American Journal of Epidemiology

Vol. 180, No. 9
DOI: 10.1093/aje/kwu224
Advance Access publication:
October 9, 2014

Practice of Epidemiology

Multiple Imputation in a Longitudinal Cohort Study: A Case Study of Sensitivity to Imputation Methods

Helena Romaniuk*, George C. Patton, and John B. Carlin

* Correspondence to Dr. Helena Romaniuk, Murdoch Childrens Research Institute, 50 Flemington Road, Parkville, Victoria 3052, Australia (e-mail: helena.romaniuk@mcri.edu.au).

Initially submitted June 26, 2013; accepted for publication July 24, 2014.

Multiple imputation has entered mainstream practice for the analysis of incomplete data. We have used it extensively in a large Australian longitudinal cohort study, the Victorian Adolescent Health Cohort Study (1992–2008). Although we have endeavored to follow best practices, there is little published advice on this, and we have not previously examined the extent to which variations in our approach might lead to different results. Here, we examined sensitivity of analytical results to imputation decisions, investigating choice of imputation method, inclusion of auxiliary variables, omission of cases with excessive missing data, and approaches for imputing highly skewed continuous distributions that are analyzed as dichotomous variables. Overall, we found that decisions made about imputation approach had a discernible but rarely dramatic impact for some types of estimates. For model-based estimates of association, the choice of imputation method and decisions made to build the imputation model had little effect on results, whereas estimates of overall prevalence and prevalence stratified by subgroup were more sensitive to imputation method and settings. Multiple imputation by chained equations gave more plausible results than multivariate normal imputation for prevalence estimates but appeared to be more susceptible to numerical instability related to a highly skewed variable.

longitudinal cohort study; missing data; multiple imputation; sensitivity analysis

Abbreviations: AC, available case; CC, complete case; CCA, complete case adult data with partial adolescent data; MICE, multiple imputation by chained equations; MVNI, multivariate normal imputation; PMM, predictive mean matching; VAHCS, Victorian Adolescent Health Cohort Study.

The method of multiple imputation for handling the analysis of incomplete data was first described in detail by Rubin (1). However, it has only achieved widespread application over the past decade or so because of the availability of suitable software tools (2–6). The method involves 2 distinct phases: first, multiple copies of the incomplete data set are completed by imputation of missing values, and second, the desired analysis is performed within each imputed data set, with final results obtained by appropriate combination of results over the imputed data sets. The second phase is relatively straightforward (1, 7), but the first phase, which involves implementation of an imputation method including the specification of an appropriate imputation model, must be done carefully to provide confidence in the validity of the resulting inferences.

In this article, we examine the effect of variation in imputation methods when imputing data for the large longitudinal Victorian Adolescent Health Cohort Study (VAHCS) (8–11). We and our colleagues have used multiple imputation extensively for VAHCS analyses. It is attractive in this context, in which analyses that use multiple waves of data may be compromised by incomplete responses and attrition. Although we have endeavored to follow best practices in performing our imputations, there has been limited systematic examination of the extent to which variations in our approach might have led to different results and, as observed elsewhere, there is little published advice to guide practice (12–14).

Most of the published articles that have used multiple imputation in the VAHCS have used the method of multivariate normal imputation (MVNI) as implemented originally by

Schafer (2) and now widely available in major statistical packages such as SAS (SAS Institute, Inc., Cary, North Carolina) and Stata (StataCorp LP, College Station, Texas). MVNI assumes that all of the variables that require imputation follow a multivariate normal distribution, and it produces imputed values by using a Bayesian Markov chain Monte Carlo (or "data augmentation") algorithm that alternates between estimating the parameters of the multivariate normal distribution and producing imputed values from the appropriate posterior predictive distributions. The multivariate normal specification is clearly unrealistic for many problems, but several studies have demonstrated that the method often works well despite this (2, 15), perhaps because most of the relevant information is contained in the means, variances, and correlations (first and second moments), which are all reproduced under the MVNI model.

The other widely used approach is multiple imputation by chained equations (MICE), which was first proposed by van Buuren et al. (16) and further developed (for SAS software) by Raghunathan et al. (17) and (for Stata software) by Royston (3, 18). In this approach, imputation is performed sequentially for each variable containing missing values, using a univariate regression model that can be tailored to the scale of the variable, so that, in particular, categorical variables can be imputed using appropriate generalized linear models. The process cycles through the univariate imputation process several times to allow imputed values on other variables to enter the imputation for each variable. MICE has the appeal of allowing considerable flexibility in the univariate model specifications; for example, models may explicitly allow for truncation or censoring, and logical dependencies may be incorporated (e.g., quantity of alcohol consumed is imputed only if person is a drinker). MICE also allows nonlinear terms or interactions between variables to be included as predictors in the univariate imputation models. Predictive mean matching (PMM) can also be used to impute continuous variables. In PMM, instead of replacing missing values with predicted values from the imputation model, an observed value is selected from data values that are close to the predicted value (19). On the other hand, the flexibility of the approach means that it is possible for inconsistencies to arise between the univariate regression models (20), with unclear potential consequences. The MICE algorithm also encounters numerical stability problems more frequently than MVNI when large numbers of variables are involved.

There are other technical choices that need to be made with multiple imputation. Here, we focus on investigating the impact of 3 of these choices: the inclusion of auxiliary variables, the omission of cases with an excessive amount of missing data, and the handling of highly skewed continuous variables.

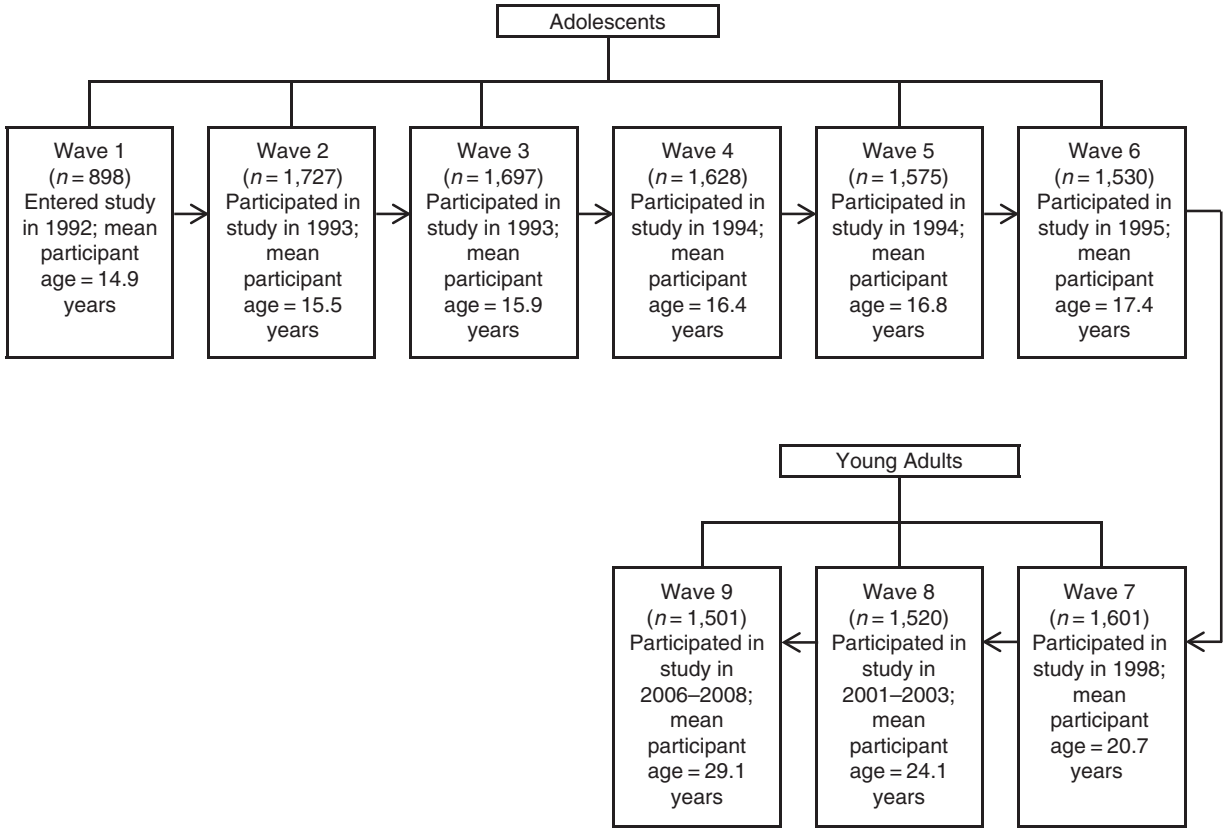

Figure 1. Sampling and ascertainment in the Victorian Adolescent Health Cohort Study, Australia, 1992–2008. There were 2 entry points (at wave 1 and wave 2). Intended sample sizes were 1,037 for wave 1 and 995 for wave 2, with a total intended sample size of 2,032. Ninety-six percent (1,943) of the sample participated at least once in waves 1–6.

Table 1. Description of Variables Used in Imputation Models and Summary of Percent Missing Data per Variable, by Imputation Data Set and Sex, in the Victorian Adolescent Health Cohort Study, Australia, 1992–2008

Variable	Waves	Variable Type	Distribution	Categories/Ranges	% Missing Values in Data Sets for Imputations[a]					
					Analyses 1, 2, 3, 4, and 6[b]		Analysis 5[c]		Analysis 7[d]	
					Among Male Participants (n = 943)	Among Female Participants (n = 1,000)	Among Male Participants (n = 772)	Among Female Participants (n = 911)	Among Male Participants (n = 813)	Among Female Participants (n = 924)
Key Variables										
Cannabis use in past year	2–9	Ordinal	Positively skewed	Nonuser, occasional user, weekly user, or daily user	13–29	10–21	10–18	7–13	11–19	8–15
Cigarette smoking in past month	2–9	Ordinal	Positively skewed	Nonsmoker, occasional smoker, or daily smoker	9–28	7–20	9–17	7–13	8–18	7–15
Alcohol use in past week[e]	2–6	Binary		Not risky drinker or risky drinker	16–34	15–23	15–24	14–19	15–25	14–18
	7–9	Continuous	Positively skewed	0–286 Standard drink units	26–36	16–26	14–23	10–19	16–27	11–21
Illicit drug use in past year[f]	7 and 8	Binary		No or yes	22 and 26	13 and 18	9 and 12	7 and 10	12 and 15	8 and 12
	9	Ordinal	Positively skewed	None, < weekly, or ≥ weekly	33	24	19	17	23	18
Sex		Binary		Male or female	0	0	0	0	0	0
Age at wave 2 (mean centered[g])	2	Continuous	Symmetrical	−3.0 to 4.6 Years	0	0	0	0	0	0
School location at study inception		Binary		Urban or rural	0	0	0	0	0	0
Highest level of parental education		Ordinal	Positively skewed	Did not complete high school, completed high school, or university degree/ further qualification	3	3	1	1	1	1
Parental divorce/ separation in adolescence		Binary		No or yes	0	0	0	0	0	0
Parental smoking		Binary		Both nonsmokers or at least 1 parent smoked	3	3	1	1	1	1
Auxiliary Variables										
Country of birth		Binary		Australia or other	8	8	5	5		
Completed high school		Binary		No or yes	4	4	1	1		
Ever had a baby	9	Binary		No or yes	19	19	13	12		
No partner	9	Binary		No or yes	20	20	13	13		

Table continues

Table 1. Continued

Variable	Waves	Variable Type	Distribution	Categories/Ranges	% Missing Values in Data Sets for Imputations[a]					
					Analyses 1, 2, 3, 4, and 6[b]		Analysis 5[c]		Analysis 7[d]	
					Among Male Participants ($n = 943$)	Among Female Participants ($n = 1,000$)	Among Male Participants ($n = 772$)	Among Female Participants ($n = 911$)	Among Male Participants ($n = 813$)	Among Female Participants ($n = 924$)
Highest level of education	9	Ordinal	Negatively skewed	Secondary, vocational, or university	13	13	6	6		
Receiving government welfare	9	Binary		No or yes	22	22	18	15		
Not currently in paid employment	9	Binary		No or yes	21	21	13	14		
CIS-R score[h]	2–7	Continuous	Positively skewed	0–55	9–28	8–15	9–17	7–11		
Mixed depression-anxiety[i]	8 and 9	Binary		No or yes	26 and 27	18 and 20	12 and 13	10 and 13		
Anxiety disorder[j]	9	Binary		No or yes	33	24	19	16		
Major depressive disorder[k]	9	Binary		No or yes	33	23	19	16		
Cannabis dependency[l]	7–9	Binary		No or yes	22–34	14–24	10–20	7–17		

Abbreviation: CIS-R, Revised Clinical Interview Schedule.

[a] Range represents minimum and maximum rates of missingness if variable was measured at multiple waves; actual rates of missingness are given if variable was measured on 1 or 2 occasions only.

[b] Rates of missingness were comparable between male and female participants for parental divorce/separation and at waves 2 and 3 for CIS-R score and all licit and illicit drug use. For all other variables with missing values, male participants had higher rates of missingness than female participants ($P \leq 0.003$).

[c] Rates of missingness were comparable between male and female participants for all variables except cigarette smoking at wave 3, cannabis use and CIS-R score at wave 6 ($P \leq 0.001$), and alcohol use at wave 6 ($P = 0.007$).

[d] Rates of missingness were comparable between male and female participants for all variables except cannabis use at waves 6 and 7; cigarette smoking at waves 5–7; alcohol use at waves 6, 7, and 9; and illicit drug use at waves 7 and 9 ($P < 0.01$).

[e] Risky drinking was defined as exceeding 14 standard drinks (1 standard drink = 10 g alcohol) in the week prior to the survey for male and female participants at all ages and in all waves.

[f] Included the use of amphetamine, ecstasy/designer drugs, or cocaine. Data on use of each illicit drug were collected separately at each wave. Amphetamine use was collected at waves 2–9. However, before wave 7, frequency of amphetamine use was judged to be too low, and rates of missingness too high, to allow imputation. Therefore, only data on amphetamine use at waves 7–9 were included in the imputation models.

[g] Participants' ages were centered around mean age of 15.4 years.

[h] The CIS-R is a psychiatric interview designed to assess symptoms of depression and anxiety in nonclinical populations (35).

[i] Per the 12-Item General Health Questionnaire (36).

[j] Per the Composite International Diagnostic Interview short form (37). Participants were classified with anxiety disorder if they were diagnosed with generalized anxiety disorder, social phobia, agoraphobia, or panic disorder.

[k] Per the Composite International Diagnostic Interview, CIDI-Auto (38).

[l] Per the Composite International Diagnostic Interview, CIDI 2.1, 12-month version (39).

By definition, auxiliary variables are not required for the analysis of interest but are included in the imputation model because they are believed to improve the quality of imputation and so potentially reduce bias and/or variance of estimation (14, 21). The "missing at random" assumption, which underlies the standard approaches to imputation, asserts that being missing is independent of any missing values given the observed data, so the inclusion of auxiliary variables may make this assumption more reasonable (21–23). Auxiliary variables may be selected because they are correlated 1) with the missingness mechanism or 2) with the variables of interest that have missing values. We investigated the effects of excluding or including auxiliary variables in the imputation model.

Similarly, there are no accepted guidelines on whether retaining cases with large proportions of missing data is worthwhile or safe. To explore this issue, we performed analyses using multiple imputation with data sets including all participants or including only those who had observed values for 50% or more of the variables included in the imputation model.

The final question of interest was how to handle highly skewed variables. The analysis of interest required dichotomizing a measure of alcohol consumption to identify high- and low-risk drinkers. For both approaches, we explored modeling the binary risk variable directly and transforming the continuous variable prior to imputation to try to achieve a more normal distribution. For MICE only, we modeled the alcohol units using a truncated normal distribution to ensure that all imputed values were plausible, with bounds of 0 and the maximum number of alcohol units at each wave (4), and we also used PMM (19). This article examines the sensitivity of analytical results to the aspects of the imputation method just described, in the context of a detailed analysis that has been published elsewhere (24).

METHODS

Sample

The VAHCS, a 9-wave cohort study of health in adolescents and young adults in Victoria, Australia, was conducted between 1992 and 2008. The cohort was defined by selection of 2 classes from a statewide sample of 44 schools (24, 25). One class entered the study in the latter part of the ninth school year (wave 1; mean participant age = 15 years), and the second class entered 6 months later (wave 2). Participants were interviewed at four 6-month intervals during their teens (waves 3–6) with 3 follow-up waves in young adulthood, when participants were aged 20–21 years (wave 7), 24–25 years (wave 8), and 29 years (wave 9). A total of 1,943 students participated at least once during the first 6 (adolescent) waves (Figure 1), defining the initial cohort. We analyzed a maximum sample size of 1,934 after omitting 9 participants who died before wave-9 assessment.

The measures we used are described briefly below and listed in Table 1. To simplify the examination of the imputation issues, we consider only a selection of the results from the published VAHCS paper (24) here. However, all 38 variables identified as being required for the analysis in the substantive paper were considered to be key variables and were treated as if they were required for the imputation-based analyses. Wave 1 measures were not included in the imputation models or analyses because, by design, each of these variables is missing 50% of values or more. However, if a participant was seen only at wave 1 in adolescence, we included these participants by bringing forward their wave-1 observations to wave 2. This admittedly ad hoc approach was judged to be reasonable because it affected only 57 participants (2.9%) in the full data set, and the waves were only 6 months apart.

Key variables included licit and illicit drug use and potential confounders. At each wave (waves 2–9), participants were asked to report their maximum frequency of cannabis use, cigarette smoking, and, if they reported drinking alcohol in the week prior to the survey, to complete a detailed diary of their alcohol consumption, which was used to calculate total number of standard drinks (1 standard drink = 10 g alcohol). Because of the high percentage of nondrinkers in adolescence (>70% in waves 2–4 and >60% in waves 5 and 6), a binary measure of high-risk alcohol use was defined as exceeding 14 standard drinks per week, following Australian guidelines (26). In the adult waves (waves 7–9), participants were additionally asked about use of amphetamines, ecstasy/designer drugs, and cocaine. Incident (new) amphetamine use was identified at waves 8 and 9 in participants who had not reported use at previous waves. Potential confounders identified as relevant to the analyses of interest were sex, age at wave 2, school location (urban or rural), and parental characteristics (highest level of education, smoking status, and divorce/separation during the participant's adolescence).

Auxiliary variables

Prior to developing the imputation model for the substantive paper (24), we identified 51 potential auxiliary variables as associated with missingness in the data. Of these, 20 were included in the final imputation model after a process of trial and error in which candidate variables were included until imputation models appeared unstable with respect to collinearities and convergence of the imputation algorithm. During the development of the imputation model for this paper, several of these variables were dichotomized prior to imputation because their distributions were heavily skewed, and normalizing transformations did not appear to improve imputations (see Diagnostics section below). Priority was given to variables with the lowest proportion of missing data and the strongest association with missingness.

Analysis

The various approaches to managing missing data were compared with respect to several key analyses from the substantive paper. Overall prevalence of cannabis use at wave 7 was estimated. Prevalence of amphetamine use at wave 9 was estimated overall and by frequency of concurrent cannabis use. Discrete-time proportional hazards models were used to estimate the association between cannabis use at the previous wave and the incidence of adult amphetamine drug use, after controlling for potential confounders (27). Cannabis users at the previous wave were classified as occasional (reference category), weekly, or daily users, and nonusers were subclassified as having never used or as being previous users.

Table 2. Summary of Analysis Data Sets and Settings for Imputation-Based Analyses in the Victorian Adolescent Health Cohort Study, Australia, 1992–2008

Analysis Data Set	Method	Auxiliary Variables	Participants Included	Adult Alcohol Variables	Analysis Label	Participants Analyzed[a]	
						No.	%[b]
Observed							
Available case[c]					AC	1,384–1,586	71–82
Complete case[d]					CC	516	27
Complete case adult data with partial adolescent data[e]					CCA	941	49
Imputed	MVNI	Yes	All	Continuous	MVNI_1	1,934	100
	MVNI	Yes	All	Binary	MVNI_4	1,934	100
	MVNI	Yes	≤50% Missing[f]	Continuous	MVNI_5	1,679	87
	MVNI	No	All	Continuous	MVNI_6	1,934	100
	MVNI	No	≤50% Missing[f]	Continuous	MVNI_7	1,731	90
	MICE	Yes	All	Continuous	MICE_1	1,934	100
	MICE	Yes	All	Continuous (truncated)[g]	MICE_2	1,934	100
	MICE	Yes	All	Predictive mean matching	MICE_3	1,934	100
	MICE	Yes	All	Binary	MICE_4	1,934	100
	MICE	Yes	≤50% Missing[f]	Continuous	MICE_5	1,679	87
	MICE	No	All	Continuous	MICE_6	1,934	100
	MICE	No	≤50% Missing[f]	Continuous	MICE_7	1,731	90

Abbreviations: MICE, multiple imputation by chained equations; MVNI, multivariate normal imputation.

[a] The following numbers of participants who had died by wave 9 were dropped from analysis data sets: 9 participants from data sets 1, 2, 3, 4, and 6; 4 participants from data set 5; and 6 participants from data set 7.

[b] Percentage of total cohort alive at wave 9.

[c] Complete data for subset of variables were used in analysis. Available case analyses were conducted only for prevalence of drug use (overall and by level of concurrent cannabis use) at waves 7–9.

[d] Complete data for all variables were used in all analyses.

[e] Complete data for 3 or more of 5 adolescent waves.

[f] Participants were included in the imputation data set if they had 50% or more observed data for all of the variables included in the imputation model.

[g] Alcohol units at each adult wave were imputed using a truncated normal distribution with bounds of 0 and the maximum number of reported units.

Imputation. A list of the imputation approaches, with a key used in figures and tables, is shown in Table 2 along with 3 approaches to analysis that did not use multiple imputation. All key variables were included in the imputation models for all approaches, with auxiliary variables included in imputation approaches 1–5 only. We imputed 20 data sets for each of the 12 multiple imputation approaches. Data from male participants were imputed separately from those of female participants, with imputed data sets combined for final analyses. After imputation, estimates were obtained by averaging results across the 20 imputed data sets with inferences under multiple imputation obtained using Rubin's rules (1). Data imputation and analysis were undertaken using Stata, version 11, software (6).

MVNI. Data were imputed using the *mi impute mvn* procedure. Default options were used for the expectation-maximization and Markov chain Monte Carlo algorithms. Prior to imputation, skewed ordinal and continuous measures were transformed to 0 skewness using shifted logs. All other variables were included in the imputation model as defined in Table 1. For analysis, log-transformed variables were transformed back to their original scales, imputed binary variables were converted to binary values using adaptive rounding (28), and ordinal variables were recorded by rounding imputed values to the nearest category.

MICE. MICE imputations were performed using the *ice* procedure (4). Prior to imputation, skewed continuous measures were transformed using shifted logs, except adult alcohol consumption for PMM (MICE_3; see Table 2 for list of analysis labels). Imputation models were set up so that binary variables were modeled using logistic regression when they were outcomes and included as single dummy variables when they were covariates. Ordinal variables were modeled using ordinal logistic regression when they were outcomes and as a series of dummy variables when they were included as covariates. Continuous variables were modeled using linear regression and included as single linear terms when they

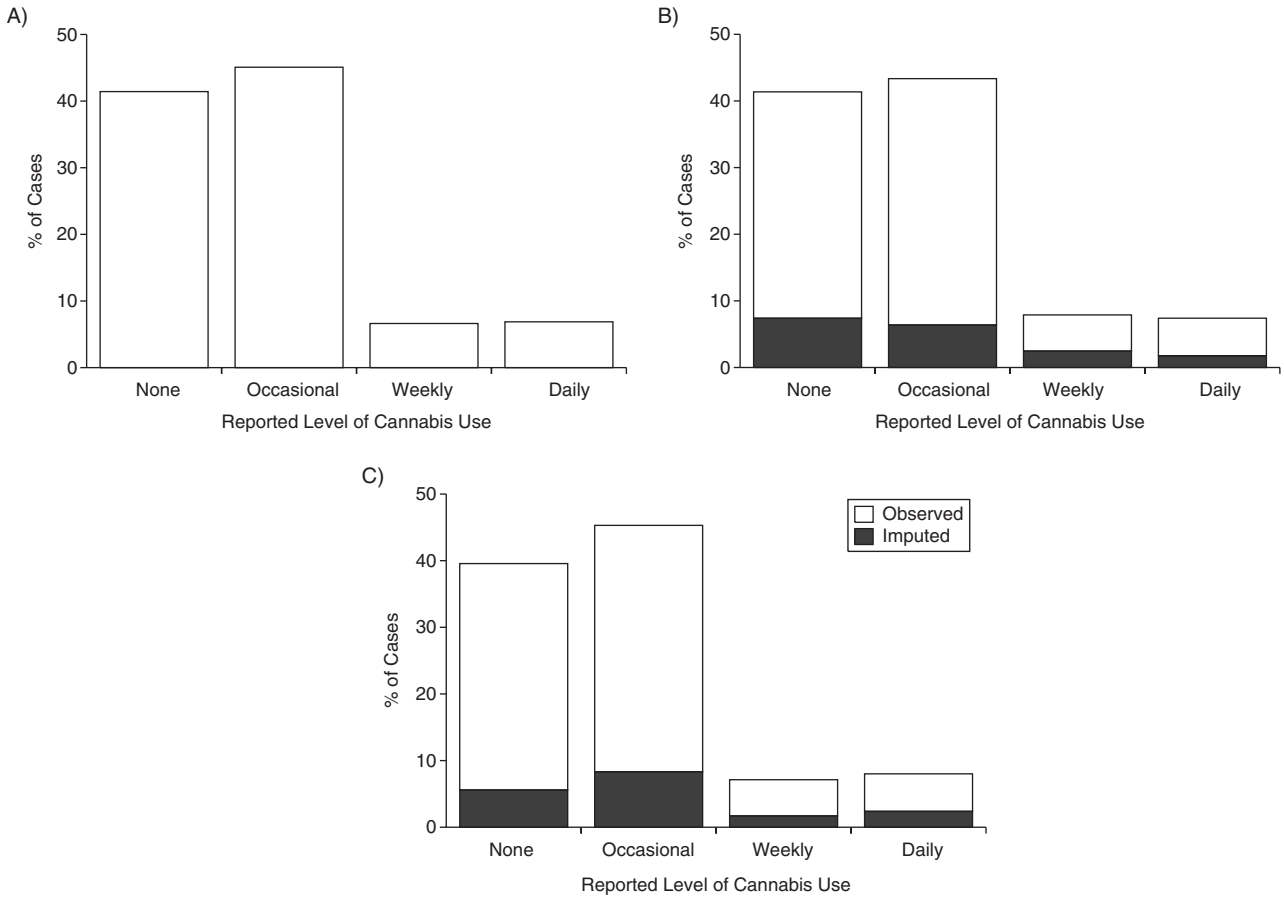

Figure 2. Comparison of distributions for available cases and imputed values for cannabis use in wave 7 of the Victorian Adolescent Health Cohort, Australia, 1992–2008. A) Available case (AC) analysis; B) multivariate normal imputation (MVNI)_4; and C) multiple imputation by chained equations (MICE)_4 (see Table 2 for list of analysis labels).

were covariates. Default options were used for other settings, including treatment of perfect prediction (29).

Available case analysis. For the prevalence estimates and cross-sectional analysis of concurrent drug use, we used available case (AC) analysis, in which all participants who had observed data for the measures being analyzed were included. This approach maximizes sample size but uses different samples for each analysis.

Complete case analysis. Two approaches were adopted for complete case (CC) analysis. The first, described herein as CC analysis, required that participants have data for all the key variables. The second, complete case adult data with partial adolescent data (CCA) analysis, required that participants have complete data from 3 or more adolescent waves, from all adult waves, and for demographic variables.

Diagnostics. Simple diagnostics were run (30, 31) to assess whether the imputed data sets were reasonable. We compared summary statistics calculated just from the observed data with those from only imputed data, as well as with those from the data sets containing both observed and imputed data. For categorical variables, we compared distributions, and for continuous variables, we compared means and variances. Large differences may (but do not necessarily) reflect inadequacies in the imputation model, so these were flagged for further consideration (32).

RESULTS

Description of missingness

In data sets that included all participants (MVNI/MICE_1, 2, 3, 4, and 6), for variables measured repeatedly across waves, the proportion missing was lowest at wave 2 and increased at each subsequent adolescent wave, with a plateauing of the rate of missing data in the adult waves (Table 1). For data sets in which participants with more than 50% missing values had been dropped (MVNI/MICE_5 and 7), the rates of missingness were lower, most noticeably for the time-constant auxiliary variables and, to some extent, for the time-varying key variables.

Convergence

MVNI. Convergence was checked prior to imputed data sets being generated by inspecting graphs of the worst linear

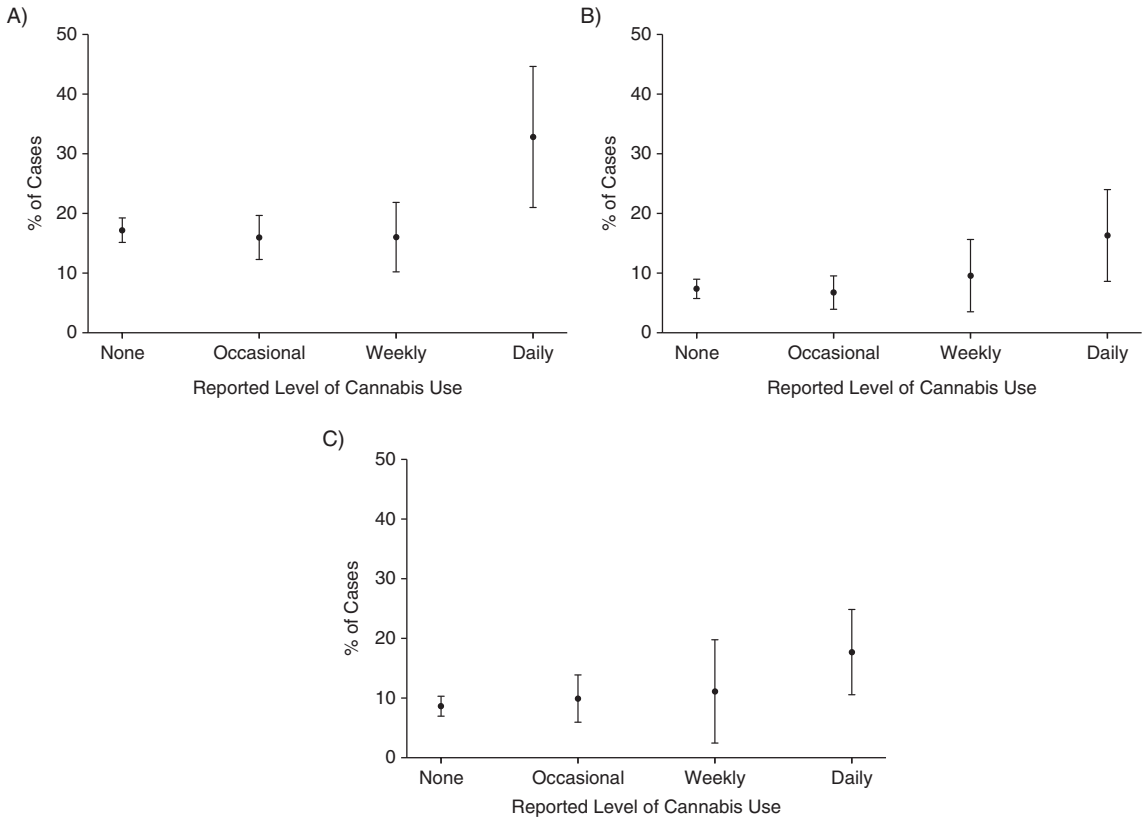

Figure 3. Percentage of participants with missing data on cannabis use at wave 7 by reported levels of cannabis use at other waves in the Victorian Adolescent Health Cohort Study, Australia, 1992–2008. A) Maximum in adolescence; B) at wave 8; and C) at wave 9. Bars, 95% confidence intervals.

function of parameter estimates (2) after 2,000 Markov chain Monte Carlo iterations had been run.

MICE. Prior to imputing data sets, convergence was checked by running 100 cycles of regression switching and saving the mean of the imputed values at each cycle. For each variable, the means were plotted against cycle number and inspected to assess convergence. A visual inspection of the convergence graphs for each approach showed that it was reasonable to assume convergence had been achieved, with 20 cycles adequate for MICE.

Diagnostics

For most variables, the imputed distributions were similar to the distributions of observed values for all imputation approaches. There were some exceptions. For example, when alcohol units were imputed as a log-shifted continuous variable, some of the imputed values were unfeasibly large. For several variables, we found higher rates of risky behavior for the imputed data compared with the observed. However, this seemed plausible given that participants with riskier behavior might be more likely to drop out of the longitudinal study, miss a wave, or be reluctant to respond to these types of questions. Figure 2 shows the distribution of cannabis use at wave 7, with observed percentages shown for AC and the imputed and observed percentages in each category for MVNI_4 and MICE_4. According to the AC analysis, the distribution of observed cannabis use is bimodal. MVNI and MICE differ in their method of imputing ordinal variables, and this is reflected in the imputed percentages, with MICE imputed values being bimodal and MVNI imputed values being unimodal. MICE imputed a higher proportion of participants to the highest level of drug use compared with MVNI. Overall, the combined distribution for cannabis use did not differ dramatically across the 3 approaches, although MICE led to a final distribution that was more different from the AC distribution than MVNI. Examination of known cannabis use at other waves for those who had missing data on cannabis use at wave 7 shows that it was plausible to expect higher rates of daily use in the imputed data sets (Figure 3).

Analysis results

Overall prevalence estimates. Both AC and CCA analyses estimated similar percentages of cannabis use for all categories (Figure 4). CC estimates were quite different from all other estimates, with wider 95% confidence intervals because of smaller numbers. For the MVNI approaches, the prevalence of cannabis use was estimated fairly consistently for all categories, although slightly lower values were seen for

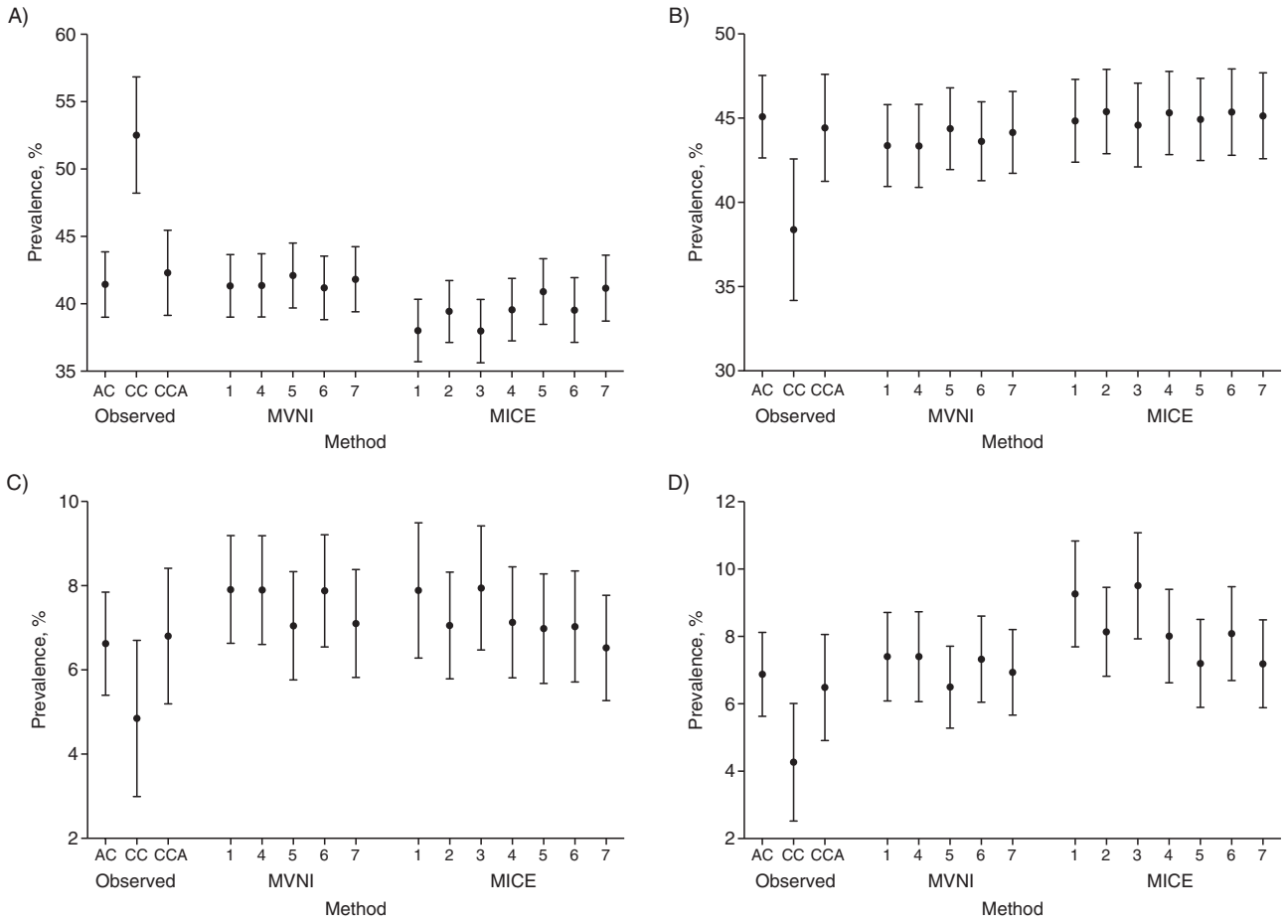

Figure 4. Estimated prevalence of cannabis use at wave 7, by missing data method, in the Victorian Adolescent Health Cohort Study, Australia, 1992–2008. A) No cannabis use; B) occasional use; C) weekly use; and D) daily use. Bars, 95% confidence intervals. AC, available case analysis; CC, complete case analysis; CCA, complete case adult data with partial adolescent data analysis; MICE, multiple imputation by chained equations; MVNI, multivariate normal imputation.

weekly and daily use for those approaches that excluded participants with more than 50% missing values (MVNI_5 and MVNI_7). MICE estimates were fairly stable for weekly and occasional use, with more variation between estimates of nonuse and daily use. In particular, estimates from MICE_1 (auxiliary variables, all participants, and adult alcohol units imputed as log-shifted continuous variables) and MICE_3 (auxiliary variables, all participants, and adult alcohol units imputed using PMM) differed substantially by margins of up to 2 standard errors from the 2 approaches that excluded participants with more than 50% missing values (MICE_5 and MICE_7). Comparing the estimates between the imputation methods, the prevalence of nonusers was higher under MVNI versus many of the MICE approaches, and the prevalence of daily use was lower. Estimates of occasional and weekly use were comparable for the 2 approaches.

Overall estimates of prevalence of amphetamine use exhibited similar inconsistencies between the observed-data methods, as were seen for cannabis use (Figure 5A). MVNI estimates showed little variation across settings. With the exception of MICE_1 and MICE_3, MICE estimates were also generally similar to each other. However, MICE tended to estimate higher levels of amphetamine use than MVNI.

Subgroup prevalence estimates. Within the imputation method, MVNI estimates varied little (Figure 5B–E). MICE estimates were also similar across approaches with the exception of MICE_1 and MICE_3, which estimated somewhat higher levels of amphetamine use for non–cannabis users. For non–, occasional, and daily cannabis users, MICE estimated consistently higher levels of amphetamine use than MVNI.

Association estimates. In contrast to the prevalence estimates, there was relatively little variation in estimated rate ratios across imputation model settings (Figure 6).

DISCUSSION

In a large longitudinal cohort study, we explored the effects of decisions made when building an imputation model. Although in a case-study analysis such as this there is no way of knowing the "truth" with respect to the parameter values of

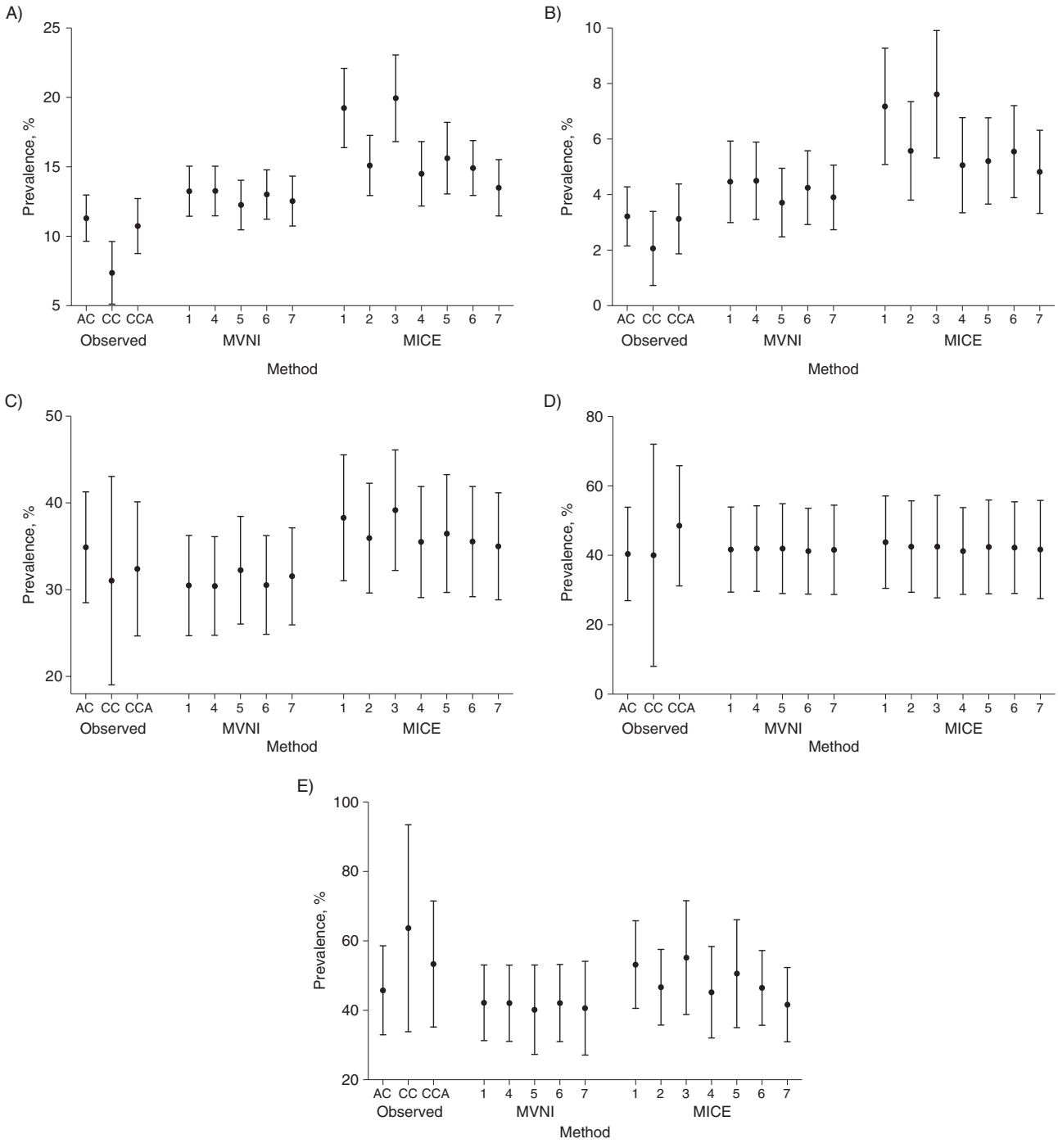

Figure 5. Estimated prevalence of amphetamine use at wave 9 by missing data method, overall and by level of concurrent cannabis use in the Victorian Adolescent Health Cohort Study, Australia, 1992–2008. A) Overall prevalence; B) no cannabis use; C) occasional use; D) weekly use; and E) daily use. Bars, 95% confidence intervals. AC, available case analysis; CC, complete case analysis; CCA, complete case adult data with partial adolescent data analysis; MICE, multiple imputation by chained equations; MVNI, multivariate normal imputation.

interest, we believe it is informative to examine the extent to which results vary across the different approaches to imputation. Minimal variation between estimates across a substantial range of methods provides prima facie evidence of validity, as long as the main assumptions of the imputation methods also seem reasonable. Overall, we found that decisions made about the imputation approach had a discernible but rarely dramatic impact on final results.

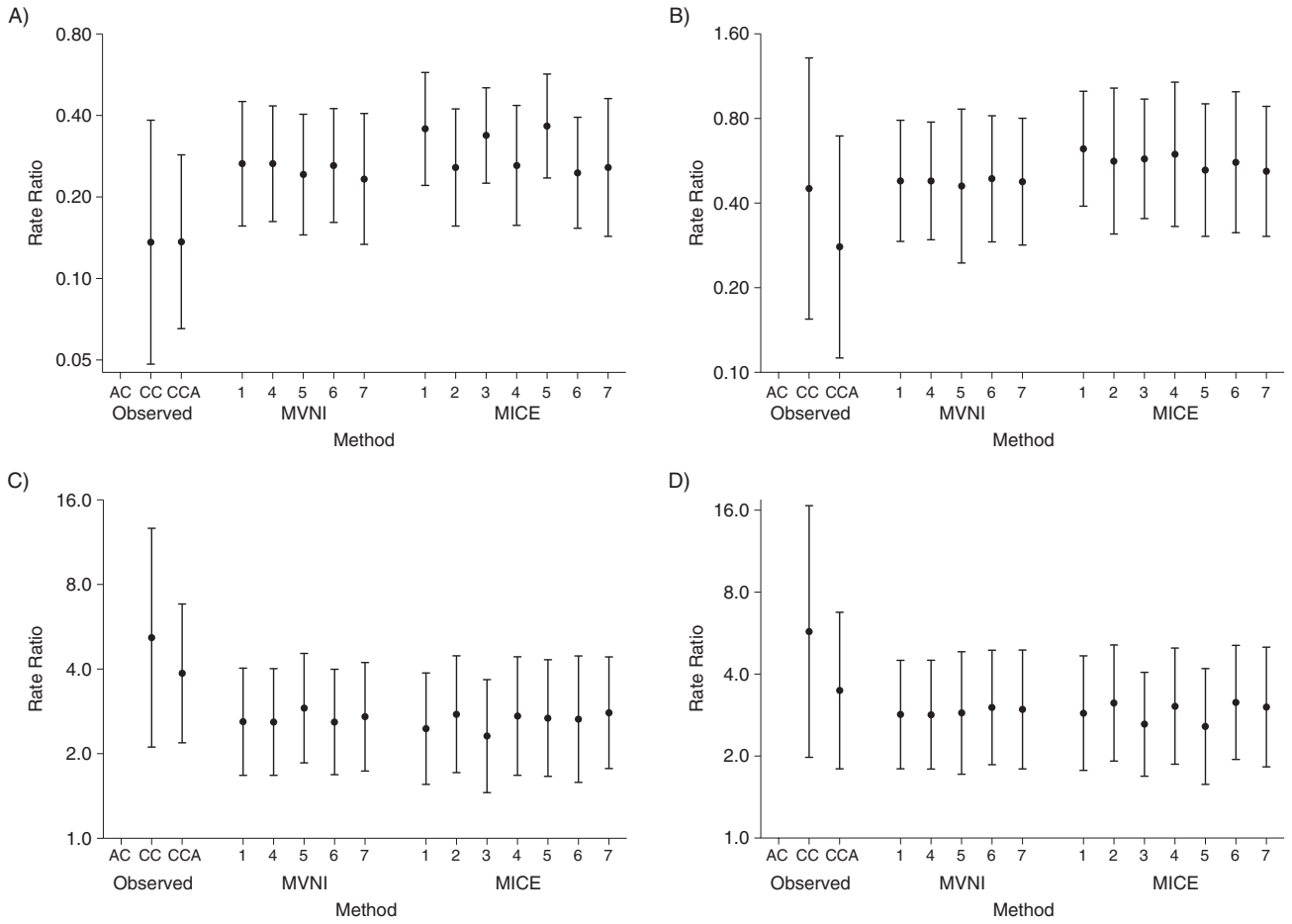

Figure 6. Rate ratio estimates for incidence of amphetamine use by level of cannabis use at the previous wave (with occasional use as the reference category) and missing data method in the Victorian Adolescent Health Cohort Study, Australia, 1992–2008. A) No cannabis use at previous wave (never user); B) no cannabis use at previous wave (past user); C) weekly cannabis use at previous wave; and D) daily cannabis use at previous wave. Bars, 95% confidence intervals. AC, available case analysis; CC, complete case analysis; CCA, complete case adult data with partial adolescent data analysis; MICE, multiple imputation by chained equations; MVNI, multivariate normal imputation.

Compared with the CC analysis, all other approaches estimated higher levels of cannabis use at wave 7 and amphetamine use at wave 9. The CC analysis included only 27% of 1,934 participants, and it is reasonable to assume that those who engaged in risky or illicit behavior were less likely to participate at every wave, so that the CC analysis would be expected to underestimate the prevalence of drug use. The AC and CCA estimates of cannabis and amphetamine use were consistent with each other. For the AC analyses conducted for this study, more than 70% of participants were included in each analysis, and for CCA analysis, just less than half of the participants were included.

For MVNI, decisions about including auxiliary variables and number of cases with complete data and how to impute a highly skewed variable had little substantial impact on any of the estimates of interest. These findings appear to support previous claims that MVNI imputation models are robust to potential model misspecification (2).

For MICE, there was greater variation in results according to decisions made when building the imputation model. In particular, considerable variation was observed between the overall prevalence estimates. In the first analysis (of cannabis use prevalence), the approach in which the highly positively skewed measure of alcohol use was imputed as continuous after transformation using shifted logs led to substantially lower estimated levels of nonuse and higher levels of daily use compared with other approaches. Simple diagnostics showed that, although all approaches in which the imputed values were unconstrained estimated a small percentage of values out of range (0.1%–0.5%), this approach imputed many alcohol values as unfeasibly large. The approach that used PMM to impute alcohol units also estimated higher levels of daily use, for reasons that were not clear, because mean and median units of alcohol were similar for the observed and imputed values. This approach imputes values that are sampled only from the observed values (19) and generally results in a distribution of imputed values that closely matches the observed data distribution (15), but it has been subject to limited systematic evaluation (33). Other MICE approaches

performed more stably. For prevalence estimates by subgroup and the association analysis, greater consistency was seen between the MICE approaches.

Comparing the MVNI and MICE approaches, apart from the apparent instability of MICE in at least 1 setting, results were different to an extent that would slightly alter some of the substantive interpretation of the study's findings. In particular, most of the MICE approaches estimated lower levels of no cannabis use and higher levels of daily cannabis use than the MVNI approaches, and most of the MICE estimates of amphetamine use were considerably higher than the MVNI estimates. These differences may reflect the way that ordinal variables, especially the bimodally distributed cannabis variables, are imputed by each approach, with the MICE approach not constrained to a unimodal distribution and therefore potentially producing more appropriate imputations.

Our examination of the effect of excluding cases with more than 50% data missing suggested that, in this setting, including cases with a high proportion of missing information made a minimal difference, no doubt in part because these exclusions amounted to only 13% and 10% of the total sample (in settings 5 and 7, respectively). Investigators are often loath to reduce their apparent sample sizes even when faced with largely incomplete information on participants, but in many settings the additional information from these participants will be negligible. On the other hand, we saw little evidence of instability or greater variance affecting the results when the most incomplete cases were included.

The inclusion of auxiliary variables did not seem to influence the results in any of our approaches. This appears to be because, in our setting, the auxiliary variables did not add substantial independent predictive information for the missing values over and above the key variables that were already included in the imputation models. It seems plausible that in the setting of repeated measures over time for all of the outcome variables, the inclusion of auxiliary variables is unnecessary. Furthermore, the inclusion of auxiliary variables in an already "large" imputation model may lead to instability of estimation in the imputation modeling. These conclusions are supported by a recent study that found that inclusion of auxiliary variables in imputation models did not improve the bias or precision of regression estimates from a logistic model (34).

Multiple imputation has proved to be a valuable technique for analysis of the VAHCS, allowing us to conduct analyses that include measures from all waves of data collection, thus recovering information from incomplete cases and avoiding potential selection biases. The sensitivity analyses presented here revealed that MVNI produced stable estimates across all variations in the method. However, stability of results is not necessarily an indicator of validity, and parameter estimates under MVNI are likely to be somewhat biased for measures that have skewed or bimodal distributions, for which the more flexible MICE approach is attractive. On the other hand, MICE may produce less stable results when a large number of variables are included, with particular sensitivity to highly skewed distributions. Further work to delineate the advantages and disadvantages of these 2 methods across a wider range of settings is needed. It was reassuring, however, that despite some sensitivity to the decisions made in building the imputation model, results were generally not affected to the extent that overall conclusions would change. A key finding is that estimates of prevalence are more sensitive to imputation modeling decisions than estimates of association, as observed elsewhere (12).

ACKNOWLEDGMENTS

Author affiliations: Population Health Studies of Adolescents, Murdoch Childrens Research Institute, Parkville, Victoria, Australia (Helena Romaniuk, George C. Patton); Clinical Epidemiology and Biostatistics Unit, Murdoch Childrens Research Institute, Parkville, Australia (Helena Romaniuk, John B. Carlin); Department of Paediatrics, University of Melbourne, Parkville, Australia (Helena Romaniuk, George C. Patton, John B. Carlin); Centre for Adolescent Health, Royal Children's Hospital, Parkville, Australia (George C. Patton); and Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, University of Melbourne, Melbourne, Australia (John B. Carlin).

Data collection for this study was supported by the National Health and Medical Research Council (NHMRC) Australia. H.R. was partially supported by the NHMRC (grant 607400). G.C.P. is supported by a NHMRC senior principal research fellowship. Research at Murdoch Childrens Research Institute is supported by the Victorian Government's Operational Infrastructure Support Program.

We thank Dr. Wendy Swift, Dr. Carolyn Coffey and Prof. Louisa Degenhardt for their contributions to the earlier paper on which this analysis is based (24).

The funding sources had no involvement in the study design; the collection, analysis, and interpretation of the data; the writing of the article; or the decision to submit for publication.

Conflict of interest: none declared.

REFERENCES

1. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York, NY: Wiley; 1987.
2. Schafer JL. *Analysis of Incomplete Multivariate Data*. London, United Kingdom: Chapman & Hall; 1997.
3. Royston P. Multiple imputation of missing values. *Stata J*. 2004;4(3):227–241.
4. Royston P. Multiple imputation of missing values: further update of ice, with an emphasis on interval censoring. *Stata J*. 2007;7(4):445–464.
5. SAS Institute Inc. *SAS/STAT 9.3 User's Guide*. Cary, NC: SAS Institute, Inc.; 2011.
6. Stata Corporation. Stata statistical software, release 11. College Station, TX: Stata Corporation; 2009.
7. Carlin JB, Galati JC, Royston P. A new framework for managing and analysing multiply imputed data in Stata. *Stata J*. 2008;8(1):49–67.
8. Patton GC, Coffey C, Carlin JB, et al. Cannabis use and mental health in young people: cohort study. *BMJ*. 2002;325(7374):1195–1198.
9. Moran P, Coffey C, Mann A, et al. Personality and substance use disorders in young adults. *Br J Psychiatry*. 2006;188:374–379.

10. Patton GC, Coffey C, Lynskey MT, et al. Trajectories of adolescent alcohol and cannabis use into young adulthood. *Addiction*. 2007;102(4):607–615.
11. Degenhardt L, Coffey C, Romaniuk H, et al. The persistence of the association between adolescent cannabis use and common mental disorders into young adulthood. *Addiction*. 2013; 108(1):124–133.
12. Spratt M, Carpenter J, Sterne JA, et al. Strategies for multiple imputation in longitudinal studies. *Am J Epidemiol*. 2010; 172(4):478–487.
13. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338:b2393.
14. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med*. 2011;30(4):377–399.
15. Lee KJ, Carlin JB. Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *Am J Epidemiol*. 2010;171(5):624–632.
16. van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med*. 1999;18(6):681–694.
17. Raghunathan TE, Lepkowski JM, Van Hoewyk J, et al. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv Methodol*. 2001; 27(1):85–95.
18. Royston P, White IR. Multiple imputation by chained equations (MICE): implementation in Stata. *J Stat Softw*. 2011;45(4):1–20.
19. Little RJA. Missing-data adjustments in large surveys. *J Bus Econ Stat*. 1988;6(3):287–296.
20. Li F, Yu Y, Rubin DB. *Imputing Missing Data by Fully Conditional Models: Some Cautionary Examples and Guidelines*. Durham, NC: Department of Statistical Science, Duke University; 2012.
21. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods*. 2001;6(4):330–351.
22. Rubin DB. Multiple imputation after 18+ years. *J Am Stat Assoc*. 1996;91(434):473–489.
23. Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychol Methods*. 2002;7(2):147–177.
24. Swift W, Coffey C, Degenhardt L, et al. Cannabis and progression to other substance use in young adults: findings from a 13-year prospective population-based study. *J Epidemiol Community Health*. 2012;66(7):e26.
25. Patton GC, Hibbert M, Rosier MJ, et al. Is smoking associated with depression and anxiety in teenagers? *Am J Public Health*. 1996;86(2):225–230.
26. National Health and Medical Research Council. *Australian Guidelines to Reduce Health Risk From Drinking Alcohol*. Canberra, Australia: National Health and Medical Research Council; 2009.
27. Carlin JB, Wolfe R, Coffey C, et al. Analysis of binary outcomes in longitudinal studies using weighted estimating equations and discrete-time survival methods: prevalence and incidence of smoking in an adolescent cohort. *Stat Med*. 1999; 18(19):2655–2679.
28. Bernaards CA, Belin TR, Schafer JL. Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Stat Med*. 2007;26(6):1368–1382.
29. White IR, Daniel R, Royston P. Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. *Comput Stat Data Anal*. 2010;54(10): 2267–2275.
30. Abayomi K, Gelman A, Levy M. Diagnostics for multivariate imputations. *Appl Stat*. 2008;57(3):273–291.
31. Graham JW. Missing data analysis: making it work in the real world. *Annu Rev Psychol*. 2009;60:549–576.
32. Stuart EA, Azur M, Frangakis C, et al. Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative. *Am J Epidemiol*. 2009;169(9):1133–1139.
33. Yu LM, Burton A, Rivero-Arias O. Evaluation of software for multiple imputation of semi-continuous data. *Stat Methods Med Res*. 2007;16(3):243–258.
34. Mustillo S. The effects of auxiliary variables on coefficient bias and efficiency in multiple imputation. *Sociol Methods Res*. 2012;41(2):335–361.
35. Lewis G, Pelosi AJ, Araya R, et al. Measuring psychiatric disorder in the community: a standardized assessment for use by lay interviewers. *Psychol Med*. 1992;22(2): 465–486.
36. Goldberg DP, Gater R, Sartorius N, et al. The validity of two versions of the GHQ in the WHO study of mental illness in general health care. *Psychol Med*. 1997;27(1): 191–197.
37. Kessler RC, Andrews G, Mroczek D, et al. The World Health Organization Composite International Diagnostic Interview short form (CIDI-SF). *Int J Methods Psychiatr Res*. 2006;7(4): 171–185.
38. World Health Organization Staff. *Composite International Diagnostic Interview, CIDI-Auto 2.1: Administrator's Guide and Reference*. Geneva, Switzerland: World Health Organization; 1997.
39. World Health Organization. Composite International Interview (CIDI) Core, version 2.1, 12-month version. Geneva, Switzerland: World Health Organization; 1997.