes Y, and values of the observed data indicators using (3) and (2).

			R				
$Y^{(1)}$	$Y^{(2)}$	Y	$R^{(1)}$	$R^{(2)}$	R^*	$R^\dagger$	Freq
1	1	1	1	1	1	1	n_1
1	1	1	1	0	0	1	n_2
1	1	1	0	1	0	1	n_3
1	1	1	0	0	0	0	n_4
1	0	1	1	1	1	1	n_5
1	0	1	1	0	0	1	n_6
1	0	1	0	1	0	0	n_7
1	0	1	0	0	0	0	n_8
0	1	1	1	1	1	1	n_9
0	1	1	1	0	0	0	n_{10}
0	1	1	0	1	1	1	n_{11}
0	1	1	0	0	0	0	n_{12}
0	0	0	1	1	1	1	n_{13}
0	0	0	1	0	0	0	n_{14}
0	0	0	0	1	0	0	n_{15}
0	0	0	0	0	0	0	n_{16}

Table 3: Monte Carlo estimated biases and mean square errors of the four estimators of the prevalence of a logically defined outcome. Sample size $n = 300$, $\pi_{01} = \pi_{10}, \pi_{00} = \pi_{11}$ for original outcomes, and $\gamma_{01} = \gamma_{10}, \gamma_{00} = \gamma_{11}$ for observed data indicators

Parameter					Estimated Bias (%)				Mean Squared Error ($\times$100)			
π_{01}	π_{11}	γ_{01}	γ_{11}	μ	$\hat{\mu}_c$	$\hat{\mu}_a$	$\hat{\mu}_m$	$\hat{\mu}_{ML}$	$\hat{\mu}_c$	$\hat{\mu}_a$	$\hat{\mu}_m$	$\hat{\mu}_{ML}$
0.10	0.40	0.10	0.40	0.60	0.1	13.4	0.2	0.1	0.189	0.785	0.162	0.151
0.10	0.40	0.20	0.30	0.60	0.3	26.9	0.3	0.3	0.256	2.715	0.171	0.142
0.10	0.40	0.25	0.25	0.60	-0.2	33.3	0.1	0.0	0.338	4.091	0.229	0.168
0.10	0.40	0.40	0.10	0.60	-0.3	53.3	0.3	0.0	0.788	10.268	0.614	0.235
0.20	0.30	0.10	0.40	0.70	-0.1	8.5	0.0	-0.1	0.175	0.473	0.154	0.140
0.20	0.30	0.20	0.30	0.70	-0.1	17.1	0.1	0.0	0.236	1.527	0.203	0.156
0.20	0.30	0.25	0.25	0.70	-0.1	21.4	-0.4	-0.3	0.285	2.327	0.245	0.171
0.20	0.30	0.40	0.10	0.70	0.1	34.3	0.3	0.2	0.789	5.806	0.701	0.335
0.25	0.25	0.10	0.40	0.75	-0.2	6.5	-0.2	-0.2	0.161	0.349	0.171	0.144
0.25	0.25	0.20	0.40	0.75	0.0	13.3	-0.1	0.0	0.211	1.086	0.213	0.157
0.25	0.25	0.25	0.25	0.75	-0.1	16.6	0.0	-0.1	0.230	1.616	0.250	0.149
0.25	0.25	0.40	0.10	0.75	0.1	26.7	-0.2	-0.1	0.666	4.029	0.644	0.304
0.40	0.10	0.10	0.40	0.90	0.0	2.2	-0.2	0.0	0.075	0.089	0.112	0.071
0.40	0.10	0.20	0.30	0.90	0.0	4.5	0.0	0.0	0.099	0.199	0.190	0.090
0.40	0.10	0.25	0.25	0.90	-0.1	5.5	-0.2	-0.1	0.126	0.280	0.219	0.109
0.40	0.10	0.40	0.10	0.90	-0.3	8.8	-0.2	-0.2	0.315	0.645	0.443	0.216

Table 4: Cross tabulation of high blood pressure and anti-hypertensive medication status for subjects from Sleep Heat Health Study

	Medication		
High BP	No (0)	Yes (1)	Missing
No (0)	2482	1310	2
Yes (1)	724	978	3
Missing	8	13	10

Table 5: Four estimates (standard deviations) of hypertension prevalence with different data missingness in hypertension and anti-hypertensive medication from Sleep Heart Health Study

γ_{00}	γ_{01}	γ_{10}	γ_{11}	$\hat{\mu}_c$	$\hat{\mu}_a$	$\hat{\mu}_m$	$\hat{\mu}_{ML}$
0.04	0.16	0.16	0.64	0.548 (0.0051)	0.618 (0.0040)	0.549 (0.0038)	0.548 (0.0035)
0.16	0.24	0.24	0.36	0.547 (0.0091)	0.696 (0.0065)	0.549 (0.0070)	0.548 (0.0062)
0.36	0.24	0.24	0.16	0.547 (0.0158)	0.784 (0.0086)	0.549 (0.0118)	0.548 (0.0096)
0.20	0.30	0.40	0.10	0.547 (0.0202)	0.871 (0.0071)	0.548 (0.0155)	0.548 (0.0110)
0.30	0.40	0.10	0.20	0.546 (0.0135)	0.773 (0.0075)	0.548 (0.0099)	0.548 (0.0085)
0.40	0.10	0.20	0.30	0.547 (0.0102)	0.686 (0.0078)	0.548 (0.0084)	0.548 (0.0076)
0.10	0.20	0.30	0.40	0.547 (0.0080)	0.686 (0.0057)	0.548 (0.0065)	0.548 (0.0055)