American Journal of Epidemiology

Vol. 191, No. 3
https://doi.org/10.1093/aje/kwab271
Advance Access publication:
November 10, 2021

Practice of Epidemiology

SuperMICE: An Ensemble Machine Learning Approach to Multiple Imputation by Chained Equations

Hannah S. Laqueur*, Aaron B. Shev, and Rose M. C. Kagawa

* Correspondence to Dr. Hannah S. Laqueur, Department of Emergency Medicine, School of Medicine, University of California, Davis, 2315 Stockton Boulevard, Sacramento, CA 95817 (e-mail: hslaqueur@ucdavis.edu).

Initially submitted October 7, 2020; accepted for publication November 8, 2021.

Researchers often face the problem of how to address missing data. Multiple imputation is a popular approach, with multiple imputation by chained equations (MICE) being among the most common and flexible methods for execution. MICE iteratively fits a predictive model for each variable with missing values, conditional on other variables in the data. In theory, any imputation model can be used to predict the missing values. However, if the predictive models are incorrectly specified, they may produce biased estimates of the imputed data, yielding inconsistent parameter estimates and invalid inference. Given the set of modeling choices that must be made in conducting multiple imputation, in this paper we propose a data-adaptive approach to model selection. Specifically, we adapt MICE to incorporate an ensemble algorithm, Super Learner, to predict the conditional mean for each missing value, and we also incorporate a local kernel-based estimate of variance. We present a set of simulations indicating that this approach produces final parameter estimates with lower bias and better coverage than other commonly used imputation methods. These results suggest that using a flexible machine learning imputation approach can be useful in settings where data are missing at random, especially when the relationships among the variables are complex.

machine learning; missing data; missingness at random; multiple imputation by chained equations; simulation

Abbreviations: BLR, Bayesian linear regression; CART, classification and regression trees; LASSO, least absolute shrinkage and selection operator; LOESS, locally estimated scatterplot smoothing; MICE, multiple imputation by chained equations; PMM, predictive mean matching.

Missing data are a common problem in quantitative research, and improper handling of missing data can lead to biased parameter estimates, decreased statistical power, and less generalizable findings (1). A range of approaches are available to address missing data. A straightforward approach, one that is often the default approach of many statistical analysis packages, is to simply remove all rows of data that have any missing values (listwise deletion). It is well-established, however, that such an approach is suboptimal: It will result in a substantial loss of power if there is a large number of missing values, and it can lead to biased parameter estimates when the data are not missing completely at random (2).

An alternative solution is to replace missing values with imputed values. Imputation maintains the full sample size and requires the less stringent assumption that data are missing at random—that is, missingness that is random after conditioning on the observed values of other variables. Multiple imputation is often preferred over single imputation, because, unlike single imputation, it accounts for the uncertainty involved in imputing data. Multiple imputation uses the distribution of the observed data to estimate a set of plausible values for the missing data and uses variability in the set of estimated values to calculate more appropriate standard errors for parameter estimates (3).

Even when the missing-at-random assumption of multiple imputation is met, consistent parameter estimates and valid confidence intervals still depend on the correct specification of the imputation model (4, 5). Given the set of modeling choices that need to be made, we propose a data-adaptive ensemble learning approach to better detect and capture complex relationships in the imputation models and mitigate

issues related to misspecification error in the imputed values (6). Specifically, we adapt the widely used Multiple Imputation by Chained Equations (MICE) package (7) in R (R Foundation for Statistical Computing, Vienna, Austria) to incorporate the Super Learner (8, 9) stacking ensemble algorithm to predict the conditional mean for each missing value and a local kernel-based estimate of variance. Super Learner uses *k*-fold cross-validation to estimate the performance of multiple base learning models and creates an optimal weighted average of the models using the test data performance. Super Learner has been proven to be asymptotically at least as accurate as the best-performing individual prediction algorithm tested in the ensemble (8). In this paper, we compare the implementation of MICE with Super Learner to 2 standard approaches: MICE with Bayesian linear regression (BLR) and MICE with predictive mean matching (PMM). We present a series of simulation results and an applied example using the National Crime Victimization Survey (10).

The present study is focused on multiple imputation, an increasingly popular technique for dealing with missing data; however, multiple imputation is just one approach among several for estimating parameters when data are missing at random. Two other common methods are inverse probability weighting (11) and likelihood-based approaches (12). In-depth reviews of these approaches and their comparative value can be found elsewhere (13, 14). In brief, inverse probability weighting involves reweighting each complete-case observation by the inverse of its estimated probability of being complete given predictors (15). Likelihood-based approaches, on the other hand, fit the statistical model of interest directly from the observed data without deleting observations. A common likelihood approach is maximum likelihood via expectation maximization (16). The expectation maximization algorithm iterates between an E-step, computing the expected value of the full data log-likelihood given the observed data and a set of initial parameter estimates, and the M-step, which performs maximum likelihood estimation of the parameters using the augmented log-likelihood obtained from the E-step.

Unlike weighting or likelihood-based approaches, multiple imputation involves generating multiple data sets in which missing values are imputed or filled in based on the distribution of observed data. There are a number of methods for implementing multiple imputation, but all involve 3 basic steps. First, the researcher generates *m* imputed data sets by replacing missing data in each variable with values randomly drawn from a posterior predictive distribution of the missing data conditional on the observed data *m* times. Second, the researcher performs the intended statistical analyses on each of the imputed data sets, thereby obtaining *m* estimates and *m* standard errors. Finally, the estimates and standard errors are combined using Rubin's rules (17), which involves pooling the *m* parameter estimates and combining the conventional sampling variance (within-imputation variance) with the additional variance generated by the missing data (between-imputation variance). By accounting for the uncertainty of the imputations, multiple imputation produces more accurate standard errors than single imputation. If there is limited information in the observed data used in the imputation model, the imputations will be highly variable, leading to high standard errors in the analyses; if the observed data are highly predictive of the missing values, the imputations will be more consistent across imputations, resulting in smaller standard errors (3).

The 2 general approaches to implementing multiple imputation are joint modeling and fully conditional specification or MICE (18, 19), the latter of which is the focus of the present paper. Joint modeling assumes a multivariate normal distribution, and imputations are generated as draws from the fitted distribution. MICE, on the other hand, imputes missing values using separate univariate conditional distributions for each incomplete variable given all the others, cycling iteratively through each incomplete variable. Specifically, MICE proceeds as follows: First, the missing values for each variable are imputed using a simple approach such as mean imputation; second, one variable at a time, the imputed values are set back to missing and the missing values are predicted using the observed and imputed values for the variables in the data set and a user-specified imputation model. One iteration consists of 1 cycle through each variable in the data set; multiple cycles are performed.

The MICE algorithm requires model specification for each incomplete variable as well as the selection of the predictors to be included in the variable imputations. Standard practice suggests that the imputation model should include all variables that will be included in the analysis model, as well as all variables thought to be predictive of each imputed variable so as to make the missing-at-random assumption more plausible (12). In practice, however, with high-dimensional data sets, the inclusion of variables beyond those used in the analysis is not always feasible (20, 21).

With respect to the imputation models, model misspecification can lead to biased estimates. The threat of bias must also be balanced with variance concerns. More complex and flexible algorithms—for example, random forest—tend to produce higher variance and less bias and are prone to overfitting the data, while simpler algorithms with more rigid structure tend to produce lower variance at the expense of accuracy because of underfitting. Because the true functional form of the imputation model is rarely known, the researcher must make a decision about how to balance bias and variance in their estimation procedure. One advantage of the Super Learner algorithm is that the researcher can supply the ensemble with a range of diverse base learners, including simple parametric models as well as flexible data-adaptive algorithms.

There have been some prior efforts to incorporate decision trees, a popular class of machine learning algorithms, into multiple imputation so as to better model nonlinear relationships (22, 23), and results have generally been promising. Implementations of classification and regression trees (CART) (24) and random forest (25) are both now available in the MICE R package (7), but more general implementations of machine learning algorithms for imputation have focused on producing only single imputations (6). To our knowledge, the present paper is the first to incorporate an ensemble machine learning approach to multiple imputation.

METHODS

The MICE algorithm is modular by design. That is, MICE requires a method of producing a random imputed value, but this method is left to the user to define. A good method for MICE would produce a sampling distribution for each missing value that reasonably reflects available information about the conditional expected value as well as the uncertainty inherent to the estimate. For example, BLR samples from a normal distribution with mean and variance determined by the regression model; PMM samples a value from a pool of neighbors who have a similar conditional expectation. In our simulations, we compare MICE imputation using Super Learner to MICE using PMM and MICE using BLR. We briefly describe each approach below.

MICE with BLR and PMM

BLR is a common and straightforward approach for multiple imputation of continuous variables with a normal distribution (6, 12). In the MICE R package, for example, the normal method of imputation ("norm" function) is based on a BLR involving first regressing z on x to estimate coefficients, β, and then drawing from the posterior predictive distribution of β at random to produce β^*. This set of coefficients is then used to predict values for missing observations of z. However, for nonnormal variables and nonlinear relationships, linear regression may fail to produce accurate predictions.

PMM is a semiparametric imputation approach. PMM samples values from the observed data to impute missing values (26). For a variable with missing values, z, and a set of predictor variables, x, PMM performs a simple linear regression of z on the set of predictor variables, x, to estimate a set of coefficients, β. It then draws from the posterior predictive distribution of β at random and uses this new set of coefficients, β^*, to predict all values of z, both missing and observed. For each missing value, PMM finds the n observed z whose predicted values are closest to the predicted value of missing z and assigns the observed value of a random draw of these values to the missing value. This is repeated for each complete data set.

One of the benefits of PMM over standard methods based on linear regression is that the distribution of the imputed data will better reflect the distribution of the observed data, be it skewed, bounded, discrete, etc. PMM does not rely on formal theory but has generally performed well in simulations (27). However, the method also has limitations. It will be less successful than a parametric approach when predicting values where the observed data are sparse or nonexistent (28). PMM relies on drawing a nearest neighbor and thus cannot extrapolate beyond the range of data or interpolate in areas of the data structure where there are no observations from which to choose (29). Similarly, small sample sizes can pose a challenge, as there are fewer observed data from which to draw (30).

Super Learner

Super Learner is an ensemble machine learning method that constructs a predictive model from a weighted combination of candidate base learner algorithms. The base algorithms are user-specified and may include any number of semiparametric and nonparametric models, such as generalized additive models (31), neural networks (32), and random forest (25), as well as parametric models such as logistic or linear regression. Super Learner uses k-fold cross-validation to estimate the predictive performance of each of the base learner algorithms. Super Learner then finds the optimal weighted combination of base learners via a second "meta-learning" step that minimizes the cross-validated error with respect to a user-defined loss function such as prediction error or negative log-likelihood. This weighted combination of algorithms is used to generate a "super" prediction function (predicted values are generated on the out-of-fold predictions from each of the base models) and combined via a weighted average applied to the full data set.

This process optimizes the bias-variance trade-off for a given prediction question and set of algorithms, and theoretical results have shown that the weighted collection of algorithms will perform asymptotically as well as or better than any single candidate algorithm (8). The recommendation is to supply diverse algorithms that represent a range of algorithm flexibility along with options to test for higher-order interactions among the variables, polynomials and other transformations, and screener algorithms (e.g., least absolute shrinkage and selection operator (LASSO) (33)) to remove variables and transformations that do not contribute meaningfully to the prediction model.

Variance estimation

Super Learner models are unable to provide standard errors for their predictions due to their semiparametric construction. To create a sampling distribution for imputed values centered around a Super Learner prediction, we use a local estimate of variance. Local imputation methods, as introduced in the multiple imputation context by Titterington and Sedransk (34) and further developed by Aerts et al. (35), like PMM, relax distributional assumptions. Local imputation methods work by sampling from a kernel-based estimate of the underlying distribution or by taking a smoothed local bootstrap sample.

Local imputation methods may assume that conditional distributions are locally normal, but more flexibly reflect nonconstant errors. For our purposes, this is important in situations where the variability of an imputed variable is nonconstant across values of another associated variable. In contrast, BLR assumes a constant variance for random error, which is utilized in its sampling distribution for imputed values, and a violation of this assumption may lead to inaccurate predictions. PMM can be interpreted as a special case of local imputation: A uniform kernel with a given bandwidth is chosen to put equal weight on some number of the closest neighbors, and then imputed values are sampled from the empirical density function.

Combining Super Learner and MICE: SuperMICE

We created a new R package for incorporating the Super Learner algorithm into MICE (SuperMICE). The package is available on GitHub (36).

We take a semiparametric approach to imputing missing data using Super Learner and MICE, generating values for missing data from a normal distribution with a mean determined from a Super Learner model and a local estimate of the variance following the strategy laid out by Aerts et al. (35).

Consider a data set, $\mathbf{X} = (x_1, x_2, \ldots, x_p)$, of length n binary and numerical vectors where values may be missing in any of the variables. The ordering of the variables is not important. In addition, define an $n \times p$ binary matrix, $\boldsymbol{\delta} = [\delta_{ij}]$, where

$$\delta_{ij} = \begin{cases} 0 & x_{ij} \text{ is missing;} \\ 1 & x_{ij} \text{ is observed.} \end{cases}$$

Furthermore, let $\mathcal{L}$ be a library of predictive algorithms with cardinality q. We denote the iteration of the algorithm by $0 \leq t \leq T$. Finally, we will denote the kth data set in the tth iteration of the algorithm by the superscript (k, t). In the remainder of this section, we describe the steps of MICE using Super Learner.

Step 1: initialize. Create m identical data sets, $\mathbf{X}^{(1,0)}$, $\mathbf{X}^{(2,0)}, \ldots, \mathbf{X}^{(m,0)}$, as follows. For each missing value, $\delta_{ij} = 0$, set $x_{ij} = 0$ and then impute using the mean (round to 0 or 1 if binary) of the observed values for the corresponding variable, $\mathbf{x}_i$. That is,

$$x_{ij}^{(k,0)} = \frac{\sum_{j=1}^{n} x_{ij}\delta_{ij}}{\sum_{j=1}^{n} \delta_{ij}}.$$

Step 2: predict. For data set $1 \leq k \leq m$ and for variable $1 \leq j \leq p$, such that j has at least 1 missing value, fit each algorithm $l \in \mathcal{L}$, predicting $\mathbf{x}_j^{(k,t)}$ from $\mathbf{X}_{(j)}^{(k,t)}$, the data with j removed. The Super Learner (SL) predictions are given by the weighted sum of predictions generated by the set of algorithms in the library

$$\hat{\mathbf{x}}_j^{(k,t)} = \hat{\Psi}_{\mathrm{SL},j}^{(k,t)}\left(\mathbf{X}_{(j)}^{(k,t)}\right) = \sum_{l=1}^{q} \hat{\alpha}_{jl}^{(k,t)} \hat{\Psi}_{jl}^{(k,t)}\left(\mathbf{X}_{(j)}^{(k,t)}\right),$$

where $\Psi(\cdot)$ is a function returning predictions from an algorithm and the set of estimated weights, $\hat{\boldsymbol{\alpha}}_j = (\hat{\alpha}_{j1}, \hat{\alpha}_{j2}, \ldots, \hat{\alpha}_{jq})$ such that $\sum_{l=1}^{q} \hat{\alpha}_{l,j} = 1$ are obtained by minimizing the cross-validated risk (for details, see Polley and van der Laan (37)).

Step 3: impute. We use a semiparametric approach to impute the missing data. In the binary case, values are sampled from a Bernoulli distribution with probabilities equal to the Super Learner predictions. For continuous variables, missing-data values are randomly sampled from a normal distribution with mean given by the Super Learner prediction and a local estimate of variance. The local variance is computed as a weighted variance giving more weight to observations with predicted values similar to the missing value and to observations in areas of greater missingness. Specifically, for a kernel function $K_h(\cdot)$ with bandwidth h, the weights are computed as $w_{ij}(x) = \delta_{ij} w_{ij}(x)/\hat{\pi}(x)$, where

$$w_{ij}(x) = \frac{K_h(x - \hat{x}_{ij}^{(k,t)})}{\sum_{i=1}^{n} K_h(x - \hat{x}_{ij}^{(k,t)})}$$

are the usual Nadaraya-Watson weights (38, 39) and $\hat{\pi}(x) = K_h(x - \hat{x}_{ij}^{(k,t)})\delta_{ij} / \sum_{i=1}^{n} K_h(x - \hat{x}_{ij}^{(k,t)})$ is a local estimate of missingness. The choice of bandwidth is important for producing a good result. Aerts et al. (35) propose a bandwidth selection method using the jackknife (40), while we have found that a bandwidth that captures 2% of observed values within 1 standard deviation under the gaussian kernel performs well. This remains an area of active research.

The weighted variance can then be calculated as $\hat{\sigma}_{ij}^{2(k,t)} = \sum_{i=1}^{n} w_{ij}(\hat{x}_{ij}^{(k,t)})(\hat{x}_{ij}^{(k,t)} - \hat{x}_{ij}^{*(k,t)})^2 / \sum_{i=1}^{n} w_{ij}(\hat{x}_{ij})$, where $\hat{x}_{ij}^{*}$ is the weighted mean. Finally, each missing value is sampled from the corresponding distribution,

$$x_{ij}^{(k,t+1)} \sim \mathcal{N}\left(\hat{x}_{ij}^{(k,t)}, \hat{\sigma}_{ij}^{2(k,t)}\right).$$

For T iterations (usually 10–20 is sufficient (7)), repeat steps 2 and 3.

Step 4: estimate. Estimation and pooling of the estimates and standard errors from each of the m imputed data sets occurs following standard MICE procedure: 1) statistical analyses (e.g., regression) are carried out on each imputed data set as would have been done if the data had been complete, and 2) estimates from the analyses are then combined via Rubin's rules (17). That is, if Q represents the quantity of interest and $[\hat{Q}^{(j)}, U^{(j)}], j = 1, \ldots, m$, represents the set of parameter estimates and standard errors from each of the imputed data sets, the overall combined parameter estimate is given by the simple average $\overline{Q} = \frac{1}{m}\sum_{j=1}^{m} \hat{Q}^{j}$ and the total variance is defined as $T = \overline{U} + B + B/m$, where $\overline{U} = \frac{1}{m}\sum_{j=1}^{m} U^{j}$ is the within-imputation variance and the between-imputation variance is calculated as $B = \frac{1}{m-1}\sum_{j=1}^{m} [\hat{Q}^{j} - \overline{Q}]^2$.

SIMULATIONS

Simulation design

We conducted 4 simulations ($n = 1{,}000$) to compare MICE with Super Learner to MICE using PMM and MICE using BLR. We conducted each of the simulations using 4 sample sizes ($n = 100$, $n = 400$, $n = 700$, and $n = 1{,}000$), and 5 levels of missingness (10%, 20%, 30%, 40%, and 50%). Our Super Learner library in the first 3 simulations included the global mean, the general linear model, locally estimated scatterplot smoothing (LOESS), the generalized additive model, and neural networks; we exchanged LOESS for linear discriminant analysis in the fourth simulation to better accommodate binary variables. We generated 30 imputed data sets ($m = 30$) for each scenario and ran MICE

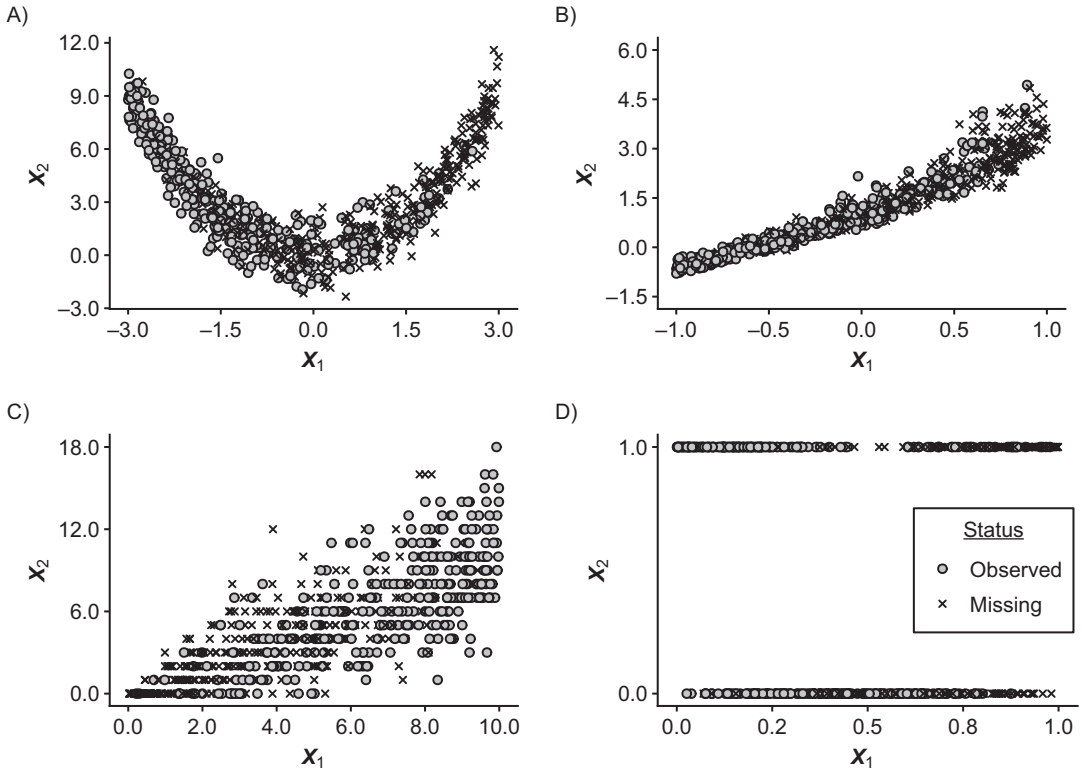

Figure 1. Examples of values simulated for X_1 and X_2 in 4 simulations comparing multiple imputation by chained equations (MICE) with Super Learner to MICE using predictive mean matching and MICE using Bayesian linear regression (i.e., quadratic relationship, lognormal relationship, zero-inflated Poisson distribution, and binary relationship). Missing values, indicated by multiplication signs (×), increase (or decrease, in the case of simulation 3) in probability with increasing values of X_1.

for 10 iterations, following the suggestion of van Buuren and Groothuis-Oudshoorn (7).

In each of the simulations, X_2 contains missing data, and the probability of missing values increases with the value of X_1. Specifically, $\Pr(\delta_{i,2} = 0) = \Pr(X_1 < x_{i,1})$. Thus, we have generated scenarios in which the data are missing at random. There are systematic differences between the missing and observed values, but these can be entirely explained by other observed variables. Figure 1 displays the relationships between X_1 and X_2 in each of the scenarios.

In the first simulation, we create a quadratic relationship between X_1 and X_2 taking the following form:

$$\begin{aligned} X_2 &= 1 + X_1 + X_1^2 + \epsilon_1 \\ Y &= X_1 + X_2 + \epsilon_2, \\ \text{where } X_1 &\sim \text{uniform}\,(-3, 3) \\ \epsilon_i &\sim \mathcal{N}(0, 1) \text{ for } i = 1, 2. \end{aligned} \tag{1}$$

In the second simulation, we create a lognormal relationship between X_1 and X_2 taking the following form:

$$\begin{aligned} X_2 &= e^{X_1 + \epsilon_1} \\ Y &= X_1 + X_2 + \epsilon_2, \\ \text{where } X_1 &\sim \text{uniform}\,(0, 1) \\ \epsilon_i &\sim \mathcal{N}(0, 1) \text{ for } i = 1, 2. \end{aligned} \tag{2}$$

The third simulation samples values from a zero-inflated Poisson distribution parameterized by the value of X_1 as follows:

$$\begin{aligned} X_2 \mid X_1 &\sim \begin{cases} 0 & \text{with } \Pr\left(e^{-X_1}\right) \\ \text{Poisson}\,(\lambda = X_1) & \text{with } \Pr\left(1 - e^{-X_1}\right) \end{cases} \\ Y &= X_1 + X_2 + \epsilon, \\ \text{where } X_1 &\sim \text{uniform}\,(0, 5) \\ \epsilon &\sim \mathcal{N}(0, 1)\,. \end{aligned} \tag{3}$$

Finally, the fourth simulation is a binary relationship and takes the following form:

$$\begin{aligned} Y &= X_1 + X_2 + \epsilon, \\ \text{where } X_1 &\sim \text{uniform}\,(0, 1) \\ X_2 &\sim \text{Bernoulli}\,(\,|2X_1 - 1|\,) \\ \epsilon &\sim \mathcal{N}(0, 1)\,. \end{aligned} \tag{4}$$

For each scenario, we imputed data using MICE with BLR, PMM, and Super Learner. We then fitted a linear regression model of the form $Y = X_1 + X_2$ using the imputed data sets and estimated the regression coefficients for X_1 and X_2. In the following section, we report the bias and coverage probability of a 95% confidence interval for the estimated coefficients of X_2, the variable for which data are missing at random. We show the mean squared error for

X_2 in Web Figure 1 (available at https://doi.org/10.1093/aje/kwab271) and the bias, coverage, and mean squared error for X_1 in Web Figure 2. A Gaussian kernel was used for all simulations. Bandwidth was selected to be the minimum value that included 1 observation within 1 standard deviation for simulations with a sample size of 100 and 7 observations within 1 standard deviation for simulations of sample size 1,000.

Simulation results

Bias. As Figure 2 shows, overall we find lower bias in the estimation of β_2 under all scenarios when imputation is done with MICE with Super Learner as compared with MICE with PMM or MICE with BLR. The bias reduction from Super Learner is most apparent in the quadratic scenario. With a sample size of 1,000, all 3 approaches are equivalent until we reach 30% missingness, at which point MICE with Super Learner has lower bias than the other 2 imputation methods. With a sample size of only 100, PMM has higher bias at all levels of missingness. MICE imputation with BLR is equivalent to MICE imputation with Super Learner with low levels of missingness; with 20% or more missingness, MICE with Super Learner has the lowest bias.

Under both the lognormal scenario and the binary scenario, MICE with Super Learner and BLR perform comparably with respect to bias in β_2, whereas PMM performs substantially less well. PMM shows particularly poor performance when missingness exceeds 30%–40%.

Finally, under the zero-inflated Poisson scenario, MICE with Super Learner has lower levels of bias than the other approaches, with levels of missingness above 30% or 40%, depending on the sample size. When n = 1,000 and missingness is below 20%, the other 2 approaches very slightly outperform MICE with Super Learner.

Coverage. Under all scenarios, with respect to coverage, MICE with Super Learner performs well relative to the other approaches. Using MICE with Super Learner, coverage for the 95% confidence interval for β_2 is close to the stated 95% for all levels of missingness with sample sizes of 100, 400, 700, and 1,000. Under the quadratic and zero-inflated Poisson scenarios, both PMM and BLR perform poorly when missingness is greater than 20%–30%. With a sample size of 1,000 and 50% of values missing, coverage using PMM is as low as 15% in the quadratic scenario and below 65% in the zero-inflated Poisson scenario; BLR coverage falls as low as 15% in both the quadratic and zero-inflated Poisson scenarios. Under the lognormal and binary scenarios, Super Learner and BLR generally perform comparably, while coverage for PMM performance begins to drop with missingness above 30%. When missingness is as high as 50%, coverage using PMM falls below 40% in the lognormal scenario and under 80% in the binary scenario.

APPLIED EXAMPLE: ANALYSIS OF NATIONAL CRIME VICTIMIZATION SURVEY

We applied SuperMICE to a real data set, the National Crime Victimization Survey (10). The National Crime Victimization Survey is a nationally representative survey on crime victimization administered by the Bureau of Justice Statistics (US Department of Justice). Our sample included all respondents who had personally experienced a violent victimization between the years 2008 and 2015 (n = 1,313).

We estimated the association between victimization with a firearm (vs. victimization with another weapon or no weapon) and the presence of self-reported psychosomatic problems (headaches, trouble sleeping, changes in eating or drinking habits, upset stomach, fatigue, high blood pressure, muscle tension or back pain, or some other physical problem). The presence and type of weapon involved in the violent victimization were reported by the respondent in response to the questions, "Did the offender have a weapon such as a firearm or knife, or something to use as a weapon, such as a bottle or wrench?" and "What was the weapon?". Psychosomatic problems were reported in response to a series of questions beginning, "Did you experience any of the following physical problems associated with being a victim of this crime for a month of more?" Separate questions were asked for each of 8 possible psychosomatic problems. For our analysis, we created a binary variable indicating whether the respondent reported experiencing any of the listed problems versus none. Covariates included sex, race, age, educational level, household income, and whether or not the perpetrator was a stranger.

The data set had some missingness. Reporting on the weapon type was missing in 10% of instances. Data on several additional covariates were missing: Educational level and household income data were both missing in 15% of instances; the relationship between the perpetrator and the victim was missing in 3% of instances. Reports of psychosomatic problems were missing in 0.3% of observations. We present a visualization of missingness patterns in Web Figure 3.

We used logistic regression to estimate the odds of experiencing a psychosomatic problem associated with victimization with a firearm. We used the 3 different MICE methods to fill in missing values: BLR, PMM, and Super Learner. We also fitted the logistic regression model without imputation, which by default applies listwise deletion to the rows of data with missing values. Results are shown in Table 1.

We included the following algorithms in our SuperMICE imputation: random forest, neural networks, LASSO, generalized linear models, and an algorithm that simply predicts the marginal mean value. Web Table 1 shows the average weights across imputations for each base learner and each imputed variable.

As Table 1 shows, we found a statistically significant relationship between victimization with a gun and self-reported psychosomatic problems. The log odds ranged from 1.73 to 1.88 across methods. Not surprisingly, the confidence interval was largest for the logistic regression model implemented without imputation, which, by default, deletes rows that have missing values. As noted above, listwise deletion will lead to a loss of power (and can lead to biased parameter estimates when the data are not missing completely at random). In this case, the coefficient remained statistically significant and comparable to the estimates calculated using multiple imputation.

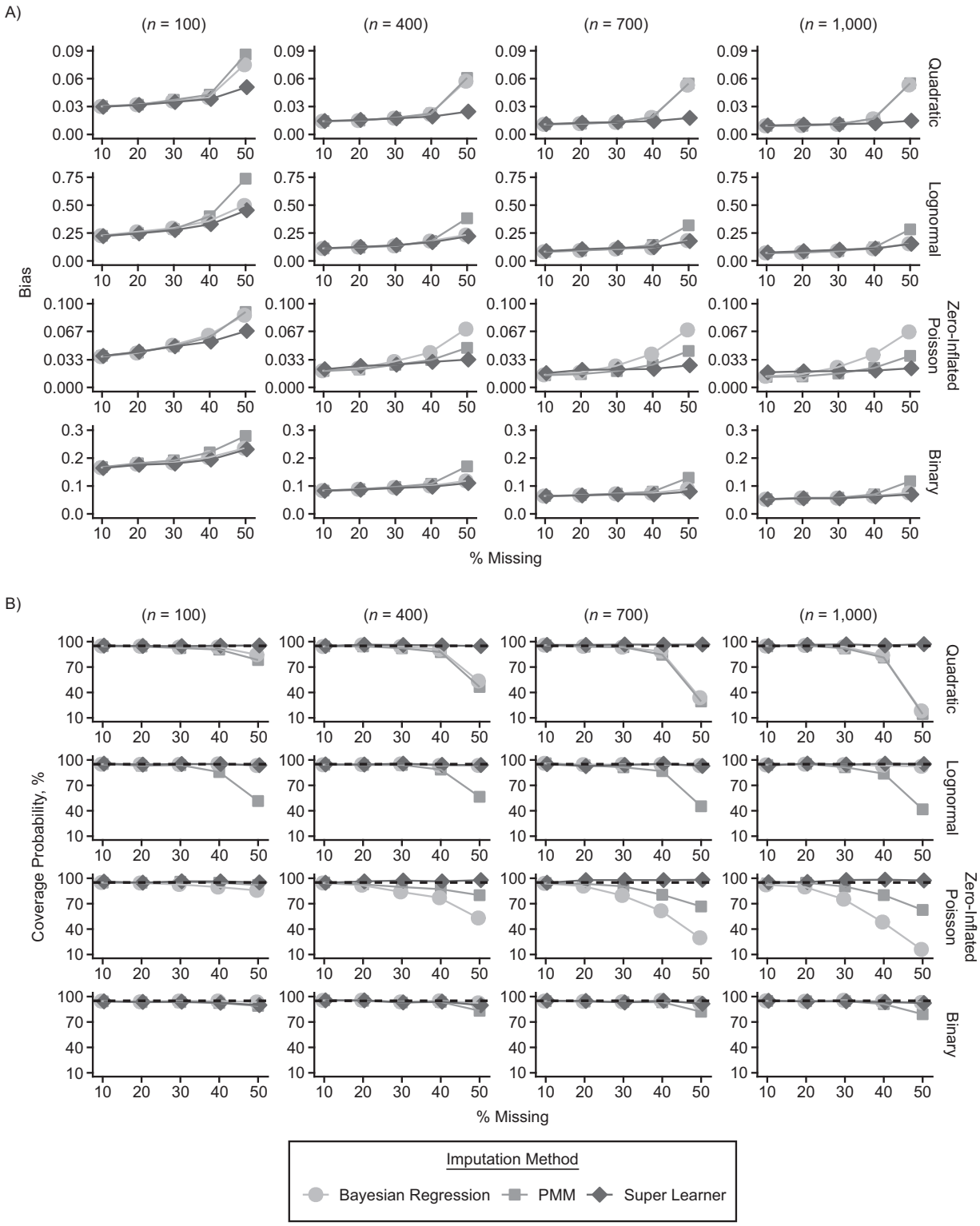

Figure 2. Simulation estimates for the bias (A) and coverage (B) of $\hat{\beta}_2$.The bias plots show the average absolute value of the bias across 1,000 simulations. The coverage plots show the percentage of 95% confidence intervals for β_2 that contain the true value, $\beta_2 = 1$, as an estimate of coverage probability.

Table 1. Odds of Experiencing a Psychosomatic Problem Associated With Firearm Victimization Derived Using Different Multiple Imputation Methods

Imputation Method	Odds Ratio	95% CI	*P* Value
Super Learner imputation	1.73	1.10, 2.73	0.01
PMM imputation	1.86	1.17, 2.94	0.01
Linear imputation	1.79	1.13, 2.84	0.01
No imputation	1.88	1.14, 3.19	0.02

Abbreviations: CI, confidence interval; PMM, predictive mean matching.

DISCUSSION

Super Learner allows for a rich and diverse user-specified library of prediction and screening algorithms, affording greater protection against misspecification of the imputation model. Our simulations showed the Super Learner method to generally have lower bias and better coverage than BLR and PMM.

The simulations demonstrated the potential for producing biased estimates when the imputation model is misspecified, especially under high levels of missingness. By supplying a flexible library of diverse algorithms spanning the bias-variance spectrum, MICE equipped with Super Learner remained more robust to problems of misspecification. Super Learner estimates are constructed as weighted averages of high-bias, low-variance methods, such as linear regression, and low-bias, high-variance methods, such as generalized additive models. The weights are chosen to minimize cross-validated risk, which is at least as small as the component method with the lowest risk. Thus, Super Learner balances bias and variance to find a more optimal estimator.

MICE equipped with Super Learner maintained near 95% coverage in all scenarios and had the smallest bias of the 3 methods. By contrast, in nearly all simulation scenarios for PMM and in the quadratic and zero-inflated Poisson scenario for BLR, the coverage probability for β_2 fell well below the stated 95%. The actual coverage fell as low as 15% in the quadratic scenario for both PMM and BLR. Additionally, bias increased with missingness in all scenarios for both PMM and BLR. While not the focus of the simulations presented, when the coefficient of β_1 was considered as well, Super Learner also greatly outperformed the other 2 methods (shown in Web Figure 2).

PMM had particularly large bias with high levels of missingness as neighbors were more distanced; BLR, on the other hand, performed poorly in cases where a linear function could not accurately approximate the true relationship. The regression approach does allow for specifying nonlinear forms a priori, and it will perform well when the model is correctly specified even under high missingness, but this requires user knowledge of all relationships to be modeled, which can be time-consuming and leaves open the potential for user error. Our Super Learner approach combines the benefits of both approaches with data-adaptive and semiparametric sampling of imputed values.

The case study analyses did not show meaningful differences in parameter estimates across imputation methods. The absence of difference across methods is probably explained by the fact that the data set we used had fairly minimal missing data. Information on our outcome was rarely missing (0.3%), and the key parameter of interest was missing in 10% of instances. Statistical guidelines generally suggest that complete-case analyses may be performed if levels of missing data are below 5% (41, 42), and our simulations confirm that differences in the performance of imputation methods arise with higher levels of missingness.

Limitations

This study had several important limitations. First, while we found that MICE equipped with Super Learner was more robust in simulations where BLM and PMM performed poorly, these scenarios do not represent all potential data structures, and there are instances in which other approaches may be preferred. If it is known, for example, that a variable is close to normally distributed, a linear regression model can be appropriately used to predict the missing values for that variable. In fact, previous research shows that a correctly specified parametric model may outperform a correctly specified singly robust nonparametric model (43). However, often the researcher does not know a priori what the appropriate model is, and one of the advantages of implementation with Super Learner is that the user may adjust the library to include models that represent a range of complexity and flexibility: The user could, for example, adopt a fully parametric approach with a single parametric base learner in the library if there was reason to believe that such a model was appropriate. Further, the user can include screener algorithms, which may be useful in contexts where there are a large number of covariates and concerns regarding multicollinearity. Additional benefits may accrue from the use of a doubly robust approach, which previous studies suggest could further reduce bias in estimates (43, 44). A second limitation is that the current Super Learner package in R, upon which our adaptation of the MICE package relies, only allows for binary and continuous outcomes. Finally, and more broadly, we note that multiple imputation may not always be the best way to deal with missing data. As we discussed above, multiple imputation is just one approach to missingness.

Depending on the nature of missingness and the parameters to be estimated, other methods might be preferable.

Conclusion

When imputing data, even if the missing-at-random assumption is met, there remains a question as to the appropriate specification for the imputation model. Incorporating an ensemble machine learning approach into MICE provides a way to more flexibly model the functional form. This is important because if the predictive models of the imputed values are incorrectly specified, the mean and variance of the imputed values may be biased, yielding inconsistent parameter estimates.

Our approach extends and generalizes efforts to incorporate decision trees into multiple imputation to better model nonlinear relationships (22, 23). Super Learner allows for a broad and diverse set of machine learning algorithms to be included in the ensemble, and it has the virtue of performing asymptotically as well as or better than any of the single constituent algorithms (e.g., random forest or CART, if included in the library).

Our simulation results show that MICE with Super Learner produces lower bias and better coverage than other commonly used methods, particularly when a large proportion of data is missing. This suggests that incorporating a flexible machine learning approach at the imputation modeling stage can be useful for complex epidemiologic data sets.

ACKNOWLEDGMENTS

Author affiliations: Department of Emergency Medicine, School of Medicine, University of California, Davis, Sacramento, California, United States (Hannah S. Laqueur, Aaron B. Shev, Rose M. C. Kagawa); Violence Prevention Research Program, University of California, Davis, Sacramento, California, United States (Hannah S. Laqueur, Aaron B. Shev, Rose M. C. Kagawa); and University of California Firearm Violence Research Center, Sacramento, California, United States (Hannah S. Laqueur, Aaron B. Shev, Rose M. C. Kagawa).

This study was supported by funding from the University of California Firearm Violence Research Center.

The data sets used in this study are available from the corresponding author.

The views expressed in this article are those of the authors and do not reflect those of the University of California Firearm Violence Research Center.

Conflict of interest: none declared.

REFERENCES

1. Schafer JL, Olsen MK. Multiple imputation for multivariate missing-data problems: a data analyst's perspective. *Multivar Behav Res.* 1998;33:545–571.
2. Demissie S, LaValley MP, Horton NJ, et al. Bias due to missing exposure data using complete-case analysis in the proportional hazards regression model. *Stat Med.* 2003;22:545–557.
3. Azur MJ, Stuart EA, Frangakis C, et al. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res.* 2011;20:40–49.
4. Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychol Methods.* 2002;7:147–177.
5. Graham JW. Missing data analysis: making it work in the real world. *Annu Rev Psychol.* 2009;60:549–576.
6. van Buuren S. *Flexible Imputation of Missing Data*. 2nd ed. Boca Raton, FL: CRC Press; 2018.
7. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. *J Stat Softw.* 2011;45:1–67.
8. van der Laan MJ, Polley EC, Hubbard AE. Super Learner. *Stat Appl Genet Mol Biol.* 2007;6:25.
9. Polley EC. SuperLearner. https://github.com/ecpolley/SuperLearner. Updated November 20, 2021. Accessed March 28, 2021.
10. Bureau of Justice Statistics, Office of Justice Programs, US Department of Justice. National Crime Victimization Survey [United States], 2015. (ICPSR 36828). Version 2. https://www.icpsr.umich.edu/web/NACJD/studies/36448. Published July 23, 2020. Accessed July 23, 2020.
11. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc.* 1994;89:846–866.
12. Schafer JL. *Analysis of Incomplete Multivariate Data*. Boca Raton, FL: CRC Press; 1997.
13. Tsiatis A. *Semiparametric Theory and Missing Data*. New York, NY: Springer Science+Business Media; 2007.
14. Ibrahim JG, Chen M-H, Lipsitz SR, et al. Missing-data methods for generalized linear models: a comparative review. *J Am Stat Assoc.* 2005;100:332–346.
15. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res.* 2013;22:278–295.
16. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B Methodol.* 1977;39:1–22.
17. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Toronto, ON, Canada: John Wiley & Sons, Inc.; 2004.
18. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res.* 2007;16:219–242.
19. Raghunathan TE, Lepkowski JM, Van Hoewyk JH, et al. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv Methodol.* 2001;27:85–96.
20. Deng Y, Chang C, Ido MS, et al. Multiple imputation for general missing data patterns in the presence of high-dimensional data. *Sci Rep.* 2016;6:1–10.
21. Zhao Y, Long Q. Multiple imputation in the presence of high-dimensional data. *Stat Methods Med Res.* 2016;25:2021–2035.
22. Shah AD, Bartlett JW, Carpenter J, et al. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *Am J Epidemiol.* 2014;179:764–774.
23. Doove LL, van Buuren S, Dusseldorp E. Recursive partitioning for missing data imputation in the presence of interaction effects. *Comput Stat Data Anal.* 2014;72:92–104.
24. Breiman L, Friedman J, Stone CJ, et al. *Classification and Regression Trees*. Boca Raton, FL: CRC Press; 1984.
25. Breiman L. Random forests. *Mach Learn.* 2001;45:5–32.

26. Little RJ. Missing-data adjustments in large surveys. *J Bus Econ Stat.* 1988;6:287–296.
27. Marshall A, Altman DG, Royston P, et al. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC Med Res Methodol.* 2010;10:7.
28. Morris TP, White IR, Royston P. Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Med Res Methodol.* 2014;14:75.
29. Kleinke K. Multiple imputation under violated distributional assumptions: a systematic evaluation of the assumed robustness of predictive mean matching. *J Educ Behav Stat.* 2017;42:371–404.
30. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med.* 2011;30:377–399.
31. Hastie TJ, Tibshirani RJ. *Generalized Additive Models*. Boca Raton, FL: CRC Press; 1990.
32. Schmidhuber J. Deep learning in neural networks: an overview. *Neural Netw.* 2015;61:85–117.
33. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc B Methodol.* 1996;58:267–288.
34. Titterington DM, Sedransk J. Imputation of missing values using density estimation. *Stat Probab Lett.* 1989;8:411–418.
35. Aerts M, Claeskens G, Hens N, et al. Local multiple imputation. *Biometrika.* 2002;89:375–388.
36. Shev AB. superMICE. https://github.com/abshev/superMICE/. Published December 16, 2019. Accessed December 16, 2019.
37. Polley EC, van der Laan MJ. *Super Learner in Prediction. (U.C. Berkeley Division of Biostatistics Working Paper Series, paper 266)*. Berkeley, CA: University of California, Berkeley; 2010.
38. Nadaraya EA. *On estimating regression. Theory Probab Appl.* 1964;9:141–142.
39. Watson GS. Smooth regression analysis. *Sankhyā: Indian J Stat Ser A.* 1964;26:359–372.
40. Miller RG. The jackknife—a review. *Biometrika.* 1974;61(1):1–15.
41. Dong Y, Peng C-YJ. Principled missing data methods for researchers. *SpringerPlus.* 2013;2:222.
42. Jakobsen JC, Gluud C, Wetterslev J, et al. When and how should multiple imputation be used for handling missing data in randomised clinical trials—a practical guide with flowcharts. *BMC Med Res Methodol.* 2017;17:162.
43. Naimi AI, Kennedy EH. Nonparametric double robustness. *arXiv*. 2017. (https://arxiv.org/pdf/1711.07137v1.pdf). Updated November 23, 2021. Accessed March 28, 2021.
44. Long Q, Hsu C-H, Li Y. Doubly robust nonparametric multiple imputation for ignorable missing data. *Stat Sin.* 2012;22:149–172.