Published in final edited form as:
Epidemiology. 2010 November ; 21(6): 863–871. doi:10.1097/EDE.0b013e3181f57571.

Adjustment for Missing Data in Complex Surveys Using Doubly Robust Estimation:

Application to Commercial Sexual Contact Among Indian Men

Kathleen E. Wirth[a], **Eric J. Tchetgen Tchetgen**[a,b], and **Megan Murray**[a,c,d]

[a]Department of Epidemiology, Harvard School of Public Health, Boston, MA

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

[c]Division of Global Health Equity, Brigham and Women's Hospital, Boston, MA

[d]Infectious Disease Unit, Massachusetts General Hospital, Boston, MA

Abstract

Background—The Demographic and Health Survey program routinely collects nationally representative information on HIV-related risk behaviors in many countries, using face-to-face interviews and a complex sampling scheme. If respondents skip questions about behaviors perceived as socially undesirable, such interviews may introduce bias. We sought to implement a doubly robust estimator to correct for dependent missing data in this context.

Methods—We applied 3 methods of adjustment for nonresponse on self-reported commercial sexual contact data from the 2005–2006 India Demographic Health Survey to estimate the prevalence of sexual contact between sexually active men and female sex workers. These methods were inverse-probability weighted regression, outcome regression, and doubly robust estimation—a recently-described approach that is more robust to model misspecification.

Results—Compared with an unadjusted prevalence of 0.9% for commercial sexual contact prevalence (95% confidence interval = 0.8%–1.0%), adjustment for nonresponse using doubly robust estimation yielded a prevalence of 1.1% (1.0%–1.2%). We found similar estimates with adjustment by outcome regression and inverse-probability weighting. Marital status was strongly associated with item nonresponse, and correction for nonresponse led to a nearly 80% increase in the prevalence of commercial sexual contact among unmarried men (from 6.9% to 12.1%–12.4%).

Conclusions—Failure to correct for nonresponse produced a bias in self-reported commercial sexual contact. To facilitate the application of these methods (including the doubly robust estimator) to complex survey data settings, we provide analytical variance estimators and the corresponding SAS and MATLAB code. These variance estimators remain valid regardless of whether the modeling assumptions are correct.

Correspondence: Kathleen E. Wirth, Department of Epidemiology, Harvard School of Public Health, 641 Huntington Ave, Boston, MA 02115. kwirth@hsph.harvard.edu.

Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).

Accurate information on high-risk sexual behavior is critical for assessing the spread and persistence of sexually transmitted diseases (STD) including HIV. Studies of sexual behavior generally rely on face-to-face interviews in which participants recall previous behaviors. Although this approach may be economically efficient and may optimize validity in certain populations (eg, low literacy), face-to-face interviews have been shown to introduce bias to the measurement of behaviors perceived as socially undesirable. Multiple studies have shown that face-to-face interviewing results in reduced disclosure compared with techniques such as audio computer-assisted self-interviewing, which do not require an interviewer to administer the survey.[1–4] Under-reporting of high-risk sexual behavior may underestimate prevalence and hinder the development of effective HIV prevention interventions.

Face-to-face interviews may also lead to systematic bias if participants avoid answering sensitive questions entirely. This “item nonresponse bias” occurs when the observed responses of persons who complete the survey item systematically differ from the unobserved responses of those who did not complete the item. Because the observed data are not a simple random sample of the complete data, conventional analytic approaches that exclude subjects with missing information (eg, complete-case analysis) can yield biased and inefficient results.[5] Such approaches are the default setting in statistical software packages such as SAS (SAS Institute, Cary, NC) and STATA (STATA Corp LP, College Station, TX).

The MEASURE DHS (Demographic and Health Surveys) project began to administer nationally-representative, household-based surveys on HIV knowledge, attitudes, and behavior in 1998. All of these surveys have employed faceto- face interviews.[6] In many of the more than 60 countries where data have been collected, the DHS constitutes the only source of data on HIV-related risk behavior, and organizations such as PEPFAR (US President’s Emergency Plan for AIDS Relief) and UNAIDS (United Nations Joint Program on HIV/AIDS) routinely use these surveys to monitor and assess national HIV programs. Despite the potential for bias, face-to-face interviewing can be advantageous in large-scale data collection efforts such as the DHS because it allows for the recruitment and participation of thousands of respondents with varying levels of literacy in a cost-efficient manner.

Precise data on HIV-related risk behavior is especially urgent in India, where an estimated 2.3 million adults are living with HIV. In absolute numbers, this disease burden is second only to South Africa and Nigeria.[7] The primary mode of transmission is heterosexual contact,[8,9] with female sex workers considered to be the core infected group from which HIV spreads to the general population.[10,11] Based on detailed cross-sectional information and mapping exercises, the National AIDS Control Organization (NACO) of India estimated that between 0.8 and 1.25 million female sex workers operated within the country in 2004.[12] However, little is known about the size of the male client population, despite its established role in India’s HIV epidemic. The only empirically- based estimate used data obtained in the 2005–2006 India DHS - also known as the National Family Health Survey 3. This study found approximately 0.9% of sexually active Indian men reported either exchanging money for sex or having a sexual partnership with a prostitute during the previous year.[13] However, recent evidence suggests that this estimate, based on face-to-face interviews, may grossly underestimate the actual number of men who exchange money for sex annually. A cross-sectional study carried out in southern India among a household-based sample of men found that disclosure of commercial sexual activity increased more than 3-fold with polling-booth surveys compared with face-to-face interviews.[14]

To examine whether the prevalence of commercial sexual contact may be underestimated due to dependent-item nonresponse, we adjusted for missing data and re-estimated the prevalence of commercial sexual contact in the Indian National Family Health Survey 3. Numerous methods have been proposed to correct for nonresponse, including inverse probability weighting and outcome regression.[5] Inverse probability weighting adjusts for nonresponse by weighting the outcomes of participants with nonmissing information by the inverse of the probability of having complete data (obtained by specifying a regression model for the missingness mechanism given fully-observed covariates). Outcome regression specifies a regression model for the outcome data given fully-observed covariates. To yield valid inferences, each of these methods relies on 2 assumptions: Both methods assume that the missingness process and the outcome are independent within levels of the fully-observed covariates; that is, the outcome data are missing at random. In addition, the outcome regression approach requires a correct regression model for the outcome given the fully-observed covariates, whereas inverse-probability weighting relies on a correctly-specified regression model for the missingness process.

The extent to which these assumptions hold is rarely known, especially in observational settings such as the DHS. In an effort to relax the second assumptions, Scharfstein et al (1999), and Bang and Robins (2005) proposed an alternative method known as doubly robust estimation which combines inverse-probability weighting and outcome regression.[15,16] A doubly robust estimator has the key advantage of providing 2 opportunities for making valid inferences, as opposed to one under either inverse-probability weighting or outcome regression alone. In other words, a doubly robust estimator will remain unbiased if the model for the missingness mechanism is correctly specified, irrespective of the outcome regression model, and vice versa.[15,17] When both models are correct, the doubly robust estimator has the advantage of being more efficient than inverse-probability weighting, as the former makes use of the entire sample while the later only uses complete cases. We use these 3 methods to adjust for missing data on commercial sexual contact in a national male respondent sample in India, and we compare the results to the conventional complete-case analysis that discards any subjects with missing outcome data.

METHODS

Study Population

The National Family Health Survey 3 was carried out in 2005–2006 to assess the general health status and family welfare among households in India. A complex sampling scheme was used to create a nationally representative, household- based sample.[18] Within each state, a rural sample was constructed by first selecting villages with probability proportional to the population size. Within each village, households were enumerated and then randomly selected for inclusion in the study. An urban sample was created for each state by first selecting urban wards with probability proportional to size. Within each ward, 1 census enumeration block was then selected, again with probability proportional to size. Finally, households were enumerated and randomly chosen for each selected census enumeration block. This sampling strategy identified 85,373 men aged 15–54 years as eligible for participation; 74,369 completed the survey, for an overall response rate of 87%.[18] Further details regarding the sampling and data collection procedures are available elsewhere.[18] For the purpose of this analysis, we restricted the sample to men who had reported recent sexual activity thereby excluding 24,649 men who reported no sexual intercourse; this left a final analytic sample of 49,720.

We defined commercial sexual contact based on men's reports about their sexual partners (up to a maximum of 3) during the previous year. We classified men who reported one or more "prostitutes" among their 3 most recent sexual partners as positive for commercial sexual contact. In addition, those men who did not report recent commercial partners were asked "In the last 12 months, did you pay anyone in exchange for having sexual intercourse?" We also classified men answering affirmatively to this question as positive for commercial sexual contact. Through this process, we identified 470 men with self-reported commercial sexual contact.

Statistical Analysis

We used 3 approaches to estimate the prevalence of commercial sexual contact adjusted for item nonresponse in the survey: inverse-probability weighting, outcome regression, and doubly robust estimation.[19] In specifying the logistic regression models required for inverse-probability weighting and outcome regression, we considered 24 potential covariates based on the published literature,[20–22] as well as publically available reports from behavioral surveillance surveys among clients of sex workers in India.[23,24] These included demographic variables (age, education, type of residence, religion, and current marital status), literacy, time spent away from the home, sexual-behavior indicators (age at first intercourse, number of lifetime sexual partners), symptoms and diagnosis of sexually transmitted infections during the past 12 months, and frequency of alcohol consumption. The survey also assessed gender norms through a series of questions on the rights of a husband to force unwanted sexual intercourse or have sexual intercourse with women other than his wife. Such questions included whether or not a husband was justified in beating his wife if she went out without telling him, neglected the children, argued with him, refused sexual intercourse, burnt the food, was unfaithful, or showed disrespect toward his parents. Because previous studies in India have documented increased levels of physical and sexual violence associated with commercial sex work, we included all of these items as potential predictors.[25] We also created 128 2-way interaction terms by taking the cross-product of each demographic covariate (age, education, type of residence, religion, current marital status, and literacy) with each behavioral covariate (rights of husband, justification for wife beatings, age at first intercourse, number of lifetime sexual partners, symptoms and diagnosis of sexually transmitted infections during the past 12 months, and frequency of alcohol consumption). We included missing indicators for each selected variable to maximize the number of cases included in the final models and to maintain a constant sample size across methods.

Finally, we conducted likelihood-ratio tests to assess nonlinearity for the following covariates: age, education, wealth, literacy, frequency of alcohol consumption, number of trips away from home, age at first intercourse, and number of lifetime sexual partners. Based on the results of these tests, nonlinear terms for age, education, wealth, literacy, and number of lifetime sexual partners were incorporated in the model-building process for inverse-probability weighting. For the outcome regression approach, nonlinear terms for age, frequency of alcohol consumption, education, number of trips away from home, and number of lifetime sexual partners were included in specifying the model.

To build the multivariate logistic regression models required by the inverse-probability weighting and outcome regression approaches, we used a stepwise, forward selection procedure to identify covariates from the list of candidate predictors described above. The entry and exit criteria were set to a $P < 0.2$.

Inverse-probability Weighted Regression

Inverse-probability weighting adjusts for item nonresponse by creating from the complete cases a pseudo-population, in which individuals are weighted by the inverse of the conditional probability of complete data given the fully observed covariates. In the resulting pseudo-population, the outcomes of participants with complete data represent themselves and those with similar characteristics who were missing data on the outcome. In other words, in the absence of model misspecification and under the missing-at-random assumption, missing outcome information in the pseudo-population is a chance mechanism unrelated to the observed or unobserved information.[26]

We began by fitting a multivariate logistic regression model for the indicator of having an observed commercialsexual- contact status, which throughout we denote C = 1, given a subset $\tilde{M}$ of observed covariates; that is we modeled $Pr[C = 1|\tilde{M}]$) as followed:

$$log\left(\frac{Pr[\,C=1|\tilde{M};\alpha]}{1-Pr[\,C=1|\tilde{M};\alpha]}\right)=log\left(\frac{\pi(\tilde{M};\alpha)}{1-\pi(\tilde{M};\alpha)}\right)=\alpha_0+\sum_{j=1}^{55}\alpha_j M_j \quad (1)$$

where, $\tilde{M} = \{M_j, j = 1,\ldots, 55\}$ includes the observed covariates identified by the stepwise-forward selection procedure. eTable 1 of eAppendix A (http://links.lww.com/EDE/A428) provides a complete description of the model.

Given the complex sampling framework, the inverse-probability weights were modified to adjust simultaneously for item nonresponse and the probability of being selected into the study population. As described previously by Moore et al[27] in a missing covariate context, the final weight W_i^* for each individual i was constructed by multiplying the inverse-probability weight $\tilde{W}_i = 1 \Big/ \pi_i(\tilde{M}_i;\hat{\alpha})$ by the survey weight $W_{i,s}$ provided in the National Family Health Survey 3 database[18] (where $\pi_i(\tilde{M}_i;\hat{\alpha})$ denotes the maximum likelihood estimated predicted probability of observing person's commercial sexual contact). That is

$W^{*}_{i,s}=\tilde{W}^{*}_{i}W_{i,s}$. The resulting inverse-probability-weighted regression estimator is given by the weighted sample average: $\hat{\mu}_{IPW}=\sum_{i=1}^{N}(W^{*}_{i,s}Y_i)\Big/\sum_{i=1}^{N}W^{*}_{i,s}$.

Outcome Regression

To implement the outcome regression approach, we specified a multivariate logistic regression model among complete cases only for commercial sexual contact, which throughout we denote by Y = 1 for those with reported commercial sexual contact and Y = 0 otherwise, given a subset $\tilde{L}$ of observed covariates. That is, we modeled $Pr[Y = 1|\tilde{L}]$ as

$$log\left(\frac{pr[\,Y=1|\tilde{L};\beta]}{1-pr[\,Y=1|\tilde{L};\beta]}\right)=log\left(\frac{b(\tilde{L};\beta)}{1-b(\tilde{L};\beta)}\right)=\beta_0+\sum_{j=1}^{50}\beta_j L_j \quad (2)$$

where, $\tilde{L} = \{L_j, j = 1,\ldots, 50\}$ includes the observed covariates identified by the stepwise-forward selection procedure. eTable 2 of eAppendix A (http://links.lww.com/EDE/A428) provides a complete description of the model.

Next, for each subject i in the sample, using the maximum-likelihood estimate $\hat{\beta}$, we obtained their predicted mean $b_i(\tilde{L}_i; \hat{\beta})$ irrespective of whether Y_i was observed. The adjusted population prevalence of commercial sexual contact was then computed by taking the sample weighted average of these predicted values:

$$\hat{\mu}_{OR}=\sum_{i=1}^{N}(W_{i,s}b_i(\tilde{L}_i;\hat{\beta}))\Big/\sum_{i=1}^{N}W_{i,s}.$$

Doubly Robust Estimation

As stated earlier, the adjusted prevalence of commercial sexual contact produced by inverse-probability weighted regression will generally be biased if model (1) is incorrect. Similarly the outcome regression estimator will generally be biased if model (2) is incorrect. As we cannot guarantee that either model (1) or model (2) was correctly specified, we implemented a doubly robust estimator.

We adopted the approach of Bang and Robins[15] and defined a pseudo-outcome $\hat{Y}_{i,DR}$ for each subject in the sample.

$$\hat{Y}_{i,DR}=\frac{C_i}{\pi_i(\tilde{M}_i;\hat{\alpha})}Y_i-\frac{C_i}{\pi_i(\tilde{M}_i;\hat{\alpha})}b_i(\tilde{L}_i;\hat{\beta})+b_i(\tilde{L}_i;\hat{\beta}) \quad (3)$$

The doubly robust estimator is defined as $\sum_{i=1}^{N}(W_{i,s}\hat{Y}_{i,DR})\Big/\sum_{i=1}^{N}W_{i,s}$. A demonstration of the doubly robust property of this estimator is instructive. Let (α^*, β^*) be the limiting value that ($\hat{\alpha}$, $\hat{\beta}$) is converging to (in probability) as sample size grows to infinity. Then, the large sample bias of the doubly robust estimator $\hat{\mu}_{DR}$ is approximately

$$E[\hat{Y}_{DR}(\alpha^*,\beta^*)-E[Y]]=E\left[\frac{C_i}{\pi(\tilde{M},\alpha^*)}Y_i-\frac{C_i}{\pi(\tilde{M},\alpha^*)}b(\tilde{L},\beta^*)+b(\tilde{L},\beta^*)-Pr[Y=1|\tilde{L}]\right]$$

by the law of iterated expectations

$$=E\left[\frac{Pr[C=1|\tilde{M}}{\pi(\tilde{M},\alpha^*)}Pr[Y=1|\tilde{L}]-\frac{Pr[C=1|\tilde{M}]}{\pi(\tilde{M},\alpha^*)}b(\tilde{L},\beta^*)+b(\tilde{L},\beta^*)-Pr[Y=1|\tilde{L}]\right]$$

by the law of iterated expectations and the missing at random assumption

$$\begin{aligned}
&=E\left[\frac{Pr[C=1|\tilde{M}]}{\pi(\tilde{M},\alpha^*)}(Pr[Y=1|\tilde{L}]-b(\tilde{L},\beta^*))+b(\tilde{L},\beta^*)-Pr[Y=1|\tilde{L}]\right]\\
&=E\left[\left(\frac{Pr[C=1|\tilde{M}]}{\pi(\tilde{M},\alpha^*)}-1(Pr[Y=1|\tilde{L}]-b(\tilde{L},\beta^*))\right]\\
&=E\left[(Pr[C=1|\tilde{M}]-\pi(\tilde{M},\alpha^*))(Pr[Y=1|\tilde{L}]-b(\tilde{L},\beta^*))\frac{1}{\pi(\tilde{M},\alpha^*)}\right]
\end{aligned}$$

The doubly robust property is obtained by noting that the last equation is equal to zero if either, but not necessarily both, of the following hold: $\pi(\tilde{M}, \alpha^*) = Pr[C = 1|\tilde{M}]$ or $b(\tilde{L}, \beta^*) = Pr[Y = 1|\tilde{L}]$.

Complete-case Analysis

The complete-case analysis, which discards subjects with missing outcome data as described earlier, will be valid only under the stringent missing-completely-at-random assumption; that is, missingness is a chance mechanism unrelated to any observed or unobserved information. We implemented this approach by fitting an intercept-only, survey-weighted logistic regression model among complete cases only, for commercial sexual contact ($Y = 1$).

The population prevalence of commercial sexual contact was then computed by taking the sample-weighted average of the observed responses: $\hat{\mu}_{CC}={\sum_{i=1}^{N}W_{i,s}C_iY_i}\Big/{\sum_{i=1}^{N}W_{i,s}C_i}$.

Finally, we empirically assessed which, if any, of our adjusted estimates were approximately unbiased (or more precisely, consistent). As described by Bang and Robins[15] and by Tchetgen Tchetgen and Robins,[28] our test statistic was motivated by the following observation: if the doubly robust estimator is correct, then either the inverse-probability weighting or outcome regression estimate should also be correct, but not necessarily both.[15,28] If the outcome regression is correct, then the resulting parameter estimate should be close to the doubly robust estimate irrespective of the missingness model being correct, and vice versa. Otherwise the 2 estimators will generally differ beyond sampling variability. To operationalize this observation, we proceed as in the paper by Tchetgen Tchetgen and Robins,[28] and construct 2 test statistics to empirically detect possible model misspecification; the squared standardized difference between the doubly robust estimate and the estimate obtained either through inverse-probability weighting (4) or outcome regression (5).[28]

$$T_{IPW}={(\hat{\mu}_{IPW}-\hat{\mu}_{DR})^2}\Big/{(\hat{\sigma}^2_{IPW}-\hat{\sigma}^2_{DR})} \quad (4)$$

$$T_{OR}={(\hat{\mu}_{OR}-\hat{\mu}_{DR})^2}\Big/{(\hat{\sigma}^2_{DR}-\hat{\sigma}^2_{OR})} \quad (5)$$

It can be shown that the denominator in equation (4) is a consistent estimate of the variance of $(\hat{\mu}_{IPW}-\hat{\mu}_{DR})$ under the null hypothesis H_1 that the inverse-probability weights are correctly specified. Similarly, the denominator of equation (5) is a consistent estimator of the variance of $(\hat{\mu}_{OR}-\hat{\mu}_{DR})$ under the null hypothesis H_2 that the outcome regression is correctly specified.[28,29]

Thus, under the null hypothesis H_1 that the model for the missingness mechanism is correct, T_{IPW} approximately follows a χ^2 distribution with 1 degree of freedom. Similarly, under the null hypothesis H_2 that the outcome regression is correct, T_{OR} approximately follows a χ^2 distribution with 1 degree of freedom. To guarantee that both T_{IPW} and T_{OR} are always positive, an analytical estimate of the variance of their respective numerators based on influence function arguments may be used, or an appropriate bootstrap estimator of the variance of the numerators may replace the difference estimator in the denominator of the test statistic. Under the alternative that the doubly robust estimator is correct but the inverse-probability weighting approach is biased, T_{IPW} is approximately scaled noncentral χ^2 distributed, with the non-centrality parameter determined by the magnitude of the large sample bias of $\hat{\mu}_{IPW}$. This is also the case for T_{OR}. Therefore, a statistical test of H_1 (respectively of H_2) that rejects the null whenever $T_{IPW}\geq\chi^2_{1,0.05}$ (respectively based on $T_{OR}\geq\chi^2_{1,0.05}$) is more likely to have increased power to reject in those settings where model misspecification results in moderate-to-large bias. If neither the outcome regression or inverse-probability weighting are correct, the test statistic may still be applied. In this scenario, the 2 approaches, along with the doubly robust estimator, will converge to different values and (in theory) both null hypotheses, H_1 and H_2, will be rejected. However, in finite samples, there may be limited power to reject, especially if both estimates result in similar biases.[28]

All statistical analyses were conducted with SAS software version 9.2 (SAS Institute, Cary, NC). Weighted logistic regression models were run using the WEIGHT, STRATA, and CLUSTER statements in PROC SURVEYLOGISTIC and PROC SURVEYREG (eAppendix B, http://links.lww.com/EDE/A428). To appropriately account for the additional uncertainty associated with the first-stage estimation of the required regression models, we used standard Taylor-series expansion arguments to derive accurate large-sample variance estimators for the inverse-probability weighted regression, outcome regression, and doubly robust estimators, which we, in turn, used to construct Wald-type 95% confidence intervals (CIs), using MATLAB 2007a (The MathWorks, Natick, MA). eAppendix C (http://links.lww.com/EDE/A428) provides both the corresponding formulae and MATLAB code. We emphasize that our variance estimators have the key robustness property of remaining valid even under model misspecification (eAppendix D, http://links.lww.com/EDE/A428). For the complete-case analyses, variance estimates were obtained in PROC SURVEYLOGISTIC.

RESULTS

The weighted population size of sexually active men, after accounting for probability of selection into National Family Health Survey 3, was 52,359. Of these, 1715 (3%) had incomplete information on self-reported commercial sexual contact (eg, nonrespondents) during the previous year. Nonrespondents were more likely to be under 20 years of age (OR = 3.2 [95% CI = 2.4–4.2]), not currently married (11.0 [9.3–12.8]), and unemployed (4.2 [3.2–5.5]) and to respond that a husband has the right to have sexual intercourse with another woman (1.6 [1.2–2.2]). Nonrespondents were also more likely to report that a husband was justified in beating his wife for 5 of the 7 reasons assessed in the survey. Missing information on the number of lifetime sexual partners (2.3 [1.2–4.6]), genital sores/ulcers (4.3 [2.6–7.2]), and genital discharge (4.2 [2.5–7.0]) was significantly associated with missingness in univariate analyses. Table 1 provides a full description of the characteristics of respondents and nonrespondents.

Table 2 presents the prevalence of commercial sexual contact before and after adjusting for dependent item nonresponse. Compared with the prevalence in complete-case analysis (0.9% [0.8%–1.0%]), adjustment through doubly robust estimation led to a slight increase in prevalence of commercial sexual contact (1.1% [1.0%–1.2%]). Adjustment using outcome regression and inverse-probability weighting resulted in indistinguishable prevalence estimates of 1.1% (1.0%–1.2%) each.

We noted a discrepancy in nonresponse by current marital status, with 2% of married men compared with nearly 18% of unmarried men having missing information on commercial sexual contact. We therefore assessed the effect of adjustment for item nonresponse for these groups separately. Among married subjects, adjustment for nonresponse did not change the prevalence estimate. However, for unmarried subjects, this adjustment led to considerably higher prevalence estimates with all 3 approaches. Compared with the complete-case analysis estimate of 6.9% (5.8%– 8.1%), the doubly robust prevalence estimate was 12.3% (10.8%–13.4%) for the unmarried, a 77% increase over the unadjusted estimate. Similar results were obtained using outcome regression (12.4% [11.3%–13.3%]) and inverse-probability weighting (12.1% [10.7%–14.0%]) (Table 2).

Table 3 presents the results for our goodness-of-fit tests organized according to parameters of interest: the overall population mean, the mean among married men, and the mean among unmarried men. Overall, the empirical evidence suggests that both the outcome regression model and missingness model used in inverse-probability weighting provided adequate estimators ($T_{OR} = 0.10$ with $P = 0.75$ and $T_{IPW} = 0.85$ with $P = 0.36$ for the overall mean outcome regression and inverse-probability weighted regression estimators, respectively). All 3 methods were in agreement in the subgroup analyses of married and unmarried men.

DISCUSSION

We adjusted for missing data using 3 statistical approaches within a complex survey data setting, and we provide the corresponding SAS and MATLAB code to obtain point estimates and confidence intervals. In our assessment of the nearly 50,000 sexually-active Indian men who participated in the National Family Health Survey 3, we found that men with missing data on commercial sexual contact were different from those with complete information. Nonrespondents were more likely to be young, unemployed, and not currently married. Furthermore, missing data on commercial sexual contact was significantly associated with missing data on lifetime sexual partners and STD-like symptoms during the previous year. Finally, we show that adjustment for dependent item nonresponse increased in the estimated overall prevalence whether we used inverse-probability weighting, outcome regression, or doubly robust estimation.

Marital status was the strongest predictor of item nonresponse in our study, and the magnitude of the apparent bias in the unadjusted estimates was largest among unmarried men. In India, where the majority of marriages are arranged by the parents, premarital sexual relationships may be considered taboo. Unmarried men may be reluctant to disclose these sexual relationships in a face-to-face interview, especially in settings such as this household survey in which other household members may be present. We note that approximately 83% of the total number of unmarried men in the sample reported no sexual activity and were therefore excluded from the present analysis. This may represent a form of nonresponse bias not addressed in the present study, as men who reported no sexual activity were not further questioned about commercial sexual contact. A recently published study conducted in Southern India found not only higher levels of disclosure of premarital sexual intercourse in polling- booth surveys compared with comparable face-to-face interviews, but the magnitude of the discrepancy was larger for unmarried men.[14]

We emphasize that the adjustment for dependent item nonresponse via any of the approaches used in the present study (excluding the complete-case approach, which relies on the data being missing completely at random) the missing-at-random assumption is fundamental. Unfortunately, this assumption cannot be empirically verified in observational settings such as the DHS, and its appropriateness depends directly on subject-matter knowledge necessary to identify, measure, and control for all factors that may be associated with nonresponse. The final multivariate models specified for the outcome regression and inverse-probability weighting considered nearly 150 covariates, including numerous nonlinear and interaction terms; the missing-at-random assumption would not hold if an unmeasured factor was associated with nonresponse but not included in this set of covariates. Alternatively, sensitivity analyses are available to explore the robustness of the results to increasingly extreme departures from the missing at random assumption (see the paper by Robins et al[30] for examples). However, these analyses do not seek to assess whether an unmeasured factor exists, but rather the impact of such a covariate if it indeed existed; such analyses are beyond the scope of the present study.

In conclusion, failure to account for missing data in this household-based sample of sexually active Indian men led to bias for self-reported commercial sexual contact. We note that these results may not be generalizable to other countries in which DHS survey data are available or to other HIV-related risk behaviors measured by the DHS. However, a missing-data analysis should be considered in the assessment of any stigmatized behavior, as there may be important differences between respondents and nonrespondents. In such analyses, doubly robust estimation may be preferable, as it is more robust to model misspecification relative to inverse-probability weighting or outcome regression.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Supported by the US National Institutes of Health (AI 007433).

REFERENCES

1. Des Jarlais DC, Paone D, Milliken J, et al. Audio-computer interviewing to measure risk behaviour for HIV among injecting drug users: a quasi-randomised trial. Lancet. 1999; 353:1657–1661. [PubMed: 10335785]
2. Tideman RL, Chen MY, Pitts MK, Ginige S, Slaney M, Fairley CK. A randomised controlled trial comparing computer-assisted with face-to-face sexual history taking in a clinical setting. Sex Transm Infect. 2007; 83:52–56. [PubMed: 17098771]
3. Turner CF, Ku L, Rogers SM, Lindberg LD, Pleck JH, Sonenstein FL. Adolescent sexual behavior, drug use, and violence: increased reporting with computer survey technology. Science. 1998; 280:867–873. [PubMed: 9572724]
4. Rogers SM, Willis G, Al-Tayyib A, et al. Audio computer assisted interviewing to measure HIV risk behaviours in a clinic population. Sex Transm Infect. 2005; 81:501–507. [PubMed: 16326855]
5. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol. 1995; 142:1255–1264. [PubMed: 7503045]
6. Measure DHS. Demographic and Health Surveys: HIV Corner. Available at: http://demo.measuredhs.com/measuredhs2/topics/hiv/start.cfm.
7. 2008 Report on the global AIDS epidemic. Geneva, Switzerland: UNAIDS; 2008.

8. Simoes EA, Babu PG, Jeyakumari HM, John TJ. The initial detection of human immunodeficiency virus 1 and its subsequent spread in prostitutes in Tamil Nadu, India. J Acquir Immune Defic Syndr. 1993; 6:1030–1034. [PubMed: 8340893]

9. Pais P. HIV and India: looking into the abyss. Trop Med Int Health. 1996; 1:295–304. [PubMed: 8673831]

10. Nagelkerke NJ, Jha P, de Vlas SJ, et al. Modelling HIV/AIDS epidemics in Botswana and India: impact of interventions to prevent transmission. Bull World Health Organ. 2002; 80:89–96. [PubMed: 11953786]

11. Gangakhedkar RR, Bentley ME, Divekar AD, et al. Spread of HIV infection in married monogamous women in India. JAMA. 1997; 278:2090–2092. [PubMed: 9403424]

12. National Institute of Medical Sciences (Indian Council of Medical Research), National AIDS Control Organization (Ministry of Health and Family Welfare). Technical report: India HIV Estimates—2006. New Delhi, India: NIMS and NACO; 2007.

13. Decker MR, Miller E, Raj A, Saggurti N, Donta B, Silverman JG. Indian men's use of commercial sex workers: prevalence, condom use, and related gender attitudes. J Acquir Immune Defic Syndr. 2010; 53:240–246. [PubMed: 19904213]

14. Lowndes, C.; Jayachandran, A.; Pradeep, B., et al. The 17th Biennial Meeting of the ISSTDR. Seattle, WA: Higher levels of HIV-related risky behavior reported in polling booth surveys compared to face-to-face interviews in a general population survey in Mysore district, Karnataka state, southern India. Abstract P-339. [July 29–Aug 1, 2007]

15. Bang H, Robins JM. Dobust estimation in missing data and causal inference models. Biometrics. 2005; 61:962–973. [PubMed: 16401269]

16. Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models. J Am Stat Assoc. 1999; 94:1096–1120.

17. Tchetgen Tchetgen EJ. A simple implementation of doubly robust estimation in logistic regression with covariates missing at random. Epidemiology. 2009; 20:391–394. [PubMed: 19363353]

18. International Institute for Population Sciences (IIPS) and Macro International. National family health survey (NFHS-3), 2005–06: India. Mumbai, India: IIPS; 2007.

19. Little, R.; Rubin, D. Statistical Analysis With Missing Data. Hoboken, NJ: Wiley John & Sons; 2002.

20. Rodrigues JJ, Mehendale SM, Shepherd ME, et al. Risk factors for HIV infection in people attending clinics for sexually transmitted diseases in India. BMJ. 1995; 311:283–286. [PubMed: 7633230]

21. Madhivanan P, Hernandez A, Gogate A, et al. Alcohol use by men is a risk factor for the acquisition of sexually transmitted infections and human immunodeficiency virus from female sex workers in Mumbai, India. Sex Transm Dis. 2005; 32:685–690. [PubMed: 16254543]

22. Bollinger RC, Brookmeyer RS, Mehendale SM, et al. Risk factors and clinical presentation of acute primary HIV infection in India. JAMA. 1997; 278:2085–2089. [PubMed: 9403423]

23. Family Health International (FHI). Behavioral surveillance survey in health highway project in India. New Delhi, India: FHI; 2001.

24. Behavioral surveillance survey in Maharashtra. New Delhi, India: FHI; 2001.

25. Sarkar K, Bal B, Mukherjee R, et al. Sex-trafficking, violence, negotiating skill, and HIV infection in brothel-based sex workers of eastern India, adjoining Nepal, Bhutan, and Bangladesh. J Health Popul Nutr. 2008; 26:223–231. [PubMed: 18686555]

26. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004; 15:615–625. [PubMed: 15308962]

27. Moore CG, Lipsitz SR, Addy CL, et al. Logistic regression with incomplete covariate data in complex survey sampling: application of reweighted estimating equations. Epidemiology. 2009; 20:382–390. [PubMed: 19289959]

28. Tchetgen Tchetgen EJ, Robins JM. The semi-parametric case-only estimator. Biometrics. In press.

29. Newey WK. Generalized method of moments specification testing. J Econom. 1985; 29:229–256.

30. Robins, JM.; Rotnitzky, A.; Scharfstein, D. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: Halloran, ME.; Berry, D., editors.

Statistical Models in Epidemiology: The Environment and Clinical Trials. New York: Springer-Verlag; 1999. p. 1-92.

TABLE 1

Distribution of Sample Characteristics of 50,644 Probability weighted[a] Sexually Active Indian Men by Response Status to Commercial Sexual Contact Survey Items, and Association of These Characteristics With Nonresponse

	Responders % (No.)[a,b]	Nonresponders % (No.)[a,b]	OR (95% CI)
Age (years)			
15–19	2 (1238)	13 (222)	3.19 (2.43–4.17)
20–24	10 (4912)	20 (351)	1.27 (1.00–1.60)
25–29	16 (8044)	12 (210)	0.46 (0.36–0.60)
30–34	18 (8930)	8 (144)	0.29 (0.22–0.38)
35–39	18 (8954)	8 (131)	0.26 (0.19–0.36)
40–44	15 (7777)	11 (191)	0.44 (0.33–0.58)
45–49	13 (6472)	13 (224)	0.61 (0.47–0.80)
50–54[c]	9 (4317)	14 (243)	1.00
Missing	0 (0)	0 (0)	N/A
Type of residence			
Capital, large city	12 (5960)	13 (222)	1.11 (0.91–1.35)
Small city	9 (4554)	9 (156)	1.02 (0.79–1.31)
Town	13 (6484)	12 (202)	0.93 (0.76–1.12)
Countryside[c]	66 (33,646)	66 (1134)	1.00
Missing	0 (0)	0 (0)	N/A
Religion			
Muslim	12 (6042)	8 (146)	0.68 (0.54–0.85)
Christian	2 (1124)	2 (35)	0.89 (0.65–1.23)
Other	3 (1682)	3 (49)	0.82 (0.58–1.15)
Hindu[c]	83 (41,789)	87 (1485)	1.00
Missing	<1 (8)	0 (0)	N/A
Literacy			
Cannot read at all	27 (13,717)	28 (475)	1.06 (0.89–1.25)
Able to read only parts of sentene[c]	6 (3176)	8 (131)	1.26 (0.98–1.63)
Able to read whole sentence[c]	66 (33,602)	64 (1103)	1.00
Missing	<1 (149)	<1 (5)	1.01[d] (0.38–2.67)
Wealth index			
Poorest	18 (9095)	17 (289)	1.01 (0.81–1.28)
Poorer	19 (9764)	20 (340)	1.11 (0.87–1.42)
Middle	20 (10,288)	22 (384)	1.19 (0.97–1.46)
Richer	21 (10,623)	21 (362)	1.09 (0.89–1.33)
Richest[c]	21 (10,873)	20 (341)	1.00
Missing	0 (0)	0 (0)	N/A
Current marital status			

	Responders % (No.)[a,b]	Nonresponders % (No.)[a,b]	OR (95% CI)
Married	93 (46,921)	46 (918)	0.09 (0.08–0.11)
Not married[c]	7 (3723)	54 (797)	1.00
Missing	0 (0)	0 (0)	N/A
Occupation			
Not working	2 (1127)	9 (159)	4.22 (3.23–5.51)
Professional, technical, manager	6 (3292)	6 (101)	0.91 (0.69–1.22)
Clerical	4 (2086)	2 (39)	0.56 (0.38–0.83)
Sales	13 (6390)	11 (184)	0.86 (0.68–1.09)
Agriculture employee	35 (17,855)	32 (554)	0.93 (0.77–1.11)
Services	5 (2495)	6 (96)	1.15 (0.86–1.53)
Skilled and unskilled manual[c]	34 (17,353)	34 (581)	1.00
Missing	<1 (47)	<1 (1)	0.84[d] (0.20–3.58)
No. trips away from home, past 12 months			
5 or more	37 (18,581)	34 (578)	0.82 (0.69–0.99)
3–4	17 (8392)	16 (268)	0.85 (0.69–1.04)
1–2	17 (8685)	18 (302)	0.92 (0.76–1.12)
None[c]	29 (14,826)	33 (560)	1.00
Missing	<1 (161)	<1 (6)	1.09[d] (0.49–2.45)
Husband has the right			
To use force for unwanted sex			
Yes	6 (3081)	7 (119)	1.16 (0.87–1.53)
No[c]	93 (46,916)	92 (1575)	1.00
Missing	1 (647)	1 (21)	0.94[d] (0.47–1.88)
To have sex with another woman			
Yes	4 (2131)	7 (113)	1.61 (1.16–2.22)
No[c]	94 (47,747)	92 (1576)	1.00
Missing	2 (766)	2 (26)	1.01[d] (0.57–1.79)
Husband justified in beating wife if			
She goes out without telling him			
Yes	24 (11,996)	26 (439)	1.11 (0.96–1.29)
No[c]	76 (38,488)	74 (1267)	1.00
Missing	<1 (160)	1 (9)	1.68[d] (0.72–3.91)
She neglects the children			
Yes	28 (14,309)	33 (561)	1.24 (1.07–1.45)
No[c]	72 (36,216)	67 (1144)	1.00
Missing	<1 (119)	1 (10)	2.38[d] (1.02–5.54)
She argues with him			
Yes	26 (13,012)	28 (480)	1.13 (0.97–1.30)
No[c]	74 (37,385)	71 (1224)	1.00

	Responders % (No.)[a,b]	Nonresponders % (No.)[a,b]	OR (95% CI)
Missing	<1 (248)	1 (11)	1.37[d] (0.69–2.70)
She refuses to have sex with him			
Yes	8 (3951)	10 (179)	1.39 (1.12–1.71)
No[c]	92 (46,347)	88 (1516)	1.00
Missing	1 (346)	1 (20)	1.73[d] (0.88–3.41)
She burns the food			
Yes	12 (6197)	16 (266)	1.32 (1.09–1.60)
No[c]	87 (44,292)	84 (1443)	1.00
Missing	<1 (155)	<1 (5)	0.94[d] (0.34–2.59)
She is unfaithful			
Yes	24 (11,928)	29 (501)	1.34 (1.14–1.58)
No[c]	75 (37,971)	69 (1190)	1.00
Missing	1 (745)	1 (24)	0.95[d] (0.57–1.60)
She shows disrespect toward in-laws			
Yes	36 (18,182)	40 (689)	1.20 (1.04–1.38)
No[c]	63 (32,129)	59 (1018)	1.00
Missing	1 (333)	<1 (8)	0.69[d] (0.30–1.57)
STD diagnosis in past 12 months			
Yes	<1 (252)	1 (10)	1.16 (0.57–2.36)
No[c]	99 (50,279)	99 (1697)	1.00
Missing	<1 (113)	<1 (8)	2.02[d] (0.82–4.96)
Genital sore/ulcer in past 12 months			
Yes	2 (1118)	3 (49)	1.31 (0.89–1.93)
No[c]	97 (49,311)	95 (1635)	1.00
Missing	<1 (216)	2 (31)	4.30[d](2.58–7.16)
Genital discharge past 12 months			
Yes	3 (1384)	4 (65)	1.42 (0.98–2.06)
No[c]	97 (49,035)	94 (1619)	1.00
Missing	<1 (226)	2 (31)	4.17[d] (2.50–6.95)
No. lifetime sexual partners			
5 or more	3 (1427)	3 (51)	1.14 (0.79–1.64)
3–4	5 (2299)	6 (105)	1.46 (1.13–1.90)
2	11 (5662)	15 (256)	1.44 (1.19–1.75)
1[c]	81 (41,057)	75 (1287)	1.00
Missing	<1 (199)	1 (15)	2.30[d] (1.15–4.58)
Frequency of alcohol consumption			
Almost every day	4 (2151)	5 (81)	1.22 (0.87–1.70)
About once a week	11 (5577)	13 (229)	1.33 (1.10–1.62)

	Responders % (No.)[a,b]	Nonresponders % (No.)[a,b]	OR (95% CI)
Less often than once a week	24 (11,913)	26 (449)	1.22 (1.04–1.43)
Never consumed alcohol[c]	61 (30,964)	56 (955)	1.00
Missing	<1 (39)	<1 (1)	0.88[d] (0.21–3.68)
Education (years); mean (no.)	7 (50,634)	6 (1714)	0.99[e] (0.98–1.00)
Missing	— (10)	— (1)	2.08[d] (0.34–12.78)
Age at first intercourse years); mean (no.)	22 (50,644)	21 (1715)	0.97[e] (0.96–0.99)
Missing	— (0)	— (0)	N/A

[a] Sample size after accounting for the probability of being selected into the sample using the survey weights provided by the National Family Health Survey 3.

[b] Column percent unless otherwise noted (Note: Values may not total 100% due to rounding).

[c] Reference category.

[d] Reference category is “not missing.”

[e] For continuous measures, OR corresponds to a 1-unit increase.

TABLE 2

Prevalence[a] of Commercial Sexual Contact During the Previous 12 Months by Current Marital Status ith and Without Adjustment for Item Nonresponse

	Prevalence % (95% CI)
All subjects	
Complete-case analysis	0.9 (0.8–1.0)
Adjusting for item nonresponse using	
Inverse-probability weighting	1.1 (1.0–1.2)
Outcome regression	1.1 (1.0–1.2)
Doubly robust	1.1 (1.0–1.2)
Married subjects	
Complete-case analysis	0.4 (0.4–0.5)
Adjusting for item nonresponse using	
Inverse-probability weighting	0.4 (0.4–0.5)
Outcome regression	0.4 (0.4–0.5)
Doubly robust	0.4 (0.4–0.5)
Unmarried subjects	
Complete-case analysis	6.9 (5.8–8.1)
Adjusting for item nonresponse using	
Inverse-probability weighting	12.1 (10.7–14.0)
Outcome regression	12.4 (11.3–13.3)
Doubly robust	12.3 (10.8–13.4)

[a] Prevalence expressed as a percentage.

TABLE 3

Model Goodness-of-fit Statistics by Current Marital Status, Using the Doubly Robust Estimator as the Standard

	Prevalence % (SD)[a]	Test Statistic	*P*
All subjects			
Doubly robust vs.	1.1 (0.1)		
Inverse-probability weighting	1.1 (0.1)	0.85	0.36
Outcome regression	1.1 (0.05)	0.10	0.75
Married subjects			
Doubly robust vs.	0.4 (0.03)		
Inverse-probability weighting	0.4 (0.03)	0.001	0.97
Outcome regression	0.4 (0.03)	0.002	0.96
Unmarried subjects			
Doubly robust vs.	12.1 (0.7)		
Inverse-probability weighting	12.4 (0.8)	0.23	0.63
Outcome regression	12.3 (0.5)	0.19	0.66

[a] Prevalence and standard deviation expressed as a percentage.