Journal of Modern Applied Statistical Methods

Volume 19 | Issue 1 Article 1

6-8-2021

How to Apply Multiple Imputation in Propensity Score Matching with Partially Observed Confounders: A Simulation Study and Practical Recommendations

Albee Ling
Stanford University, yling@stanford.edu

Maria Montez-Rath
Stanford University

Maya Mathur
Stanford University

Kris Kapphahn
Stanford University

Manisha Desai
Stanford University

Recommended Citation

Ling, A., Montez-Rath, M., Mathur, M., Kapphahn, K., & Desai, M. (2020). How to Apply Multiple Imputation in Propensity Score Matching with Partially Observed Confounders: A Simulation Study and Practical Recommendations. Journal of Modern Applied Statistical Methods, 19(1), eP3439. https://doi.org/10.22237/jmasm/1608552120

How to Apply Multiple Imputation in Propensity Score Matching with Partially Observed Confounders: A Simulation Study and Practical Recommendations

Cover Page Footnote

This work was supported by a Sanofi iDEA Award. We are particularly grateful for the excellent insights and guidance on this research provided by the Sanofi research team members: Robert LoCasale, Karen Chandross, Liz Zhou and Cliona Molony. We also want to thank Joseph Rigdon, Ariadna Garcia, and Christopher Gardner for providing us the DIETFITS trial data as our real data analysis example.

Journal of Modern Applied Statistical Methods
May 2020, Vol. 19, No. 1, eP3439
doi: 10.22237/jmasm/1608552120

בס"ד

ISSN 1538 − 9472

How to Apply Multiple Imputation in Propensity Score Matching with Partially Observed Confounders: A Simulation Study and Practical Recommendations

Albee Ling
Stanford University
Stanford, CA

Maria Montez-Rath
Stanford University
Stanford, CA

Maya Mathur
Stanford University
Stanford, CA

Kris Kapphahn
Stanford University
Stanford, CA

Manisha Desai
Stanford University
Stanford, CA

Propensity score matching (PSM) has been widely used to mitigate confounding in observational studies, although complications arise when the covariates used to estimate the PS are only partially observed. Multiple imputation (MI) is a potential solution for handling missing covariates in the estimation of the PS. However, it is not clear how to best apply MI strategies in the context of PSM. We conducted a simulation study to compare the performances of popular non-MI missing data methods and various MI-based strategies under different missing data mechanisms. We found that commonly applied missing data methods resulted in biased and inefficient estimates, and we observed large variation in performance across MI-based strategies. Based on our findings, we recommend 1) estimating the PS after applying MI to impute missing confounders; 2) conducting PSM within each imputed dataset followed by averaging the treatment effects to arrive at one summarized finding; 3) a bootstrapped-based variance to account for uncertainty of PS estimation, matching, and imputation; and 4) inclusion of key auxiliary variables in the imputation model.

Keywords: propensity score matching, missing covariates, multiple imputation, confounders, observational studies, causal inference

doi: 10.22237/jmasm/1608552120 | Accepted: Oct. 14, 2020; Published: June 8, 2021.
Correspondence: Albee Ling, yling@stanford.edu

Introduction

Randomized clinical trials serve as the gold standard for providing strong evidence for the effects of new and existing treatments for disease and disease prevention (Concato, Shah, & Horwitz, 2000). For numerous reasons including ethical and financial costs, however, such trials are not always feasible to conduct. Alternatively, observational studies have a long history of providing evidence for comparative effectiveness of treatments and interventions, and can also serve as justification for conducting a definitive randomized clinical trial (Fleurence, Naci, & Jansen, 2010; Holcomb et al., 2013; Lauer & Collins, 2010). The presence of confounding, however, can threaten the ability of an observational study to draw causal inference (Cochran & Rubin, 1973). Methods based on the propensity score (PS), defined as the conditional probability of being assigned a particular treatment given the subject's observed baseline covariates, can be used to mitigate such issues (D'Agostino, 1998; Goodman, Schneeweiss, & Baiocchi, 2017; Rosenbaum, 1987; Rosenbaum & Rubin, 1983). While the true PS is not typically known, it can be estimated using a variety of techniques (Austin, 2011a). Traditionally, PS-based methods include matching, inverse probability of treatment weighting (IPTW), stratification or subclassification, and covariate adjustment (Austin, 2011a; D'Agostino, 1998; Lunceford & Davidian, 2004; Rosenbaum & Rubin, 1984). Matching approaches are typically used to estimate the average treatment effect in the treated (ATT), whereas weighting, stratification, and covariate adjustment are more commonly used to estimate the population average treatment effect (ATE). More recently, several new PS-based approaches have emerged (F. Li, Morgan, & Zaslavsky, 2018; L. Li & Greene, 2013). Although they often produce results similar to those of regression adjustment (Vable et al., 2019), PS-based methods have several notable advantages. They allow separation of study design and analysis, enable assessment of overlap in covariates and their balance after adjustment, and are especially useful when the outcomes are rare and treatment is common (Austin, 2011a). PS matching (PSM) – where individuals with comparable PSs and discordant exposures are matched to achieve balance in covariates across the comparator groups of interest – is one of the more common tools used among the PS-based techniques and, thus the primary focus of the study presented here.

Under the potential outcomes framework (Imbens, 2004; Rubin, 1974), PSM produces unbiased estimates of the ATT with the assumptions of strongly ignorable treatment assignment (SITA) and Stable Unit Treatment Value Assumption (SUTVA) (Rosenbaum & Rubin, 1983). SITA requires 1) the exposure to be independent of potential outcomes given a set of covariates (unconfoundedness) 2)

the probability of receiving each treatment conditional on any set of covariates to be strictly between zero and one (positivity) (Rubin, 1986). SUTVA states that the outcome of a subject is not affected by the treatment assignment of other subjects (Rubin, 1986). Once these assumptions are met, researchers need to make a series of decisions involving matching methods including but not limited to: greedy or optimal matching, matching with or without replacement, one-to-one or many-to-one matching, and the use of calipers (Stuart, 2010). For example, introducing a caliper to calibrate the required distance between matched observations can aid the quality of matches by discarding units outside the area of common support, further reducing the bias; however, caliper matching can also induce potential bias and reduce efficiency as a result of incomplete matching (Crump, Hotz, Imbens, & Mitnik, 2009; Rosenbaum & Rubin, 1985a, 1985b; Stuart, 2010). In this paper, we focus on 1:1 nearest neighbor matching, a commonly used greedy matching algorithm, without replacement with caliper.

Once balance of covariates has been achieved in the matched samples, an analysis can be conducted to estimate the treatment effect and its variance. In contrast to a simple comparison between the treatment groups within the matched samples, a regression-based treatment effect estimator removes residual imbalance in covariates between treatment groups by adjusting for confounders in the model after matching (Ho, Imai, King, & Stuart, 2007; Schafer & Kang, 2008; Stuart, 2010; Wan, 2019). The variance estimation of the treatment effect in the context of PSM is not straightforward and remains controversial despite the large body of literature devoting attention to this issue (Abadie & Imbens, 2006, 2008, 2016; Abadie & Spiess, 2016; Austin, 2008; Austin & Small, 2014; Hill & Reiter, 2006; Ho et al., 2007; Lechner, 2002; Schafer & Kang, 2008; Stuart, 2008, 2010). In addition to the uncertainty in the treatment effect estimation, researchers disagree on how to account for uncertainty in the PS estimation (Abadie & Imbens, 2016; Stuart, 2010) or in the matching process (Abadie & Spiess, 2016; Hill & Reiter, 2006; Lechner, 2002), if at all. Based on the current literature, we considered two variance estimators as relevant choices: a robust cluster variance estimator (Abadie & Spiess, 2016) to account for the clustering induced by matched observations as well as a bootstrapped-based estimator (Austin & Small, 2014; Efron & Tibshirani, 1994) as it takes into account uncertainties in both the PS estimation and the matching process.

The statistical validity of PSM is threatened in the presence of missing data (D'Agostino Jr, 2004; D'Agostino Jr & Rubin, 2000; Ibrahim, Lipsitz, & Chen, 1999; Rosenbaum & Rubin, 1984). For example, if systematic missingness exists among measured confounders, the estimated ATT may be biased. The most

common approaches to handling partially observed confounders in PSM include complete-case analyses (CC), complete-variable analysis (CVA), and single imputation methods (Choi, Dekkers, & le Cessie, 2019). In the former, subjects missing at least one confounder are excluded from the analysis (Malla et al., 2018; White & Carlin, 2010). Importantly, CC produces unbiased estimates when data are missing completely at random (MCAR), i.e. missingness is not related to either observed or unobserved data. In contrast to CC, CVA is a different method that involves excluding confounders with missingness from the analysis. Single imputation methods have also been applied in this context, although less frequently than CC and CVA (Choi et al., 2019). Multiple imputation (MI) is a reasonably flexible method for handling missing data with good statistical properties that leads to unbiased and efficient estimators of parameters of interest when the data are missing at random (MAR), i.e., when the missingness is related to observed data only and specifically not unobserved data conditional on the observed (Little & Rubin, 2014). MI may also be applicable when data are missing not at random (MNAR), i.e., when missingness is related to unobserved variables, although researchers need to explicitly model the missing data mechanisms under MNAR (Collins, Schafer, & Kam, 2001). The implementation of MI even in the simplest of contexts and particularly in the context of PSM, however, requires that the user makes numerous decisions which can greatly impact the results (Van Buuren, 2018). Among the two modelling approaches of MI, our study focuses on fully conditional specification instead of joint modeling for its flexibility to accommodate multiple data types and its increase in application (Azur, Stuart, Frangakis, & Leaf, 2011).

As alluded to above, MI presents unique issues in the context of PSM. To incorporate the PS when using MI, one has to (1) estimate the PS and (2) integrate the PS into the analysis to obtain the treatment effect. There are multiple options for applying MI in the estimation step. Specifically, it is not clear whether one should impute the confounders first and then estimate the PS, referred to as a *passive* approach (Van Buuren, 2018), or whether one should impute the PS as if it were any other variable, referred to as an *active* approach (Von Hippel, 2009). The question of imputing in the presence of derived variables is not new and has been discussed in previous contexts, including for imputing interaction terms and higher-order terms (Desai, Mitani, Bryson, & Robinson, 2016; Mitani, Kurian, Das, & Desai, 2015; S. R. Seaman, Bartlett, & White, 2012; Von Hippel, 2009; White, Royston, & Wood, 2011). However, the approach utilized in the context of PSM has been limited (B. B. L. P. de Vries & Groenwold, 2017; Granger, Sergeant, & Lunt, 2019; Hill, 2004; Mitra & Reiter, 2016). *Active* approaches have been

promoted as bias-reducing because all variables and their interrelationships are considered in the imputation process, reflecting principles behind a proper and congenial imputation approach (Meng, 1994; Rubin, 2004; Rubin & Thomas, 1996; Van Buuren, 2018). In contrast, *passive* approaches have been supported because they result in internally consistent imputations (where the PS for subjects will perfectly correspond to its estimation as a function of their underlying confounders). Regarding the integration of PS, one can apply PSM within each imputed dataset and then arrive at an overall treatment effect estimate by averaging the effects obtained across imputed datasets (known as *within* integration). Alternatively, one can average the PSs across the imputed datasets to obtain one PS before estimating treatment effect from PSM (known as *across* integration) (B. B. L. P. de Vries & Groenwold, 2017; Granger et al., 2019; Hill, 2004; Leyrat et al., 2019; Mitra & Reiter, 2016).

We are not the first to consider MI methods when using PSM for causal inference (B. B. L. P. de Vries & Groenwold, 2017; Granger et al., 2019; Hill, 2004; Mitra & Reiter, 2016). However, significant gaps in methods remain, as work to date has been limited and has consisted of only one form of *passive* imputation (where confounders are first imputed without consideration of the PS, which is subsequently estimated) along with *within* and *across* integration strategies (B. B. L. P. de Vries & Groenwold, 2017; Granger et al., 2019; Hill, 2004; Mitra & Reiter, 2016). We build upon this excellent body of literature by evaluating *active* imputations and variations of *passive* imputations that allow the consideration of auxiliary terms in the imputation model. Further, there is no consensus on how to best estimate the uncertainty of the treatment effect within this framework. This paper presents a novel simulation study to comprehensively evaluate MI imputation and integration approaches in the context of PSM for the purpose of causal inference. We detail gaps in the current literature that examined MI for PSM, describe our methods for conducting a simulation study, present our findings, and discuss interpretation of our findings that inform best statistical practice in the final section.

Background

MI is a simulation-based statistical tool to handle missing data, which involves three main steps. In Step 1, multiple sets of plausible values of the missing variables are generated based on the posterior predictive distribution of observed variables to reflect the uncertainties of the imputation process. In Step 2, analyses are performed within each imputed dataset, before their results are combined with the application

of Rubin's Rules in Step 3 (Carpenter & Kenward, 2012). It has been well established in the MI literature that the outcome should always be included in the imputation process when regression parameters are of interest (B. B. L. de Vries & Groenwold, 2016; Little & Rubin, 2014; Moons, Donders, Stijnen, & Harrell Jr, 2006; Sterne et al., 2009; Van Buuren, 2018). In the context of PSM, the various strategies we consider involve Steps 1 and 3, are described below and summarized in the Glossary.

With respect to Step 1, there are two broad categories of MI strategies that have been introduced in the literature for derived variables or variables that are functions of other variables: active (*MI-active*) and passive (*MI-passive*) (Figure 1a). Such derived variables include interaction terms, higher order terms, ratios of two variables (e.g. body mass index), and rates of change (Desai et al., 2016; Mitani et al., 2015; S. R. Seaman et al., 2012; Von Hippel, 2009; White et al., 2011). In *MI-active*, the derived variable is imputed as if it were any other variable (Von Hippel, 2009). The simplest, regular form of *MI-active*, *MI-regActive*, involves calculating the derived variable in complete cases and imputing it together with all other missing variables in the imputation process, with no consideration of its known relationship to the variables involved in its derivation. *MI-regActive* is a proper imputation method, as all the relationships specified in the scientific model are included in the imputation models, i.e. the imputation model is congenial with the scientific model (Meng, 1994; Rubin, 2004; Rubin & Thomas, 1996; Van Buuren, 2018). Although *MI-active* is advantageous given its consideration of the entire covariance structure, some argue that it undermines the imputation process by creating internally inconsistent values. This motivated a re-derived version of *MI-active* where the derived variable is recalculated post-imputation (*MI-redActive*) (Von Hippel, 2009).

In contrast to *MI-active* approaches, *MI-passive* approaches maintain the internal consistency between variables used in the derivation and the derived variable itself (Van Buuren, 2018). In this case, the derived term is not to be imputed but derived after imputing the variables involved in the term's construction. The simplest form of *MI-passive* is *MI-derPassive*, where all variables involved in the derivation are imputed prior to deriving the term from the imputed data (Von Hippel, 2009). However, because the derived variable is not included in the imputation process, *MI-derPassive* may introduce bias. Another form of *MI-passive*, *MI-regPassive*, was developed to partially address this issue by including the derived variable in the imputation process of those variables that are not involved in its derivation (Royston, 2009). The latter includes auxiliary variables, which can enhance the imputation process but do not provide any useful

information for the scientific model (Collins et al., 2001). Examples include variables associated with the pattern of missingness or the missing variable itself (Collins et al., 2001). Previous work in MI for PSM has been limited to *MI-derPassive*. Neither *MI-regPassive*, which involves an auxiliary variable, fully or partially observed, nor any active approaches (*MI-regActive* and *MI-redActive*) have been considered previously for handling missingness in PSM.

PS estimates need to be integrated in the analysis to estimate the treatment effect. There has been considerable work in examining integration methods for *MI-derPassive*. Specifically, the PS can be estimated and incorporated within each imputed dataset (*INT-within*) prior to obtaining the treatment effect through summarization in Step 3, or the PS can be averaged across the imputed datasets after completing Step 1 and applied to the original dataset to obtain the treatment effect (*INT-across*) (B. B. L. P. de Vries & Groenwold, 2017; Granger et al., 2019; Hill, 2004; Mitra & Reiter, 2016). An additional variation on the latter has been previously applied in the context of IPTW (*INT-across2*) and involves averaging both the estimated regression coefficients corresponding to the covariates used to estimate the PS model and the covariates values themselves to arrive at one PS that can be applied to obtain the treatment effect (Leyrat et al., 2019). The rationale is that the PS coefficients are more suitable for combination using Rubin's Rules given their distributional properties than the PSs themselves, which are confined to be between 0 and 1 (Figure 1b). We will comprehensively evaluate the different combinations of MI imputation and integration strategies described.

How to best estimate the variance of the treatment effect in the context of PSM when applying MI is an open research topic (B. B. L. P. de Vries & Groenwold, 2017; Hill, 2004; Mitra & Reiter, 2016). In addition to the complications in variance estimation in PSM mentioned above in the absence of missing data, the uncertainty introduced by the MI process needs to be considered. The application of Rubin's Rules in *INT-within* accomplishes this goal, but it is unclear how to capture this uncertainty when applying *INT-across* and *INT-across2* (B. B. L. P. de Vries & Groenwold, 2017; Hill, 2004; Mitra & Reiter, 2016). Bootstrap methods have been proposed in the context of MI (Brand, van Buuren, le Cessie, & van den Hout, 2019; Schomaker & Heumann, 2018) and specifically with respect to PS-based methods (B. B. L. P. de Vries & Groenwold, 2017; Qu & Lipkovich, 2009). For example, Austin & Small evaluated two potential estimators for PSM in the absence of missing data, where the variance was obtained by either resampling matched pairs or the original observations (Austin & Small, 2014).

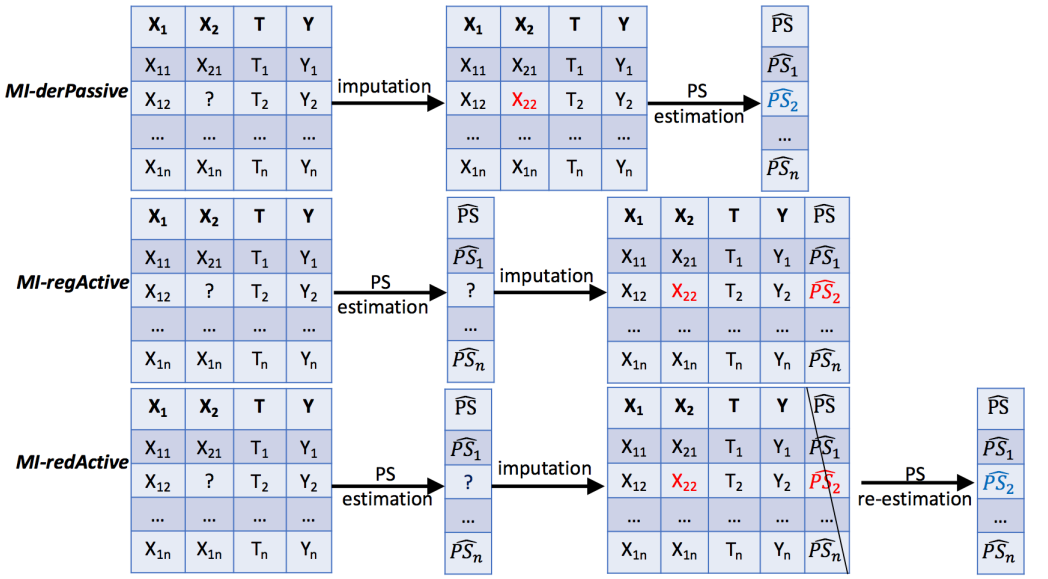

Figure 1a

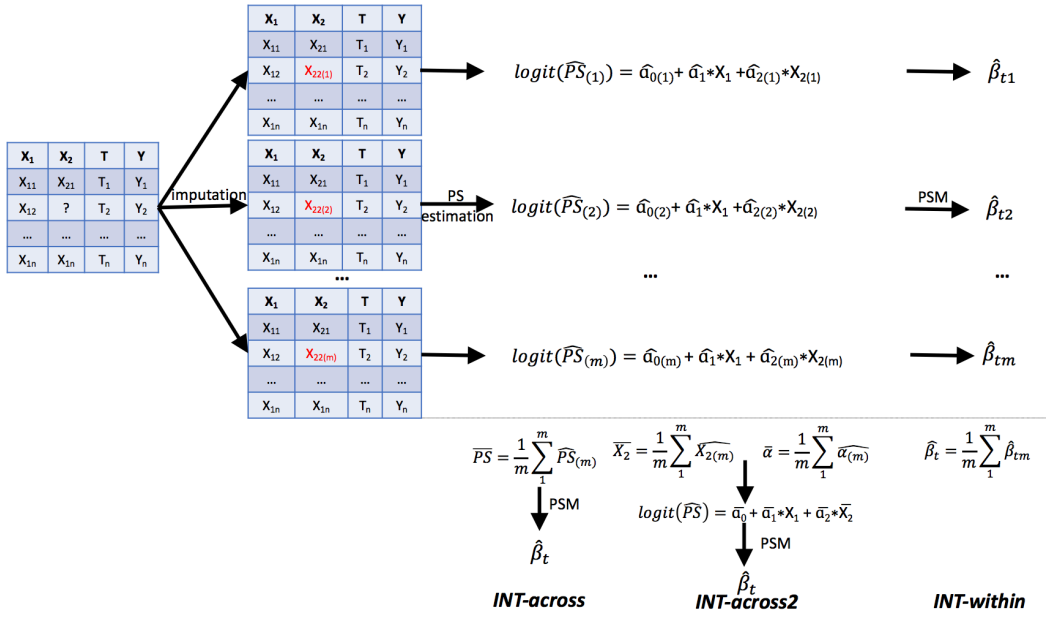

Figure 1b

Figure 1. *Figure 1a*: Illustration of three imputation strategies demonstrated with one imputed dataset; *Figure 1b*: Illustration of three strategies to integrate the propensity score for estimation of the treatment effect.

Although the former performed well relative to the empirical variance, the latter was comparable and enabled extension to the MI context. Schomaker & Heumann evaluated four bootstrapped-based approaches in the context of MI when PSM was not considered (Schomaker & Heumann, 2018). One of these approaches, Boot MI, is applicable to the PSM context and overlaps with the ideas described by Austin & Small. We therefore compare two competing variance estimators in this study to better inform those applying MI in the PSM setting: a bootstrapped-based variance estimator and a robust cluster variance estimator (with Rubin's Rules when applying *INT-within*) to account for various sources of variation when MI is applied in the context of PSM.

Simulation Study Design

We conducted an extensive simulation study to assess the performance of various MI-based strategies and commonly applied missing data methods when estimating ATT using PSM. In all scenarios, we included two binary confounders (X_1 and X_2) of the relationship between treatment and outcome, a binary variable representing the treatment or exposure of interest (T), and a continuous outcome (Y). Two auxiliary variables (Z_2 and Z_{ps}) were generated to aid the imputation process. Missing values were present in X_2 whereas X_1, T and Y were always fully observed. For each scenario, 1,000 simulated datasets were generated, each consisting of n = 2000 subjects. All data analyses were conducted in R version 3.5.1 (R Core Team, 2018). MI and PSM were implemented using the `mice` and `Matching` packages respectively (Sekhon, 2008; Van Buuren & Groothuis-Oudshoorn, 2010). The R code to replicate this study is publicly available in a Github repository at https://github.com/yling2019/psm_mi. Below we provide details on the data generation, missing data mechanisms, missing data methods considered, and metrics for performance evaluation.

Data generation

To motivate our simulation study, we want to see if treatment variable T, an electronic text message intervention, has any effect on the outcome, treatment adherence, where variables X_1 and X_2 are confounders such as sex and race. More details of the simulated variables can be found in the rest of the section, whereas more details of the motivating study can be found in a later section, Case Study.

Confounders. Two binary variables $X = (X_1, X_2)$ that confound the relationship between treatment and outcome were generated, by first creating two variables from a bivariate normal distribution (correlation of 0.5) each with mean 0 and variance 1, which were then dichotomized at the mean. The resulting distributions of X_1, X_2 are binomial distributions ($p \approx 0.5$) with correlation around 0.33.

Treatment indicator. A binary treatment variable T was generated from a binomial distribution such that:

$$logit\,(p\,(T = 1 \mid X_1, X_2)) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2$$

where $\alpha_1 = \alpha_2 = 2$ so both covariates contributed equally to the treatment assignment. The intercept of the treatment α_0 was selected such that roughly 30% of subjects were treated, to reflect real-world datasets where there are often many more control subjects than treated.

Outcome. A continuous outcome variable Y was generated as a linear function of the treatment and both covariates.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_t T + \varepsilon$$

where $\beta_1 = \beta_2 = 2$ so both covariates were equally and positively associated with outcome and $\varepsilon \sim N(0,10^2)$. The intercept β_0 was set to zero and the true treatment effect β_t was set to 2.

Auxiliary variables. Auxiliary variables Z_2 and Z_{ps} were generated to be highly correlated with X_2 and the estimated PS score respectively (based on full observed data without missing data, with correlation of 0.98). More specifically, setting $\delta_{02} = 1, \delta_{0ps} = 0$, $\delta_{12} = \delta_{1ps} = 10$, the auxiliary variables were generated as:

$$Z_2 = \delta_{02} + \delta_{12} X_2 \; + \varepsilon_2; \;\; Z_{ps} \; = \delta_{0ps} + \delta_{1p} \hat{PS} + \varepsilon_{ps}$$

where $\varepsilon_2 \sim N(0,1^2)$ and $\varepsilon_{ps} \sim N(0,1^2)$ denote errors generated from standard normal distributions independent from all other variables in the data generating models.

Missing data mechanisms (MDMs)

Missingness was always induced in X_2, whereas X_1 was fully observed. Missingness occurred under five different mechanisms: MCAR, MAR1, MAR2A, missing not at random, and an MAR2B scenario. Whereas MAR1 represented a simple MAR scenario, MAR2A and MAR2B were more sinister scenarios that captured the complexity of MDM in real-world datasets. In MAR1, MAR2A, and MNAR, missingness was related to treatment and outcome. In MAR2A, missingness was also related to Z_2, the auxiliary variable associated with X_2. In MNAR, missingness was also related to X_2. Let R_2 be an indicator variable denoting whether X_2 is missing ($R_2 = 1$) or not ($R_2 = 0$). We set Y_b and Z_{2b} to be dichotomizations at the median of the outcome variable Y and auxiliary variable Z_2 respectively. Under each MDM, the intercept γ_0 was selected such that 50% of the observations were missing. Let $\gamma_{11} = 5$, and $\gamma_{00} = 1$, and I be an indicator variable. Missingness in X_2 was induced as follows:

$$\text{MAR1: } logit(R_2 = 1 \mid complete\ data) = \gamma_0 + \gamma_{11} I(t = 1, y_b = 1) + \gamma_{00} I(t = 0, y_b = 0)$$

$$\begin{aligned}\text{MAR2A: } logit(R_2 = 1 \mid complete\ data) &= \gamma_0 + \gamma_{11} I(t = 1, y_b = 1, Z_{2b} = 1) \\ &\quad + \gamma_{00} I(t = 0, y_b = 0, Z_{2b} = 0)\end{aligned}$$

$$\begin{aligned}\text{MNAR: } logit(R_2 = 1 \mid complete\ data) &= \gamma_0 + \gamma_{11}(t = 1, y_b = 1, X_2 = 1) \\ &\quad + \gamma_{00}(t = 0, y_b = 0, X_2 = 0)\end{aligned}$$

To study the impact of having a partially observed auxiliary variable, we also induced missingness in X_2 according to a second MAR2 missing mechanism, MAR2B, based on treatment, outcome, and PS. Letting Z_{psb} be the dichotomizations of Z_{ps}, missingness in X_2 was induced as follow:

$$\begin{aligned}\text{MAR2B: } logit(R_2 = 1 \mid complete\ data) &= \gamma_0 + \gamma_{11} I(t = 1, y_b = 1, Z_{psb} = 1) \\ &\quad + \gamma_{00} I(t = 0, y_b = 0, Z_{psb} = 0)\end{aligned}$$

Additionally, we induced missingness in the auxiliary variable, Z_{ps}, under three scenarios that assumed MAR2B for X_2: aux_MCAR, aux_MAR1, and aux_MAR2. In both aux_MAR1 and aux_MAR2, missingness was related to T and PS_b, where PS_b is the dichotomization at the median of the PS estimated using full data prior to inducing missingness. Let R_z be an indicator variable denoting whether Z_{ps} is missing ($R_z = 1$) or not ($R_z = 0$). The intercept term ε_0 was selected to ensure 20% missingness in Z_{ps} and missingness can be expressed as:

$$logit(R_z = 1 \mid complete\ data) = \varepsilon_0 + \varepsilon_{11}I(t = 1, PS_b = 1) + \varepsilon_{10}I(t = 1, PS_b = 0) + \varepsilon_{01}I(t = 0, PS_b = 1) + \varepsilon_{00}I(t = 0, PS_b = 0)$$

where

$$\varepsilon_{11} = 5,\ \varepsilon_{10} = 0,\ \varepsilon_{01} = 0,\ \varepsilon_{00} = 5 \text{ in aux\_MAR1}$$

and

$$\varepsilon_{11} = 0,\ \varepsilon_{10} = 5,\ \varepsilon_{01} = 5,\ \varepsilon_{00} = 0 \text{ in aux\_MAR2.}$$

Missing data methods

Common missing data methods. We applied various missing data methods that are widely used in the medical research literature including CC, CVA, mean imputation, and the use of missing data indicators.

Multiple imputation strategies. Figure 1 displays the MI strategies considered for PS estimation and integration. *MI-derPassive*, *MI-regActive*, and *MI-redActive* were applied under MCAR, MAR1, MAR2A, and MNAR conditions with or without auxiliary variable Z_2 in the imputation model, where Z_2 had no missing values (Appendix Table A1). Note that since missingness in MAR2A is associated with an auxiliary variable, when the auxiliary variable was not included in the imputation, the MI imputation model is misspecified and the MDM becomes an MNAR scenario. Under the MAR2B MDM, we included an additional partially observed auxiliary variable Z_{ps} in the imputation model when *MI-regPassive* and *MI-derPassive* were applied. Under this scenario we also examined performance by order of inclusion of the variables in the imputation model (i.e., whether X_2 was imputed before Z_{ps} or not). Integration approaches considered were *INT-within*, *INT-across*, and *INT-across2*. Note that *INT-across2* cannot be combined with *MI-regActive*, as the PS is directly imputed. In `MICE`, 50 multiply imputed datasets ($m = 50$) (White et al., 2011), five iterations (*maxit* = 5) and default settings for the imputation method (predictive mean matching, or PMM, for continuous variable and logistic regression for binary) were used. The treatment and outcome were included in all imputation models (Van Buuren, 2018).

PSM and treatment effect estimation

We estimated coefficients α_1 and α_2 using a correctly specified logistic regression model, $logit(p(T = 1 \mid X)) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2$. PS scores were estimated as the fitted values of the regression model on the response scale. One-to-one nearest neighbor matching without replacement was applied. Subjects were matched by PS scores with calipers of width that is 0.2 of the standard deviation of the logit of PS (Austin, 2011b; Rosenbaum & Rubin, 1985a). After matching subjects, the treatment effect or ATT was estimated using standard linear regression methods (Abadie & Spiess, 2016), by regressing Y on T and confounders X_1 and X_2 to obtain the estimate for the beta coefficient representing T. Adjustment not done in *INT-across* strategies because of the presence of multiple sets of X_1 and X_2.

Variance estimation

In the absence of missing data, we used two approaches to estimate the uncertainty of the treatment effect: (1) a robust cluster variance estimate (McCullagh, 2018) that accounts for the matched design and (2) a bootstrapped variance calculated as the standard deviation of treatment effects in 1,000 bootstrapped samples to account for both PS estimation and the matching process. For the latter, the detailed procedure is described as follows:

1. Sample with replacement $n = 2000$ rows from the observed dataset $D = (X,T,Y,Z,R)$ to obtain a bootstrapped dataset D_{boot} which contains missing values;
2. Impute m datasets for D_{boot} using the imputation strategy (*MI-derPassive*, *MI-regPassive*, *MI-regActive*, or *MI-redActive*), for $k = 1, 2, ..., m$, denoted as $D_{boot}(k)$;
3. Apply the integration approach (*INT-within*, *INT-across* or *INT-across2*) to obtain a single effect estimate for D_{boot};
4. Repeat steps 1-3 B times to obtain B bootstrap replicates from which treatment effect β_{tboot} can be estimated for a given bootstrap sample D_{boot};
5. Calculate bootstrapped standard error as the standard deviation of B treatment effects estimated from each bootstrap sample: $SE_{bootstrap} = sd(\beta_{tboot})$ for $b = 1, 2, ..., B$

When commonly applied missing data methods were considered, the robust cluster variance estimator was used. When MI was applied in the context of PSM, we compared 1) the robust cluster variance estimator and 2) a bootstrapped variance. For the former, when the integration strategy was *INT-within*, a robust cluster variance was estimated within each of the m imputed datasets, before application of Rubin's Rules to yield one final variance. For both *INT-across* and *INT-across2*, Rubin's Rules do not apply; instead we obtain only one robust cluster variance.

Sensitivity simulation study

To test the robustness of our main simulation study, we conducted a second simulation study by reducing coefficients in the treatment and outcome generating models while keeping all other aspects of the simulations the same. To be specific, we set $\alpha_1 = \alpha_2 = \log(2)$ and $\beta_1 = \beta_2 = 1$ to mimic more realistic data examples encountered by applied researchers when data is missing MCAR, MAR1, MAR2A, and MNAR. All statistical analyses were performed in the exact same way as the main simulation study as described above. In order to test the sensitivity of the optimal MI strategy found in the main simulation study, a third set of simulations was conducted by varying the study population size (n = 1000, 500, 250), missing rate (25%, 10%), or the number of multiply imputed datasets (m = 10).

Performance metrics

After PSM, we examined the percentage of treated subjects matched and the standardized differences of covariates. For mean imputation, standardized differences were calculated in the original full data and the imputed data. For missing indicator variables, standardized differences were calculated in the full data without missingness, as well as its observed and missing part. For *INT-within*, standardized differences were calculated in 1) each of the imputed datasets, and 2) the full dataset, before averaging over all multiply imputed datasets. For *INT-across* and *INT-across2*, standardized differences were calculated in 1) the average of m imputed dataset and its observed and imputed parts respectively 2) the full dataset (Leyrat et al., 2019; Moons et al., 2006). For each missing data method, we report on bias, variance, mean squared error (MSE), relative MSE (relative to PSM in the full dataset), and coverage probability summarized over 1,000 simulations per scenario for estimating treatment effect β_t. The robust cluster variance and bootstrapped variance were compared to their corresponding empirical variance for each MI strategy. Coverage was estimated as the proportion of 1,000 simulations

such that the interval $[\hat{\beta}_t - 1.96 \times SE, \hat{\beta}_t + 1.96 \times SE]$ contained the true treatment effect of $\beta_t = 2$ (*SE*: robust cluster standard error or bootstrapped standard error). We used the normal theory estimator because the percentile based method did not perform well in simulations by Austin & Small (Austin & Small, 2014), and calculating accelerated and bias-corrected confidence intervals (BCa) (Efron & Tibshirani, 1994) proved too computationally intensive. Monte Carlo standard errors were calculated for bias, empirical standard error, MSE, and coverage (Morris, White, & Crowther, 2019). Reference metrics for missing data methods were based on applying PSM to the full data (PSM_full).

Results

We first quantified the confounding effects introduced by our data generation by regressing outcome on treatment only in the full data set without missingness (data not shown in tables). A large bias was present (12.23), indicating a strong confounding effect. Next. we compared the resulting bias and standard error from two methods that adjusted for confounding: (1) fitting the true data generating model or regression adjustment, where both confounders were included as covariates in the regression model and (2) applying PSM to the full data (PSM_full). Both methods yielded unbiased treatment effect estimates (bias = −0.006 in both cases). PSM yielded a higher standard error as expected due to discarding unmatched samples (0.313 using regression in the full dataset and 0.380 using PSM_full). Coverage reached the nominal level of 95% using both methods. These results matched well with their corresponding empirical standard error (0.306 and 0.376 respectively). In PSM, the robust cluster standard error and bootstrapped estimators were comparable (0.380 in both cases) and close to the empirical (0.376). MSE in PSM_full was 0.141, which was used as the denominator for calculating all rMSEs later.

Commonly applied missing data methods

Of the commonly applied approaches, CC had the most favorable MSE relative to that the of PSM_full (rMSE ranged 1.857 to 48.658 in various MDMs, Appendix Table A2). CC produced biased treatment estimates (bias = −2.489, −0.815, and −1.084 in MAR1, MAR2A, and MNAR respectively) and less efficient estimates relative to PSM_full (robust standard error = 0.537, 0.838, 0.682, 0.682 in MCAR, MAR1, MAR2A and MNAR respectively vs 0.380 in PSM_full). CVA, mean imputation, and the use of missing indicators yielded greater bias relative to CC

(5.058 to 5.059 for CVA; 2.985 to 5.759 for mean imputation; and 2.973 to 5.534 for missing indicator), although their robust cluster variances were smaller than that of CC. Comparisons of statistical properties obtained when not adjusting for X_1 and X_2 were similar.

Variance estimation in MI-based strategies

While the robust cluster variance estimator and the bootstrapped-based variance estimator were comparable in the absence of missing data, differences were observed in the presence of missingness and when MI was applied. Specifically, the robust cluster variance estimator consistently underestimated the empirical variance in *INT-across* integration strategy. Among the *INT-across* approaches, the variance ratios (robust/empirical) were much smaller than 1 and only exceeded 0.8 when an auxiliary variable was included (Figure 2). The worst performance for the robust estimator was observed in *MI-regActive INT-across* approaches, where the variance ratios can be lower than 0.1. In contrast, the variance was consistently overestimated in *INT-within* approaches, where the variance ratio surpassed 5 under *MI-regActive* and *MI-redActive* approaches, especially when an auxiliary term was used. The ratio of the robust estimator for the variance relative to the empirical under *INT-across2* was close to 1 across all MI methods and MDMs. On the other hand, the bootstrapped-based variance was more comparable to the empirical variance across all MI integration strategies; the ratio of bootstrapped variance to the empirical ranged from 0.675 to 1.875, with mean 1.01. We did not observe any trend specific to imputation methods, integration methods, or the inclusion of auxiliary variable in the imputation model. All subsequent results were therefore calculated using the bootstrapped-based variance estimator.

Comparing various MI strategies

For simplicity, performance of MI strategies under MAR1 is highlighted here (when auxiliary variable was not included in the imputation model). For the majority of the MI strategies, balance was achieved such that the absolute standardized difference in X_1 and X_2 between treated and controls based on the imputed dataset was below 0.1 with the exception of *MI-regActive* (Appendix Table A3). When considering both bias and efficiency, *MI-derPassive* approaches achieved the lowest rMSE, followed by *MI-redActive* and *MI-regActive* (Table 1 and Figure 3). Among the three integration strategies under *MI-derPassive* approaches, *INT-within*, *INT-across*, *INT-across2* were ranked from the lowest to

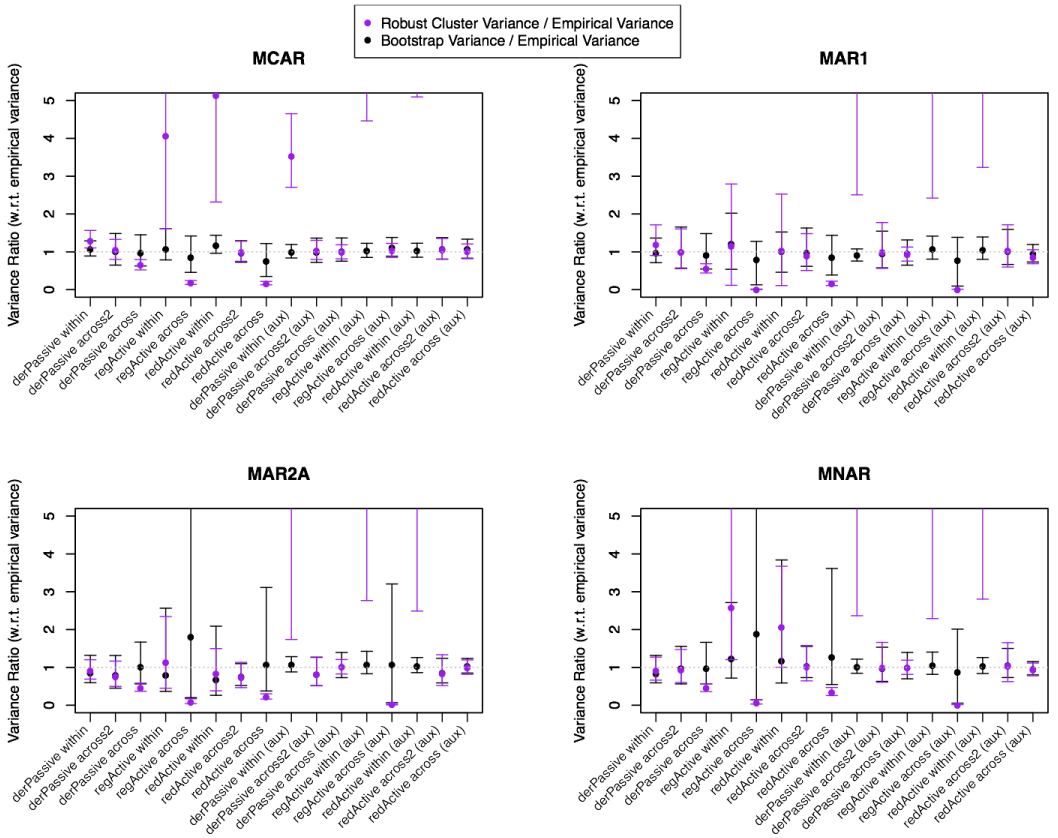

Figure 2. Distribution (mean, 2.5th and 97.5th quantile across 1,000 simulations) of the ratio of robust cluster and bootstrap variances with respect to empirical variance by multiple imputation estimation and integration strategies for propensity score matching. Confidence intervals are trimmed at the value of 5.

the highest with respect to rMSE, which was largely driven by the bias (Table 1). Similar trend in MI integration strategies was observed for *MI-regActive*, where *INT-within* also outperformed *INT-across*. In contrast, the performance of various integration strategies varied under *MI-redActive*, where *MI-redActive INT-within* was the worst performer (rMSE = 60.454, bias = 2.557) and *MI-redActive INT-across* was the best performer (rMSE = 8.784, bias = 0.465) (Table 1).

The impact of auxiliary terms

Under MAR1, when a fully observed auxiliary term, Z_2, was included in the imputation model, the statistical properties of most MI strategies were comparable

Table 1. Main simulation results: bias, standard error, mean squared error (MSE), relative mean squared error (rMSE)*, and coverage results of various multiple imputation strategies under MAR1 (Monte Carlo standard errors in parentheses).

MI Strategy			Standard Error				
Imputation	Integration	Bias	Empirical	Bootstrap	MSE	rMSE*	Coverage
Auxiliary variable not included in imputation model							
MI-derPassive	INT-within	0.038 (0.014)	0.454 (0.01)	0.446 (0.001)	0.207 (0.011)	1.466	1 (0)
	INT-across	−0.176 (0.016)	0.497 (0.011)	0.472 (0.002)	0.278 (0.012)	1.969	1 (0)
	INT-across2	−2.859 (0.024)	0.770 (0.017)	0.757 (0.003)	8.764 (0.14)	62.083	0 (0)
MI-regActive	INT-within	2.471 (0.043)	1.345 (0.03)	1.463 (0.007)	7.913 (0.174)	56.055	0.823 (0.012)
	INT-across	5.275 (0.173)	5.473 (0.122)	4.719 (0.033)	57.746 (2.605)	409.068	0.915 (0.009)
MI-redActive	INT-within	2.557 (0.045)	1.414 (0.032)	1.405 (0.006)	8.534 (0.183)	60.454	0.719 (0.014)
	INT-across	0.465 (0.032)	1.013 (0.023)	0.926 (0.005)	1.240 (0.076)	8.784	1 (0)
	INT-across2	−2.633 (0.025)	0.792 (0.018)	0.772 (0.003)	7.558 (0.132)	53.54	0 (0)
Auxiliary variable included in imputation model							
MI-derPassive	INT-within	0.319 (0.012)	0.395 (0.009)	0.376 (0.001)	0.258 (0.01)	1.828	1 (0)
	INT-across	0.067 (0.013)	0.398 (0.009)	0.385 (0.001)	0.163 (0.008)	1.155	1 (0)
	INT-across2	−2.904 (0.024)	0.771 (0.017)	0.744 (0.003)	9.028 (0.141)	63.954	0 (0)
MI-regActive	INT-within	0.502 (0.014)	0.449 (0.01)	0.465 (0.001)	0.453 (0.022)	3.209	1 (0)
	INT-across	4.096 (0.176)	5.572 (0.125)	4.702 (0.040)	47.793 (2.478)	338.561	0.948 (0.007)
MI-redActive	INT-within	0.461 (0.013)	0.409 (0.009)	0.421 (0.001)	0.380 (0.016)	2.692	1 (0)
	INT-across	0.059 (0.013)	0.418 (0.009)	0.406 (0.001)	0.178 (0.008)	1.261	1 (0)
	INT-across2	−2.860 (0.024)	0.770 (0.017)	0.768 (0.003)	8.773 (0.14)	62.147	0 (0)

** rMSE: the ratio of MSE calculated from each missing data method over MSE obtained from propensity score matched results in the absence of missing data.*

to those when auxiliary variable was not in the imputation model (Table 1 and Figure 3). Large improvement in performance was observed for *MI-regActive INT-within*, *MI-redActive INT-within*, and *MI-redActive INT-across* (Table 1). Under MAR2A, we observed an improvement of performance across a wider range of MI-strategies: *MI-derPassive*/*MI-redActive INT-within*/*INT-across* and *MI-regActive INT-within* (Appendix Table A4 and Figure 3). For example, the rMSE for *MI-derPassive INT-within* in the absence and presence of auxiliary variable is 17.341 and 1.672 respectively (Appendix Table A4). The auxiliary term improved efficiency for most MI strategies except for *MI-regActive INT-across*. The auxiliary term improved the absolute bias for most MI strategies except for *MI-derPassive INT-across2* and *MI-redActive INT-across2*. MNAR results were similar to MAR2A, and inclusion of the auxiliary variable was required to obtain nominal level of coverage probability in both MAR2A and MNAR (Appendix Table A4 and Appendix Figure A3).

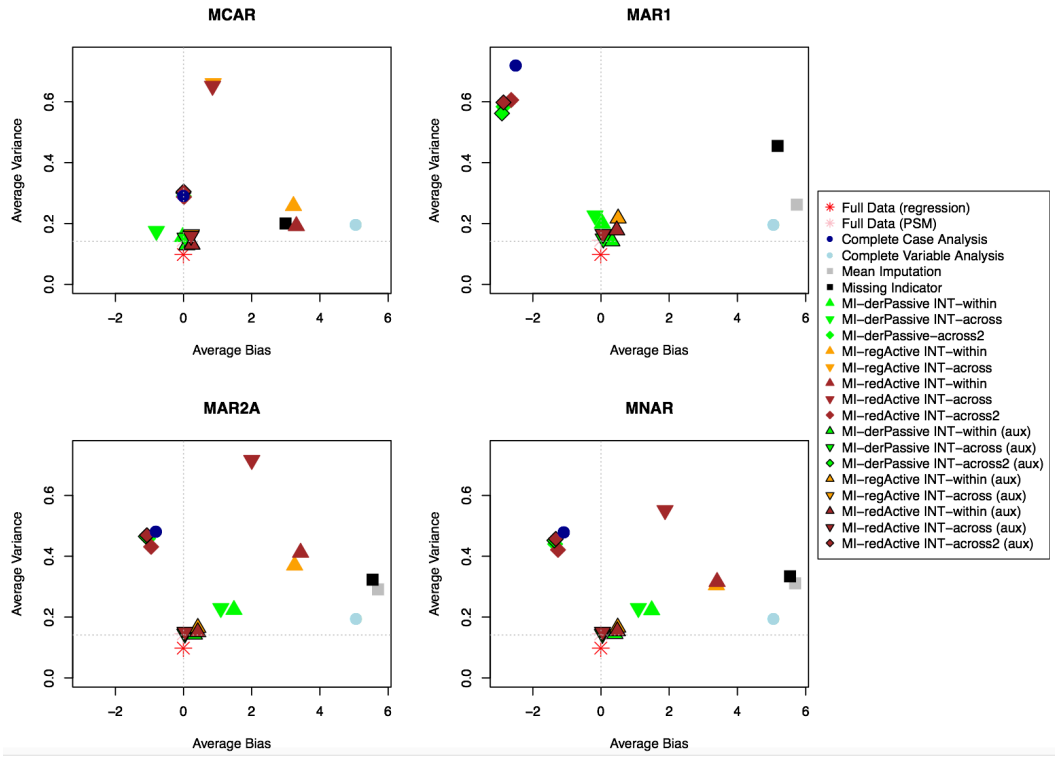

Figure 3. Average bootstrap variance vs. average bias computed over 1,000 simulated datasets, by multiple imputation estimation and integration strategies for propensity score matching where X_2 was missing under MCAR, MAR1, MAR2A, and MNAR. MCAR, missing completely at random; MAR, missing at random; MNAR, missing not at random; MAR1, simple MAR; MAR2A, complex MAR.

The performance of MI strategies is shown in Appendix Table A5 and Appendix Figure A1 under a modified MAR2 scenario (MAR2B) where missingness was a function of a different auxiliary variable, Z_{ps} (dichotomized), treatment, and outcome. For reference, performance of *passive* approaches when Z_{ps} was fully observed was first evaluated. *MI-derPassive INT-within* achieved the lowest rMSE and bias (rMSE = 1.032, bias = −0.019), followed by *INT-across* (rMSE = 1.394, bias = −0.205), and *INT-across2* (rMSE = 10.319, bias = −1.016). The bootstrapped standard error was the largest in *INT-across2* (0.682) and comparable for *INT-within* (0.378) and *INT-across* (0.381). When Z_{ps} was partially observed, *MI-derPassive* and *MI-regPassive* were largely comparable. Although *INT-across* methods yielded the smallest bootstrap standard error, *INT-within* methods resulted in smaller bias and MSE. While imputing X_2 before or after Z_{ps}

affected the resulting bias, bootstrap standard error and MSE, the order did not change our conclusions on the best performing imputation and integration MI strategies as mentioned above.

Comparison across MDMs

Summarizing results across all MDMs, *MI-derPassive* outperformed *MI-redActive*, followed by *MI-regActive* in terms of rMSE, bias, and efficiency (Table 1, Appendix Table A4, Figure 3, and Appendix Figure A3). *MI-derPassive* INT-within demonstrated strong performance regardless of the presence of an auxiliary term. Even though *MI-derPassive INT-across* achieved the smallest rMSE in MAR2A and MNAR when auxiliary variable was included, *MI-derPassive INT-within* had comparable rMSE. Inclusion of an auxiliary variable did not greatly improve properties of the top performers under MCAR and MAR1. In contrast, the performance of MI-strategies was much improved under MAR2A and MNAR, where missingness was related to the auxiliary variable.

Sensitivity simulation results

The results from the second set of simulations, where there was a reduced confounding effect — shown in Appendix Tables A8-A10 — are largely consistent with the main simulation results in sections above. Next, applying the optimal MI strategy from the main simulation study, *MI-derPassive INT-within*, we found that both bias and efficiency suffered with higher missing rate and inadequately specified number of multiply imputed datasets, *m* (Appendix Table A11). When the missing rate decreased from 50% in the main simulation study to 25% and 10% under MAR1, the bias and bootstrap standard error also decreased (bias = 0.038, 0.021, and 0.005, bootstrap standard error = 0.446, 0.389, and 0.365 respectively). When only 10 multiply imputed datasets were used at 50% missingness, both bias (0.096) and bootstrap standard error (0.454) increased compared to when $m = 50$ (bias = 0.038, bootstrap standard error = 0.446 under MAR1) (Appendix Table A11). In the absence of missingness, when the study sample size was 1000, 500, and 250, the bias (−0.186, 0.300, -0.315) and robust standard error (0.524, 0.715, 1.073) increased compared to when the sample size was 2,000 in the main simulation study (bias = −0.006, standard error = 0.380) (data not shown in tables). Nevertheless, *MI-derPassive INT-within* performed reasonably well in terms of both bias and efficiency. For example, under MAR1, the biases were 0.11, 0.22, 0.374 and bootstrap standard errors were 0.63, 0.878, 1.138 with sample sizes 1000, 500, and 250 respectively (Appendix Table A11).

Case Study

The various MI strategies discussed in our simulation study were illustrated in a real-world example. The goal of the exemplified study was to assess whether an electronic text message (e-message) intervention had an impact on treatment adherence, measured by proportion of class attendance, in the Diet Intervention Examining The Factors Interacting with Treatment Success (DIETFITS) trial (Oppezzo et al., 2019). DIETFITS investigated the effects of a healthy low-fat diet vs a healthy low-carbohydrate diet on weight change at 12 months in 609 overweight adults (Gardner et al., 2018). Throughout the study period, participants in both diet groups were given education classes designed to enhance participant adherence. The investigators found a decline in the education class attendance in the first four study cohorts after 6 months (Oppezzo et al., 2019). Thus, an e-message intervention was deployed to both arms of the fifth cohort with the goal of increasing adherence to their diet plan (Oppezzo et al., 2019). In the original study, PSM was performed to match the patients in Cohort 5 who received e-message intervention to historical controls in Cohort 1-4 and the effect was measured through a two-sampled t-test (Oppezzo et al., 2019).

Instead of replicating the original study, the statistical analysis was modified to better match our simulation design. A PS model was estimated in all 609 participants, among whom 97 received e-messages, and 512 did not (31 from Cohort 5 declined and 481 from Cohorts 1-4 did not have the chance to receive this e-messaging intervention). PS was estimated in the same way as the original study -- using logistic regression with confounding variables age, sex, race, weight change at 6 months, and proportion of attendance at 6 months (Oppezzo et al., 2019). 1:1 nearest neighbor matching was used without replacement with caliper 0.2 to match each participant who received e-messages to a control. The effect of e-messages on outcome variable, proportion of class attendance between the 6- and 12-month study endpoints, was estimated in a linear regression model adjusting for all confounding variables. To handle missing data, commonly used methods (CC, CVA, mean imputation, and missing indicator) were applied as well as MI strategies. A robust cluster variance was used for non-MI missing data methods and bootstrap variance was used for all MI-strategies.

Trial participants' characteristics are shown in Appendix Table A6. There were three PS variables with missing data: age (5.15% and 24.80% in exposure and control groups respectively), race (1.03% and 0.98% in exposure and control groups respectively) and weight change at 6 months (8.25% and 26.37% in exposure and control groups respectively). Overall, 20% patient-level data was

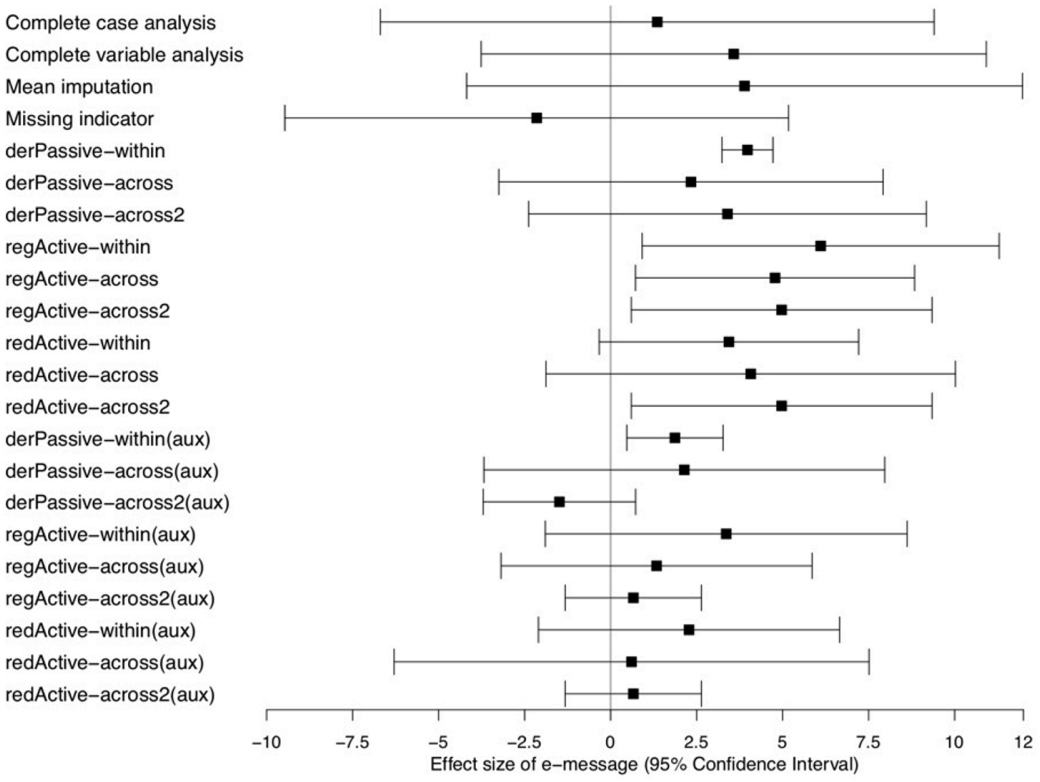

Figure 4. Application of the various missing data methods to a real-world example to assess the impact of an electronic text message (e-message) intervention on treatment adherence in the Diet Intervention Examining The Factors Interacting with Treatment Success (DIETFITS) trial, showing effect sizes of e-message with 95% confidence interval estimated using propensity score matching.

missing and $m = 20$ was used in MI algorithms. Additionally, the following auxiliary variables: baseline weight, weight at 6 months, weight at 12 months, and weight change after 6 months, were included into our imputation model, as they were associated with PS variables but not of interest in our scientific model. As observed in our simulation study, *MI-derPassive* and *MI-redActive* coupled with *INT-within* balanced more covariates among all missing data methods (Appendix Table A7). Figure 4 shows the estimated effects of e-message and their 95% confidence intervals. The effect size of e-message intervention varied greatly from one missing data method to another, as observed in our simulations. Informed by the simulation study, findings based on *MI-derPassive INT-within* (both with and

without auxiliary variable), indicated that e-messaging had a significant impact on adherence.

Discussion

We investigated several pragmatic research questions concerning how to optimally apply MI when utilizing PSM in the presence of a partially observed confounder. We compared the performance of non-MI missing data methods that are commonly applied along with various MI-based strategies that vary both in how the PS is estimated or imputed and in how the PS is integrated into the analysis. In addition, we evaluated the impact of inclusion of an auxiliary term in the imputation model on the ranked performance of the MI strategies as well as the impact of the order of inclusion when there is more than one variable with missing data. Among the commonly applied missing data methods, CVA and single imputation methods (mean imputation and missing indicator) led to large bias in our simulation study. In contrast, CC was not as biased due to the use of a caliper that ensured only those subjects with closely matched PSs were included. CC did, however, suffer from loss of efficiency. There was large heterogeneity among the MI strategies considered. Based on our results, we caution applied researchers against adopting the aforementioned commonly applied missing data methods and recommend: 1) adopt *MI-derPassive* approaches; and 2) consider *INT-within* 3) use of the bootstrap to estimate variance; and 4) inclusion of key auxiliary variables in the imputation model.

Our study is important in identifying the limitations of commonly applied methods. Considerable bias and inefficiency were observed among all commonly applied methods relative to that yielded by applying PSM to the dataset without missingness. At least one of the MI-based strategies always outperformed the commonly applied methods. It is well established that balancing diagnostics are useful for PSM, but this proves difficult with commonly applied methods. For example, while CC can be applied using those matched pairs where balance is achieved, bias may still occur because of the observations not included due to missingness. For missing indicators, we were unable to achieve balance in the variable with missing data, as revealed in our simulations when we parsed the data by observed and missing data to evaluate balance in these respective parts. Thus, in practice, one may have a false sense of the balance as the user is only privy to assessing the balance in the observed data. Similarly, application of mean imputation distorts the distributional properties of the variable with missing data, potentially yielding a distorted view of balance when the imputed values are

utilized in calculating the standardized differences (Appendix Table A2). Finally, for CVA, a false sense of security may be given when evaluating balance in only one variable, when exclusion of the other variable could lead to bias.

We have primarily examined the differences between *passive* and *active* MI methods when PS, a derived variable, was considered in the analysis and only partially observed. *MI-derPassive* methods surpassed *MI-redActive* approaches in almost all performance metrics across all MDMs, and *MI-regActive* had the worst statistical properties. Since *MI-redActive* can be thought of as a hybrid between *MI-derPassive* and *MI-regActive* and partly mitigated the issues of *MI-regActive* by re-deriving PS post-imputation, we hereby mainly discuss the rationale behind the poor performance of *MI-regActive*. To start with, *MI-regActive* was proposed so the entire covariance structure of all variables in the analysis, including the PS itself, could be retained. However, the covariance structure between PS and the PS variables is complex and difficult to learn using complete cases only. Unlike usual derived measures (e.g. interactions and higher order terms) that are derived as a deterministic function of other variables, PS requires estimation and its exact function will vary depending on the data considered. Thus, imputing PS together with missing covariates introduced bias into the imputation procedure and consequently the estimated treatment effect. Such bias was also reflected in the difficulty to achieve balance both in the imputed and fully observed data under *MI-regActive* (Appendix Table A3). Further, the poor estimation of the treatment effect had implications for estimates of uncertainty. The bias introduced in the estimation of the treatment effect highly varied across the bootstrap samples, leading to an increased estimate of the variation (Appendix Figure A2). Although adding an auxiliary variable reduced bias in most of the bootstrap samples, it did not help reduce bias for the extreme cases where the bias without auxiliary variable was unexpectedly high. Second, PSM is a two-stage analysis, in which after matching, PS itself is not directly used in estimating the treatment effect. In contrast, other derived variables (e.g. interactions) are usually directly involved in the regression model for treatment effect estimation. As a result, capturing the covariance structure between PS and other variables in the imputation process did not yield the same benefits as seen in the derived variables that were studied before. Third, although we only simulated two PS variables, researchers are likely to include a larger number of PS covariates with complex missing pattern. This will lead to a high overall level of missingness for PS estimation in the first step of *MI-regActive*, which increases the difficulty of the subsequent imputation step. Fourth, unless we specify the true relationship between PS and PS variables using an inverse logistic function (PMM was used in our simulations), which is cumbersome to implement

in reality, the imputation model is technically misspecified, potentially leading to bias as well.

We recommend *INT-within* as the optimal MI integration strategy to combine with imputation strategy *MI-derPassive*, for its superior performance in all evaluating metrics as well as its ability to balance covariates post-matching. While Mitra & Reiter recommended *INT-across*, de Vries & Groenwold argued that such findings were due to a combination of omitting outcome in the imputation model and a violation of positivity assumption in the PSM process (B. B. L. de Vries & Groenwold, 2016; Mitra & Reiter, 2016). Further, they found INT-within to yield estimators with better statistical properties (B. B. L. P. de Vries & Groenwold, 2017). In a different setting (IPTW instead of PSM), Leyrat et. al. also demonstrated superior properties of *INT-within* methods over *INT-across* and *INT-across2* (Leyrat et al., 2019). Our results were consistent with de Vries & Groenwold (B. B. L. P. de Vries & Groenwold, 2017) and Leyrat and others (Leyrat et al., 2019). We also share the perspective of Leyrat and others (Leyrat et al., 2019) that it was more straightforward to assess balance in *INT-within* strategies and observed that the covariates were mostly balanced in both the imputed and full datasets (Appendix Table A3). Further, Leyrat and others pointed out that the *INT-across* and *INT-across2* produced consistent estimators only when both the observed and imputed data were balanced (Leyrat et al., 2019). We therefore paid close attention to balance diagnostics for *INT-across* and *INT-across2* methods but did not observe balance in both parts of the data (observed and missing) under all MDMs other than under MCAR.

An important contribution of our paper is resolution of how to estimate the variance when doing PSM and applying MI. There has been extensive but conflicting research on this topic in the context of MI and IPTW where Rubin's Rules have been recommended for *INT-within* by some authors (Leyrat et al., 2019; S. Seaman & White, 2014) and a bootstrapped-based estimator was recommended by others (Qu & Lipkovich, 2009). Relative to IPTW, however, we are faced with the additional issue of capturing the uncertainty of matching in PSM. Prior studies of MI applications in the context of PSM acknowledged this issue (Hill, 2004; Mitra & Reiter, 2016), but only one study has explicitly stated their recommendation of a bootstrapped-based variance (B. B. L. P. de Vries & Groenwold, 2017), although the choice was not studied comparatively or discussed fully. In our study, we found that application of Rubin's Rules when the robust cluster estimator was used for each imputed dataset overestimated the variance under *INT-within* approaches and underestimated it under *INT-across* approaches. We therefore agree with de Vries & Groenwold (B. B. L. P. de Vries & Groenwold, 2017) in recommending the

bootstrapped variance, as it captures the uncertainty of PS estimation, matching procedure, and imputation process. Further, it demonstrated good performance with respect to the empirical variance. We acknowledge the lack of theoretical support for this choice, which comes with challenges, as the estimator for the treatment effect based on PSM and MI is not a smooth function. Although Abadie & Imbens (Abadie & Imbens, 2008) proved that the bootstrap variance is not valid in matching *with* replacement, their results may not be applicable in our study when matching was done *without* replacement, where one control unit can only be used for matching at most once (Austin & Small, 2014). Other alternative non-parametric solutions with stronger theoretical justification, such as subsampling, has their own limitations (e.g. the need for a sufficiently large sample size and a burden on the user to appropriately select a sub-sample and replication size) (Politis & Romano, 1994).

Auxiliary variables are often useful for adhering to a MAR MDM, but not always possible in the context of PSM. As shown in our simulation study, when missingness is related to an auxiliary variable in MAR2A, inclusion of the auxiliary variable ensured a truly MAR scenario, which would be an MNAR scenario otherwise, making it difficult to obtain statistically valid results using MI. Having the auxiliary variable in the imputation model indeed improved all statistical properties of our recommended optimal MI strategy, *MI-derPassive INT-within*. In reality, however, variables related to partially observed confounders may be considered confounders themselves and thus, may not exist outside of estimation of the PS. Our team has worked on studies, however, where auxiliary terms may be available. For example, in a comparative effectiveness study of anticoagulants among kidney transplant patients, a PS that balances patient characteristics may include body mass index (BMI) at treatment initiation but not BMI at transplant listing. The latter is an excellent candidate for an auxiliary variable that can aid in imputing BMI at treatment initiation as well as other PS covariates. By including a strong auxiliary variable in the imputation process, we showcased the maximal performance improvement given any auxiliary variable. In practice, the strength of auxiliary variable varies and consequently the improvement in performance may be moderate.

There are several limitations to our study. As with any simulation study, we recognize that the limited scope of our simulations may compromise generalizability. Specifically, only two binary confounders were generated, which might not reflect a real-world scenario. We adopted a simple design, following the lead of others studying similar topics, so that we could hone in on the properties of the various MI methods without extra layers of complexities (Choi et al., 2019; B.

B. L. P. de Vries & Groenwold, 2017; Hill, 2004; Mitra & Reiter, 2016; S. Seaman & White, 2014). Even though we only included a continuous outcome in our simulation, we believe our findings on the MI imputation and integration strategies are not specific to the type of outcome. In fact, we found consistencies between our findings and those from simulation studies with binary outcomes studying a similar topic, IPTW (Leyrat et al., 2019; S. Seaman & White, 2014). Regardless, future studies using time-to-event outcomes are important, especially given the added complexity of applying MI with right-censored outcomes (Barzi & Woodward, 2004; Desai et al., 2019; Van Buuren, Boshuizen, & Knook, 1999; White & Royston, 2009). We also only considered the scenario when one confounder was partially observed, whereas missingness of covariates that are not confounders, treatment or outcome was not considered. However, we recommend the same MI strategy in cases of any missingness in potential confounders (covariates that are associated with the outcome but not exposure), as potential confounders should also be included in the PS model (Austin, Grootendorst, Normand, & Anderson, 2007; Brookhart et al., 2006). One specific caveat in our comparison among different MI integration strategies is that the two confounders were not included in the regression models in the post-matching outcome analysis for *INT-across*. This was infeasible under this strategy since we had multiple sets of confounders but only one set of PS. Nonetheless, the lack of inclusion likely led to some of the observed bias in the comparative performance of our MI methods. Finally, there were only main effects in the data generating mechanism and we did not explore the impact of mis-specifying the correct PS models.

Overall, we have addressed an important topic – how to apply MI strategies in the presence of missing values in confounders in the context of PSM. Our work will facilitate future applied researchers' choice of optimal missing data methods in all kinds of statistical analyses that involve PSM. In addition to classical causal inference settings, our results are applicable to other types of studies that utilize PS, including those that generalize randomized clinical trial findings to real-world target populations captured in observational databases (Cole & Stuart, 2010; Stuart, Cole, Bradshaw, & Leaf, 2011).

Glossary

Various multiple imputation (MI) strategies for the propensity score matching (PSM) context. PS, propensity score; MICE, multivariate imputation by chained equations in R. The recommended approach is marked by *.

Step 1. Choosing an imputation strategy (how to obtain PS through imputation)

1. *MI-passive* (PS is derived after missing PS model variables are imputed)
    a. * *MI-derPassive* (PS is derived after completion of MICE algorithm, within which missing PS model variables are imputed)
    b. *MI-regPassive* (PS is derived within MICE algorithm following the step where missing PS model variables are imputed)
2. MI-active (PS is imputed together with other missing PS model variables)
    a. *MI-regActive* (PS is imputed together with other missing PS model variables)
    b. *MI-redActive* (PS is imputed together with other missing PS model variables and redrived post imputation)

Step 2. Choosing an integration strategy (how to integrate PS in PSM)

1. * *INT-within* (PSM is conducted m times within each multiply imputed dataset and results are summarized using Rubin's Rules)
2. *INT-across* (PSM is conducted after averaging PS across the PSs obtained from the m imputed datasets)
3. *INT-across2* (PSM is conducted after calculating the PS from a model where coefficients for the PS model are averaged across the m imputed datasets)

Step 3. Choosing a variance estimator

1. Robust cluster variance estimator
2. * Bootstrap-based variance estimator

References

Abadie, A., & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, *74*(1), 235–267. https://doi.org/10.1111/j.1468-0262.2006.00655.x

Abadie, A., & Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica, 76*(6), 1537–1557. https://doi.org/10.3982/ecta6474

Abadie, A., & Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica, 84*(2), 781–807. https://doi.org/10.3982/ecta11293

Abadie, A., & Spiess, J. (2016). Robust post-matching inference. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.2020.1840383

Austin, P. C. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine, 27*(12), 2037–2049. https://doi.org/10.1002/sim.3150

Austin, P. C. (2011a). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46*(3), 399–424. https://doi.org/10.1080/00273171.2011.568786

Austin, P. C. (2011b). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics, 10*(2), 150–161. https://doi.org/10.1002/pst.433

Austin, P. C., Grootendorst, P., Normand, S.-L. T., & Anderson, G. M. (2007). Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. *Statistics in Medicine, 26*(4), 754–768. https://doi.org/10.1002/sim.2618

Austin, P. C., & Small, D. S. (2014). The use of bootstrapping when using propensity-score matching without replacement: A simulation study. *Statistics in Medicine, 33*(24), 4306–4319. https://doi.org/10.1002/sim.6276

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research, 20*(1), 40–49. https://doi.org/10.1002/mpr.329

Barzi, F., & Woodward, M. (2004). Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies. *American Journal of Epidemiology, 160*(1), 34–45. https://doi.org/10.1093/aje/kwh175

Brand, J., van Buuren, S., le Cessie, S., & van den Hout, W. (2019). Combining multiple imputation and bootstrap in the analysis of cost-effectiveness trial data. *Statistics in Medicine, 38*(2), 210–220. https://doi.org/10.1002/sim.7956

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology, 163*(12), 1149–1156. https://doi.org/10.1093/aje/kwj149

Carpenter, J., & Kenward, M. (2012). *Multiple imputation and its application*. John Wiley & Sons. https://doi.org/10.1002/9781119942283

Choi, J., Dekkers, O. M., & le Cessie, S. (2019). A comparison of different methods to handle missing data in the context of propensity score analysis. *European Journal of Epidemiology, 34*(1), 23–36. https://doi.org/10.1007/s10654-018-0447-z

Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhya: The Indian Journal of Statistics, Series A, 35*, 417–446.

Cole, S. R., & Stuart, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *American Journal of Epidemiology, 172*(1), 107–115. https://doi.org/10.1093/aje/kwq084

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*(4), 330. https://doi.org/10.1037/1082-989x.6.4.330

Concato, J., Shah, N., & Horwitz, R. I. (2000). Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine, 342*(25), 1887–1892. https://doi.org/10.1056/nejm200006223422507

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika, 96*(1), 187–199. https://doi.org/10.1093/biomet/asn055

D'Agostino Jr, R. B. (2004). Propensity score estimation with missing data. In: A. Gelman & X.-L. Meng, Eds. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family* (Wiley Series in Probability and Statistics). John Wiley & Sons., pp. 163–174. https://doi.org/10.1002/0470090456.ch15

D'Agostino Jr, R. B., & Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association, 95*(451), 749–759. https://doi.org/10.1080/01621459.2000.10474263

D'Agostino, R. B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine, 17*(19), 2265–2281. https://doi.org/10.1002/(sici)1097-0258(19981015)17:19<2265::aid-sim918>3.0.co;2-b

de Vries, B. B. L., & Groenwold, R. H. H. (2016). Comments on propensity score matching following multiple imputation. *Statistical Methods in Medical Research, 25*(6), 3066-3068. https://doi.org/10.1177/0962280216674296

de Vries, B. B. L. P., & Groenwold, R. H. H. (2017). A comparison of two approaches to implementing propensity score methods following multiple imputation. *Epidemiology, Biostatistics and Public Health, 14*(4), e12630-1-e12630-21.

Desai, M., Mitani, A. A., Bryson, S. W., & Robinson, T. (2016). Multiple Imputation When Rate of Change Is The Outcome of Interest. *Journal of Modern Applied Statistical Methods, 15*(1), 160-192. https://doi.org/10.22237/jmasm/1462075740

Desai, M., Montez-Rath, M. E., Kapphahn, K., Joyce, V. R., Mathur, M. B., Garcia, A., … Owens, D. K. (2019). Missing data strategies for time-varying confounders in comparative effectiveness studies of non-missing time-varying exposures and right-censored outcomes. *Statistics in Medicine, 38*(17), 3204-3220. https://doi.org/10.1002/sim.8174

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC Press. https://doi.org/10.1201/9780429246593

Fleurence, R. L., Naci, H., & Jansen, J. P. (2010). The critical role of observational evidence in comparative effectiveness research. *Health Affairs, 29*(10), 1826–1833. https://doi.org/10.1377/hlthaff.2010.0630

Gardner, C. D., Trepanowski, J. F., Del Gobbo, L. C., Hauser, M. E., Rigdon, J., Ioannidis, J. P. A., … King, A. C. (2018). Effect of low-fat vs low-carbohydrate diet on 12-month weight loss in overweight adults and the association with genotype pattern or insulin secretion: the DIETFITS randomized clinical trial. *JAMA, 319*(7), 667–679. https://doi.org/10.1001/jama.2018.0245

Goodman, S. N., Schneeweiss, S., & Baiocchi, M. (2017). Using design thinking to differentiate useful from misleading evidence in observational research. *JAMA, 317*(7), 705–707. https://doi.org/10.1001/jama.2016.19970

Granger, E., Sergeant, J. C., & Lunt, M. (2019). Avoiding pitfalls when combining multiple imputation and propensity scores. *Statistics in Medicine, 38*(26), 5120-5132. https://doi.org/10.1002/sim.8355

Hill, J. (2004). *Reducing Bias in Treatment Effect Estimation in Observational Studies Suffering from Missing Data*. Columbia University Institute for Social Social and Economic Research and Policy (ISERP) Working Paper 04-01, (January). https://doi.org/10.7916/D8B85G11

Hill, J., & Reiter, J. P. (2006). Interval estimation for treatment effects using propensity score matching. *Statistics in Medicine, 25*(13), 2230–2256. https://doi.org/10.1002/sim.2277

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15*(3), 199–236. https://doi.org/10.1093/pan/mpl013

Holcomb, J. B., Del Junco, D. J., Fox, E. E., Wade, C. E., Cohen, M. J., Schreiber, M. A., … Others. (2013). The prospective, observational, multicenter, major trauma transfusion (PROMMTT) study: comparative effectiveness of a time-varying treatment with competing risks. *JAMA Surgery, 148*(2), 127–136. https://doi.org/10.1001/2013.jamasurg.387

Ibrahim, J. G., Lipsitz, S. R., & Chen, M.-H. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61*(1), 173–190. https://doi.org/10.1111/1467-9868.00170

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics, 86*(1), 4–29. https://doi.org/10.1162/003465304323023651

Lauer, M. S., & Collins, F. S. (2010). Using science to improve the nation's health system: NIH's commitment to comparative effectiveness research. *JAMA, 303*(21), 2182–2183. https://doi.org/10.1001/jama.2010.726

Lechner, M. (2002). Some practical issues in the evaluation of heterogeneous labour market programmes by matching methods. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 165*(1), 59–82. https://doi.org/10.1111/1467-985x.0asp2

Leyrat, C., Seaman, S. R., White, I. R., Douglas, I., Smeeth, L., Kim, J., … Williamson, E. J. (2019). Propensity score analysis with partially observed covariates: How should multiple imputation be used? *Statistical Methods in Medical Research, 28*(1), 3–19. https://doi.org/10.1177/0962280217713032

Li, F., Morgan, K. L., & Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association, 113*(521), 390–400. https://doi.org/10.1080/01621459.2016.1260466

Li, L., & Greene, T. (2013). A weighting analogue to pair matching in propensity score analysis. *The International Journal of Biostatistics, 9*(2), 215–234. https://doi.org/10.1515/ijb-2012-0030

Little, R. J. A., & Rubin, D. B. (2014). *Statistical analysis with missing data* (Vol. 333). John Wiley & Sons. https://doi.org/10.1002/9781119013563

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine, 23*(19), 2937–2960. https://doi.org/10.1002/sim.1903

Malla, L., Perera-Salazar, R., McFadden, E., Ogero, M., Stepniewska, K., & English, M. (2018). Handling missing data in propensity score estimation in comparative effectiveness evaluations: a systematic review. *Journal of Comparative Effectiveness Research, 7*(3), 271–279. https://doi.org/10.2217/cer-2017-0071

McCullagh, P. (2018). *Generalized linear models*. Routledge. https://doi.org/10.1201/9780203753736

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science, 9*(4), 538–558. https://doi.org/10.1214/ss/1177010269

Mitani, A. A., Kurian, A. W., Das, A. K., & Desai, M. (2015). Navigating choices when applying multiple imputation in the presence of multi-level categorical interaction effects. *Statistical Methodology, 27*, 82–99. https://doi.org/10.1016/j.stamet.2015.06.001

Mitra, R., & Reiter, J. P. (2016). A comparison of two methods of estimating propensity scores after multiple imputation. *Statistical Methods in Medical Research, 25*(1), 188–204. https://doi.org/10.1177/0962280212445945

Moons, K. G. M., Donders, R. A. R. T., Stijnen, T., & Harrell Jr, F. E. (2006). Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology, 59*(10), 1092–1101. https://doi.org/10.1016/j.jclinepi.2006.01.009

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine, 38*(11), 2074-2102. https://doi.org/10.1002/sim.8086

Oppezzo, M. A., Stanton, M. V, Garcia, A., Rigdon, J., Berman, J. R., & Gardner, C. D. (2019). To Text or Not to Text: Electronic Message Intervention to Improve Treatment Adherence Versus Matched Historical Controls. *JMIR MHealth and UHealth, 7*(4), e11720. https://doi.org/10.2196/11720

Politis, D. N., & Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics, 22*(4), 2031-2050. https://doi.org/10.1214/aos/1176325770

Qu, Y., & Lipkovich, I. (2009). Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. *Statistics in Medicine, 28*(9), 1402–1414. https://doi.org/10.1002/sim.3549

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association, 82*(398), 387–394. https://doi.org/10.1080/01621459.1987.10478441

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*(387), 516–524. https://doi.org/10.1080/01621459.1984.10478078

Rosenbaum, P. R., & Rubin, D. B. (1985a). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*(1), 33–38. https://doi.org/10.1080/00031305.1985.10479383

Rosenbaum, P. R., & Rubin, D. B. (1985b). The bias due to incomplete matching. *Biometrics, 41*(1), 103–116. https://doi.org/10.2307/2530647

Royston, P. (2009). Multiple imputation of missing values: further update of ice, with an emphasis on categorical variables. *The Stata Journal, 9*(3), 466–477. https://doi.org/10.1177/1536867x0900900308

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688. https://doi.org/10.1037/h0037350

Rubin, D. B. (1986). Comment: Which ifs have causal answers. *Journal of the American Statistical Association, 81*(396), 961–962. https://doi.org/10.2307/2289065

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys* (Vol. 81). John Wiley & Sons. https://doi.org/10.1002/9780470316696

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: relating theory to practice. *Biometrics, 52*(1), 249–264. https://doi.org/10.2307/2533160

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological Methods, 13*(4), 279-313. https://doi.org/10.1037/a0014268

Schomaker, M., & Heumann, C. (2018). Bootstrap inference when using multiple imputation. *Statistics in Medicine, 37*(14), 2252–2266. https://doi.org/10.1002/sim.7654

Seaman, S. R., Bartlett, J. W., & White, I. R. (2012). Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical

methods. *BMC Medical Research Methodology, 12*(1), 46. https://doi.org/10.1186/1471-2288-12-46

Seaman, S., & White, I. (2014). Inverse probability weighting with missing predictors of treatment assignment or missingness. *Communications in Statistics-Theory and Methods, 43*(16), 3499–3515. https://doi.org/10.1080/03610926.2012.700371

Sekhon, J. S. (2008). Multivariate and propensity score matching software with automated balance optimization: the matching package for R. *Journal of Statistical Software, 42*(7). https://doi.org/10.18637/jss.v042.i07

Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., … Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ, 338*(jun29 1), b2393. https://doi.org/10.1136/bmj.b2393

Stuart, E. A. (2008). Developing practical recommendations for the use of propensity scores: Discussion of 'A critical appraisal of propensity score matching in the medical literature between 1996 and 2003' by Peter Austin. *Statistics in Medicine. Statistics in Medicine, 27*(12), 2062–2065. https://doi.org/10.1002/sim.3207

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. Statistical Science: A Review. *Journal of the Institute of Mathematical Statistics, 25*(1), 1. https://doi.org/10.1214/09-sts313

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 174*(2), 369–386. https://doi.org/10.1111/j.1467-985x.2010.00673.x

Vable, A. M., Kiang, M. V, Glymour, M. M., Rigdon, J., Drabo, E. F., & Basu, S. (2019). Performance of matching methods as compared with unmatched ordinary least squares regression under constant effects. *American Journal of Epidemiology, 188*(7), 1345–1354. https://doi.org/10.1093/aje/kwz093

van Buuren, S, & Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1–68. https://doi.org/10.18637/jss.v045.i03

Van Buuren, Stef. (2018). *Flexible imputation of missing data*. Chapman and Hall/CRC. https://doi.org/10.1201/b11826

Van Buuren, Stef, Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine, 18*(6), 681–694. https://doi.org/10.1002/(sici)1097-0258(19990330)18:6<681::aid-sim71>3.0.co;2-r

Von Hippel, P. T. (2009). 8. How to impute interactions, squares, and other transformed variables. *Sociological Methodology, 39*(1), 265–291. https://doi.org/10.1111/j.1467-9531.2009.01215.x

Wan, F. (2019). Matched or unmatched analyses with propensity-score--matched data? *Statistics in Medicine, 38*(2), 289–300. https://doi.org/10.1002/sim.7976

White, I. R., & Carlin, J. B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine, 29*(28), 2920–2931. https://doi.org/10.1002/sim.3944

White, I. R., & Royston, P. (2009). Imputing missing covariate values for the Cox model. *Statistics in Medicine, 28*(15), 1982–1998. https://doi.org/10.1002/sim.3618

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine, 30*(4), 377–399. https://doi.org/10.1002/sim.4067

Appendix A

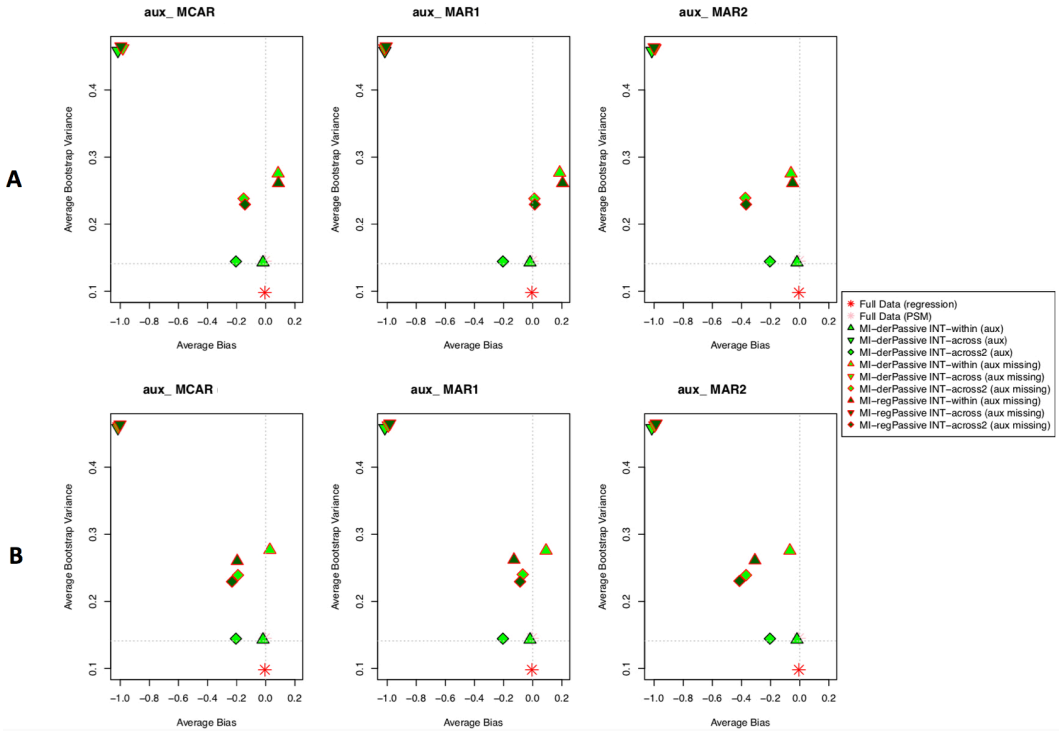

Figure A1. Average bootstrap variance vs. average bias computed over 1,000 simulated datasets, by *passive* multiple imputation estimation and integration strategies for propensity score matching where X_2 was missing under a complex MAR (MAR2B) involving auxiliary variable Z_{ps} which was missing under simple MCAR (aux_MCAR), simple MAR (aux_MAR1), or complex MAR (aux_MAR2). A. Default imputation: X_2 imputed before imputation of Z_{ps}; B. Reverse imputation: Z_{ps} imputed before imputation of X_2. MCAR, missing completely at random; MAR, missing at random.

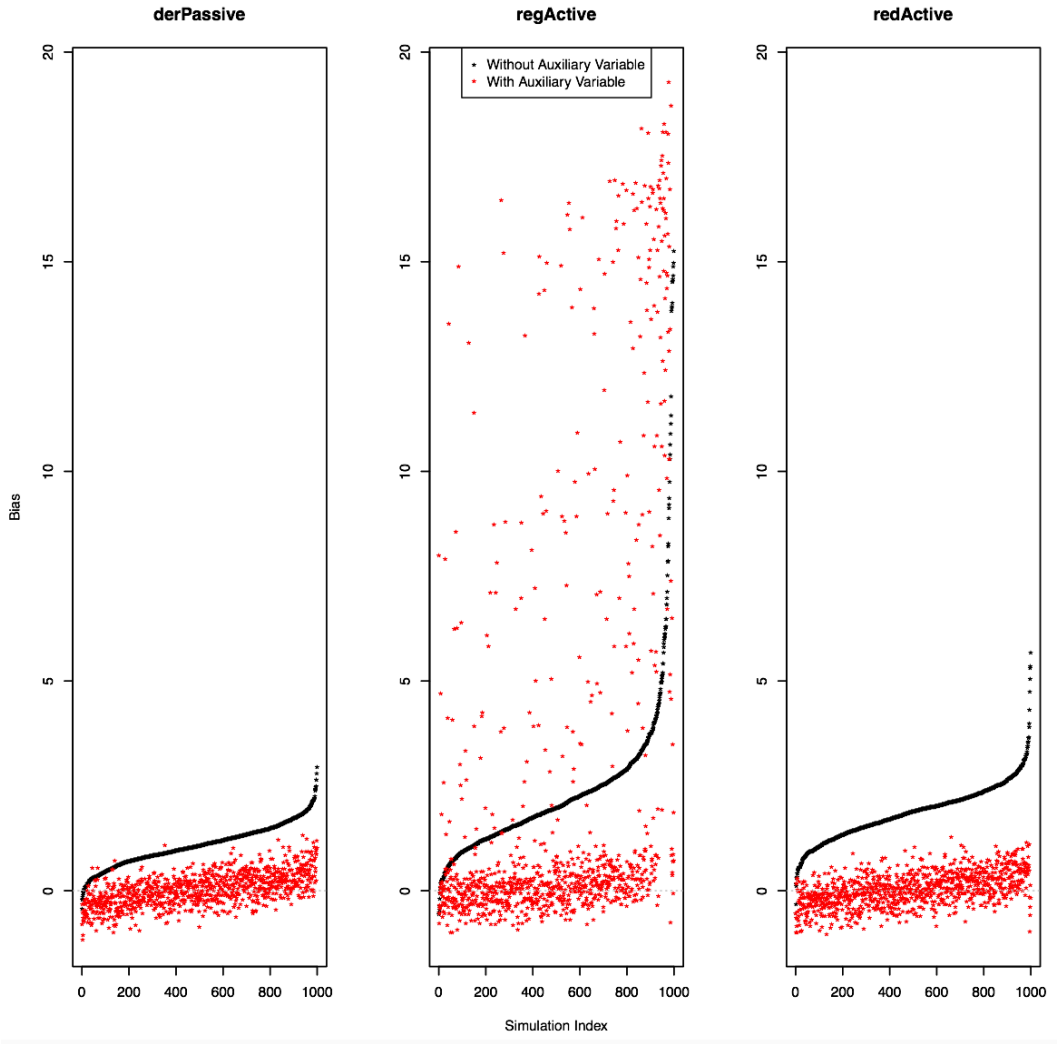

Figure A2. Illustration of bias in estimated treatment effects in 1,000 simulations *MI-regActive* when X_2 was missing under MNAR. The bias from each simulation when auxiliary variable was not included (black dots) was ranked from the lowest (left) to the highest (right). The red dots indicate the bias from the same simulation but when the auxiliary variable was included. We can see that, including an auxiliary variable reduces bias in most cases, except in the simulations with extremely high bias (right hand side of the plot). This explains why including an auxiliary variable increased the bootstrap variance as seen Appendix Table A4.

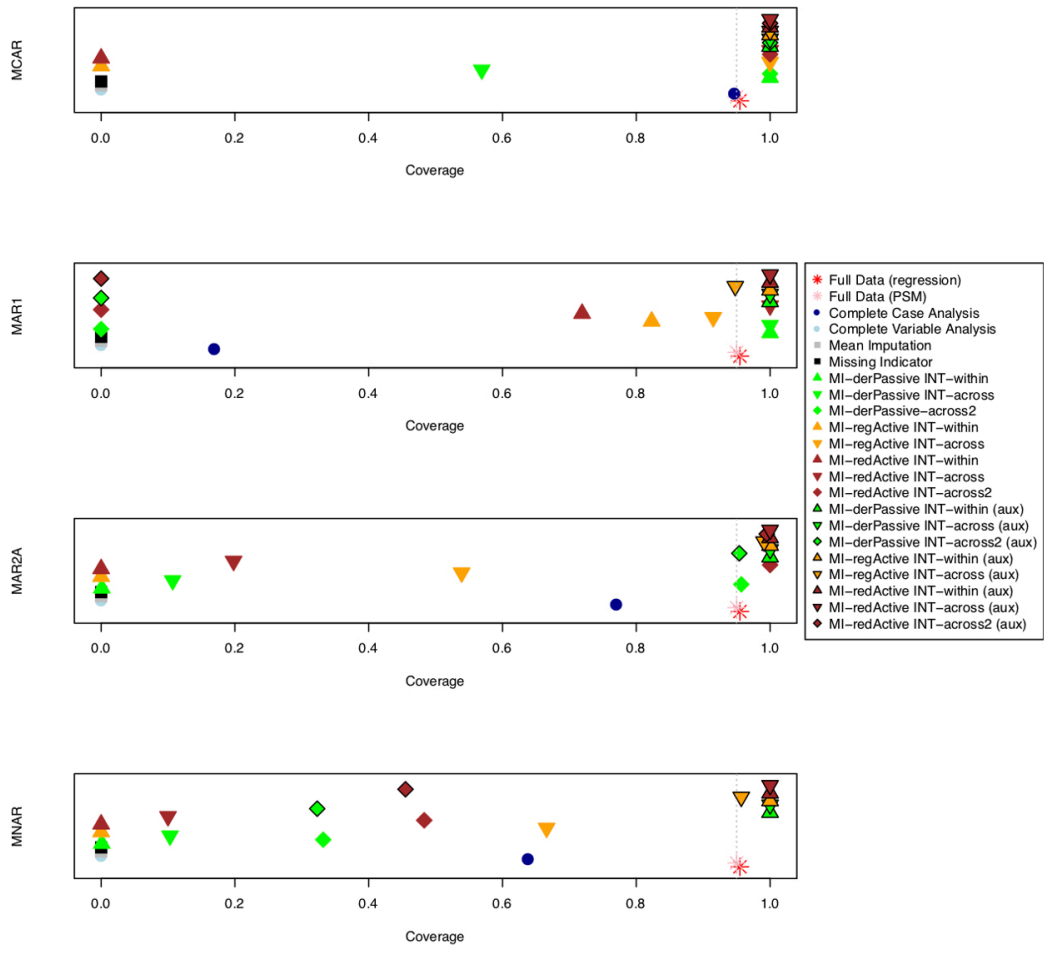

Figure A3. Coverage of various missing data methods by missing data mechanisms.

Table A1. Comparison of the different predictors included in the imputation model for the multiple imputation approaches included.

	Scientific model: $Y \sim T + X_1 + X_2$			
Missing Variable	**Predictors in imputation model**			
	MI-regActive	***MI-redActive****	***MI-derPassive*****	***MI-regPassive*****
X_2	$X_1, T, Y, PS\ (Z_2\ )$	$X_1, T, Y, PS\ (Z_2)$	$X_1, T, Y, (Z_2\ or\ Z_{ps})$	X_1, T, Y, Z_{ps}
PS	$X_1, X_2, T, Y\ (Z_2)$	$X_1, X_2, T, Y (Z_2)$	-	-
Z**	-	-	X_1, X_2, T, Y	X_1, X_2, T, Y, PS

*PS is re-derived from X_1 and X_2 after MI procedure

** Z_2 was used in *MI-derPassive*, *MI-regActive*, and *MI-redActive* when it was fully observed. Z_{ps} was used in *MI-derPassive* and *MI-regPassive* when 20% of values were missing. Under MAR2B, the order of imputation (whether Z_{ps} was imputed before or after X_2) was also evaluated.

*** We compare and contrast *MI-derPassive* and *MI-regPassive* when X_2 and Z_{ps} are both missing below (the case where X_2 is imputed before Z_{ps}):

In *MI-derPassive*:

1) Within Multivariate Imputation via Chained Equations (MICE) algorithm:
    a. Impute X_2 using X_1, T, Y, and Z_{ps}
    b. Impute Z_{ps} using X_1, X_2 , T, and Y
    c. Repeat until algorithm converges
2) Estimate PS using X_1 and X_2

In *MI-regPassive*:

1) Within MICE algorithm
    a. Impute X_2 using X_1, T, Y, and Z_{ps}
    b. Estimate PS from X_1 and X_2
    c. Impute Z_{ps} using X_1, X_2, T, Y, and PS
    d. Repeat until algorithm converges

Table A2. Main simulation results: balance diagnosis and treatment effect estimation results using commonly applied missing data method before propensity score matching (PSM), in reference to applying PSM to full data without missingness (Monte Carlo standard errors in parentheses). MDM= missing data mechanism, CC = complete-case analysis, CVA = complete-variable analysis.

			Standard Error						Post-Matching Standardized Difference		
MDM	Method	Bias	Empirical	Robust	MSE	rMSE	Coverage	Percentage Matched	Underlying data	X1	X2
Full Data	NA	-0.006 (0.012)	0.376 (0.008)	0.380 (0.001)	0.141 (0.006)	1.000	0.950 (0.006)	0.653	-	0.000	0.000
MCAR	CC	0.002 (0.011)	0.513 (0.011)	0.537 (0.001)	0.262 (0.007)	1.857	0.957 (0.006)	0.653*	-	-0.001	-0.004
	CVA	5.058 (0.013)	0.436 (0.01)	0.442 (0)	25.772 (0.132)	182.565	0 (0)	0.960	-	0	0.947
	Mean Imputation	2.985 (0.014)	0.461 (0.01)	0.447 (0.001)	9.122 (0.084)	64.617	0 (0)	0.803	original data	0	0.546
									imputed data	0	0
	Missing Indicator	2.973 (0.016)	0.440 (0.01)	0.446 (0.001)	9.032 (0.098)	63.981	0 (0)	0.535	full data	0	0.546
									observed part	0	0
									missing part	0	0.946
MAR1	CC	-2.489 (0.017)	0.821 (0.018)	0.838 (0.002)	6.869 (0.088)	48.658	0.157 (0.012)	1.000*	-	0.002	-0.001
	CVA	5.059 (0.013)	0.434 (0.01)	0.443 (0)	25.782 (0.131)	182.634	0 (0)	0.960	-	0	0.948
	Mean Imputation	5.759 (0.017)	0.573 (0.013)	0.497 (0.004)	33.498 (0.195)	237.298	0 (0)	0.735	original data	0	1.206
									imputed data	0	0.095
	Missing Indicator	5.201 (0.025)	0.686 (0.015)	0.674 (0.001)	27.517 (0.267)	194.929	0 (0)	0.271	full data	0	0.936
									observed part	0	0.034
									missing part	0	1.517
MAR2A	CC	-0.815 (0.015)	0.726 (0.016)	0.682 (0.002)	1.191 (0.03)	8.435	0.764 (0.013)	1*	-	-0.001	0.001
	CVA	5.059 (0.013)	0.443 (0.01)	0.442 (0)	25.788 (0.134)	182.680	0 (0)	0.960	-	0	0.947
	Mean Imputation	5.706 (0.017)	0.574 (0.013)	0.535 (0.002)	32.886 (0.196)	232.962	0 (0)	0.602	original data	0	1.080
									imputed data	0	0.044
	Missing Indicator	5.528 (0.021)	0.578 (0.013)	0.568 (0.001)	30.890 (0.237)	218.819	0 (0)	0.352	full data	-0.001	0.954
									observed part	-0.001	0

									missing part	-0.001	1.704
MNAR	**CC**	-1.084 (0.014)	0.675 (0.015)	0.682 (0.002)	1.629 (0.033)	11.538	0.631 (0.015)	1*	-	-0.005	-0.004
	CVA	5.059 (0.013)	0.435 (0.01)	0.443 (0)	25.783 (0.132)	182.646	0 (0)	0.960	-	0	0.947
	Mean Imputation	5.689 (0.018)	0.592 (0.013)	0.559 (0.001)	32.711 (0.202)	231.723	0 (0)	0.557	original data	0	1.078
									imputed data	0	0.045
	Missing Indicator	5.534 (0.021)	0.582 (0.013)	0.577 (0.001)	30.958 (0.239)	219.305	0 (0)	0.345	full data	0	0.962
									observed part	0	0
									missing part	0	1.725

*The percentage of treated patients being matched using CC is calculated from complete cases.

Table A3. Main simulation results: percentage of treated subjects matched and standardized difference of covariates after matching in various multiple imputation (MI) strategies. For *INT-across* and *INT-across2*, standardized differences were also calculated on the observed and missing part of the imputed data separately. MDM=missing data mechanism.

				Post-matching standardized difference							
	MI Strategy			Full Data		Imputed Data		Observed Part		Missing Part	
MDM	Imputation	Integration	Percentage Matched	X1	X2	X1	X2	X1	X2	X1	X2
MCAR											
Auxiliary variable not included in imputation model											
	MI-derPassive	**INT-within**	0.652	0	0.126	0	0	-	-	-	-
		INT-across	0.633	-0.002	0.008	-0.002	0.006	0.036	-0.114	-0.046	0.125
		INT-across2	0.636	-0.002	0.012	-0.002	0.010	0.039	-0.108	-0.048	0.127
	MI-regActive	**INT-within**	0.791	0.128	0.600	0.128	-0.127	-	-	-	-
		INT-across	0.722	-0.004	0.345	-0.004	0.010	0.010	0.003	-0.019	0.022
	MI-redActive	**INT-within**	0.792	0	0.656	0	0	-	-	-	-
		INT-across	0.738	0	0.345	0	0.009	0.004	-0.002	-0.004	0.021
		INT-across2	0.782	0	0.430	0	0.045	0.001	0.001	-0.002	0.079
Auxiliary variable included in imputation model											
	MI-derPassive	**INT-within**	0.655	0	0.024	0	0	-	-	-	-
		INT-across	0.652	0.006	-0.008	0.006	-0.001	-0.115	-0.088	0.098	0.066
		INT-across2	0.653	0.010	-0.011	0.010	-0.003	-0.123	-0.087	0.110	0.062
	MI-regActive	**INT-within**	0.659	0.006	0.040	0.006	-0.011	-	-	-	-
		INT-across	0.653	-0.028	0.025	-0.028	0.010	-0.122	0.053	0.049	-0.039
	MI-redActive	**INT-within**	0.659	0	0.051	0	0	-	-	-	-
		INT-across	0.653	-0.031	0.029	-0.031	0.014	-0.203	0.086	0.110	-0.059
		INT-across2	0.652	-0.043	0.040	-0.043	0.026	-0.214	0.112	0.100	-0.061
MAR1											
Auxiliary variable not included in imputation model											
	MI-derPassive	**INT-within**	0.656	0	0.114	0	0	-	-	-	-
		INT-across	0.617	0.002	-0.028	0.002	-0.015	-0.748	-1.105	0.205	0.647

	INT-across2	0.621	0.008	-0.034	0.008	-0.021	-0.746	-1.109	0.201	0.645
MI-regActive	**INT-within**	0.877	0.368	0.752	0.368	0.322	-	-	-	-
	INT-across	0.666	0.562	0.531	0.562	0.344	-0.160	-0.397	0.714	0.708
MI-redActive	**INT-within**	0.772	0	0.483	0	-0.006	-	-	-	-
	INT-across	0.628	0.002	0.031	0.002	-0.178	-0.746	-1.013	0.241	0.466
	INT-across2	0.670	0.003	0.219	0.003	-0.029	-0.780	-0.84	0.260	0.517
Auxiliary variable included in imputation model										
MI-derPassive	**INT-within**	0.666	0	0.065	0	-0.001	-	-	-	-
	INT-across	0.647	0.026	-0.031	0.026	-0.069	-0.895	-1.010	0.467	0.404
	INT-across2	0.652	0.022	-0.027	0.022	-0.060	-0.953	-0.974	0.519	0.375
MI-regActive	**INT-within**	0.785	0.351	0.420	0.351	0.338	-	-	-	-
	INT-across	0.745	0.386	0.382	0.386	0.321	-0.672	-0.663	0.856	0.741
MI-redActive	**INT-within**	0.669	0	0.085	0	-0.001	-	-	-	-
	INT-across	0.643	0.012	-0.015	0.012	-0.069	-0.921	-1.008	0.460	0.405
	INT-across2	0.651	0.013	-0.016	0.013	-0.061	-0.972	-0.982	0.506	0.384
MAR2A										
Auxiliary variable not included in imputation model										
MI-derPassive	**INT-within**	0.754	0	0.379	0	0	-	-	-	-
	INT-across	0.686	0	0.146	0	-0.017	-0.390	-1.182	0.226	0.558
	INT-across2	0.695	0	0.171	0	-0.016	-0.392	-1.185	0.212	0.563
MI-regActive	**INT-within**	0.881	0.076	0.698	0.076	0.078	-	-	-	-
	INT-across	0.571	0.026	0.428	0.026	0.021	-0.038	-0.472	0.020	0.204
MI-redActive	**INT-within**	0.863	0	0.656	0	0	-	-	-	-
	INT-across	0.734	0	0.348	0	-0.095	-0.331	-0.990	0.107	0.221
	INT-across2	0.780	0	0.396	0	-0.090	-0.403	-1.061	0.140	0.245
Auxiliary variable included in imputation model										
MI-derPassive	**INT-within**	0.669	0	0.071	0	-0.001	-	-	-	-
	INT-across	0.652	-0.014	0.013	-0.014	-0.030	-0.441	-1.283	0.294	0.546
	INT-across2	0.653	-0.025	0.023	-0.025	-0.019	-0.473	-1.257	0.308	0.541

MI-regActive	**INT-within**	0.729	0.122	0.248	0.122	0.166	-	-	-	-
	INT-across	0.701	0.131	0.177	0.131	0.126	-0.374	-1.061	0.510	0.614
MI-redActive	**INT-within**	0.672	0	0.084	0	-0.001	-	-	-	-
	INT-across	0.652	-0.018	0.017	-0.018	-0.032	-0.476	-1.230	0.341	0.480
	INT-across2	0.652	-0.026	0.024	-0.026	-0.023	-0.506	-1.203	0.361	0.465
MNAR										
Auxiliary variable not included in imputation model										
MI-derPassive	**INT-within**	0.754	0	0.377	0	0	-	-	-	-
	INT-across	0.686	0	0.143	0	-0.017	-0.413	-1.245	0.233	0.571
	INT-across2	0.694	0	0.168	0	-0.017	-0.416	-1.247	0.219	0.576
MI-regActive	**INT-within**	0.888	0.117	0.733	0.117	0.107	-	-	-	-
	INT-across	0.524	0.041	0.437	0.041	0.032	0.066	-0.410	-0.045	0.179
MI-redActive	**INT-within**	0.863	0	0.652	0	0	-	-	-	-
	INT-across	0.729	0	0.329	0	-0.107	-0.346	-1.080	0.093	0.229
	INT-across2	0.776	0	0.381	0	-0.099	-0.418	-1.142	0.132	0.256
Auxiliary variable included in imputation model										
MI-derPassive	**INT-within**	0.671	0	0.079	0	-0.001	-	-	-	-
	INT-across	0.652	-0.012	0.011	-0.012	-0.039	-0.457	-1.339	0.301	0.542
	INT-across2	0.652	-0.022	0.021	-0.022	-0.029	-0.488	-1.310	0.317	0.531
MI-regActive	**INT-within**	0.747	0.204	0.300	0.204	0.208	-	-	-	-
	INT-across	0.709	0.240	0.246	0.240	0.184	-0.294	-1.027	0.623	0.636
MI-redActive	**INT-within**	0.675	0	0.096	0	-0.001	-	-	-	-
	INT-across	0.652	-0.012	0.011	-0.012	-0.047	-0.494	-1.281	0.364	0.455
	INT-across2	0.652	-0.020	0.019	-0.020	-0.037	-0.528	-1.251	0.387	0.441

Table A4. Main simulation results: bias, standard error mean squared error (MSE), relative mean squared error (rMSE), and coverage (calculated using bootstrap standard error) results using various multiple imputation (MI) strategies in MCAR, MAR1, and MNAR. (Monte Carlo standard errors in parentheses).

MI Strategy			Standard Error				
Imputation	**Integration**	**Bias**	**Empirical**	**Bootstrap**	**MSE**	**rMSE**	**Coverage**
MCAR							
Auxiliary variable not included in imputation model							
MI-derPassive	INT-within	-0.024 (0.012)	0.381 (0.009)	0.395 (0.001)	0.146 (0.006)	1.034	1 (0)
	INT-across	-0.796 (0.013)	0.424 (0.009)	0.418 (0.001)	0.814 (0.022)	5.766	0.569 (0.016)
	INT-across2	0.008 (0.017)	0.534 (0.012)	0.539 (0.002)	0.285 (0.013)	2.019	1 (0)
MI-regActive	INT-within	3.224 (0.015)	0.489 (0.011)	0.508 (0.002)	10.636 (0.103)	75.345	0 (0)
	INT-across	0.870 (0.028)	0.882 (0.020)	0.813 (0.004)	1.534 (0.055)	10.867	1 (0)
MI-redActive	INT-within	3.309 (0.013)	0.405 (0.009)	0.438 (0.001)	11.113 (0.087)	78.724	0 (0)
	INT-across	0.851 (0.029)	0.926 (0.021)	0.807 (0.004)	1.580 (0.057)	11.193	1 (0)
	INT-across2	0.012 (0.017)	0.541 (0.012)	0.536 (0.001)	0.293 (0.013)	2.076	1 (0)
Auxiliary variable included in imputation model							
MI-derPassive	INT-within	0.095 (0.011)	0.359 (0.008)	0.359 (0.001)	0.138 (0.006)	0.978	1 (0)
	INT-across	0.046 (0.012)	0.388 (0.009)	0.390 (0.001)	0.152 (0.007)	1.077	1 (0)
	INT-across2	-0.005 (0.017)	0.55 (0.012)	0.547 (0.001)	0.303 (0.013)	2.146	1 (0)
MI-regActive	INT-within	0.241 (0.011)	0.358 (0.008)	0.363 (0.001)	0.186 (0.008)	1.318	1 (0)
	INT-across	0.231 (0.012)	0.384 (0.009)	0.405 (0.001)	0.200 (0.009)	1.417	1 (0)
MI-redActive	INT-within	0.247 (0.011)	0.356 (0.008)	0.361 (0.001)	0.188 (0.008)	1.332	1 (0)
	INT-across	0.221 (0.012)	0.387 (0.009)	0.398 (0.001)	0.198 (0.008)	1.403	1 (0)
	INT-across2	0.003 (0.017)	0.534 (0.012)	0.551 (0.001)	0.285 (0.013)	2.019	1 (0)
MAR2A							
**Auxiliary variable not included in imputation model*							
MI-derPassive	INT-within	1.479 (0.016)	0.511 (0.011)	0.470 (0.002)	2.448 (0.051)	17.341	0.001 (0.001)
	INT-across	1.094 (0.015)	0.476 (0.011)	0.473 (0.002)	1.422 (0.036)	10.073	0.107 (0.010)
	INT-across2	-1.037 (0.024)	0.761 (0.017)	0.671 (0.003)	1.654 (0.053)	11.717	0.957 (0.006)

MI-regActive	INT-within	3.258 (0.022)	0.684 (0.015)	0.582 (0.007)	11.081 (0.128)	78.497	0 (0)
	INT-across	2.252 (0.050)	1.587 (0.036)	1.769 (0.048)	7.587 (0.663)	53.746	0.539 (0.016)
MI-redActive	INT-within	3.433 (0.025)	0.781 (0.017)	0.605 (0.008)	12.397 (0.181)	87.819	0 (0)
	INT-across	2.001 (0.026)	0.818 (0.018)	0.805 (0.009)	4.674 (0.146)	33.110	0.198 (0.013)
	INT-across2	-0.950 (0.024)	0.756 (0.017)	0.654 (0.002)	1.473 (0.049)	10.435	1 (0)
Auxiliary variable included in imputation model							
MI-derPassive	INT-within	0.319 (0.012)	0.366 (0.008)	0.378 (0.001)	0.236 (0.009)	1.672	1 (0)
	INT-across	0.039 (0.012)	0.379 (0.008)	0.380 (0.001)	0.145 (0.006)	1.027	1 (0)
	INT-across2	-1.104 (0.024)	0.756 (0.017)	0.677 (0.003)	1.789 (0.057)	12.673	0.954 (0.007)
MI-regActive	INT-within	0.419 (0.012)	0.394 (0.009)	0.406 (0.001)	0.331 (0.012)	2.345	1 (0)
	INT-across	1.642 (0.120)	3.807 (0.085)	3.516 (0.057)	17.172 (1.684)	121.645	0.991 (0.003)
MI-redActive	INT-within	0.418 (0.012)	0.381 (0.009)	0.389 (0.001)	0.320 (0.011)	2.267	1 (0)
	INT-across	0.045 (0.012)	0.383 (0.009)	0.388 (0.001)	0.149 (0.007)	1.056	1 (0)
	INT-across2	-1.073 (0.023)	0.741 (0.017)	0.682 (0.002)	1.700 (0.056)	12.043	0.995 (0.002)
MNAR							
Auxiliary variable not included in imputation model							
MI-derPassive	INT-within	1.487 (0.016)	0.515 (0.012)	0.470 (0.002)	2.476 (0.052)	17.540	0.001 (0.001)
	INT-across	1.096 (0.015)	0.485 (0.011)	0.473 (0.002)	1.437 (0.037)	10.180	0.103 (0.01)
	INT-across2	-1.345 (0.021)	0.675 (0.015)	0.658 (0.003)	2.264 (0.06)	16.038	0.332 (0.015)
MI-regActive	INT-within	3.384 (0.016)	0.498 (0.011)	0.541 (0.004)	11.701 (0.111)	82.889	0 (0)
	INT-across	2.371 (0.062)	1.953 (0.044)	2.255 (0.050)	9.430 (0.793)	66.801	0.666 (0.015)
MI-redActive	INT-within	3.404 (0.016)	0.519 (0.012)	0.539 (0.007)	11.860 (0.119)	84.015	0 (0)
	INT-across	1.878 (0.021)	0.661 (0.015)	0.719 (0.007)	3.963 (0.092)	28.074	0.1 (0.009)
	INT-across2	-1.261 (0.02)	0.637 (0.014)	0.646 (0.002)	1.996 (0.054)	14.139	0.483 (0.016)
Auxiliary variable included in imputation model							
MI-derPassive	INT-within	0.408 (0.012)	0.377 (0.008)	0.380 (0.001)	0.309 (0.011)	2.189	1 (0)
	INT-across	0.042 (0.012)	0.382 (0.009)	0.378 (0.001)	0.148 (0.007)	1.048	1 (0)
	INT-across2	-1.367 (0.021)	0.679 (0.015)	0.667 (0.003)	2.33 (0.062)	16.506	0.323 (0.015)
MI-regActive	INT-within	0.490 (0.013)	0.398 (0.009)	0.406 (0.001)	0.398 (0.015)	2.819	1 (0)

	INT-across	2.505 (0.159)	5.024 (0.112)	4.295 (0.054)	31.492 (2.419)	223.087	0.957 (0.006)
MI-redActive	INT-within	0.480 (0.012)	0.387 (0.009)	0.392 (0.001)	0.380 (0.014)	2.692	1 (0)
	INT-across	0.045 (0.013)	0.396 (0.009)	0.389 (0.001)	0.159 (0.007)	1.126	1 (0)
	INT-across2	-1.325 (0.021)	0.661 (0.015)	0.673 (0.002)	2.191 (0.058)	15.521	0.455 (0.016)

*Note that since missingness in MAR2A is associated with an auxiliary variable, when the auxiliary variable was not included in the imputation, the imputation model is misspecified and the missing data mechanism becomes an MNAR scenario.

Table A5. Main simulation results: bias, standard error mean squared error (MSE), relative mean squared error (rMSE), and coverage calculated using *MI-regPassive* and *MI-derPassive* when X_2 was missing MAR2B and Z_{ps} was missing under various missing data mechanisms (aux_MCAR, aux_MAR1, and aux_MAR2). Monte Carlo standard errors are in parentheses.

MI Strategy			Standard Error				
Imputation	**Integration**	**Bias**	**Empirical**	**Bootstrap**	**MSE**	**rMSE**	**Coverage**
Fully observed aux							
MI-derPassive	INT-within	-0.019 (0.012)	0.381 (0.009)	0.378 (0.001)	0.146 (0.006)	1.032	1 (0)
	INT-across	-0.205 (0.012)	0.394 (0.009)	0.381 (0.001)	0.197 (0.008)	1.394	1 (0)
	INT-across2	-1.016 (0.021)	0.652 (0.015)	0.682 (0.003)	1.457 (0.048)	10.319	0.993 (0.003)
aux_MCAR							
Default imputation order							
MI-derPassive	INT-within	0.085 (0.012)	0.381 (0.009)	0.550 (0.006)	0.152 (0.007)	1.079	1 (0)
	INT-across	-0.152 (0.013)	0.402 (0.009)	0.499 (0.004)	0.184 (0.008)	1.304	1 (0)
	INT-across2	-0.982 (0.021)	0.671 (0.015)	0.683 (0.002)	1.415 (0.048)	10.024	1 (0)
MI-regPassive	INT-within	0.087 (0.012)	0.378 (0.008)	0.533 (0.006)	0.150 (0.007)	1.065	1 (0)
	INT-across	-0.143 (0.013)	0.413 (0.009)	0.489 (0.003)	0.191 (0.008)	1.352	1 (0)
	INT-across2	-0.996 (0.021)	0.666 (0.015)	0.685 (0.002)	1.436 (0.046)	10.173	1 (0)
Reverse imputation order							
MI-derPassive	INT-within	0.028 (0.012)	0.382 (0.009)	0.550 (0.006)	0.147 (0.006)	1.040	1 (0)
	INT-across	-0.191 (0.012)	0.391 (0.009)	0.500 (0.004)	0.189 (0.008)	1.340	1 (0)
	INT-across2	-1.01 (0.021)	0.671 (0.015)	0.682 (0.002)	1.471 (0.047)	10.418	0.999 (0.001)
MI-regPassive	INT-within	-0.196 (0.012)	0.372 (0.008)	0.532 (0.006)	0.177 (0.008)	1.251	1 (0)
	INT-across	-0.232 (0.013)	0.405 (0.009)	0.489 (0.003)	0.218 (0.009)	1.543	1 (0)
	INT-across2	-1.002 (0.021)	0.668 (0.015)	0.684 (0.002)	1.450 (0.045)	10.273	1 (0)
aux_MAR1							
Default imputation order							
MI-derPassive	INT-within	0.185 (0.012)	0.393 (0.009)	0.550 (0.006)	0.189 (0.008)	1.337	1 (0)
	INT-across	0.011 (0.013)	0.414 (0.009)	0.499 (0.004)	0.171 (0.008)	1.211	1 (0)
	INT-across2	-1.017 (0.021)	0.651 (0.015)	0.684 (0.002)	1.458 (0.045)	10.329	1 (0)

MI-regPassive	INT-within	0.205 (0.013)	0.396 (0.009)	0.533 (0.006)	0.199 (0.008)	1.410	1 (0)
	INT-across	0.014 (0.013)	0.414 (0.009)	0.488 (0.003)	0.171 (0.008)	1.213	1 (0)
	INT-across2	-1.007 (0.021)	0.656 (0.015)	0.685 (0.002)	1.444 (0.045)	10.231	1 (0)
Reverse imputation order							
MI-derPassive	INT-within	0.091 (0.012)	0.391 (0.009)	0.550 (0.006)	0.161 (0.007)	1.140	1 (0)
	INT-across	-0.069 (0.013)	0.408 (0.009)	0.500 (0.004)	0.171 (0.007)	1.213	1 (0)
	INT-across2	-0.993 (0.021)	0.661 (0.015)	0.684 (0.002)	1.422 (0.045)	10.071	1 (0)
MI-regPassive	INT-within	-0.129 (0.012)	0.379 (0.008)	0.534 (0.006)	0.160 (0.007)	1.135	1 (0)
	INT-across	-0.087 (0.013)	0.403 (0.009)	0.489 (0.003)	0.170 (0.008)	1.201	1 (0)
	INT-across2	-0.982 (0.021)	0.650 (0.015)	0.685 (0.002)	1.385 (0.043)	9.813	1 (0)
aux_MAR2							
Default imputation order							
MI-derPassive	INT-within	-0.060 (0.012)	0.390 (0.009)	0.550 (0.006)	0.155 (0.006)	1.100	1 (0)
	INT-across	-0.374 (0.013)	0.412 (0.009)	0.500 (0.004)	0.310 (0.012)	2.194	1 (0)
	INT-across2	-0.994 (0.021)	0.660 (0.015)	0.683 (0.002)	1.422 (0.044)	10.076	1 (0)
MI-regPassive	INT-within	-0.049 (0.013)	0.396 (0.009)	0.533 (0.006)	0.159 (0.006)	1.128	1 (0)
	INT-across	-0.369 (0.013)	0.408 (0.009)	0.489 (0.003)	0.302 (0.012)	2.141	1 (0)
	INT-across2	-1.001 (0.021)	0.656 (0.015)	0.685 (0.002)	1.431 (0.043)	10.137	0.999 (0.001)
Reverse imputation order							
MI-derPassive	INT-within	-0.068 (0.012)	0.390 (0.009)	0.549 (0.006)	0.157 (0.006)	1.111	1 (0)
	INT-across	-0.369 (0.013)	0.412 (0.009)	0.500 (0.004)	0.305 (0.012)	2.163	1 (0)
	INT-across2	-0.996 (0.021)	0.652 (0.015)	0.683 (0.002)	1.416 (0.043)	10.032	0.999 (0.001)
MI-regPassive	INT-within	-0.308 (0.012)	0.391 (0.009)	0.533 (0.006)	0.248 (0.010)	1.754	1 (0)
	INT-across	-0.414 (0.013)	0.411 (0.009)	0.489 (0.003)	0.340 (0.013)	2.410	1 (0)
	INT-across2	-0.987 (0.021)	0.665 (0.015)	0.685 (0.002)	1.417 (0.043)	10.039	1 (0)

Table A6. Summary statistics, missing values, and standardized differences (pre-matching and post-matching in commonly applied methods) in DIETFITS trial participants (N=609).

Variable	Summary Statistics				Standardized Difference						
	Intervention (N=97)		Control (N=512)		Pre-matching	Complete Case Analysis		Mean Imputation	Mean Imputation		Missing Indicator
	Summary	Missing	Summary	Missing		original	complete cases	original	original	imputed	original
Age (years)	39.38±6.06	5 (5.15%)	41.23±6.78	127 (24.80%)	-0.29	**-0.4**	0.08	-	0.04	0.01	**0.11**
Sex (female)	62 (63.92%)	0	284 (55.47%)	0	0.17	**-0.21**	0.02	0	**0.11**	**0.11**	**-0.11**
Race		1 (1.03%)		5 (0.98%)	-						
Caucasian	52 (53.61%)	-	302 (58.98%)	-	-0.11	**-0.23**	0.07	-	**-0.23**	**0.21**	**-0.25**
Hispanic	27 (27.84%)	-	103 (20.12%)	-	0.18	**0.16**	0	-	**0.22**	-0.02	**0.19**
Black	3 (3.09%)	-	20 (3.91%)	-	-0.04	**0.22**	0.09	-	-0.06	**-0.15**	**-0.1**
Asian/Pacific Islander	7 (7.22%)	-	57 (11.13%)	-	-0.14	**-0.11**	**-0.15**	-	-0.07	**-0.21**	0.03
Other	7 (7.22%)	-	25 (4.88%)	-	0.1	**0.14**	0	-	0	0	0
Weight (kg) change 6 months	-6.19±5.19	8 (8.25%)	-7.45±6.11	135 (26.37%)	0.22	**0.38**	-0.02	-	-0.02	0	**-0.11**
Class attendance 6 months (%)	76.86±17.01	0	72.88±22.47	0	0.2	**-0.17**	-0.06	0	**-0.19**	**-0.19**	-0.06

*post-matching standardized differences were measured both in the original dataset with missingness and the analytical datasets (e.g. complete cases for complete cases analysis and imputed dataset for mean imputation)

Table A7. Post-matching standardized differences for all multiple imputation (MI) methods in DIETFITS trial participants (N=609). Note that those whose absolute values are above 0.1 are bolded.

				Race						
MI Method	**Dataset**	**Age (years)**	**Sex (female)**	**Caucasian**	**Hispanic**	**Black**	**Asian/Pacific Islander**	**Other**	**Weight (kg) change 6 months**	**Class attendance 6 months (%)**
Auxiliary variable not included in imputation model										
derPassive-within	original	-0.02	0.03	**-0.31**	**0.21**	0.04	0.02	0.02	**0.11**	0
	imputed	0.02	0.03	0.04	-0.01	-0.01	-0.01	0	0.01	0
derPassive-across	original	**0.19**	**-0.16**	**-0.23**	**0.12**	0	0.07	-0.04	0.07	**0.13**
	imputed	**0.17**	**-0.16**	-0.06	0.02	-0.06	0	**0.13**	-0.05	**0.13**
derPassive-across2	original	0.09	-0.09	**-0.38**	**0.3**	0.07	-0.07	0.1	**0.13**	0.09
	imputed	0.07	-0.09	-0.08	-0.05	0.07	0.09	**0.13**	0.04	0.09
regActive-within	original	-0.02	**0.12**	**-0.31**	**0.31**	-0.02	0.03	-0.09	0.04	**0.37**
	imputed	-0.04	**0.12**	0.05	-0.01	0.04	**-0.13**	0.05	-0.03	**0.37**
regActive-across	original	0.07	0.02	**-0.17**	**0.24**	-0.06	0	**-0.19**	**0.3**	**0.28**
	imputed	0.07	0.02	0.23	-0.05	-0.1	**-0.29**	0.09	0.08	**0.28**
regActive-across2	original	0.03	0.06	**-0.32**	**0.22**	**0.26**	0	-0.09	0.03	0.02
	imputed	0.03	0.06	0.06	-0.05	**0.15**	-0.08	-0.04	0.03	0.02
redActive-within	original	-0.08	0.03	**-0.31**	**0.24**	-0.03	0.06	-0.04	0.09	-0.02
	imputed	0	0.03	0.02	-0.03	-0.01	-0.01	0.05	0	-0.02
redActive-across	original	0.08	**0.19**	**-0.28**	**0.33**	0	0	**-0.2**	**0.2**	-0.09
	imputed	0.06	**0.19**	0	-0.05	0	**0.15**	0	**0.19**	-0.09
redActive-across2	original	0.03	0.06	**-0.32**	**0.22**	**0.26**	0	-0.09	0.03	0.02
	imputed	0.03	0.06	0.06	-0.05	**0.15**	-0.08	-0.04	0.03	0.02
Auxiliary variable included in imputation model										
derPassive-within	original	-0.1	-0.01	**-0.29**	**0.2**	-0.02	0.04	0.02	**-0.11**	-0.02
	imputed	-0.07	-0.01	-0.07	0.08	0	-0.01	0.03	-0.06	-0.02
derPassive-across	original	**-0.14**	0.02	**-0.29**	**0.27**	-0.1	-0.03	0.05	-0.06	**-0.11**
	imputed	**-0.12**	0.02	-0.02	0.09	0	0.09	**-0.17**	-0.01	**-0.11**

derPassive-across2	original	**-0.23**	-0.02	**-0.27**	**0.24**	**-0.15**	-0.03	0.1	**-0.18**	-0.06
	imputed	**-0.2**	-0.02	0.02	0.05	-0.1	0.09	**-0.11**	-0.05	-0.06
regActive-within	original	**-0.17**	-0.01	**-0.33**	**0.23**	-0.04	0.05	0.05	-0.04	**0.34**
	imputed	**-0.11**	-0.01	0.01	**-0.13**	0.07	**0.15**	0.03	**0.37**	**0.34**
regActive-across	original	**-0.11**	-0.07	**-0.25**	0.05	-0.06	**0.15**	0.1	0.02	**0.32**
	imputed	**-0.13**	-0.07	-0.02	**-0.16**	0.07	**0.17**	**0.13**	**0.33**	**0.32**
regActive-across2	original	**-0.12**	**-0.11**	-0.4	**0.21**	**0.15**	**0.11**	0	-0.01	-0.06
	imputed	-0.09	**-0.11**	**-0.19**	**0.14**	**0.15**	0	0.04	-0.06	-0.06
redActive-within	original	-0.07	-0.01	**-0.29**	**0.19**	0.01	0.03	-0.01	-0.08	-0.07
	imputed	-0.04	-0.01	-0.07	0.09	0.02	0	0	-0.04	-0.07
redActive-across	original	-0.08	**-0.13**	**-0.36**	**0.12**	**0.15**	**0.19**	0	-0.06	-0.06
	imputed	-0.08	**-0.13**	**-0.15**	**0.14**	0.07	-0.04	0.04	-0.07	-0.06
redActive-across2	original	**-0.12**	**-0.11**	**-0.4**	**0.21**	**0.15**	**0.11**	0	-0.01	-0.06
	imputed	-0.09	**-0.11**	**-0.19**	**0.14**	**0.15**	0	0.04	-0.06	-0.06

Table A8. Sensitivity simulation results (with reduced confounding effects): balance diagnosis and treatment effect estimation results using commonly applied missing data method before propensity score matching (PSM), in reference to applying PSM to full data without missingness (Monte Carlo standard errors in parentheses). MDM=missing data mechanism, CC = complete-case analysis, CVA = complete-variable analysis.

			Standard Error						Post-Matching Standardized Difference		
MDM	**Method**	**Bias**	**Empirical**	**Robust**	**MSE**	**rMSE**	**Coverage**	**Percentage Matched**	**Underlying data**	**X1**	**X2**
Full Data	**NA**	0.001	0.298	0.310	0.089	1.000	0.960	1.000	-	0	0
MCAR	**CC**	-0.009 (0.014)	0.429 (0.01)	0.439 (0.001)	0.184 (0.008)	2.076	0.955 (0.007)	0.999	-	0	0
	CVA	1.732 (0.014)	0.44 (0.01)	0.435 (0.001)	3.194 (0.048)	35.975	0.028 (0.005)	1	-	0	0.313
	Mean Imputation	0.855 (0.012)	0.377 (0.008)	0.378 (0)	0.873 (0.021)	9.838	0.387 (0.015)	1	original data	0	0.156
									imputed data	0	0
	Missing Indicator	0.858 (0.012)	0.373 (0.008)	0.379 (0)	0.876 (0.021)	9.867	0.383 (0.015)	1	full data	0	0.157
									observed part	0	0.14
									missing part	NA	0.138
MAR1	**CC**	-1.69 (0.02)	0.632 (0.014)	0.649 (0.002)	3.256 (0.07)	36.672	0.269 (0.014)	1	-	0	0
	CVA	1.729 (0.014)	0.439 (0.01)	0.434 (0.001)	3.181 (0.048)	35.831	0.023 (0.005)	1	-	0	0.314
	Mean Imputation	2.763 (0.015)	0.477 (0.011)	0.484 (0.001)	7.863 (0.084)	88.566	0 (0)	0.722	original data	-0.003	0.483
									imputed data	-0.003	-0.003
	Missing Indicator	2.75 (0.015)	0.477 (0.011)	0.484 (0.001)	7.792 (0.084)	87.769	0 (0)	0.722	full data	-0.001	0.485
									observed part	-0.002	0.26
									missing part	NA	0.26
MAR2A	**CC**	-0.679 (0.017)	0.529 (0.012)	0.531 (0.001)	0.741 (0.027)	8.349	0.751 (0.014)	1	-	0.001	0
	CVA	1.74 (0.014)	0.431 (0.01)	0.434 (0.001)	3.213 (0.047)	36.187	0.029 (0.005)	1	-	0	0.314
	Mean Imputation	3.203 (0.014)	0.446 (0.01)	0.443 (0.001)	10.457 (0.092)	117.786	0 (0)	0.803	original data	-0.001	0.57
									imputed data	-0.001	-0.001
	Missing Indicator	3.208 (0.014)	0.437 (0.01)	0.444 (0.001)	10.484 (0.089)	118.086	0 (0)	0.803	full data	0	0.571
									observed part	0	0.312

									missing part	NA	0.312
MNAR	**CC**	-0.778 (0.017)	0.527 (0.012)	0.531 (0.001)	0.882 (0.028)	9.94	0.663 (0.015)	1	-	0	0
	CVA	1.739 (0.014)	0.43 (0.01)	0.434 (0.001)	3.209 (0.047)	36.143	0.023 (0.005)	1	-	0	0.314
	Mean Imputation	3.213 (0.014)	0.453 (0.01)	0.446 (0.001)	10.528 (0.092)	118.59	0 (0)		original data	0	0.573
								0.797	imputed data	0	0
	Missing Indicator	3.214 (0.014)	0.445 (0.01)	0.446 (0.001)	10.526 (0.091)	118.57	0 (0)	0.797	full data	0	0.573
									observed part	0	0.312
									missing part	NA	0.31

*The percentage of treated patients being matched using CC is calculated from complete cases.

Table A9. Sensitivity simulation results (with reduced confounding effects): percentage of treated subjects matched and standardized difference of covariates after matching in various multiple imputation (MI) strategies. For *INT-across* and *INT-across2*, standardized differences were also calculated on the observed and missing part of the imputed data separately. MDM=missing data mechanism.

				Post-matching standardized difference							
	MI Strategy			Full Data		Imputed Data		Observed Part		Missing Part	
MDM	Imputation	Integration	Percentage Matched	X1	X2	X1	X2	X1	X2	X1	X2
MCAR											
Auxiliary variable not included in imputation model											
	MI-derPassive	**INT-within**	0.652	0	0.126	0	0	-	-	-	-
		INT-across	0.633	-0.002	0.008	-0.002	0.006	0.036	-0.114	-0.046	0.125
		INT-across2	0.636	-0.002	0.012	-0.002	0.01	0.039	-0.108	-0.048	0.127
	MI-regActive	**INT-within**	0.791	0.128	0.6	0.128	-0.127	-	-	-	-
		INT-across	0.722	-0.004	0.345	-0.004	0.01	0.01	0.003	-0.019	0.022
	MI-redActive	**INT-within**	0.792	0	0.656	0	0	-	-	-	-
		INT-across	0.738	0	0.345	0	0.009	0.004	-0.002	-0.004	0.021
		INT-across2	0.782	0	0.43	0	0.045	0.001	0.001	-0.002	0.079
Auxiliary variable included in imputation model											
	MI-derPassive	**INT-within**	0.655	0	0.024	0	0	-	-	-	-
		INT-across	0.652	0.006	-0.008	0.006	-0.001	-0.115	-0.088	0.098	0.066
		INT-across2	0.653	0.01	-0.011	0.01	-0.003	-0.123	-0.087	0.11	0.062
	MI-regActive	**INT-within**	0.659	0.006	0.04	0.006	-0.011	-	-	-	-
		INT-across	0.653	-0.028	0.025	-0.028	0.01	-0.122	0.053	0.049	-0.039
	MI-redActive	**INT-within**	0.659	0	0.051	0	0	-	-	-	-
		INT-across	0.653	-0.031	0.029	-0.031	0.014	-0.203	0.086	0.11	-0.059
		INT-across2	0.652	-0.043	0.04	-0.043	0.026	-0.214	0.112	0.1	-0.061
MAR1											
Auxiliary variable not included in imputation model											
	MI-derPassive	**INT-within**	0.656	0	0.114	0	0	-	-	-	-

	INT-across	0.617	0.002	-0.028	0.002	-0.015	-0.748	-1.105	0.205	0.647
	INT-across2	0.621	0.008	-0.034	0.008	-0.021	-0.746	-1.109	0.201	0.645
MI-regActive	**INT-within**	0.877	0.368	0.752	0.368	0.322	-	-	-	-
	INT-across	0.666	0.562	0.531	0.562	0.344	-0.16	-0.397	0.714	0.708
MI-redActive	**INT-within**	0.772	0	0.483	0	-0.006	-	-	-	-
	INT-across	0.628	0.002	0.031	0.002	-0.178	-0.746	-1.013	0.241	0.466
	INT-across2	0.67	0.003	0.219	0.003	-0.029	-0.78	-0.84	0.26	0.517
Auxiliary variable included in imputation model										
MI-derPassive	**INT-within**	0.666	0	0.065	0	-0.001	-	-	-	-
	INT-across	0.647	0.026	-0.031	0.026	-0.069	-0.895	-1.01	0.467	0.404
	INT-across2	0.652	0.022	-0.027	0.022	-0.06	-0.953	-0.974	0.519	0.375
MI-regActive	**INT-within**	0.785	0.351	0.42	0.351	0.338	-	-	-	-
	INT-across	0.745	0.386	0.382	0.386	0.321	-0.672	-0.663	0.856	0.741
MI-redActive	**INT-within**	0.669	0	0.085	0	-0.001	-	-	-	-
	INT-across	0.643	0.012	-0.015	0.012	-0.069	-0.921	-1.008	0.46	0.405
	INT-across2	0.651	0.013	-0.016	0.013	-0.061	-0.972	-0.982	0.506	0.384
MAR2A										
Auxiliary variable not included in imputation model										
MI-derPassive	**INT-within**	0.754	0	0.379	0	0	-	-	-	-
	INT-across	0.686	0	0.146	0	-0.017	-0.39	-1.182	0.226	0.558
	INT-across2	0.695	0	0.171	0	-0.016	-0.392	-1.185	0.212	0.563
MI-regActive	**INT-within**	0.881	0.076	0.698	0.076	0.078	-	-	-	-
	INT-across	0.571	0.026	0.428	0.026	0.021	-0.038	-0.472	0.02	0.204
MI-redActive	**INT-within**	0.863	0	0.656	0	0	-	-	-	-
	INT-across	0.734	0	0.348	0	-0.095	-0.331	-0.99	0.107	0.221
	INT-across2	0.78	0	0.396	0	-0.09	-0.403	-1.061	0.14	0.245
Auxiliary variable included in imputation model										
MI-derPassive	**INT-within**	0.669	0	0.071	0	-0.001	-	-	-	-
	INT-across	0.652	-0.014	0.013	-0.014	-0.03	-0.441	-1.283	0.294	0.546

	INT-across2	0.653	-0.025	0.023	-0.025	-0.019	-0.473	-1.257	0.308	0.541
MI-regActive	**INT-within**	0.729	0.122	0.248	0.122	0.166	-	-	-	-
	INT-across	0.701	0.131	0.177	0.131	0.126	-0.374	-1.061	0.51	0.614
MI-redActive	**INT-within**	0.672	0	0.084	0	-0.001	-	-	-	-
	INT-across	0.652	-0.018	0.017	-0.018	-0.032	-0.476	-1.23	0.341	0.48
	INT-across2	0.652	-0.026	0.024	-0.026	-0.023	-0.506	-1.203	0.361	0.465
MNAR										
Auxiliary variable not included in imputation model										
MI-derPassive	**INT-within**	0.754	0	0.377	0	0	-	-	-	-
	INT-across	0.686	0	0.143	0	-0.017	-0.413	-1.245	0.233	0.571
	INT-across2	0.694	0	0.168	0	-0.017	-0.416	-1.247	0.219	0.576
MI-regActive	**INT-within**	0.888	0.117	0.733	0.117	0.107	-	-	-	-
	INT-across	0.524	0.041	0.437	0.041	0.032	0.066	-0.41	-0.045	0.179
MI-redActive	**INT-within**	0.863	0	0.652	0	0	-	-	-	-
	INT-across	0.729	0	0.329	0	-0.107	-0.346	-1.08	0.093	0.229
	INT-across2	0.776	0	0.381	0	-0.099	-0.418	-1.142	0.132	0.256
Auxiliary variable included in imputation model										
MI-derPassive	**INT-within**	0.671	0	0.079	0	-0.001	-	-	-	-
	INT-across	0.652	-0.012	0.011	-0.012	-0.039	-0.457	-1.339	0.301	0.542
	INT-across2	0.652	-0.022	0.021	-0.022	-0.029	-0.488	-1.31	0.317	0.531
MI-regActive	**INT-within**	0.747	0.204	0.3	0.204	0.208	-	-	-	-
	INT-across	0.709	0.24	0.246	0.24	0.184	-0.294	-1.027	0.623	0.636
MI-redActive	**INT-within**	0.675	0	0.096	0	-0.001	-	-	-	-
	INT-across	0.652	-0.012	0.011	-0.012	-0.047	-0.494	-1.281	0.364	0.455
	INT-across2	0.652	-0.02	0.019	-0.02	-0.037	-0.528	-1.251	0.387	0.441

Table A10. Sensitivity simulation results (with reduced confounding effects): bias, standard error, mean squared error (MSE), relative mean squared error (rMSE), and coverage (calculated using bootstrap standard error) results using various multiple imputation (MI) strategies in MCAR, MAR1, MAR2A, and MNAR. (Monte Carlo standard errors in parentheses).

MI Strategy			Standard Error				
Imputation	Integration	Bias	Empirical	Bootstrap	MSE	rMSE	Coverage
MCAR							
Auxiliary variable not included in imputation model							
MI-derPassive	INT-within	-0.001 (0.01)	0.316 (0.007)	0.315 (0)	0.099 (0.004)	1.115	1 (0)
	INT-across	-0.355 (0.011)	0.348 (0.008)	0.348 (0.001)	0.247 (0.009)	2.782	1 (0)
	INT-across2	0.013 (0.013)	0.424 (0.009)	0.433 (0.001)	0.18 (0.008)	2.028	1 (0)
MI-regActive	INT-within	0.973 (0.011)	0.344 (0.008)	0.36 (0.003)	1.064 (0.024)	11.985	0.042 (0.006)
	INT-across	-0.423 (0.014)	0.45 (0.01)	0.548 (0.002)	0.381 (0.014)	4.292	1 (0)
MI-redActive	INT-within	0.967 (0.011)	0.343 (0.008)	0.343 (0.001)	1.052 (0.024)	11.85	0.01 (0.003)
	INT-across	-1.386 (0.018)	0.584 (0.013)	0.525 (0.002)	2.261 (0.055)	25.468	0.004 (0.002)
	INT-across2	0.015 (0.014)	0.44 (0.01)	0.436 (0.001)	0.194 (0.009)	2.185	1 (0)
Auxiliary variable included in imputation model							
MI-derPassive	INT-within	0.031 (0.008)	0.266 (0.006)	0.279 (0)	0.072 (0.003)	0.811	1 (0)
	INT-across	-0.044 (0.01)	0.317 (0.007)	0.356 (0.001)	0.103 (0.005)	1.16	1 (0)
	INT-across2	0.004 (0.014)	0.458 (0.01)	0.444 (0.001)	0.209 (0.01)	2.354	1 (0)
MI-regActive	INT-within	0.169 (0.009)	0.27 (0.006)	0.284 (0)	0.101 (0.005)	1.138	1 (0)
	INT-across	0.231 (0.01)	0.329 (0.007)	0.355 (0.001)	0.162 (0.007)	1.825	1 (0)
MI-redActive	INT-within	0.17 (0.009)	0.269 (0.006)	0.282 (0)	0.101 (0.005)	1.138	1 (0)
	INT-across	-0.14 (0.01)	0.325 (0.007)	0.413 (0.001)	0.125 (0.006)	1.408	1 (0)
	INT-across2	0.013 (0.014)	0.456 (0.01)	0.448 (0.001)	0.208 (0.01)	2.343	1 (0)
MAR1							
Auxiliary variable not included in imputation model							
MI-derPassive	INT-within	0.149 (0.012)	0.374 (0.008)	0.372 (0.001)	0.162 (0.007)	1.825	1 (0)
	INT-across	0.203 (0.012)	0.385 (0.009)	0.396 (0.002)	0.189 (0.008)	2.129	1 (0)
	INT-across2	-2.102 (0.019)	0.609 (0.014)	0.607 (0.002)	4.791 (0.082)	53.966	0 (0)

MI-regActive	INT-within	1.495 (0.012)	0.381 (0.009)	0.406 (0.001)	2.381 (0.037)	26.82	0 (0)
	INT-across	-0.197 (0.033)	1.057 (0.024)	1.114 (0.008)	1.154 (0.054)	12.999	1 (0)
MI-redActive	INT-within	1.496 (0.012)	0.385 (0.009)	0.397 (0.001)	2.386 (0.038)	26.876	0 (0)
	INT-across	0.273 (0.077)	2.436 (0.054)	2.089 (0.016)	6.001 (0.232)	67.595	1 (0)
	INT-across2	-1.972 (0.019)	0.591 (0.013)	0.608 (0.001)	4.236 (0.075)	47.714	0 (0)
Auxiliary variable included in imputation model							
MI-derPassive	INT-within	0.198 (0.009)	0.283 (0.006)	0.291 (0)	0.119 (0.005)	1.34	1 (0)
	INT-across	-0.059 (0.012)	0.392 (0.009)	0.339 (0.001)	0.157 (0.008)	1.768	1 (0)
	INT-across2	-2.139 (0.02)	0.617 (0.014)	0.609 (0.002)	4.956 (0.085)	55.824	0 (0)
MI-regActive	INT-within	0.241 (0.01)	0.316 (0.007)	0.31 (0.001)	0.158 (0.007)	1.78	1 (0)
	INT-across	0.097 (0.013)	0.42 (0.009)	0.579 (0.014)	0.186 (0.015)	2.095	1 (0)
MI-redActive	INT-within	0.251 (0.01)	0.316 (0.007)	0.31 (0)	0.163 (0.007)	1.836	1 (0)
	INT-across	-0.062 (0.014)	0.433 (0.01)	0.406 (0.003)	0.191 (0.011)	2.151	1 (0)
	INT-across2	-2.082 (0.02)	0.617 (0.014)	0.616 (0.002)	4.715 (0.082)	53.11	0 (0)
MAR2A							
Auxiliary variable not included in imputation model							
MI-derPassive	INT-within	1.356 (0.012)	0.393 (0.009)	0.367 (0.001)	1.994 (0.037)	22.46	0 (0)
	INT-across	1.248 (0.018)	0.559 (0.013)	0.567 (0.005)	1.87 (0.058)	21.064	0.352 (0.015)
	INT-across2	-0.801 (0.017)	0.526 (0.012)	0.511 (0.001)	0.918 (0.029)	10.34	0.994 (0.002)
MI-regActive	INT-within	2.427 (0.012)	0.366 (0.008)	0.351 (0)	6.025 (0.058)	67.866	0 (0)
	INT-across	3.362 (0.049)	1.559 (0.035)	1.408 (0.005)	13.732 (0.358)	154.677	0.035 (0.006)
MI-redActive	INT-within	2.426 (0.012)	0.365 (0.008)	0.351 (0)	6.017 (0.058)	67.775	0 (0)
	INT-across	4.829 (0.024)	0.757 (0.017)	0.873 (0.011)	23.891 (0.233)	269.108	0 (0)
	INT-across2	-0.705 (0.017)	0.529 (0.012)	0.523 (0.001)	0.776 (0.027)	8.741	1 (0)
Auxiliary variable included in imputation model							
MI-derPassive	INT-within	0.232 (0.009)	0.293 (0.007)	0.289 (0)	0.139 (0.006)	1.566	1 (0)
	INT-across	-0.018 (0.011)	0.341 (0.008)	0.327 (0.001)	0.116 (0.005)	1.307	1 (0)
	INT-across2	-0.886 (0.017)	0.543 (0.012)	0.524 (0.001)	1.08 (0.032)	12.165	0.979 (0.005)
MI-regActive	INT-within	0.525 (0.01)	0.327 (0.007)	0.327 (0)	0.382 (0.014)	4.303	1 (0)

	INT-across	0.731 (0.052)	1.644 (0.037)	1.322 (0.008)	3.235 (0.154)	36.439	1 (0)
MI-redActive	INT-within	0.524 (0.01)	0.327 (0.007)	0.327 (0)	0.381 (0.014)	4.292	1 (0)
	INT-across	-0.077 (0.012)	0.385 (0.009)	0.435 (0.005)	0.154 (0.008)	1.735	1 (0)
	INT-across2	-0.815 (0.017)	0.526 (0.012)	0.512 (0.001)	0.94 (0.029)	10.588	0.999 (0.001)
MNAR							
Auxiliary variable not included in imputation model							
MI-derPassive	INT-within	1.369 (0.013)	0.404 (0.009)	0.368 (0.001)	2.038 (0.038)	22.956	0 (0)
	INT-across	1.266 (0.019)	0.598 (0.013)	0.571 (0.005)	1.96 (0.067)	22.077	0.344 (0.015)
	INT-across2	-0.96 (0.016)	0.498 (0.011)	0.511 (0.001)	1.17 (0.031)	13.179	0.682 (0.015)
MI-regActive	INT-within	2.47 (0.012)	0.365 (0.008)	0.353 (0)	6.236 (0.059)	70.242	0 (0)
	INT-across	3.146 (0.047)	1.492 (0.033)	1.408 (0.005)	12.123 (0.321)	136.553	0.119 (0.01)
MI-redActive	INT-within	2.46 (0.012)	0.368 (0.008)	0.353 (0)	6.189 (0.059)	69.713	0 (0)
	INT-across	4.852 (0.024)	0.763 (0.017)	0.84 (0.009)	24.124 (0.232)	271.732	0 (0)
	INT-across2	-0.801 (0.017)	0.529 (0.012)	0.525 (0.001)	0.922 (0.029)	10.385	1 (0)
Auxiliary variable included in imputation model							
MI-derPassive	INT-within	0.259 (0.009)	0.282 (0.006)	0.289 (0)	0.147 (0.006)	1.656	1 (0)
	INT-across	-0.014 (0.011)	0.36 (0.008)	0.328 (0.001)	0.129 (0.007)	1.453	1 (0)
	INT-across2	-1.044 (0.016)	0.516 (0.012)	0.522 (0.001)	1.356 (0.036)	15.274	0.367 (0.015)
MI-regActive	INT-within	0.557 (0.01)	0.317 (0.007)	0.329 (0.001)	0.411 (0.013)	4.629	1 (0)
	INT-across	0.942 (0.053)	1.68 (0.038)	1.367 (0.008)	3.707 (0.163)	41.756	1 (0)
MI-redActive	INT-within	0.559 (0.01)	0.316 (0.007)	0.328 (0.001)	0.412 (0.013)	4.641	0.999 (0.001)
	INT-across	-0.103 (0.013	0.416 (0.009)	0.449 (0.005)	0.184 (0.011)	2.073	1 (0)
	INT-across2	-0.967 (0.016)	0.5 (0.011)	0.509 (0.001)	1.184 (0.032)	13.337	0.672 (0.015)

Table A11. Sensitivity simulation results: bias, standard error, mean squared error (MSE), and coverage (using bootstrap standard error) results using ***MI-derPassive INT-within*** strategy (auxiliary variable was not included in the imputation model) when varying the number of subjects, missing rate, or number of multiply imputed datasets compared to main simulation study (Monte Carlo standard errors in parentheses). (MDM = missing data mechanism, m indicates the number of multiply imputed datasets). The results from the main simulations were highlighted in grey as reference.

MDM	m	Missing Rate (%)	Number of Subjects	Bias	Standard Error		MSE	Coverage
					Empirical	Bootstrap		
Varying the number of subjects								
MCAR	50	50	2000	-0.001 (0.01)	0.316 (0.007)	0.315 (0)	0.099 (0.004)	1 (0)
MCAR	50	50	1000	-0.005 (0.122)	0.545 (0.012)	0.557 (0.001)	0.297 (0.095)	1 (0)
MCAR	50	50	500	-0.005 (0.173)	0.772 (0.017)	0.786 (0.002)	0.596 (0.19)	1 (0)
MCAR	50	50	250	0.042 (0.253)	1.131 (0.025)	1.087 (0.004)	1.282 (0.417)	1 (0)
MAR1	50	50	2000	0.038 (0.014)	0.454 (0.01)	0.446 (0.001)	0.207 (0.011)	1 (0)
MAR1	50	50	1000	0.11 (0.14)	0.626 (0.014)	0.63 (0.002)	0.404 (0.135)	1 (0)
MAR1	50	50	500	0.22 (0.192)	0.859 (0.019)	0.878 (0.005)	0.786 (0.255)	1 (0)
MAR1	50	50	250	0.374 (0.289)	1.294 (0.029)	1.138 (0.005)	1.814 (0.653)	1 (0)
MAR2A	50	50	2000	1.356 (0.012)	0.393 (0.009)	0.367 (0.001)	1.994 (0.037)	0 (0)
MAR2A	50	50	1000	1.478 (0.148)	0.659 (0.015)	0.676 (0.004)	2.619 (0.469)	1 (0)
MAR2A	50	50	500	1.516 (0.208)	0.931 (0.021)	0.939 (0.007)	3.165 (0.757)	0 (0)
MAR2A	50	50	250	1.624 (0.303)	1.353 (0.03)	1.195 (0.007)	4.466 (1.284)	0 (0)
MNAR	50	50	2000	1.369 (0.013)	0.404 (0.009)	0.368 (0.001)	2.038 (0.038)	0 (0)
MNAR	50	50	1000	1.471 (0.146)	0.653 (0.015)	0.675 (0.004)	2.591 (0.459)	0 (0)
MNAR	50	50	500	1.531 (0.209)	0.933 (0.021)	0.947 (0.007)	3.214 (0.784)	0.995 (0.002)
MNAR	50	50	250	1.651 (0.303)	1.354 (0.03)	1.214 (0.007)	4.56 (1.295)	0 (0)
Varying missing rate								
MCAR	50	50	2000	-0.001 (0.01)	0.316 (0.007)	0.315 (0)	0.099 (0.004)	1 (0)
MCAR	25	25	2000	0.006 (0.057)	0.361 (0.008)	0.37 (0.001)	0.13 (0.029)	1 (0)
MCAR	10	10	2000	0.002 (0.036)	0.361 (0.008)	0.363 (0.001)	0.13 (0.018)	1 (0)
MAR1	50	50	2000	0.038 (0.014)	0.454 (0.01)	0.446 (0.001)	0.207 (0.011)	1 (0)

MAR1	25	25	2000	0.021 (0.062)	0.39 (0.009)	0.389 (0.001)	0.153 (0.035)	1 (0)
MAR1	10	10	2000	0.005 (0.036)	0.362 (0.008)	0.365 (0.001)	0.131 (0.019)	1 (0)
MAR2A	50	50	2000	1.356 (0.012)	0.393 (0.009)	0.367 (0.001)	1.994 (0.037)	0 (0)
MAR2A	25	25	2000	0.779 (0.061)	0.383 (0.009)	0.379 (0.001)	0.753 (0.101)	0.05(0.007)
MAR2A	10	10	2000	0.202 (0.035)	0.355 (0.008)	0.358 (0.001)	0.167 (0.023)	1 (0)
MNAR	50	50	2000	1.369 (0.013)	0.404 (0.009)	0.368 (0.001)	2.038 (0.038)	0 (0)
MNAR	25	25	2000	0.759 (0.06)	0.379 (0.008)	0.378 (0.001)	0.721 (0.097)	0.066 (0.008)
MNAR	10	10	2000	0.198 (0.035)	0.352 (0.008)	0.358 (0.001)	0.163 (0.023)	1 (0)
Varying the number of multiply imputed dataset, m								
MCAR	50	50	2000	-0.001 (0.01)	0.316 (0.007)	0.315 (0)	0.099 (0.004)	1 (0)
MCAR	10	50	2000	0.032 (0.039)	0.391 (0.009)	0.399 (0.001)	0.154 (0.022)	1 (0)
MAR1	50	50	2000	0.038 (0.014)	0.454 (0.01)	0.446 (0.001)	0.207 (0.011)	1 (0)
MAR1	10	50	2000	0.094 (0.046)	0.461 (0.01)	0.454 (0.001)	0.221 (0.034)	1 (0)
MAR2A	50	50	2000	1.356 (0.012)	0.393 (0.009)	0.367 (0.001)	1.994 (0.037)	0 (0)
MAR2A	10	50	2000	1.495 (0.048)	0.482 (0.011)	0.481 (0.002)	2.467 (0.152)	0 (0)
MNAR	50	50	2000	1.369 (0.013)	0.404 (0.009)	0.368 (0.001)	2.038 (0.038)	0 (0)
MNAR	10	50	2000	1.49 (0.048)	0.48 (0.011)	0.481 (0.002)	2.449 (0.151)	0.811 (0.012)