Title

Comparison of Imputation Strategies for Incomplete Longitudinal Data in Life-Course Epidemiology

Journal

American Journal of Epidemiology, 192(12)

ISSN

0002-9262

Authors

Shaw, Crystal
Wu, Yingyan
Zimmerman, Scott C
et al.

Publication Date

2023-11-10

DOI

10.1093/aje/kwad139

Peer reviewed

American Journal of Epidemiology

Vol. 192, No. 12
https://doi.org/10.1093/aje/kwad139
Advance Access publication:
June 20, 2023

Practice of Epidemiology

Comparison of Imputation Strategies for Incomplete Longitudinal Data in Life-Course Epidemiology

Crystal Shaw, Yingyan Wu, Scott C. Zimmerman, Eleanor Hayes-Larson, Thomas R. Belin, Melinda C. Power, M. Maria Glymour, and Elizabeth Rose Mayeda*

* Correspondence to Dr. Elizabeth Rose Mayeda, Department of Epidemiology, Fielding School of Public Health, University of California, Los Angeles, 650 Charles E. Young Drive South, Los Angeles, CA 90095-1772 (e-mail: ermayeda@ph.ucla.edu).

Initially submitted August 2, 2022; accepted for publication June 13, 2023.

Incomplete longitudinal data are common in life-course epidemiology and may induce bias leading to incorrect inference. Multiple imputation (MI) is increasingly preferred for handling missing data, but few studies explore MI-method performance and feasibility in real-data settings. We compared 3 MI methods using real data under 9 missing-data scenarios, representing combinations of 10%, 20%, and 30% missingness and missing completely at random, at random, and not at random. Using data from Health and Retirement Study (HRS) participants, we introduced record-level missingness to a sample of participants with complete data on depressive symptoms (1998–2008), mortality (2008–2018), and relevant covariates. We then imputed missing data using 3 MI methods (normal linear regression, predictive mean matching, variable-tailored specification), and fitted Cox proportional hazards models to estimate effects of 4 operationalizations of longitudinal depressive symptoms on mortality. We compared bias in hazard ratios, root mean square error, and computation time for each method. Bias was similar across MI methods, and results were consistent across operationalizations of the longitudinal exposure variable. However, our results suggest that predictive mean matching may be an appealing strategy for imputing life-course exposure data, given consistently low root mean square error, competitive computation times, and few implementation challenges.

fully conditional specification; Health and Retirement Study; joint modeling; longitudinal data; missing not at random; multiple imputation; multiple imputation by chained equations; predictive mean matching

Abbreviations: BMI, body mass index; CES-D, Center for Epidemiological Studies Depression scale; HRS, Health and Retirement Study; LMM, linear mixed model; MAR, missing at random; MCAR, missing completely at random; MI, multiple imputation; MNAR, missing not at random; NORM, normal linear regression; PMM, predictive mean matching; VTS, variable-tailored specification.

Missing data is common in longitudinal studies and can reduce precision and bias estimates, leading to invalid inferences. Missing data challenges are particularly salient in life-course epidemiology, which focuses on understanding the health effects of social, behavioral, and biological factors over the lifespan. Life-course studies often require repeated measures on the same person at specific ages (e.g., ages 50, 60, and 70 years) to define exposures. Missing data at points of interest necessitates decision-making on how to best use available data.

Several analytical approaches exist for handling missing data. Complete-case analysis—discarding observations with missing data—is the simplest and most common approach (1–4). However, complete-case analyses are inefficient and results may be biased (5–8). Imputation strategies include single imputation (e.g., mean imputation, last observation carried forward), which typically produces incorrectly small variance estimates (4), and multiple imputation (MI), which incorporates uncertainty in imputed values and may use relationships between measured quantities to impute missing values. MI methods are preferred and are available in most statistical software; however, researchers must decide which MI method to use.

Packages like mice (9) in R (R Foundation for Statistical Computing, Vienna, Austria) supply several MI methods. Methods vary predominantly in statistical models used to

estimate relationships between variables and are often classified by data type (numeric, binary, ordered, unordered, or any) and structure (e.g., longitudinal). Additional factors influencing MI method performance include missing data mechanism, amount of missing data, distributions of variables to be imputed, and MI method modeling assumptions. Choosing an MI method also involves practical considerations, including computational time. However, there is limited guidance on how to choose an MI method, especially for longitudinal/life-course data.

Existing MI method comparisons in longitudinal settings use simulated data sets with simplified joint distributions among variables (10), and are often conducted using unrealistically few covariates (11–13). A few comparisons of imputation approaches use real data sets (6, 7, 14), but there is a need for systematic evaluation of MI method performance across missing data characteristics (missing data mechanism and proportion of missing data) and operationalizations of exposure variables. We compared performance of 3 MI methods in a real-data setting under varied induced missing-data mechanisms and proportions.

METHODS

Overview

We studied MI method performance in a real data set using the example of estimating effects of elevated depressive symptoms on mortality among middle-aged and older adults, with depressive symptoms measured repeatedly, allowing for different definitions of elevated depressive symptomatology (e.g., cross-sectional, cumulative). We acknowledge the ambiguity in defining depressive symptoms as an exposure; regardless of concerns about the consistency assumption (15), depressive phenotypes are commonly used as independent variables in research on myriad health outcomes (1, 16–20).

The main steps of our analysis were: 1) construct a longitudinal data set without missing data (the "complete sample"); 2) estimate effect of elevated depressive symptoms on mortality in the complete sample ("true" effect estimates); 3) induce missingness under several missing data mechanisms and proportions; 4) impute missing data using 3 MI methods; and 5) compare results from MI methods with "true" effect estimates from the complete sample. Details of each step follow.

Step 1. Construct a longitudinal data set without missing data (the "complete sample"). The Health and Retirement Study (HRS) is a population-representative cohort study of US adults aged 50 years or older with biennial follow-up (21). We identified a sample aged 50–90 years in 1998 with complete records of depressive symptoms for all 6 study waves from 1998–2008, adequate data for covariates (details below), and 2008–2018 mortality data.

Depressive symptoms were measured using the 8-item Center for Epidemiological Studies Depression scale (CES-D) (22). Scores ranged from 0 to 8, with higher scores indicating more depressive symptoms. We defined elevated depressive symptoms as CES-D score $\geq$4 (23). Baseline covariates included sex/gender, age, race/ethnicity, and educational attainment. Wave-updated covariates included marital status, body mass index (BMI), self-rated health, smoking status, drinking behavior, and self-reported diagnosis of individual chronic conditions: hypertension, diabetes, cancer, stroke, heart disease, lung disease, or memory problems. We used dichotomous variables for each chronic condition in analytical models and derived chronic condition count by summing the number of self-reported chronic conditions per participant at each wave for models used to induce missingness (described below).

BMI was calculated as height (kg)/weight (m)^2 (details in Web Appendix 1). Educational attainment was categorized based on self-reported years of education ($<$12, 12, 13–15, 16, $>$16). Marital status was coded as "married/partnered" versus "not married/partnered." Drinking behavior (no drinking, moderate drinking, heavy/high-risk drinking) was classified according to the 2020 Dietary Guidelines for Americans (24). To preserve sample size, we carried forward the last observation for individuals with missing data on marital status, drinking behavior, and self-reported chronic conditions, as there was little intra-individual variation in these variables over time.

HRS conducts mortality follow-up during regular biennial waves by contacting all HRS participants from the previous wave; if participants are unreachable, family members are contacted. If a participant is reported dead, family members complete exit interviews that include dates and causes of death (25).

Web Figure 1 shows the flow diagram for obtaining the complete sample. Of those with fully observed CES-D measurements from 1998–2008 and mortality data through 2018, we dropped those aged $<$50 or $>$90 in 1998 (n = 524), those missing race/ethnicity data (n = 1), and those missing drinking behavior for all waves (n = 1), BMI data at any wave (n = 370), self-rated health data at any wave (n = 27), or all ever/never chronic conditions (n = 6) for all waves, resulting in n = 9,445 participants in the complete sample.

Step 2. Estimate effect of elevated depressive symptoms on mortality in the complete sample. In practice, "elevated depressive symptoms" is operationalized different ways (1, 16–20). We used 4 operationalizations: 1) elevated (CES-D scores $\geq$4) (23) depressive symptoms at baseline (1998, dichotomous yes/no); 2) elevated depressive symptoms at end of exposure period (2008, dichotomous yes/no); 3) proportion of waves (1998–2008) with elevated depressive symptoms; and 4) elevated average depressive symptom scores, derived by averaging the CES-D measures across waves (1998–2008) and applying the CES-D cutoff to the average (dichotomous yes/no). Percent agreement between exposure operationalizations demonstrated variability in composition of exposed and unexposed groups (Web Table 1, available at https://doi.org/10.1093/aje/kwad139).

We fitted Cox proportional hazards models in the complete sample to estimate effects of each of the 4 approaches to operationalizing elevated depressive symptoms on mortality. We considered effect estimates from the complete sample to be the "truth," that is, the effect estimates in the absence of missing data. We controlled for variables we conceptualized as potential confounders of depressive

symptoms and mortality: age, sex, race/ethnicity, educational attainment, marital status, BMI, drinking behavior, smoking status, and self-reported diagnosis of hypertension, diabetes, cancer, stroke, heart disease, lung disease, and memory problems. We used baseline (1998) measurements of potential confounders in all models except the model for elevated depressive symptoms at end of exposure period, which used end-of-exposure-period (2008) measurements.

Step 3. Induce missingness under several missing data mechanisms and proportions. We induced record-level missingness (i.e., missingness for all wave j measures) in the complete sample using 9 scenarios that combined 3 missing data mechanisms (missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR)) and 3 proportions missing (10%, 20%, and 30%).

Under MCAR (probability of missingness is independent of all covariates), probability of missing wave j was set at the proportion missing in the scenario for all participants in all waves. Following examples in literature (7, 10), probability of missingness under MAR (probability of missingness depends on observed covariates) and MNAR (probability of missingness depends on unobserved values) (5) were modeled using logistic regression for each wave j. We used the predicted probability of missingness for individual i at wave j as the success probability for a Bernoulli random draw (1 = missing, 0 = not missing). An example of MNAR missingness would be a participant missing wave j because they were too depressed to participate (i.e., higher CES-D scores increased probability of missingness).

MAR and MNAR missingness were induced using models 1 and 2, respectively. Predictors in each missing data model were based on conceptual models of what could cause participants to miss study visits (Web Figure 2). The MAR model (model 1) used covariates that would still be observed after inducing missing data for wave j. The MNAR model (model 2) used covariates that would not be observed after inducing missing data for wave j, including depressive symptoms at wave j and an unobserved common cause of missingness and mortality. To simplify simulations in MNAR scenarios, we induced missingness consistent with an unobserved U influencing both missingness and mortality by allowing mortality to influence missingness directly. Thus, model 2 is not causal since subsequent mortality affects missing data at previous waves (1998–2008). We fixed effect sizes in both models for inducing missing data and obtained intercepts by optimizing for proportion of missingness in each scenario (Web Table 2). To ensure that MAR and MNAR missingness scenarios were distinct, we did not allow values that were ultimately set to missing at wave $j-1$ to impact missingness at wave j.

$$\begin{aligned}\text{logit}\left(P\left(missing_j\right)\right) &= \beta_0 + \beta_1 CESD_{j-1} \\ &+ \beta_2(\textit{chronic condition count})_{j-1} + \beta_3 CESD_{j-1} \\ &\times (\textit{chronic condition count})_{j-1} \qquad (1)\end{aligned}$$

$$\begin{aligned}\text{logit}\left(P\left(missing_j\right)\right) &= \beta_0 + \beta_1 CESD_j + \beta_2 death2018 \\ &+ \beta_3 CESD_j \times death2018 \qquad (2)\end{aligned}$$

Step 4. Impute missing data using 3 MI methods. All chosen MI methods use fully conditional specification, a strategy that specifies separate models for each variable to be imputed, conditional on other variables in the data set. The analyses here used the mice (multiple imputation with chained equations) package in R (R Foundation for Statistical Computing) (9). Each variable was imputed using the specified model type; this process was repeated multiple times, cycling through all variables to be imputed (26). We evaluated 3 common MI methods: 1) normal linear regression (NORM) for all conditional models (equivalent to joint multivariate normal modeling) (26); 2) predictive mean matching (PMM) for all conditional models; and 3) conditional models tailored to the outcome type (variable-tailored specification (VTS)). Across methods, we treated ordinal variables (education, drinking behavior, and CES-D scores) as having underlying continuous scales and created indicator variables for levels of nominal categorical variables (marital status and race/ethnicity). We excluded participants with missing data at all 6 waves after inducing missingness; this resulted in dropping <20 participants (0.2%) from analyses in scenarios with 30% missingness induced.

NORM treats all variables as continuous and uses normal linear regression for each conditional model. This can be implemented in the R (R Foundation for Statistical Computing) package mice by specifying the "norm" method for each imputed variable and was implemented in this analysis using the miceFast package version of the "norm" option (27). We treated marital status indicators as if continuous, with imputed values thus interpretable as predicted probabilities of category membership. Predicted probabilities were then truncated at (0, 1) and used to impute marital status indicators with draws from Bernoulli distributions with success probability equal to predicted probabilities. In rare cases where this procedure resulted in a participant imputed as belonging to multiple categories, a category was drawn from the imputed categories at random.

PMM calculates a predicted value of the variable to be imputed for observed and missing participants using linear regression. Missing values are imputed by randomly choosing an observed value from a pool of "nearest neighbor" donors based on proximity between predicted values for observed participants and predicted value of the missing data point and user-specified pool size. Thus, PMM guarantees imputations within range of observed data. PMM can be implemented in the mice package by specifying the "pmm" method for each imputed variable and was implemented in this analysis using the miceFast package version of the "pmm" option with 10 donors (27).

For VTS, we used the R package mice and imputed binary variables using logistic regression ("logreg" method) and continuous variables using linear regression ("norm" method).

Although not designed for longitudinal data, NORM, PMM, and VTS can be adapted for longitudinal settings and preserve within-person correlation by imputing wide data sets and including all other variables at all other waves in imputation models (28). We imputed wave-specific values using all other variables at all other waves and used imputed

values to derive 4 operationalizations of elevated depressive symptoms.

Imputation models are iterative and require a burn-in period to produce convergence (stable imputed values). The burn-in period required for convergence varies by imputation method, data set complexity, and missing data proportion. We monitored plots of means and standard deviations across imputation runs for each imputed variable and chose appropriate burn-in periods for each MI method and missing data proportion combination (26). Burn-in ranged from 10 to 15 iterations across imputation methods and scenarios. We set number of imputations equal to missing data percent (e.g., 10 imputations for 10% missing data) as suggested by Bodner (29) and White et al. (30).

Step 5. Compare results from MI methods with "true" effect estimates from the complete sample. After inducing missing data and imputing data sets using 3 MI methods, we fitted the Cox proportional hazards analytical models described in Step 2 (estimate effect of elevated depressive symptoms on mortality in the complete sample) in each imputed data set, pooled estimates across imputed data sets for each scenario using Rubin's rules (31), and compared results with the "truth" in the complete sample. For comparison, we included complete-case analyses where missing observations were dropped instead of imputed. Complete-case analyses for cumulative exposures (elevated average CES-D, proportion elevated CES-D) included participants with at least 4 (out of 6 possible) CES-D measures to mimic real-data approaches.

We repeated steps 3–5 (induce missingness, impute missing data, run analyses) 1,000 times for each scenario. We summarized effect estimates for each exposure by taking means across 1,000 simulation runs and calculated 95% confidence intervals by taking means of upper and lower confidence interval values across runs. We measured performance of each MI method using bias $\left(\bar{\hat{\beta}}-\beta\right)$; root mean square error $\left(RMSE = \sqrt{bias^2 + variance\left(\hat{\beta}\right)}\right)$; and computation time. We assessed sensitivity of results to prevalence of elevated depressive symptoms by repeating analyses using CES-D $\geq$1 to define elevated depressive symptoms.

Analyses were conducted in R (R Foundation for Statistical Computing) version 4.0.2. MI was implemented using mice version 3.13.18 (9) and miceFast version 0.7.1 (27) packages. We used computational and storage services associated with the Hoffman2 Shared Cluster provided by UCLA Institute for Digital Research and Education's Research Technology Group.

Linear mixed effects model MI-method case study

NORM, PMM, and VTS were adapted for this longitudinal setting by using a wide data set and including observations from other waves in imputation models. However, the mice package includes the linear mixed models (LMM) MI method for longitudinal data. This MI method posed several implementation challenges and is therefore presented as an isolated case study. We used LMM to assess whether modeling within-person correlation would reduce bias or increase precision in MI estimates by transforming our data set from wide (one row per participant) to long (multiple rows per participant, one for each observed time point) and specifying the "2l.lmer" method for each imputed variable. LMM was implemented only for MAR and MNAR missing data mechanisms at 30% missingness since it required significantly more computation time and resources compared with other MI methods. The burn-in period for LMM was fixed at 5 iterations for feasibility; convergence was not achieved for all variables.

RESULTS

Table 1 shows characteristics of the complete sample. Compared with those without elevated depressive symptoms at baseline (n = 8,209), the group with elevated depressive symptoms at baseline (n = 1,236) had a higher proportion of people who identified as female or as Hispanic or Black, lower levels of education, a lower proportion of married/partnered participants, and higher prevalence of chronic conditions.

We encountered no implementation challenges with NORM and PMM; however, VTS models did not run with all variables originally included in imputation models or when we correctly specified multinomial logistic regression for nominal categorical outcomes and proportional odds models for ordered categorical outcomes. Thus, we dropped auxiliary variables (self-rated health), used indicator variables for nominal categories, and treated ordered categorical outcomes as continuous in all MI models. However, there were still 12 VTS imputation runs (out of 9,000) that did not complete.

Figure 1 shows estimated effects of elevated depressive symptoms on mortality according to scenario and MI method compared with "true" effect estimates in the complete sample. Bias was minimal in MCAR and MAR scenarios regardless of MI method. For example, the "true" effect estimate for elevated average CES-D on mortality was hazard ratio = 1.27 (ln (hazard ratio) = 0.24); with 30% missing data under an MAR mechanism, the NORM estimate was hazard ratio = 1.27 (95% confidence interval: 1.09, 1.49), the PMM estimate was hazard ratio = 1.27 (95% confidence interval: 1.07, 1.48), and the VTS estimate was hazard ratio = 1.26 (95% confidence interval: 1.08, 1.46) (Figure 1G). Complete-case analysis was nearly unbiased for the MAR mechanism but was less precise (hazard ratio = 1.26, 95% confidence interval: 1.04, 1.54). Although bias was smaller than complete-case analyses, no MI method recovered "true" effect estimates in MNAR scenarios—most analyses incorrectly suggested null or protective effects of elevated depressive symptoms on mortality. Across simulation scenarios, PMM estimates were least biased (Figure 1, Web Table 3).

Bias and RMSEs are presented by scenario and MI method in Figures 2 and 3, respectively. Lower RMSEs indicate better performance for recovering "true" effect estimates. RMSEs were consistently highest for complete-case analyses and consistently lowest for PMM in MAR and MNAR scenarios.

Figure 4 shows MI method computation time across 1,000 runs by percent missing data, aggregated across missing

Table 1. Baseline Sample Characteristics, Overall and Stratified by Elevated Scores on the Center for Epidemiological Studies Depression Scale, in the Health and Retirement Study, United States, 1998–2008

Characteristic	Overall (*n* = 9,445)		Baseline CES-D			
			Elevated CES-D[a] (*n* = 1,236)		Not Elevated CES-D[a] (*n* = 8,209)	
	No.	%	No.	%	No.	%
Baseline age, years[b]	63 (8)		62 (8)		63 (8)	
Female sex	5,758	61.0	917	74.2	4,841	59.0
Race/ethnicity						
Hispanic	659	7.0	194	15.7	465	5.7
Non-Hispanic Black	1,144	12.1	236	19.1	908	11.1
Non-Hispanic White	7,480	79.2	777	62.9	6,703	81.7
Other	162	1.7	29	2.3	133	1.6
Educational level						
<12 years	2,149	22.8	501	40.5	1,648	20.1
12 years	3,347	35.4	420	34.0	2,927	35.7
13–15 years	1,908	20.2	183	14.8	1,725	21.0
16 years	936	9.9	63	5.1	873	10.6
>16 years	1,105	11.7	69	5.6	1,036	12.6
Marital status						
Married/partnered	6,883	72.9	713	57.7	6,170	75.2
Not married/partnered	1,253	13.3	256	20.7	997	12.1
Widowed	1,309	13.9	267	21.6	1,042	12.7
Self-reported chronic conditions						
Hypertension	3,570	37.8	565	45.7	3,005	36.6
Diabetes	937	9.9	184	14.9	753	9.2
Heart disease	1,239	13.1	216	17.5	1,023	12.5
Stroke	316	3.3	67	5.4	249	3.0
Cancer	758	8.0	89	7.2	669	8.1
Lung disease	376	4.0	95	7.7	281	3.4
Memory problems	44	0.5	24	1.9	20	0.2
Chronic condition count[b]	0.77 (0.88)		1.00 (1.02)		0.73 (0.86)	
BMI[b,c]	28.7 (5.4)		29.7 (6.2)		28.6 (5.3)	
Alcohol intake						
Heavy alcohol use	951	10.1	110	8.9	841	10.2
Moderate alcohol use	2,234	23.7	177	14.3	2,057	25.1
No alcohol use	6,260	66.3	949	76.8	5,311	64.7
Smoking status						
Ever smoked	1,754	18.6	319	25.8	1,435	17.5
Never smoked	7,691	81.4	917	74.2	6,774	82.5

Abbreviations: BMI, body mass index; CES-D: Center for Epidemiological Studies Depression scale.
[a] Elevated CES-D score defined as ≥4.
[b] Values are expressed as mean (standard deviation).
[c] Weight (kg)/height (m)^2.

data mechanisms. NORM was consistently fastest and VTS consistently slowest.

Our alternative cutoff for elevated depressive symptoms increased prevalence of elevated depressive symptoms at baseline (1998) in the complete sample from 13.1% to 56.4%. MI methods performed similarly in sensitivity analyses using the higher prevalence definition of elevated depressive symptoms (Web Figures 3–6, Web Table 4). In

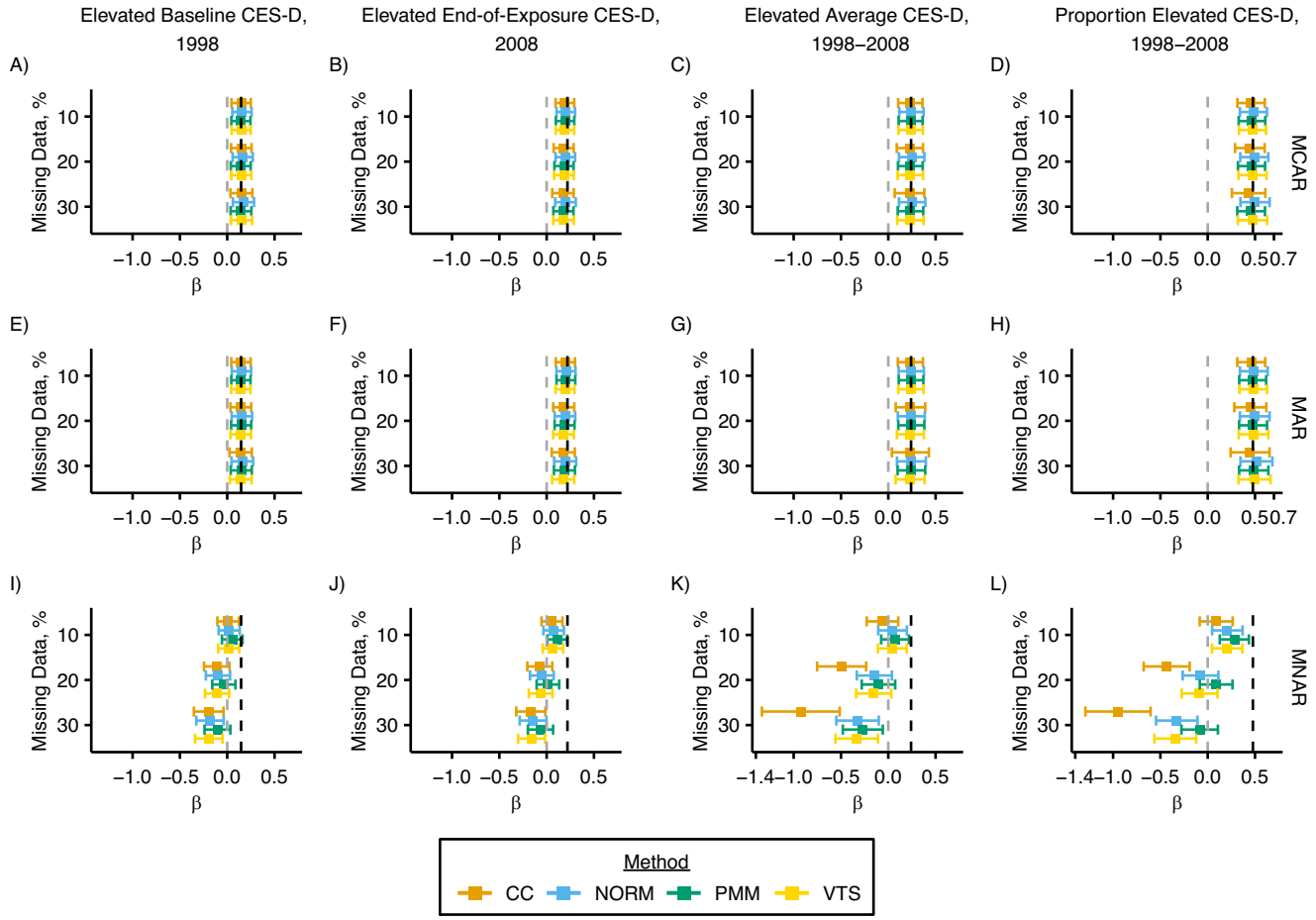

Figure 1. Mean point estimates β (ln(hazard ratio)) and 95% confidence intervals for estimated effects of each operationalization of elevated depressive symptoms on mortality across 1,000 simulation runs according to missing data mechanism, percent missing data, and multiple imputation method, Health and Retirement Study, United States, 1998–2008. A–D) Missing completely at random (MCAR) mechanism; E–H) missing at random (MAR) mechanism; I–L) missing not at random (MNAR) mechanism; columns 1 through 4 indicate different operationalizations of elevated Center for Epidemiological Studies Depression (CES-D) scores. Dashed black lines indicate "true" effect estimates in the complete sample prior to inducing missing data. Dashed gray lines indicate the null. CC, complete-case analysis; NORM, normal linear regression; PMM, predictive mean matching; VTS, variable-tailored specification.

simulation scenarios where LMM was assessed (MAR 30%, MNAR 30%), bias and precision were similar to other MI methods (Web Figures 7–8, Web Table 3). Root mean square error was not consistently higher or lower for LMM vs. other MI methods. (Web Figure 9), but LMM required nearly quadruple the computation time of other MI methods (Web Figure 10).

DISCUSSION

Incomplete data is a challenge in life-course epidemiology, but relative performance of imputation approaches is not well characterized. We compared performance of 3 MI methods in a real data set using the example of estimating the effect of longitudinally measured elevated depressive symptoms on mortality. This example represents a realistic setting where missing data are plausibly influenced by the exposure and outcome through complex missing data processes and inferences may be affected if missing data are handled improperly. As relevant timing of elevated depressive symptoms has been debated (20, 32), we considered multiple operationalizations of elevated depressive symptoms.

We assessed MI method bias, precision, and feasibility. We observed minimal bias in complete-case analyses under MCAR and MAR mechanisms, but the precision loss would be particularly important when true effects are small, and imprecise estimates could lead to incorrect inference. Thus, MI methods outperformed the complete-case approach across all simulation scenarios and performed similarly to each other in terms of bias, but with wide variability in feasibility. Broadly, PMM imputation had consistently low root mean square error, fast computation times, and few implementation challenges.

Researchers must make several choices in analyses involving missing data. First, they must choose between

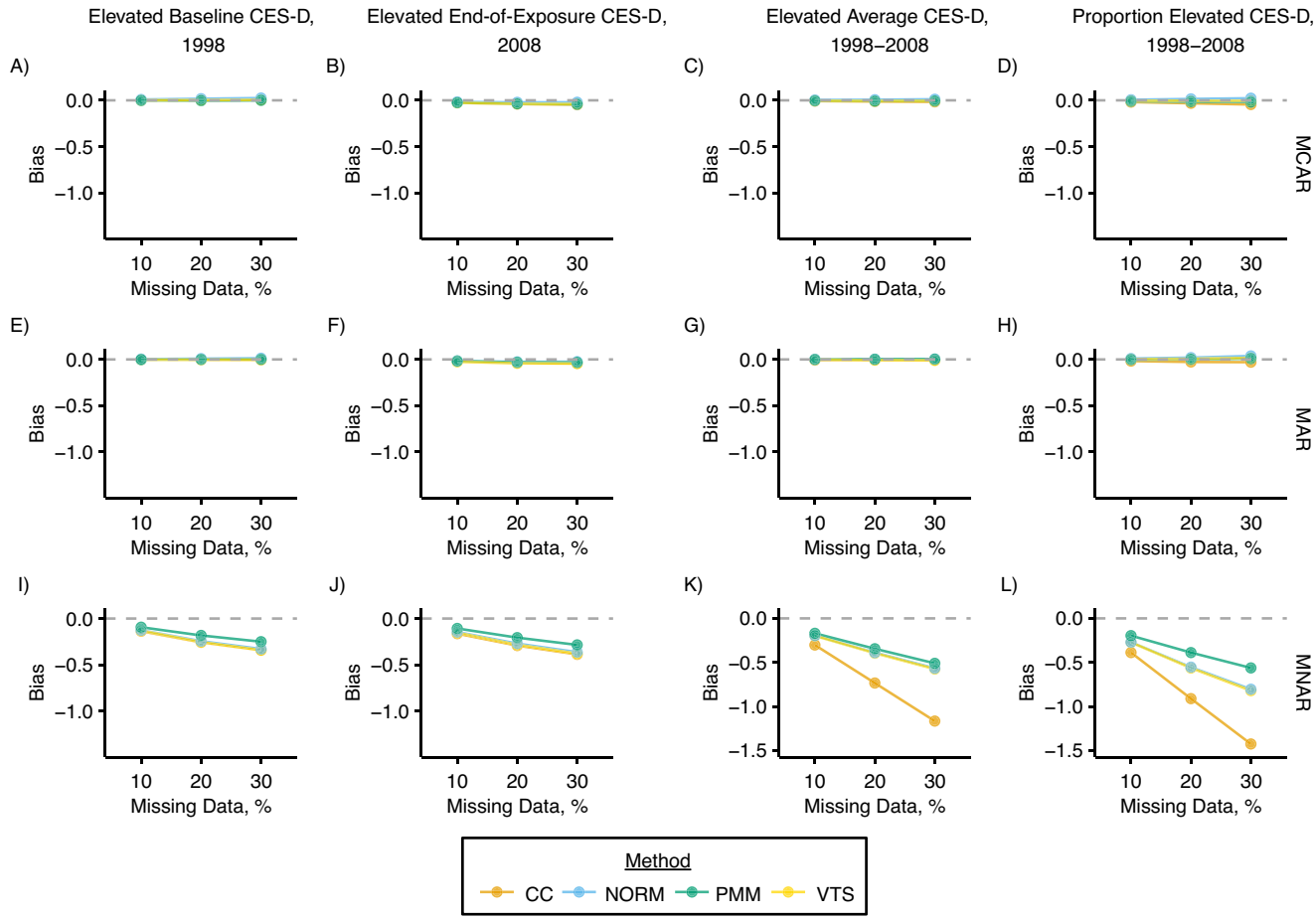

Figure 2. Bias $(\bar{\hat{\beta}} - \beta)$ in the estimated effect of each operationalization of elevated depressive symptoms on mortality across 1,000 simulation runs by missing data mechanism, percent missing data, and multiple imputation method, Health and Retirement Study, United States, 1998–2008. A–D) Missing completely at random (MCAR) mechanism; E–H) missing at random (MAR) mechanism; I–L) missing not at random (MNAR) mechanism; columns 1 through 4 indicate different operationalizations of elevated Center for Epidemiological Studies Depression (CES-D) scores. Dashed gray line at 0 indicates no bias. CC, complete-case analysis; NORM, normal linear regression; PMM, predictive mean matching; VTS, variable-tailored specification.

easy-to-implement but potentially misspecified imputation models and correctly specified models that are computationally intensive or may fail to converge. In our study, MI model misspecification did not have a negative impact on inferences. For example, the true data structure was longitudinal, but researchers commonly use MI methods that impute data cross-sectionally (NORM, PMM, VTS), using wide data sets to preserve longitudinal information (33, 34). Explicitly modeling data longitudinally with LMM imputation required nearly quadruple the computation time of VTS, the most computationally demanding cross-sectional method, despite restricting the burn-in to 5 iterations to save computational time, and did not have smaller root mean square error to the other MI methods. Additionally, PMM and NORM treat all variables as continuous, misspecifying binary, categorical, and ordinal variables; however, both methods performed well across scenarios and in some cases outperformed VTS, which correctly specified all variable types.

When data is missing on variables with deterministic relationships with other variables (e.g., BMI, calculated from height and weight), researchers can use active imputation and impute the missing variable directly (e.g., impute BMI) or use passive imputation and impute missing data in component parts and then derive the variable of interest (e.g., impute height and weight and then calculate BMI) (26). Neither strategy performs uniformly better than the other (35, 36). We chose to passively impute the 4 operationalizations of elevated depressive symptoms because it guarantees congeniality (i.e., so there would be no discrepancy between a participant's imputed CES-D values and their elevated depressive symptoms classification).

Variable selection for imputation models is another challenge. The "kitchen sink" approach—including every variable from analytical models, the outcome variable, and any auxiliary variables correlated with variables to be imputed—is often recommended (28). However, this strategy combined with the recommendation above to use multiple waves

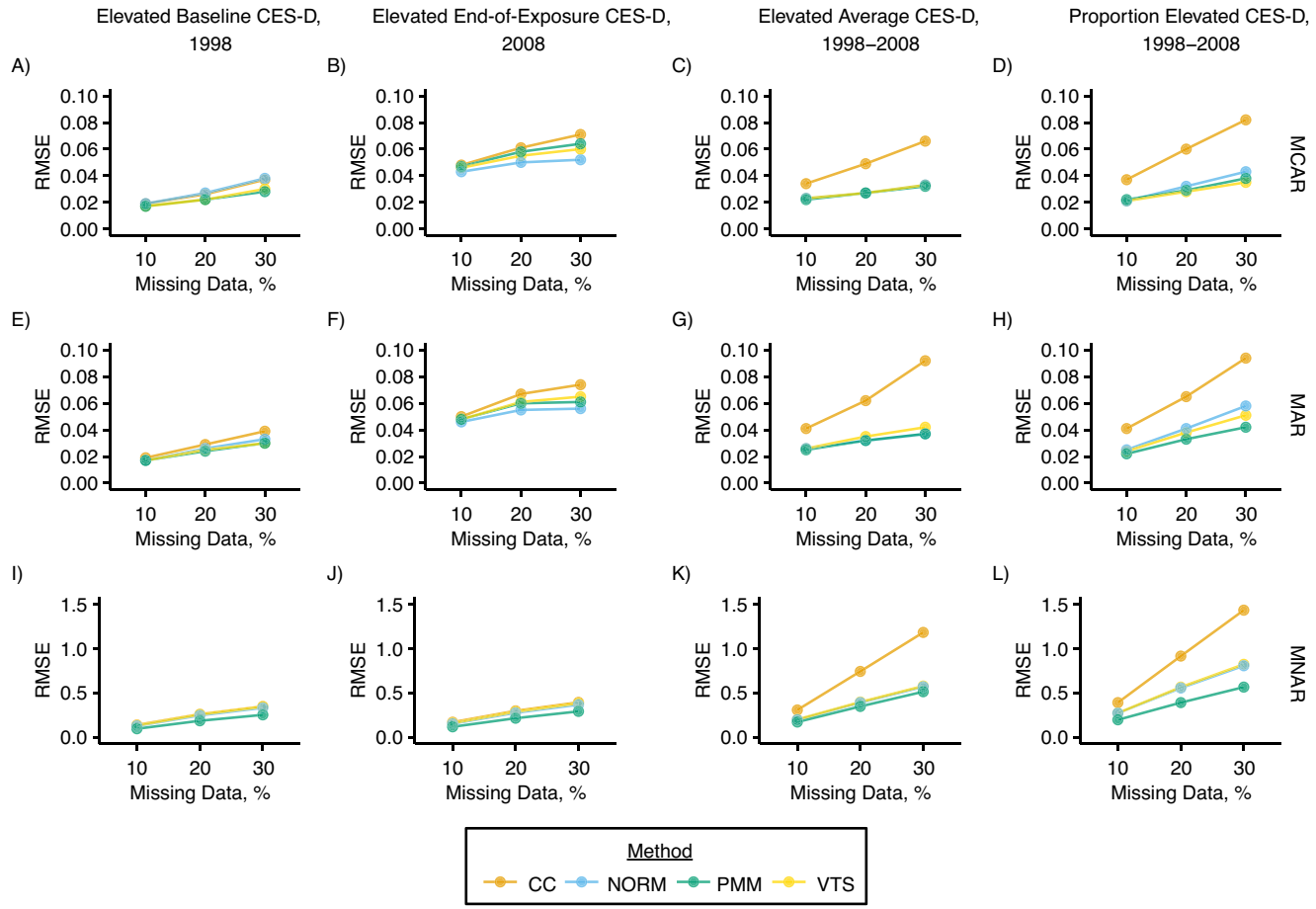

Figure 3. Root mean square error (RMSE) of estimated effects for each operationalization of elevated depressive symptoms on mortality across 1,000 runs by missing data mechanism, percent missing data, and multiple imputation method, Health and Retirement Study, United States, 1998–2008. A–D) Missing completely at random (MCAR) mechanism; E–H) missing at random (MAR) mechanism; I–L) missing not at random (MNAR) mechanism; columns 1 through 4 indicate different operationalizations of elevated Center for Epidemiological Studies Depression (CES-D) scores. CC, complete-case analysis; NORM, normal linear regression; PMM, predictive mean matching; VTS, variable-tailored specification.

of data in a longitudinal setting can multiplicatively increase the number of variables, may lead to convergence issues, and increases computational time.

The MAR assumption embedded in MI methods (5) is untestable, and missing data encountered in practice likely arise from a combination of MAR and MNAR missingness. Evaluating each missing data mechanism separately provides insight for what to expect when multiple mechanisms contribute to missingness; results from a scenario where missing data arise from a combination of MAR and MNAR mechanisms would fall between results presented in this work and would more closely resemble results from whichever mechanism contributed more heavily to the missing data structure. As expected, we were not able to obtain valid inferences using MI methods in MNAR scenarios. However, inferences were closer to the truth when using MI rather than a complete-case analysis, and PMM consistently performed the best under MNAR missingness. This finding aligns with a 2010 simulation study showing that PMM performed at least as well as theoretically superior imputation methods (37) and corroborates results from 2005 and 2017 studies that concluded PMM was robust to model misspecification (7, 38).

A key strength of our study was comparing MI method performance in a real data set with complex joint distributions among covariates. Our simulation design required a completely observed data set from which we induced missingness under known mechanisms. Thus, analyses were restricted to HRS participants with fully observed data who survived past 2008, which may have distorted effect estimates for CES-D on mortality in this sample. Although this sample would be inappropriate for estimating population-level effects of CES-D on mortality, it is effective for assessing whether MI methods can recover inferences we would have drawn had the sample been completely observed. Another strength is inclusion of MNAR mechanisms. Simulation studies comparing MI methods are usually limited to MCAR and MAR mechanisms (7, 10–13, 39, 40), which match MI assumptions and therefore represent ideal MI method performance. Results from

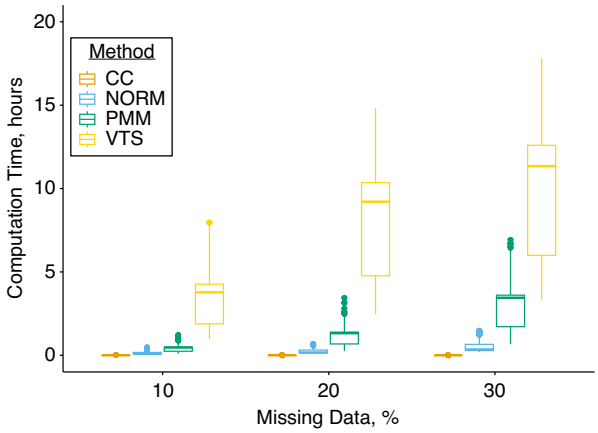

Figure 4. Computation time across 1,000 runs according to percent missing data and multiple imputation method, aggregated across missing data mechanism and analytical models. CC, complete-case analysis; NORM, normal linear regression; PMM, predictive mean matching; VTS, variable-tailored specification.

MNAR scenarios demonstrate MI method performance when assumptions are violated, as they likely are in practice.

Performance of a given imputation method may differ according to data set characteristics and research question. Our results may not generalize to all other settings; however, we have increased confidence in their robustness because we systematically tested combinations of amount and patterns of missingness, considered multiple exposure operationalizations, conducted sensitivity analyses with increased exposure prevalence, and drew similar conclusions about MI method performance across these scenarios.

Researchers may encounter more than 30% missingness in practice. We attempted scenarios with higher missing data proportions and encountered implementation issues with VTS that could only be remedied by significantly reducing the number of variables in imputation models. This suggests that NORM and PMM may be especially preferred in scenarios with large amounts of missing data.

We conducted analyses using estimated effects of elevated longitudinally measured depressive symptoms on mortality as a life-course epidemiology analysis example. Our study can inform recommendations for researchers facing inference challenges in the presence of missing data. MI is efficient, widely available across software, and thus generally considered superior to complete-case analyses even when MI assumptions are likely violated due to MNAR missing data mechanisms. In our study, MI methods were robust to model misspecifications made to improve feasibility, but more work is needed on systematic comparisons of modeling violations (e.g., imputing highly skewed variables) in a real data setting. Understanding MI method performance when exposures are defined based on age at assessment (versus assessment wave in this paper) is also an important area. Because study participants are recruited at different ages, using age as the timescale often increases missing data proportion, leading to MI implementation challenges, and potentially introduces cohort effects. Missing data methods are essential in life-course epidemiology; our results suggest many established MI methods are computationally feasible and perform well.

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, Fielding School of Public Health, University of California, Los Angeles, Los Angeles, California, United States (Crystal Shaw, Yingyan Wu, Eleanor Hayes-Larson, Elizabeth Rose Mayeda); Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles, Los Angeles, California, United States (Crystal Shaw, Thomas R. Belin); Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, California, United States (Scott C. Zimmerman, M. Maria Glymour); and Department of Epidemiology, Milken Institute School of Public Health, The George Washington University, Washington, DC, United States (Melinda C. Power).

This work was funded by National Institutes of Health, National Institute on Aging (grants R01AG057869, R00AG053410, K99AG075317, and F31AG071191), and National Institutes of Health, National Center for Advancing Translational Sciences (grant number UL1 TR001881).

Health and Retirement Study (HRS) data are publicly available at https://hrs.isr.umich.edu/data-products. Data set construction and analytical code is available on Github: https://github.com/Mayeda-Research-Group/exposure-trajectories.

Presented as a poster at the 2022 Society for Epidemiologic Research annual meeting, June 14–17, 2022, Chicago, Illinois (poster P624).

The views expressed in this article are those of the authors and do not reflect those of the National Institutes of Health.

Conflict of interest: none declared.

REFERENCES

1. Pedersen J, Thorsen SV, Andersen MF, et al. Impact of depressive symptoms on worklife expectancy: a longitudinal study on Danish employees. *Occup Environ Med.* 2019; 76(11):838–844.
2. Colman I, Kingsbury M, Sucha E, et al. Depressive and anxious symptoms and 20-year mortality: evidence from the Stirling County Study. *Depress Anxiety.* 2018;35(7):638–647.
3. Li H, Qian F, Hou C, et al. Longitudinal changes in depressive symptoms and risks of cardiovascular disease and all-cause mortality: a nationwide population-based cohort study. *J Gerontol A Biol Sci Med Sci.* 2020;75(11):2200–2206.
4. Okpara C, Edokwe C, Ioannidis G, et al. The reporting and handling of missing data in longitudinal studies of older adults is suboptimal: a methodological survey of geriatric journals. *BMC Med Res Methodol.* 2022;22(1):122.
5. Rubin DB. Inference and missing data. *Biometrika*. 1976; 63(3):581–592 Accessed April 14, 2022.

6. Dahal P, Stepniewska K, Guerin PJ, et al. Dealing with indeterminate outcomes in antimalarial drug efficacy trials: a comparison between complete case analysis, multiple imputation and inverse probability weighting. *BMC Med Res Methodol.* 2019;19(1):215.
7. Tang L, Song J, Belin TR, et al. A comparison of imputation methods in a longitudinal randomized clinical trial. *Stat Med.* 2005;24(14):2111–2128.
8. Little RJA, Rubin DB. *Statistical Analysis With Missing Data*. 3rd ed. Hoboken, NJ: Wiley; 2019.
9. van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. *J Stat Softw*. 2011; 45(3):1–67.
10. Huque MH, Carlin JB, Simpson JA, et al. A comparison of multiple imputation methods for missing data in longitudinal studies. *BMC Med Res Methodol.* 2018;18(1):168–116.
11. Drechsler J. Multiple imputation of multilevel missing data—rigor versus simplicity. *J Educ Behav Stat.* 2015;40(1): 69–95.
12. Grund S, Lüdtke O, Robitzsch A. Multiple imputation of missing data for multilevel models: simulations and recommendations. *Organ Res Methods.* 2017;21(1): 111–149.
13. Plumpton CO, Morris T, Hughes DA, et al. Multiple imputation of multiple multi-item scales when a full imputation model is infeasible. *BMC Res Notes.* 2016; 9(1):45.
14. Hron K, Templ M, Filzmoser P. Imputation of missing values for compositional data using classical and robust methods. *Comput Stat Data Anal.* 2010;54(12):3095–3107.
15. Hernán MA, Robins JM. *Causal Inference: What If?* 1st ed. Boca Raton, FL: Chapman & Hall/ CRC; 2020.
16. Cui Y, Zheng W, Steinwandel M, et al. Associations of depressive symptoms with all-cause and cause-specific mortality by race in a population of low socioeconomic status: a report from the Southern Community Cohort Study. *Am J Epidemiol.* 2021;190(4):562–575.
17. Harshfield EL, Pennells L, Schwartz JE, et al. Association between depressive symptoms and incident cardiovascular diseases. *JAMA.* 2020;324(23):2396–2405.
18. Han FF, Wang HX, Wu JJ, et al. Depressive symptoms and cognitive impairment: a 10-year follow-up study from the Survey of Health, Ageing and Retirement in Europe. *Eur Psychiatry.* 2021;64(1):e55.
19. Lu W, Pai M, Scholes S, et al. Do depressive symptoms link chronic diseases to cognition among older adults? Evidence from the Health and Retirement Study in the United States. *J Affect Disord.* 2021;294:357–365.
20. Li Y, Wang X, Wang W, et al. 6-year trajectories of depressive symptoms and incident stroke in older adults: results from the Health and Retirement Study. *J Affect Disord.* 2022;309:229–235.
21. Sonnega A, Faul JD, Ofstedal MB, et al. Cohort profile: the Health and Retirement Study (HRS). *Int J Epidemiol.* 2014; 43(2):576–585.
22. Survey Research Center. *Documentation of Affective Functioning Measures in the Health and Retirement Study*. Ann Arbor, MI: Survey Research Center; 2000. https://hrs.isr.umich.edu/sites/default/files/biblio/dr-005.pdf. Accessed December 7, 2020.
23. Murchland AR, Eng CW, Casey JA, et al. Inequalities in elevated depressive symptoms in middle-aged and older adults by rural childhood residence: the important role of education. *Int J Geriatr Psychiatry.* 2019;34(11):1633–1641.
24. US Department of Agriculture, US Department of Health and Human Services. *Dietary Guidelines for Americans 2020-2025*. Washington, DC: U.S. Department of Agriculture; 2020. https://www.dietaryguidelines.gov/sites/default/files/2020-12/Dietary_Guidelines_for_Americans_2020-2025.pdf. Accessed July 17, 2022.
25. Survey Research Center. *Health and Retirement Study 2018 Tracker Final, Version 1.0, April 2022, Data Description and Usage*. Ann Arbor, MI: Survey Research Center; 2022. https://hrsdata.isr.umich.edu/sites/default/files/documentation/data-descriptions/trk2018v2a.pdf. Accessed June 1, 2022.
26. van Buuren S. *Flexible Imputation of Missing Data*. 2nd ed. 2019: Accessed May 19, 2020.
27. Nasinski M. miceFast: fast imputations using "Rcpp" and "armadillo." 2021. https://CRAN.R-project.org/package=miceFast. Accessed August 2, 2022.
28. Rubin DB. Multiple imputation after 18+ years. *J Am Stat Assoc.* 1996;91(434):473–489.
29. Bodner TE. What improves with increased missing data imputations? *Struct Equ Modeling.* 2008;15(4):651–675.
30. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med.* 2011;30(4):377–399.
31. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. 1st ed. Hoboken, NJ: John Wiley & Sons; 1987.
32. Gilman SE, Sucha E, Kingsbury M, et al. Depression and mortality in a longitudinal study: 1952–2011. *CMAJ.* 2017; 189(42):E1304–E1310.
33. Ferro MA. Missing data in longitudinal studies: cross-sectional multiple imputation provides similar estimates to full-information maximum likelihood. *Ann Epidemiol.* 2014;24(1):75–77.
34. van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, et al. Fully conditional specification in multivariate imputation. *J Stat Comput Simul.* 2006;76(12):1049–1064.
35. Wagstaff DA, Kranz S, Harel O. A preliminary study of active compared with passive imputation of missing body mass index values among non-Hispanic White youths. *Am J Clin Nutr.* 2009;89(4):1025–1030.
36. Austin PC, White IR, Lee DS, et al. Missing data in clinical research: a tutorial on multiple imputation. *Can J Cardiol.* 2021;37(9):1322–1331.
37. Marshall A, Altman DG, Royston P, et al. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC Med Res Methodol.* 2010;10(1):Article 7.
38. Kleinke K. Multiple imputation under violated distributional assumptions: a systematic evaluation of the assumed robustness of predictive mean matching. *J Educ Behav Stat.* 2017;42(4):371–404.
39. de Silva AP, Moreno-Betancur M, de Livera AM, et al. A comparison of multiple imputation methods for handling missing values in longitudinal data in the presence of a time-varying covariate with a non-linear association with time: a simulation study. *BMC Med Res Methodol.* 2017; 17(1):Article 114.
40. Kim S, Sugar CA, Belin TR. Evaluating model based imputation methods for missing covariates in regression models with interactions. *Stat Med.* 2015;34(11):1876–1888.