AIDS Behav (2012) 16:1382–1393
DOI 10.1007/s10461-011-0125-6

SUBSTANTIVE REVIEW

Are We Missing the Importance of Missing Values in HIV Prevention Randomized Clinical Trials? Review and Recommendations

Ofer Harel · Jennifer Pellowski · Seth Kalichman

Published online: 6 January 2012

Abstract Missing data in HIV prevention trials is a common complication to interpreting outcomes. Even a small proportion of missing values in randomized trials can cause bias, inefficiency and loss of power. We examined the extent of missing data and methods in which HIV prevention randomized clinical trials (RCT) have managed missing values. We used a database maintained by the HIV/AIDS Prevention Research Synthesis (PRS) Project at the Centers for Disease Control and Prevention (CDC) to identify related trials for our review. The PRS cumulative database was searched on June 15, 2010 and all citations that met the following criteria were retrieved: All RCTs which reported HIV/STD/HBV/HCV behavioral interventions with a biological outcome from 2005 to present. Out of the 57 intervention trials identified, all had some level of missing values. We found that the average missing values per study ranged between 3 and 97%. Averaging over all studies the percent of missing values was 26%. None of the studies reported any assumptions for managing missing data in their RCTs. Under some relaxed assumptions discussed below, we expect only 12% of studies to report unbiased results. There is a need for more detailed and thoughtful consideration of the missing data problem in HIV prevention trials. In the current state of managing missing data we risk major biases in interpretations. Several viable alternatives are available for improving the internal validity of RCTs by managing missing data.

O. Harel (✉)
Department of Statistics, University of Connecticut,
215 Glenbrook Road, Unit 4120, Storrs, CT 06269-4120, USA
e-mail: ofer.harel@uconn.edu

J. Pellowski · S. Kalichman
Department of Psychology, University of Connecticut,
Storrs, CT, USA

Keywords Incomplete data · Missing data · Bias · HIV prevention · RCT

Introduction

The validity of statistical inferences is at risk when analyzed data are incomplete, especially if missing data are handled incorrectly. It has been shown that even very small proportions of incomplete cases (in RCTs) can lead to substantial missing information, and misleading inferences [1]. Although the statistical tools to deal with incomplete data are available in statistics and biostatistics literature [2–4], the degree to which HIV prevention scientists are applying them to their studies is unknown. For example, although drop-out is a common complication in longitudinal studies of health and health behavior, it is still the convention to use only the available data [2, 5, 6]. It has been shown repeatedly that ignoring the problems caused by missing data could lead to biased results, flawed interpretation, loss of statistical power and inefficiency [2, 5].

Many studies show that incomplete data may differ by key variables such as treatment group, gender, age, race, and education level [7–9]. Hence, we expect a higher probability of nonresponse for some subgroups compared with others. Differential missing data can lead to differences between those with complete data and those with incomplete data, causing a lack of generalizability to nonresponders. Despite this fact, one of the most commonly used missing data techniques is list-wise deletion, which makes use of complete case data to the exclusion of cases with incomplete data. When study completers differ substantively from non-completers, statistical conclusions drawn from the selected data will be particularly misleading. Although deleting cases with incomplete data is straight-forward and is the default in many statistical packages (e.g. SAS, SPSS, MINITAB), this technique may lead to important biases and loss of statistical power. Fortunately, methods have been developed to handle missing data with significant advantages over case deletion. The purpose of this commentary is to review the techniques used for managing missing data and assumptions for managing missing data for recent published HIV prevention trials.

In this review we examine the missing data assumptions, their applications, and their solutions. Our focus is on the extent of missing data in HIV prevention trials and the implications for interpreting findings. We conclude with some recommendations for managing missing data in future prevention trials.

Missing Data Assumptions

Prior to examining the methods used for managing missing data in HIV prevention trials, we review the underlying assumptions for managing missing data. Assumptions for managing missing values are built upon some conceptual mechanisms. These mechanisms can be thought of as the reasons for missing values. These assumptions are important to understand in order to choose the correct analysis procedures. It is also very important to report the assumptions so researchers reading manuscripts will know the exact assumptions made. The main mechanisms for missing values are: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR) [3, 10].

MCAR is what most people would think if told that data was randomly missing. Under the MCAR mechanism, the observed data are a random subset of the hypothetical, (but unobserved) complete data set, and are representative of the hypothetically complete set and population. This happens when missingness is unrelated to values in the data set, either missing or observed [3]. Consider an HIV prevention study where HIV status is missing due to a random error in data entry, this condition of "nonresponse" would be MCAR; the missingness is unrelated to the response variables. Another analogue is to think of MCAR as scenario in which a lightning strike destroyed certain parts of the data completely by chance.

MAR should be thought as conditional missingness. Under the MAR assumption, missingness can be related to an observed part of the data. For example, in the same HIV prevention trial, if HIV status is missing as a function of age and gender alone, having complete data of the variables age and gender will constitute a MAR mechanism. Also, consider a case that missing values are more prevalent in the treatment arm relative to control arm, as long as we have the treatment assignment, this will be considered MAR mechanism.

When MAR cannot be assumed, we have to assume the data are missing at some non-random mechanism, MNAR. Under this assumption missing values can be due to unobserved (missing) values, even after controlling for other variables. For example, if HIV status is more likely to be missing for individuals whose unobserved HIV status is positive, then unobserved HIV status values are MNAR. In this case, the observed data represent a subgroup of participants whose HIV status is more likely to be negative. Clearly, statistical inferences derived from the available data would be unrepresentative of non-responders.

The caveat is that the distinction between MAR and MNAR assumptions cannot be verified with unplanned missingness without follow up with non-responders, i.e. getting more information about the missing values. Obviously, following up non-respondents does not occur in prevention trials. The distinction between MAR and MNAR is important in order to define ignorable missingness.

When data are complete, researchers have to come up with a substantive model (e.g. regression model) in order to explain the data. When data are incomplete, researchers have to model not only the available data, but the missingness as well unless they are willing to assume some ignorability assumption. Ignorable Missingness refers to whether or not the mechanism accounting for the missingness must be explicitly modeled together with the substantive model [3, 10]. *Often, researchers mistakenly confuse the ignorability concept with the claim that the missing data can be ignored*. However, when making statistical inferences the missing data should never be ignored. The difference between ignorable and non-ignorable missingness is whether a separate model for missingness must be included together with the substantive model. Under the ignorability assumption only the substantive model need to be specified (e.g. regression model), while under non-ignorable set-ups joint models for the substantive and missingness models need to be specified using either selection models, [3] pattern-mixture models [3] or shared parameter models [3].

We can assume ignorable missingness when data are MAR and when the missingness has no bearing on the substantive model parameters [5]. When using likelihood-based or Bayesian estimation techniques MAR and MCAR can reasonably be treated as ignorable [11] which means that no additional information is required about the distribution of the nonresponse [10]. However, when using semi-parametric techniques such as generalized estimating equations (GEE) [12], *only* the MCAR condition is ignorable. Always, if data are MNAR, the condition is non-ignorable.

Possible Solutions to Missing Data

There are several statistical procedures that deal with incomplete data. We introduce a few of them here with relative advantages and disadvantages.

Complete Case Analysis

The most common and straightforward approach to dealing with incomplete data is to omit those subjects with incomplete data from the analysis. This is often the default method of handling incomplete data by statistical procedures in commonly-used statistical software packages, such as Stata [13], SAS [14], and SPSS [15]. The advantage of case deletion is that it can be used for any kind of statistical analysis and no special computational methods are required, when data are MCAR this approach may yield results that are unbiased [3, 16]. However the disadvantages are loss of power, inefficiency, and possible bias. Reduced sample size may impose limitations on the types of analyses that can be conducted, and may preclude the use of large-sample techniques. In particular, consider data trial with 10 variables and each variable is missing 5% of it values randomly (MCAR). Using case deletion will reduce the data to around 60%, larger data sets and larger rates of missing values can have even bigger impacts (e.g. 20 variable and 10% missing values will result in 12% of the data).

Generalized Estimating Equation (GEE)

For correlated data, generalized estimating equation (GEE) [12], became one of the most used procedures in practice. GEE Procedures are used regularly is large studies when clustering or longitudinal structures are desired or unavoidable. Using this procedure, the researcher specifies a working correlation structure but this structure does not have to hold exactly. This constitutes as a semi-parametric procedure as the model for the data has to be evaluated but the correlations are not of main interest. Unfortunately, GEE of incomplete data is unbiased only under the MCAR assumption. However, an infrequently used extension, weighted GEE [17], allows missing data under the MAR condition.

Maximum Likelihood

Maximum likelihood is a large sample technique looking for the parameter estimates that have the greatest likelihood of producing the observed data, given a specified model. These parameters are called the maximum likelihood estimates (MLE) [3, 16, 18, 19]. Maximum likelihood estimation does not require observations to be balanced; individuals may have differing numbers of observations spaced at different intervals. All complete and partially-observed cases contribute to the maximum likelihood estimation of model parameters, and the missing data values are treated as random variables to be averaged over [20].

Bayesian Estimation and Multiple Imputation

Bayesian estimation techniques use prior information (distribution) together with the likelihood distribution to produce a posterior distribution. The estimates drawn from the posterior distribution takes into account prior knowledge and the distribution of the data [3, 21]. Usually this is being done using Markov Chain Monte Carlo (MCMC) estimation which allows the analysis of the data without dropping cases.

Multiple imputation (MI) [2] replaces missing observations with $m>1$ plausible values to complete multiple alternative completed data sets [3, 4, 11]. The complete data sets are analyzed individually, and multiple parameter estimates are combined. MI provides the advantage of allowing complete-data analytical routines while accounting for uncertainty of estimates due to imputation. In the past a small number of imputations were considered adequate for efficient parameter estimation [5], but many more may be needed to improve efficiency [22–24].

Missing Data in HIV Prevention Trials

Our review was performed with assistance of the HIV/AIDS Prevention Research Synthesis (PRS) Project at the Centers for Disease Control and Prevention (CDC) [25]. This is database of prevention studies maintained by the CDC for monitoring evidence-based HIV behavioral interventions. The review process is conducted using well-established systematic procedures for searching and reviewing the intervention research literature. Our search was based on automated strategies in four electronic bibliographic databases (EMBASE, MEDLINE, PsycINFO, and Sociological Abstracts) together with a manual search which involved reviewing approximately 35 journals to identify articles not yet indexed in the electronic databases. More detailed information about the CDC PRS database can be found in the CDC literature [25].

The PRS cumulative database was searched on June 15, 2010 and all citations that met the following criteria were retrieved.

1. Reports of HIV/STD/HBV/HCV behavioral interventions. Rationale, the interpretation of biological and behavioral interventions differs along multiple

dimensions. We chose to exclude biological interventions because the conditions under which data are missing vary considerably from those of behavioral interventions.
2. Based on a randomized control trial (RCT) research design. The RCT is the most rigorous design for testing clinical methods and procedures and findings can be blurred by missing data.
3. Reported a biological outcome. In particular, any sexually transmitted infection/disease endpoint, such as HIV incidence/seroconversion, STD incidence/re-infection, or Hepatitis B or C infection. Biological endpoints in HIV prevention trials avoid self report biases and represent a clinical disease outcome.
4. From 2005 to present. Studies prior to 2005 do not represent the current state of HIV prevention science and the implications for missing data are fewer for these studies.

The search resulted in ($n = 57$) citations that met the inclusion criteria. A reviewer with background in HIV prevention assessed each study and extracted pre-specified needed information (See Table 1). A second reviewer with background in biostatistics assessed a sample of studies and arrived at 100% agreement with the first reviewer. The results of the reviewing process plus descriptive information on the studies were entered into a computer database and are summarized in Table 1 [26–82].

The trials were conducted all over the world with regions/countries ranging from the United States and Mexico in North America, Jamaica in Central America, United Kingdom and Belgium in Western Europe, Russia and Bulgaria in Eastern Europe, Thailand, Philippines and China in Asia and several countries in Sub-Saharan Africa such as Tanzania, Zimbabwe, South Africa, Kenya, Uganda and Madagascar.

All of the trials had some level of missing values. Although not all studies reported them in the same manner, we found that the averaged missing values per study ranged between 3 and 97%. Averaging over all studies the percent of missing values was 26% (median 23%). In this (missing values) sample, values greater then 50% were considered outliers. We had four outliers in our sample. We extracted the information about the missing data levels from the participant flow charts reported in each trial. It is clear from the range of cell sizes that many studies varied in available data for different analyses. We speculate that the majority of missing values are due to missing outcomes, but cannot know it for certain for all studies due to the differences in reporting. However, due to the fact that both types of missing values may bias the results we do not distinguish between them.

None (0%) of the studies reported any information on what missing data assumptions were used in their analyses. In most cases this implies that only analyses under the MCAR the results will be unbiased. The majority of studies (42, 74%) used complete case analysis (CCA) and reduced the sample only for those with complete data (Table 1, references [26–67]). Eight studies (14%) used some variation of GEE analysis which used the whole data (observed and missing) but is potentially biased under MAR assumption (Table 1, references [68–75]). There are few studies that used maximum likelihood estimation (7, 12%) and therefore their results will be unbiased under the MAR assumption (Table 1, references [76–82]). Collins et al. [21] showed that if one collects enough auxiliary information, one can get close to the MAR assumption. Assuming all studies collected enough information so that the MAR assumption is reasonable, and since we know that MCAR rarely happens in practice, only seven studies out of 57 (12%) had some of their analyses done so we can expect them to be unbiased.

The studies reviewed that used complete case analysis ($n = 42$, 74%; references [26–67] did so for many different types of analyses. For example, parametric tests such as t test, F test, and χ^2 tests; non-parametric tests such as Rank tests; regression analyses such as linear regression, multiple regression, logistic regression, Poisson and binomial regressions; Analysis of variance (ANOVA) with its derivatives MANOVA and ANCOVA. All these analyses are in danger of being biased under MAR and MNAR, and have a chance of being unbiased under MCAR.

The studies using GEE in our review ($n = 8$, 14%; references [68–75]) reported conventional (unweighted) GEE, which implies possible biased results unless missing data were missing completely at random (MCAR).

Studies using maximum likelihood estimation ($n = 7$, 12%; references [76–82]) used Generalized multilevel models and linear mixed models. These procedures are also called generalized linear mixed model, mixed effect linear regression, random effect regression, and multilevel random effect model. These procedures are expected to be unbiased under both MAR and MCAR.

Bayesian and multiple imputation procedures are well equipped to deal with incomplete data. Unfortunately none of the trials we reviewed used these procedures. Both of these procedures (with adequate modeling) can be unbiased under MAR, MCAR and MNAR scenarios.

Recommendations

With any applied research and in particular RCTs, the best thing to do with regard to missing data is to avoid it. The second best thing is to plan for it, understand it and address it with appropriate modeling techniques. (1) *Plan for missingness*. Researchers should anticipate unavoidable

Table 1 Summary of behavioral HIV prevention trials reporting biological outcomes

Author (s)	Year	Country	Target pop	Outcome	Min missing values (%)	Max missing values (%)	Mean missing values (%)	Mentioned missing in analysis	Main analysis	Missingness assumption
CCA										
Abou-Saleh et al.[26]	2008	United Kingdom	Injection drug users, users in treatment	Hepatitis C: blood test 12 months	13.95	44.23	25.65	No	Chi square	N.R.
Artz et al. [27]	2005	United States	Clinic patients, Urban/ Inner City, female	STD: test, Pregnancy: test; 6 months	16.39	16.39	16.39	Yes	Mantel–Haenszel rate ratio	N.R.
Becker et al. [28]	2010	Tanzania	Clinic patients, pregnancy, female	HIV: test 9 months	7.49	61.32	38.12	No	ANOVA, *f* test, Chi square	N.R.
Boyer et al. [29]	2005	United States	Military, female	STD: test, Pregnancy: test 14 months	10.96	43.93	24.39	No	Multivariate logistic regression, Huber-White sandwich estimator, Chi square	N.R.
Brems et al. [30]	2009	United States	Alcohol users, users in treatment	HIV: self-report, STD: test 3–4 months	0	23.08	12.59	No	MANOVA	N.R.
Chacko et al. [31]	2010	United States	Youth and young adults, clinic patients, HIV negative, Urban, female, high sexual risk behaviors	STD: test 12 months	17.93	39.06	29.99	No	Multiple logistic regression	N.R.
Corbett et al. [32]	2007	Zimbabwe	HIV negative	HIV: test 3 month	26.90	69.73	48.36	Yes	Standardized Incidence Ratio, Wilcoxon Rank Sum, *t* test	N.R.
Crosby et al. [33]	2009	United States	Heterosexual, Black/ African American, male, Young adult, clinic patient, STD diagnosis	STD: chart abstraction 6 month	25.53	26.40	25.97	Yes	Logistic regression, linear regression, sensitivity analysis	N.R.
DiClemente et al. [34]	2009	United States	Female, Urban, Black/ African American, youth, clinic patients	STD: test 12 months	0	16.95	10.85	No	Logistic regression, linear regression	N.R.
Gregson et al. [35]	2007	Zimbabwe	Commercial sex worker clients, community members	HIV: test, STI: self-report 3 years	19.18	28.31	23.69	No	Adjusted incidence rate ratio, *t* test, coefficient of variation	N.R.
Grimley et al. [36]	2009	United States	Clinic patients, Urban, Lower income/indigent	STD: test 6 months	24.63	41.85	33.24	Yes	Chi square, Fisher exact test, logistic regression	N.R.
Hoke et al. [37]	2007	Madagascar	Female, Commercial sex workers	STD: test 18 months	1.76	9.71	5.10	Yes	Logistic regression	N.R.
Jemmott et al. [38]	2005	United States	Youth, Clinic Patients, Black/African American/ Spanish/Hispanic/Latino, Urban, Female, Low income/indigent	STD: test 12 months	4.26	14.04	8.11	No	Chi square, ANOVA, Poission regression, *t* test	N.R.
Kalichman et al. [39]	2005	United States	Clinic Patients	STD: chart abstraction 12 months	6.38	29.08	20.48	Yes	ANOVA, Multinomial logistic regression	N.R.

Table 1 continued

Author (s)	Year	Country	Target pop	Outcome	Min missing values (%)	Max missing values (%)	Mean missing values (%)	Mentioned missing in analysis	Main analysis	Missingness assumption
Koblin et al. [40]	2009	United States	Crack and other substance users, HIV negative, Female, High sexual risk behavior	HIV: test 12 months	10.83	19.48	14.64	No	McNemar discordant pair analysis	N.R.
Lau et al. [41]	2008	China	Males, Homosexual, other high risk sexual behavior	STD: self-report	59.00	59.00	59.00	Yes	Chi square	N.R.
Lau et al. [42]	2010	China	Adult, Truck Drivers, Male, Commercial sex worker clients, other high risk sexual behaviors	STD: self-report, HIV: test 6 months	3.90	4.55	4.15	No	Chi square	N.R.
Low et al. [43]	2006	United Kingdom	Clinic patients, STD diagnosis, sex partners of those with other high risk	STD: test 6 weeks	0	30.88	20.57	Yes	Regression models	N.R.
Matovu et al. [44]	2007	Uganda	Uncircumcised, HIV negative, male, high risk	STD: test 24 month	7.97	60.55	26.15	No		N.R.
McNulty et al. [45]	2008	United Kingdom	Service providers	STD: test 10 months	0	9.52	6.94	Yes	Logarithmic link function,	N.R.
Metcalf et al. [46]	2005	United States	Clinic patients, HIV negative	STD: test 12 months	1.26	29.04	14.02	Yes	Relative Risk	N.R.
Morisky et al. [47]	2005	Philippines	Female, commercial sex worker	STD: chart abstraction 2 years				No	Chi square, ANOVA, regression models	N.R.
Ngugi et al. [48]	2007	Kenya	HIV negative, Female, commercial sex worker	HIV: test, STD: test 12 months				Yes	Wilcoxon signed rank test, Chi square	N.R.
Patterson et al. [49]	2008	Mexico	HIV negative, High risk residential area, commercial sex worker, other high risk sexual behavior	HIV: test, STD: test 6 months				No	Repeated measures ANCOVA, Cohen d test, Chi square, Poisson regression	N.R.
Peipert et al. [50]	2008	United States	Female, heterosexual, high risk	STD: test, Pregnancy: test 24 month	38.97	50.00	44.49	No	Logistic regression, log rank statistic	N.R.
Petersen et al. [51]	2007	United States	Clinic patients, female, high sexual risk behavior	STD: test, Pregnancy: test 12 month	3.53	3.53	3.53	Yes	McNemar's Chi square	N.R.
Pronyk et al. [52]	2006	Southern African [Limpopo province)	Rural, general population, low income/indigent	HIV: test 2 years	.93	42.29	19.78	Yes	Coefficient of variance	N.R.
Ross et al. [53]	2007	Tanzania	Youth, rural residence	HIV: test, STD: test, Pregnancy: test 3 years	26.37	27.64	27.01	Yes	Unadjusted risk ratio, ANOVA, *t* statistic, logisitic and Poisson regression	N.R.

Table 1 continued

Author (s)	Year	Country	Target pop	Outcome	Min missing values (%)	Max missing values (%)	Mean missing values (%)	Mentioned missing in analysis	Main analysis	Missingness assumption
Roye et al. [54]	2006	United States	Youth, HIV negative, Black/African American/Spanish/Hispanic/Latino, Urban, females, high risk sexual behavior	Chlamydia: test, Other STD: self-report 12 months	27.16	64.20	41.14	No	Chi square, logistic regression, *t* test	N.R.
Samet et al. [55]	2008	Russia	Drug users in treatment	HIV: test 6 months	10.50	20.44	15.47	No	Logistic regression, median regression, Chi square	N.R.
Scholes et al. [56]	2007	United States	Young adult, males	STD: test	92.18	99.30	97.24	No		N.R.
Semaan et al. [57]	2009	United States	Adult, crack users, injection drug users, other substance users, clinic patients, HIV negative, heterosexuals	STD: test 12 months				No	Unadjusted odds ratio, multivariate logistic regression, multiple logistic regression	N.R.
Spielberg et al. [58]	2005	United States	Injection drug users, high risk behavior	HIV: test	77.16	97.08	87.76	No	Multiple logistic regression, Wald test, univariate logistic model; Chi square	N.R.
Steiner et al. [59]	2006	Jamaica	Clinic patients, male	STD: test 6 months	38.46	41.26	39.86	Yes	Wilcoxon rank sum test	N.R.
Thurman et al. [60]	2008	United States	STD diagnosis, black or African American, Spanish/Hispanic/Latino, female	HIV: test, STD: test 12 months	0	15.38	5.01	No	Pearson's Chi square, multivariate logistic regression analysis	N.R.
Trenholm et al. [61]	2008	United States	Youth, black or African American, Spanish/Hispanic/Latino, white, rural, urban, low income/indigent	STD: self-report, Pregnancy: self-report 42–78 months	35	43	39	Yes	Weighted regression model	N.R.
Verhoeven et al. [62]	2005	Belgium	Service providers	STD: test	20	33.33	26.67	Yes	*t* test, Chi square, ANOVA, Intracluster correlation coefficient (ICC)	N.R.
Wechsberg et al. [63]	2006	Southern Africa [Pretoria]	Crack users, people of color, females, commercial sex workers, other high sexual risk behaviors	STD- self-report				No	Bivariate analysis, logistic regression analysis; McNemar; *t* test;	N.R.
Wechsberg et al. [64]	2008	South Africa	Alcohol users, other substance users, people of color, females, low income/indigent	STD: test 1 month				Yes	*t* test, ANCOVA	N.R.
Wolitski et al. [65]	2005	United States	HIV positive, Male, Bisexual and Homosexual	STD: test 6 months	9.20	50.92	23.03	Yes	Chi square, logistic regression for dichotomous data, cumulative logit models for ordinal data, odds ratio (OR),	N.R.

Table 1 continued

Author (s)	Year	Country	Target pop	Outcome	Min missing values (%)	Max missing values (%)	Mean missing values (%)	Mentioned missing in analysis	Main analysis	Missingness assumption
Wu et al. [66]	2005	China	High risk	STD: test 1 year				No	Multivariate analysis	N.R.
Wu et al. [67]	2007	China	Injection drug users	Hepatitis: test, HIV: test 12 months	87.61	87.61	87.61	Yes	Chi square, *t* test	N.R.
GEE										
Jemmott et al. [68]	2007	United States	Clinic Patients, Black/ African American, Urban, Female	STD: test 12 months	5.65	17.28	13.08	Yes	GEE, binomial-error model with logit link functions, Poisson error models with log-link functions, Wald Chi square test	N.R.
Kissinger et al. [69]	2005	United States	Clinic patients, STD diagnosis, Male, Heterosexual, High risk sexual behavior	STD: test 1 month	21.19	21.19	21.19	No	Bivariate GEE, multivariate GEE	N.R.
Latkin et al. [70]	2009	United States and Thailand	Injection Drug Users, Users out of treatment, HIV negative, Urban, Living in a High Risk area, Sex partners of IDU	HIV: test 24 months	17	18	17.5	No	GEE modeling methods	N.R.
Lutchers et al. [71]	2008	Kenya	Adult, clinic patients, HIV positive	STD: test 12 months	2.99	15.38	9.19	No	Bivariate cross-sectional logistic regression model, GEE	N.R.
Sherman et al. [72]	2008	Thailand	Young adults, substance users, sex partners of those with other high risk	Hepatitis: test, HIV: test, STD: test 12 months	4.85	13.48	9.98	No	Logistic regression, GEE, Wald test,	N.R.
Stephenson et al. [73]	2008	United Kingdom	Youth, Rural and urban, high sexual risk behavior	STD: self-report, Pregnancy: self-report 7 years	6.64	58.09	23.66	Yes	GEE, logistic regression model	N.R.
Wilson et al. [74]	2009	United States	Clinic patients with STD diagnosis, People of color, living in high risk area, urban, sex partners of those with other high risk	STD: test 6 months	3.71	13.49	7.89	Yes	Multiple logistic regression model; GEE	N.R.
Wingood et al. [75]	2006	United States	Youth, Black or African American, Other high risk sexual behavior	STD: test 12 months	10.69	16.44	13.02	Yes	*t* test and Chi square, logistic regression GEE, linear regression GEE	N.R.
ML										
Carey et al. [76]	2010	United States	Clinic Patients, High Sexual Risk Behavior	STD: test 12 months	0	47.60	27.97	Yes	Generalized multilevel models using logit link, GEE, linear mixed model	N.R.

Table 1 continued

Author (s)	Year	Country	Target pop	Outcome	Min missing values (%)	Max missing values (%)	Mean missing values (%)	Mentioned missing in analysis	Main analysis	Missingness assumption
Garfein et al. [77]	2007	United States	Young Adults, Injection Drug Users, HIV negative	Hepatitis: test, HIV: test 6 months	22.74	39.01	29.94	Yes	Negative binomial fit of repeated count data; zero-inflated negative binomial model; cumulative logit models for repeated ordinal data; Poisson regression	N.R.
Jeweks et al. [78]	2008	South Africa	Rural, Youth and young adults	HIV: test, STD: test, Pregnancy: self-report 24 months	21.40	28.08	24.16	Yes	Fitting generalizes linear mixed models, GEE model	N.R.
Kelly et al. [79]	2006	Bulgaria	Low income/indigent, Male, Social group leaders	STD: test 12 months	0	10.27	3.81	Yes	Mixed effects linear regression, logistic regression	N.R.
Kershaw et al. [80]	2009	United States	Young adults, Clinic patients, Pregnancy, Female	STD: test, Pregnancy: self-report 12 months	9.75	28.06	16.33	Yes	Random effects regression, post hoc analysis, generalized linear mixed models, logistic regression	N.R.
Marion et al. [81]	2009	United States	STD diagnosis, Black/African American, High risk area, Female, Low income/indigent	STD: test 15 months	23.53	51.74	38.75	Yes	Multilevel random effects model	N.R.
Tebb et al. [82]	2005	United States	Youth, Clinic patients, Male, High sexual risk behavior	STD: test 18 months				No	Repeated measures ANCOVA, Mixed effects model analysis, ANOVA, *f* test	N.R.

NR None reported, *CCA* Complete case analysis, *GEE* Generalized estimating equations, *ML* Maximum likelihood missing data. Variables determined to relate to nonresponse should be identified and measured. (2) *Minimize nonresponse*. Incorporate procedures into the study plan to reduce missed assessments and ensure regular review of data. (3) *Determine the mechanism of missingness*. Researchers should test the assumption of MCAR, and carefully consider the plausibility of ignorable missingness. (4) *Apply appropriate techniques*. Techniques such as ML, GEE, Bayesian, and MI are effective when applied appropriately under proper assumptions, but will provide misleading results when implemented incorrectly. (5) *Report missingness and techniques used*. Researchers should fully describe missing data methods; the incomplete data structure, missing data assumptions, and the techniques selected to handle them. (6) *Sensitivity analysis*. Researchers should analyze their data under different missing data assumptions and report the differences the missing data assumptions make on conclusions.

Conclusions

In this review, we examined the past 5 years of behavioral HIV prevention RCTs reporting biological outcomes. We found that all the reviewed publications had varying degrees of missing data, and yet none reported assumptions regarding the management of missing data. Most studies used statistical methods which are most probably biased to most common missing data assumptions. In particular, most studies reviewed used complete case analysis ($n = 42$, 74%; references [26–67]), eight studies (14%) used some GEE type procedures [68–75], seven studies (12%) used maximum likelihood procedures [76–82], while none used Bayesian or multiple imputation procedures. Although we cannot comment on the direction and magnitude of the bias, the fact that approximately 88% (74 + 14%) of the studies reported possibly biased results (under the MAR assumption) is alarming. We touched on some available methodology more appropriate to deal with incomplete data and gave some general recommendations of how to deal with incomplete data.

The idea that missing data can impact the results of clinical trials is not new. Researchers in many fields have shown the risk of ignoring the missing data complications [3, 5, 11, 16]. Recently there were several reviews which examined the problem from different directions. One study, for example, reports on the use and abuse of missing data procedures in longitudinal data settings in developmental psychology [83], while another discuss issues of noncompliance in randomized trials [84].

We hope researchers will attend more closely to the missing data in HIV prevention trials. Methods for incomplete data are available and offer the potential for unbiased and efficient estimation. Not thinking of the missing data problem does not mean the problem goes away. Leaving the problem to the pre-specified statistical software will, in most cases, reduce the data to complete set, an unsatisfactory solution to missing data.

We entreat researchers to disclose missing data rates, missing data assumptions, and the methods used to address them in published work. We hope that this practice will promote the application of proper techniques and a greater understanding of the methodological and statistical issues involved in handling incomplete data.

Acknowledgments This project was partially supported by Award Number K01MH087219 from the National Institute of Mental Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Mental Health or the National Institutes of Health.

References

1. Belin TR. Missing data: what a little can do, and what researchers can do in response. Am J Ophthalmol. 2009;148(6):820–2.
2. Rubin DB. Multiple imputation for nonresponse in surveys. New York: Wiley; 1987.
3. Little RJA, Rubin DB. Statistical analysis with missing data. New York: Wiley; 2002.
4. Harel O, Zhou XH. Multiple imputation: review of theory, implementation and software. Stat Med. 2007;26:3057–77.
5. Schafer JL. Imputation of missing covariates under a multivariate linear mixed model; 1997.
6. Molenberghs, G, Kenward MG. Missing data in clinical studies. 1st ed. New York: Wiley; 2007.
7. Cranford JA, McCabe SE, Boyd CJ, Slayden J, Reed MB, Ketchie JM, et al. Reasons for nonresponse in a web-based survey of alcohol involvement among first-year college students. Addict Behav. 2008;33:206–10.
8. Kupek E. Determinants of item nonresponse in a large national sex survey. Arch Sex Behav. 1998;27:581–94.
9. Ngo-Metzger Q, Kaplan SH, Sorkin DH, Clarridge BR, Phillips RS. Surveying minorities with limited-English proficiency: does data collection method affect data quality among Asian Americans? Med Care. 2004;42(9):893–900.
10. Rubin DB. Inference and missing data. Biometrika. 1976;63: 581–92.
11. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods. 2002;7:147–77.
12. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73:1322.
13. StataCorp. Stata Statistical Software: Release Special Edition 7.0. College Station TX: Stata Corporation; 2002.
14. SAS, SAS User's Guide: Statistics, SAS Institute, Cary, NC (1982).
15. Norusis MJ. SPSS Advanced Statistics 6.1. Chicago: SPSS Inc; 1998.
16. Little RJA, Rubin DB. The analysis of social science data with missing values. Sociol Methods Res. 1989;18:292–326.
17. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc. 1994;89:846–66.
18. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J Royal Stat Soc, Series B. 1977;39:1–38.

19. Hox J. Multilevel analysis: techniques and applications. Mahwah: Erlbaum; 2002.
20. Collins LM, Schafer JL, Kam C. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Methods. 2001;6:330–51.
21. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. 2nd ed. London: Chapman & Hall/CRC; 2003.
22. Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prev Sci. 2007;8:206–13.
23. Harel O. Inferences on missing information under multiple imputation and two-stage multiple imputation. Stat Method. 2007;4:75–89.
24. Hershberger SL, Fisher DG. A note on determining the number of imputations for missing data. Struct Equ Model. 2003;10:648–50.
25. The HIV/AIDS Prevention Research Synthesis (PRS) Project, Centers for Disease Control and Prevention (CDC), February, 10 2011. Available at: http://www.cdc.gov/hiv/topics/research/prs/.
26. Abou-Saleh M, Davis P, Rice P, et al. The effectiveness of behavioural interventions in the primary prevention of Hepatitis C amongst injecting drug users: a randomised controlled trial and lessons learned. Harm Reduct J. 2008;31(5):25.
27. Artz L, Macaluso M, Meinzen-Derr J, et al. A randomized trial of clinician-delivered interventions promoting barrier contraception for sexually transmitted disease prevention. Sex Transm Dis. 2005;32:672–9.
28. Becker S, Mlay R, Schwandt HM, Lyamuya E. Comparing couples' and individual voluntary counseling and testing for HIV at antenatal clinics in Tanzania: a randomized trial. AIDS Behav. 2010;14(3):558–66.
29. Boyer CB, Shafer MA, Shaffer RA, et al. Evaluation of a cognitive-behavioral, group, randomized controlled intervention trial to prevent sexually transmitted infections and unintended pregnancies in young women. Prev Med. 2005;40:420–31.
30. Brems C, Dewane SL, Johnson ME, Eldridge GD. Brief motivational interventions for HIV/STI risk reduction among individuals receiving alcohol detoxification. AIDS Educ Prev. 2009; 21:397–414.
31. Chacko MR, Wiemann CM, Kozinetz CA, von Sternberg K, Velasquez MM, Smith PB. Efficacy of a motivational behavioral intervention to promote chlamydia and gonorrhea screening in young women: a randomized controlled trial. J Adolesc Health. 2010;46:152–61.
32. Corbett EL, Makamure B, Cheung YB, et al. HIV incidence during a cluster-randomized trial of two strategies providing voluntary counselling and testing at the workplace, Zimbabwe. AIDS. 2007;21:483–9.
33. Crosby R, DiClemente RJ, Charnigo R, Snow G, Troutman A. A brief, clinic-based, safer sex intervention for heterosexual African American men newly diagnosed with an STD: a randomized controlled trial. Am J Public Health. 2009;99:S96–103.
34. DiClemente RJ, Wingood GM, Rose ES, et al. Efficacy of sexually transmitted disease/human immunodeficiency virus sexual risk-reduction intervention for African American adolescent females seeking sexual health services: a randomized controlled trial. Arch Pediatr Adolesc Med. 2009;163:1112–21.
35. Gregson S, Adamson S, Papaya S, et al. Impact and process evaluation of integrated community and clinic-based HIV-1 control: a cluster-randomised trial in eastern Zimbabwe. PLoS Medicine. 2007;4(3):e102.
36. Grimley DM, Hook EW. III. A 15-minute interactive, computerized condom use intervention with biological endpoints. Sex Transm Dis. 2009;36:73–8.
37. Hoke TH, Feldblum PJ, Van Damme K, et al. Temporal trends in sexually transmitted infection prevalence and condom use following introduction of the female condom to Madagascar sex workers. Int J STD AIDS. 2007;18:461–6.
38. Jemmott JB III, Jemmott LS, Braverman PK, Fong GT. HIV/STD risk reduction interventions for African American and Latino adolescent girls at an adolescent medicine clinic: a randomized controlled trial. Arch Pediatr Adolesc Med. 2005;159:440–9.
39. Kalichman SC, Cain D, Weinhardt L, et al. Experimental components analysis of brief theory-based HIV-AIDS risk reduction counseling for sexually transmitted infection patients. Health Psychol. 2005;24:198–208.
40. Koblin BA, Bonner S, Hoover DR, et al. A randomized trial of enhanced HIV risk-reduction and vaccine trial education interventions among HIV-negative, high-risk women who use non injection drugs: the UNITY Study. JAIDS J Acquir Immune Defic Syndr. 2009;53:378–87.
41. Lau JT, Lau M, Cheung A, Tsui HY. A randomized controlled study to evaluate the efficacy of an Internet-based intervention in reducing HIV risk behaviors among men who have sex with men in Hong Kong. AIDS Care. 2008;20:820–8.
42. Lau JT, Tsui HY, Cheng S, Pang M. A randomized controlled trial to evaluate the relative efficacy of adding voluntary counseling and testing (VCT) to information dissemination in reducing HIV-related risk behaviors among Hong Kong male cross-border truck drivers. AIDS Care. 2010;22:17–28.
43. Low N, McCarthy A, Roberts TE, et al. Partner notification of chlamydia infection in primary care: randomised controlled trial and analysis of resource use. BMJ. 2006;332:14–8.
44. Matovu JK, Ssempijja V, Makumbi FE, et al. Sexually transmitted infection management, safer sex promotion and voluntary HIV counselling and testing in the male circumcision trial, Rakai, Uganda. Reprod Health Matters. 2007;15:68–74.
45. McNulty CA, Thomas M, Bowen J, et al. Interactive workshops increase chlamydia testing in primary care–A controlled study. Fam Pract. 2008;25:279–86.
46. Metcalf CA, Douglas JM Jr, Malotte CK, et al. Relative efficacy of prevention counseling with rapid and standard HIV testing: a randomized, controlled trial (RESPECT-2). Sex Transm Dis. 2005;32:130–8.
47. Morisky DE, Chiao C, Stein JA, Malow RM. Impact of social and structural influence interventions on condom use and sexually transmitted infections among establishment-based female bar workers in the Philippines. J Psychol Hum Sex. 2005;17:45–63.
48. Ngugi EN, Chakkalackal M, Sharma A, et al. Sustained changes in sexual behavior by female sex workers after completion of a randomized HIV prevention trial. J Acquir Immune Defic Syndr. 2007;45:588–94.
49. Patterson TL, Mausbach B, Lozada R, et al. Efficacy of a brief behavioral intervention to promote condom use among female sex workers in Tijuana and Ciudad Juarez, Mexico. Am J Public Health. 2008;98:2051–7.
50. Peipert JF, Redding CA, Blume JD, et al. Tailored intervention to increase dual-contraceptive method use: a randomized trial to reduce unintended pregnancies and sexually transmitted infections. Am J Obstet Gynecol. 2008;198(6):630. e1-630.e8.
51. Petersen R, Albright J, Garrett JM, Curtis KM. Pregnancy and STD prevention counseling using an adaptation of motivational interviewing: a randomized controlled trial. Perspect Sex Reprod Health. 2007;39:21–8.
52. Pronyk PM, Hargreaves JR, Kim JC, et al. Effect of a structural intervention for the prevention of intimate-partner violence and HIV in rural South Africa: a cluster randomised trial. Lancet. 2006;368:1973–83.
53. Ross DA, Changalucha J, Obasi AI, et al. Biological and behavioural impact of an adolescent sexual health intervention in Tanzania: a community-randomized trial. AIDS. 2007;21:1943–55.

54. Roye C, Perlmutter Silverman P, Krauss B. A brief, low-cost, theory-based intervention to promote dual method use by black and Latina female adolescents: a randomized clinical trial. Health Educ Behav. 2006;34:608–21.
55. Samet JH, Krupitsky EM, Cheng DM, et al. Mitigating risky sexual behaviors among Russian narcology hospital patients: The PREVENT (Partnership to Reduce the Epidemic Via Engagement in Narcology Treatment) randomized controlled trial. Addiction. 2008;103:1474–83.
56. Scholes D, Heidrich FE, Yarbro P, Lindenbaum JE, Marrazzo JM. Population-based outreach for chlamydia screening in men: results from a randomized trial. Sex Trans Dis. 2007;34:837–9.
57. Semaan S, Neumann MS, Hutchins K, D'Anna LH, Kamb ML, for the Project RESPECT Study Group. Brief counseling for reducing sexual risk and bacterial STIs among drug users-results from project RESPECT. Drug Alcohol Depend. 2009;106:7–15.
58. Spielberg F, Branson BM, Goldbaum GM, et al. Choosing HIV counseling and testing strategies for outreach settings: a randomized trial. J AIDS J Acquir Immune Defic Syndr. 2005;38: 348–55.
59. Steiner MJ, Hylton-Kong T, Figueroa JP, et al. Does a choice of condoms impact sexually transmitted infection incidence? A randomized, controlled trial. Sex Transm Dis. 2006;33:31–5.
60. Thurman AR, Holden AE, Shain RN, Perdue S, Piper JM. Preventing recurrent sexually transmitted diseases in minority adolescents: a randomized controlled trial. Obstet Gynecol. 2008;111:1417–25.
61. Trenholm C, Devaney B, Fortson K, Clark M, Bridgespan LQ, Wheeler J. Impacts of abstinence education on teen sexual activity, risk of pregnancy, and risk of sexually transmitted diseases. J Policy Anal Manage. 2008;27:255–76.
62. Verhoeven V, Avonts D, Vermeire E, Debaene L, Van Royen P. A short educational intervention on communication skills improves the quality of screening for chlamydia in GPs in Belgium: a cluster randomised controlled trial. Patient Educ Couns. 2005;57:101–5.
63. Wechsberg WM, Luseno WK, Lam WK, Parry CD, Morojele NK. Substance use, sexual risk, and violence: HIV prevention intervention with sex workers in Pretoria. AIDS Behav. 2006;10: 131–7.
64. Wechsberg WM, Luseno WK, Karg RS, et al. Alcohol, cannabis, and methamphetamine use and other risk behaviours among black and coloured South African women: a small randomized trial in the Western Cape. Int J Drug Policy. 2008;19:130–9.
65. Wolitski RJ, Gómez CA, Parsons JT, SUMIT Study Team. Effects of a peer-led behavioral intervention to reduce HIV transmission and promote serostatus disclosure among HIV-seropositive gay and bisexual men. AIDS. 2005;19:S99–109.
66. Wu Z, Rou K, Xu C, Lou W, Detels R. Acceptability of HIV/AIDS counseling and testing among premarital couples in China. AIDS Educ Prev. 2005;17:12–21.
67. Wu Z, Luo W, Sullivan SG, et al. Evaluation of a needle social marketing strategy to control HIV among injecting drug users in China. AIDS. 2007;21:S115–22.
68. Jemmott LS, Jemmott JB III, O'Leary A. Effects on sexual risk behavior and STD rate of brief HIV/STD prevention interventions for African American women in primary care settings. Am J Public Health. 2007;97:1034–40.
69. Kissinger P, Mohammed H, Richardson-Alston G, et al. Patient-delivered partner treatment for male urethritis: a randomized, controlled trial. Clin Infect Dis. 2005;41:623–9.
70. Latkin CA, Donnell D, Metzger D, et al. The efficacy of a network intervention to reduce HIV risk behaviors among drug users and risk partners in Chiang Mai, Thailand and Philadelphia, USA. Soc Sci Med. 2009;68:740–8.
71. Luchters S, Sarna A, Geibel S, et al. Safer sexual behaviors after 12 months of antiretroviral treatment in Mombasa, Kenya: a prospective cohort. AIDS Patient Care STDs. 2008;22:587–94.
72. Sherman SG, Sutcliffe C, Srirojn B, Latkin CA, Aramratanna A. Evaluation of a peer network intervention trial among young methamphetamine users in Chiang Mai, Thailand. Soc Sci Med. 2008;68:69–79.
73. Stephenson J, Strange V, Allen E, et al. The long-term effects of a peer-led sex education programme (RIPPLE): a cluster randomised trial in schools in England. PLoS Med. 2008;5(11):e224. 1-e224.12.
74. Wilson TE, Hogben M, Malka ES, et al. A randomized controlled trial for reducing risks for sexually transmitted infections through enhanced patient-based partner notification. Am J Public Health. 2009;99:S104–10.
75. Wingood GM, DiClemente RJ, Harrington KF, et al. Efficacy of an HIV prevention program among female adolescents experiencing gender-based violence. Am J Public Health. 2006;96: 1085–90.
76. Carey MP, Senn TE, Vanable PA, Coury-Doniger P, Urban MA. Brief and intensive behavioral interventions to promote sexual risk reduction among STD clinic patients: results from a randomized controlled trial. AIDS Behav. 2010;14:504–17.
77. Garfein RS, Golub ET, Greenberg AE, et al. A peer-education intervention to reduce injection risk behaviors for HIV and hepatitis C virus infection in young injection drug users. AIDS. 2007;21:1923–32.
78. Jewkes R, Nduna M, Levin J, et al. Impact of Stepping Stones on incidence of HIV and HSV-2 and sexual behaviour in rural South Africa: cluster randomised controlled trial. BMJ. 2008;337:a506. a506.11.
79. Kelly JA, Amirkhanian YA, Kabakchieva E, et al. Prevention of HIV and sexually transmitted diseases in high risk social networks of young Roma (Gypsy) men in Bulgaria: randomised controlled trial. BMJ. 2006;333:1098.
80. Kershaw TS, Magriples U, Westdahl C, Schindler Rising S, Ickovics J. Pregnancy as a window of opportunity for HIV prevention: effects of an HIV intervention delivered within prenatal care. Am J Public Health. 2009;99:2079–86.
81. Marion LN, Finnegan L, Campbell RT, Szalacha LA. The well woman program: a community-based randomized trial to prevent sexually transmitted infections in low-income African American women. Res Nurs Health. 2009;32:274–85.
82. Tebb KP, Pantell RH, Wibbelsman CJ, et al. Screening sexually active adolescents for chlamydia trachomatis: what about the boys? Am J Public Health. 2005;95:1806–10.
83. Jeličić H, Phelps E, Lerner RM. Use of missing data methods in longitudinal studies: the persistence of bad practices in developmental psychology. Dev Psychol. 2009;4:1195–9.
84. Jo B, Ginexi EM, Ialongo NS. Handling missing data in randomized experiments with noncompliance. Prev Sci. 2010;11(4):384–96.