American Journal of Epidemiology

Vol. 172, No. 9
DOI: 10.1093/aje/kwq260
Advance Access publication:
September 14, 2010

Practice of Epidemiology

Multiple Imputation for Missing Data via Sequential Regression Trees

Lane F. Burgette* and Jerome P. Reiter

* Correspondence to Dr. Lane F. Burgette, Department of Statistical Science, Box 90251, Duke University, Durham, NC 27708 (e-mail: lb131@stat.duke.edu).

Initially submitted January 19, 2010; accepted for publication July 12, 2010.

Multiple imputation is particularly well suited to deal with missing data in large epidemiologic studies, because typically these studies support a wide range of analyses by many data users. Some of these analyses may involve complex modeling, including interactions and nonlinear relations. Identifying such relations and encoding them in imputation models, for example, in the conditional regressions for multiple imputation via chained equations, can be daunting tasks with large numbers of categorical and continuous variables. The authors present a nonparametric approach for implementing multiple imputation via chained equations by using sequential regression trees as the conditional models. This has the potential to capture complex relations with minimal tuning by the data imputer. Using simulations, the authors demonstrate that the method can result in more plausible imputations, and hence more reliable inferences, in complex settings than the naive application of standard sequential regression imputation techniques. They apply the approach to impute missing values in data on adverse birth outcomes with more than 100 clinical and survey variables. They evaluate the imputations using posterior predictive checks with several epidemiologic analyses of interest.

diagnostic check; imputation; missing data; pregnancy outcome; regression tree

Abbreviations: CART, classification and regression trees; MICE, multiple imputation by chained equations.

In large epidemiologic studies, data collection almost inevitably is plagued by missing data, for example, due to item nonresponse. One approach for handling missing data in such contexts is multiple imputation (1). Multiple imputation is appealing because it allows a team of researchers to address the missing data, after which any number of analyses may be performed by standard complete-data techniques. To carry out multiple imputation, the team fills in the missing values with draws from some predictive model *m* times, resulting in *m* completed data sets that can be used for the analysis. The analyst computes point and variance estimates of interest with each data set and combines these estimates (1). These formulas serve to propagate the uncertainty introduced by missing values through analysts' inferences (2). For reviews of multiple imputation, refer to several previously published articles (3–8).

A popular approach for implementing multiple imputation is sequential regression modeling, also called multiple imputation by chained equations (MICE) (9–11). The basic idea is to impute missing values in Y_1 from a regression of the observed elements of Y_1 on (Y_2, Y_3, etc.), impute missing values in Y_2 from a regression of Y_2 on (Y_1, Y_3, etc.), impute missing values in Y_3 from a regression of Y_3 on (Y_1, Y_2, etc.), and so on. It is generally easier to specify these conditional models than a plausible joint distribution of all the data. However, in general, there need not exist a joint distribution that corresponds to the set of specified conditional distributions, so it is possible that this imputation method produces logically inconsistent imputation models (12). Despite this deficiency, the method is widely used because of its flexibility and relative ease of implementation.

With MICE, the imputer has to specify conditional models for all variables with missing data. With dozens or hundreds of variables, as is often the case in large epidemiologic studies, specifying these models is no easy task. Relations among the variables may be interactive and nonlinear, and identifying these complexities can be a laborious task with no guarantee of success. Furthermore, often variables have distributions that are not easily captured with standard parametric models.

Motivated by these challenges, we present a MICE approach that uses classification and regression trees (CART) (13–15) as the conditional models for imputation. CART has several features that suggest it can be a useful imputation engine. It is flexible enough to fit interactions, nonlinear relations, and complex distributions without parametric assumptions or data transformations. Further, it does so automatically: There is little tuning needed by the imputer. Using simulation studies, we show that the CART imputation engine can result in more reliable inferences compared with naive applications of MICE based on main-effects generalized linear models. We also apply sequential CART to impute missing values in a study of adverse birth outcomes, which includes a wide array of psychological, health, and environmental variables. The study team expects that interactions among the variables in these domains, rather than main effects alone, are likely to be predictors of adverse birth outcomes. Yet, the nature of these interactions is not known a priori. Hence, the imputations of missing data must be flexible enough to capture the most important interactions in the data. Finally, we check the plausibility of our imputation models using posterior predictive checks (16).

MICE AND CART

Multiple imputation through chained equations

Suppose that we have an $n \times p$ data matrix Y arranged so that $Y = (Y_P, Y_C)$, where Y_P is composed of the p_1 columns of Y that are partially observed, and Y_C is composed of the remaining columns that are completely observed. Let Y_{obs} be the set of observed elements in Y, and let Y_{mis} be the set of missing elements in Y. Finally, we assume that the columns of Y_P are arranged such that, moving from left to right, the number of missing elements in each column is nondecreasing.

To implement MICE, the imputer specifies a set of conditional distributions $p(Y_i|Y_{-i})$, where Y_i is the ith column of Y_P, and Y_{-i} is the matrix Y with its ith column removed. The imputed values can be produced with a 4-step strategy.

Step 1. Fill in initial values for the missing values as follows:

a. Define a matrix Z equal to Y_C.

b. For $i = 1, \ldots p_1$, impute missing values in Y_i with draws from the predictive distribution conditional on Z, and append the completed version of Y_i to Z prior to incrementing i.

Step 2. For $i = 1, \ldots p_1$, replace the originally missing values of Y_i with draws from the predictive distribution conditional on Y_{-i}.

Step 3. Repeat step 2 so as to have performed it l times.

Step 4. Repeat steps 1–3 m times, yielding m imputed sets.

We order the columns of Y_P to have increasing numbers of missing values so that we build the models in step 1b with as much information as possible. Although one can formally check stochastic convergence with a diagnostic tool such as the scale reduction factor (17), using $l = 10$ typically yields satisfactory results (18). It is standard to use generalized linear models as the basis of the predictive draws in steps 1b and 2, but in this paper we adapt CART for this purpose.

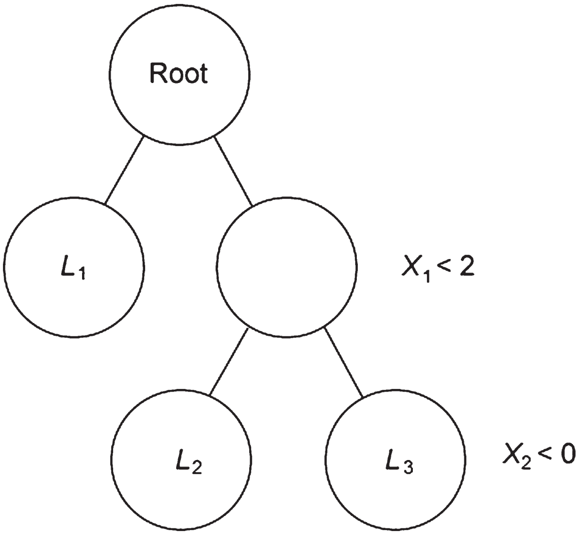

Figure 1. Example of a tree structure.

Classification and regression trees

CART models seek to approximate the conditional distribution of a univariate outcome from multiple predictors. The CART algorithm partitions the predictor space so that subsets of units formed by the partitions have relatively homogeneous outcomes. The partitions are found by recursive binary splits of the predictors. The series of splits can be effectively represented by a tree structure, with leaves corresponding to the subsets of units. The values in each leaf represent the conditional distribution of the outcome for units in the data with predictors that satisfy the partitioning criteria that define the leaf. For further discussion of CART, refer to previous publications (13, 14).

An example of a tree structure for a univariate outcome Y and 2 predictors, X_1 and X_2, is displayed in Figure 1. Units with $X_1 \geq 2$ fall in the leaf labeled L_1, regardless of their value of X_2. Units with $X_1 < 2$ and $X_2 \geq 0$ fall in the leaf labeled L_2, and units with $X_1 < 2$ and $X_2 < 0$ fall in the leaf labeled L_3. Thus, if we wanted to approximate the distribution of Y for units with $X_1 < 2$ and $X_2 < 0$, we would use the values of Y in L_3. Because CART provides distributions for units defined by various combinations of X, it effectively can result in models with many interaction effects.

The primary disadvantages of CART relative to parametric models include decreased efficiency when the parametric models are adequate and discontinuities at partition boundaries (19, 20). Additionally, large trees can be difficult to interpret, but this is not a major concern when using CART for imputations. Categorical predictors with many levels can cause computational difficulties for CART, as it examines all possible partitions of predictors when selecting splits. For example, a categorical predictor with 32 levels—which is the hard-coded maximum number of levels in the "tree" routine for fitting CART in the software package R—results in over 2 billion potential partitions (15).

After growing a tree, one can prune it by removing branches. When trees are used as an analytical tool, pruning is desirable because smaller trees are easier to interpret, and they are less prone to overfitting the data. When trees are used as an imputation engine, interpretation is not a primary concern; we primarily seek plausible imputations. Furthermore, it is generally advisable to use large imputation models so as to minimize bias (1). Therefore, we recommend pruning weakly if at all. In our applications of the technique, we do not prune the trees. Rather, we modulate the size of trees by requiring a minimum number of observations in each leaf and by controlling the minimum heterogeneity in the values in the leaf in order to consider it for further splitting.

To implement sequential CART, we use steps 1–4 with CART models. In step 1b, we use a CART of each Y_i on Z, and in step 2 we use CARTs of each Y_i on Y_{-i}. We take draws from the predictive distribution by sampling elements from the leaf that corresponds to the covariate values of the record of interest. Using Figure 1 as an example, for a record with ($X_1 < 2, X_2 < 0$) and missing Y, we sample a value of Y from L_3. In order to reflect uncertainty about the population distributions in the leaves, we actually perform a Bayesian bootstrap (21) within each leaf before sampling. For continuously valued variables, it is also possible to draw predictions from a smoothed distributional estimator.

CART models can be used with continuous and categorical variables, as both dependent and independent variables. Users must specify nominal variables to ensure that they are not treated as continuous. Because CART imputations come from the observed values, certain restrictions, for example, variables that must be between 0 and 1 or that must be positive, are automatically enforced. Skip patterns can be handled in ways akin to those for existing multiple imputation packages, such as IVEware (18).

CART has been suggested previously as the basis of imputation algorithms, somewhat outside of the standard MICE framework. It has been called "an ideal choice for this imputation 'engine'" (14, p. 333). Rather than filling in initial values and using $l > 1$ iterations, these authors suggest using surrogate splits to deal with the issue of missing values in more than 1 column. This was implemented by Dai et al. (22). Others have used trees as an imputation engine, but only to obtain a single imputation and without the multiple iterations (i.e., $l > 1$) of typical MICE algorithms (23). Our approach is most like that of Reiter (20), which uses a sequential CART approach to generate replacement values for observed confidential data.

APPLICATION TO SIMULATED DATA

To assess the performance of a CART-based MICE algorithm, we compare it with a naïve application of the generalized linear model-based "mi" package in R (10, 24) using simulation studies. The data-generating model is as follows:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_4 x_{8,i} + \beta_5 x_{9,i} + \beta_6 x_{3,i}^2 + \beta_7 x_{1,i} x_{2,i} + \beta_8 x_{8,i} x_{9,i} + \varepsilon_i,$$

where the true value of the regression parameters $\beta = (0, 0.5, 0.5, 0.5, 0.5, 0.5, 1, 1)$. The errors ε_i have independent, standard normal distributions. The explanatory variables are drawn from a multivariate normal distribution such that the first 4 columns have a correlation of 0.5 and the last 6 columns have a correlation of 0.3. We simulate 1,000 observations from this design and delete observations from Y and X_1 through X_8 via a missing-at-random mechanism that depends on X_9 and X_{10}, which are completely observed. On average, this results in around 17% missing values in every variable except X_9 and X_{10}; on average, fewer than 25% of the records are complete.

We perform multiple imputation using the "mi" package default settings and its adaptive choice of l (10, 24). We also carry out the CART-based method with $l = 5$, 10, and 20, using the "tree" package in R to fit the CART models (15). A basic implementation of this procedure—along with the predictive diagnostic check described below—is available at http://www.stat.duke.edu/~lb131/software.html. We find that the performance of the sequential CART imputations is insensitive to the number of iterations l for this application; we present only the $l = 10$ results.

We use a minimum leaf size of 5, and a leaf is not considered for further splitting if the deviance of its values is less than 0.0001. This combination results in relatively large trees, which exhibit less bias than those grown with a larger leaf size or deviance criterion. We generate $m = 10$ imputed sets for each method, although in some situations using $m = 20$ or 40 may be warranted for added accuracy (25). We include Y as a predictor in the imputation models for each X (26, 27).

Using the rules described by Rubin (1), we estimate the parameters in the model along with their standard errors using the correct model specification. We then evaluate the root mean-squared errors and biases for these estimates of β. Table 1 displays averages of these quantities over 1,000 generated sets. For each repetition, we also simulate 500 additional records and use the fitted models to predict Y for these new cases. We evaluate the root mean-squared prediction errors on the evaluation sample using the parameter estimates from the fitted models. The averages of the root mean-squared prediction errors are in the last row of Table 1.

For the quadratic and interaction terms, CART-based MICE results in notably lower mean-squared errors and biases. Even the estimated main effects are somewhat closer to the truth. This combines to make out-of-sample prediction much more accurate. The models fit on the CART imputations were uniformly better in this regard. Because the residual standard deviation equals 1.0, the excess prediction error from standard MICE is more than 3 times higher than that of CART on average.

Both CART-based and standard MICE result in many intervals that do not cover the corresponding truths, because

Table 1. Average Root Mean-squared Error and Bias for β Estimates[a]

	True β	Root Mean-squared Error		Bias	
		CART-MICE	Default "mi"	CART-MICE	Default "mi"
β_0	0.0	0.168	0.379	0.156	0.373
β_1	0.5	0.061	0.077	−0.020	−0.018
β_2	0.5	0.061	0.078	−0.015	−0.015
β_3	0.5	0.059	0.076	−0.010	−0.018
β_4	0.5	0.120	0.149	−0.108	−0.132
β_5	0.5	0.054	0.067	0.006	0.016
β_6	0.5	0.053	0.132	−0.035	−0.125
β_7	1.0	0.144	0.315	−0.134	−0.310
β_8	1.0	0.198	0.314	−0.190	−0.309
ARMSPE		1.106	1.348		

Abbreviations: ARMSPE, average root mean-squared prediction error; CART, classification and regression trees; default "mi," multiple imputation with diagnostics in R language; MICE, multiple imputation by chained equations.

[a] The columns correspond to default "mi" package behavior and CART-based MICE with $l = 10$. The last row gives out-of-sample ARMSPE based on parameter estimates from the various imputed sets. All of the model fits use the true model.

they are based on imperfect imputation models. For example, the 95% intervals from the CART imputations cover only the true values of β_7 and β_8 (the interaction terms) in approximately 42% and 9% of the simulated runs, respectively; these percentages are 0.2% and 0.0% for standard MICE. Across all β elements, approximately 70% of the intervals cover the truth when using CART-based MICE, compared with 53% for standard MICE.

We also compared CART-based and standard MICE using the complex data-generating model of van der Laan et al. (28), in which the continuous outcome is a function of 10 binary predictors including 3- and 4-way interactions. The relative performances of the 2 approaches are materially unchanged.

APPLICATION TO ADVERSE BIRTH OUTCOMES

We now apply the sequential CART imputation algorithm to a prospective study of adverse birth outcomes, for example, low birth weight and preterm birth. The data comprise 115 variables measured on 1,054 non-Hispanic white and black mothers who gave singleton births in Durham, North Carolina. The variables include mothers' demographics, such as age, race, education, and income; mothers' medical history variables, such as the existence of chronic hypertension, anemia, and previous birth outcomes; mothers' environmental variables, such as levels of cadmium, nicotine, cotinine, mercury, and lead in the mothers' blood; mothers' psychological factors, such as Interpersonal Support Evaluation List (ISEL) measurements (29) and the NEO Personality Inventory (Psychological Assessment Resources, Inc., Lutz, Florida); and social factors, such as perceived racism and availability of social support. These variables are a mix of categorical and numerical data, many with irregular distributions.

The study team was successful in recruiting and retaining mothers in the study; retention rates among eligible women exceed 95%. However, many variables have modest amounts of missing data. All but 21 of the variables have less than 10% missing values; 18 of the variables have between 10% and 45% missing; and 3 variables have between 58% and 61% missing. Although the missing rates are mostly modest, they are scattered among the variables such that only 7 mothers have complete data on all variables. There is weak evidence that low birth weights are associated with lower rates of missingness. We include the outcome variables in the imputation models to account for a missing-at-random mechanism consistent with such a pattern (26).

A large research team composed of social, environmental, and medical scientists plans to use the data for a variety of analyses, many of which will involve interactions among predictors of adverse birth outcomes. Hence, the team decided to create $m = 10$ completed data sets using MICE via sequential CART. Imputations were done separately for black mothers and white mothers, because cross-racial comparisons are of primary interest to several team members.

We order the variables from least amount to largest amount of missing data and proceed as in steps 1–4 of the imputation algorithm. As in the simulated example, we use a minimum leaf size of 5 and the splitting criteria of a deviance greater than 0.0001. We use $l = 10$ iterations of step 3; the results did not change systematically with $l > 10$, and $l = 5$ would have been acceptable.

Some of the variables have logical constraints that we enforce in imputations. For instance, if Y_1 records the number of previous preterm pregnancies, and Y_2 is the number of previous pregnancies, we require that $0 \leq Y_1 \leq Y_2$. Whenever a constraint of equality exists among the columns (e.g., $Y_1 + Y_2 = Y_3$), we exclude one of the algebraically dependent columns from the imputation process and then determine its value from the other imputed values in the constraint. Before eliminating columns, we fill in any values that can be logically deduced through different missing patterns in the relevant variables. In the case of constraints of inequality, of which we had only a few, we simply make a post hoc adjustment to ensure that the inequality is satisfied. In data sets that are characterized by many such constrained relations, it may be necessary to incorporate the restrictions explicitly into the conditional models of a chained equation imputation. Because CART draws values from the collection of observed values in a given column, marginal constraints (such as positivity) are automatic.

As suggested in the work of Abayomi et al. (30), we check the appropriateness of the imputation models with graphical diagnostics that compare the marginal distributions of observed and imputed values. These did not raise red flags. However, these diagnostics may not tell us enough about joint distributions to identify problems in the imputation models (8). To illustrate, if one were to impute missing values in a column of Y by sampling at random from the observed elements in that column, associations involving that variable would be attenuated, but the univariate diagnostics would not raise any red flags.

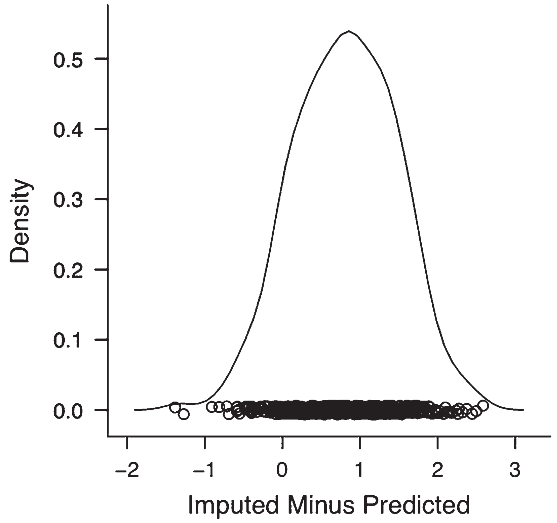

Figure 2. Density histogram of differences between regression coefficients calculated on imputed and predicted sets for the NEO-Openness/NEO-Conscientiousness interaction. A total of 56 of 500 of the differences are negative, so we have the 2-sided estimated $P = 2.56/500 = 0.224$, which does not indicate a deficiency in the imputation model for this parameter. If 12 or fewer of these differences were negative (or positive), however, we would have a 2-sided P value below 0.05, which would indicate a possible problem. "NEO" is part of the name of the test and no longer considered an acronym for "neuroticism–extroversion–openness."

Therefore, we also examined posterior predictive checks (31), as suggested by He et al. (16). These are implemented as follows. First, we form 500 imputed sets using the imputation models under consideration. At the same time, we also use CART to create 500 data sets with Y_P completely replaced (not merely completed) with approximate draws from the distribution of $Y_P|Y_C$. We call these the *predicted sets*. To obtain these sets, we create a copy of Y (call it Y_{new}) and consider all the observed elements of Y_P in the new copy to be missing. Then, using the fitted model that was used to impute the missing values in Y_i, we draw replacements for all elements in the ith column of Y_{new}. We do this by tracing down the branches of the imputation tree using the other columns of Y_{new} as predictors. These draws are not used for the imputation; they are additional and used only for imputation diagnostics.

Second, we identify some statistic with epidemiologic relevance, which we refer to as T. For example, T could be a regression coefficient of a particular interaction in a linear regression of birth weight on several covariates. Let $T_{\text{imp},i}$ be the value of the statistic computed with the ith imputed set, and let $T_{\text{pred},i}$ be the statistic computed with the ith predicted set. We then compute a 2-sided posterior predictive P value,

$$P = 2/500 \cdot \min\Big(\sum I\Big\{\Big(T_{\text{imp},i} - T_{\text{pred},i}\Big) > 0\Big\},\ \sum I\big\{\big(T_{\text{pred},i} - T_{\text{imp},i}\big) > 0\big\}\Big),$$

where $I\{\cdot\}$ is the indicator function that equals 1 if the argument is true and 0 otherwise (16). If $T_{\text{imp},i}$ and $T_{\text{pred},i}$ consistently deviate from each other in one direction—which would be indicated by a small P value—the imputation model may be distorting the relation implicit in the test statistic. To illustrate, suppose that a regression coefficient is consistently larger in the imputed sets than it is in the predicted sets. If this coefficient is estimated to be positive, the association involving this coefficient might be attenuated by the imputed values. Essentially, if the imputation models do not recreate important features in the observed data, they may not generate plausible values for the missing data.

From a practical standpoint, posterior predictive checks are well-suited for use in large studies with many investigators. The imputation team can create and store many imputed and predicted sets. Researchers interested in using the imputed data sets for their particular model can compute posterior predictive P values for their model to check the suitability of the imputations for their analyses. This process takes only seconds of computer time (whereas generating the 500 predicted sets can take several days of computer time depending on the number of imputations), and it can be automated in software that is distributed with the imputed data sets.

If evidence of serious imputation deficiencies arises, the analyst can inform the imputation team about the significant P values, and the team can adjust the imputation procedure

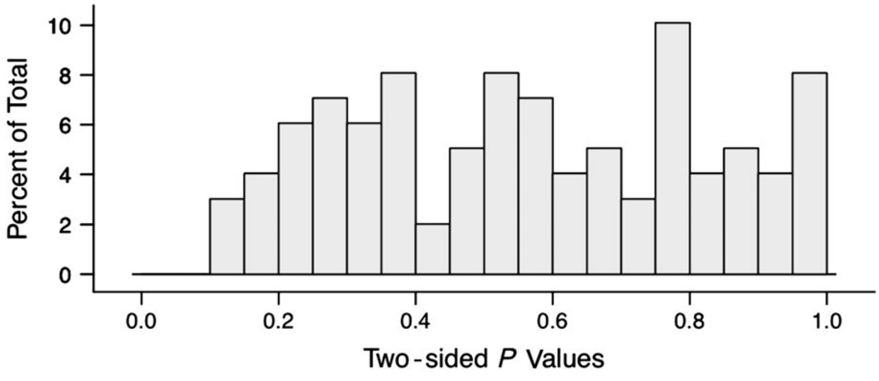

Figure 3. Histogram of the 99 two-sided posterior predictive P values related to the coefficients of interest.

with the aim of remedying the problems if necessary. This might involve, for example, reducing the minimum leaf size or the minimum deviance value for splitting. It also might involve using different imputation models for the offending variables, for example, parametric models based on an exhaustive search for complex interactions. If the imputation team cannot remedy the problems, analysts are left with the options of generating their own imputations in ways tailored to their specific models—which may not necessarily improve the quality of the imputations—or reporting potential sensitivity to the imputations in the analysis.

In the adverse birth outcomes imputation project, we focus on posterior predictive checks of linear and logistic regression coefficients in models of interest to the scientific team, where T is the value of the maximum likelihood estimate of the regression coefficient. Each model includes a particular response (birth weight, low/normal birth weight, gestational age, preterm/term birth, maternal hypertension); standard control variables for race, age, education, and an indicator of the mother's first pregnancy; and additional covariates selected from the remaining variables. For example, one of the regressions is a linear model of birth weight as a function of NEO Openness and Conscientiousness scores and their interaction, along with the standard control variables. Figure 2 displays the 500 values of $T_{\text{imp},i} - T_{\text{pred},i}$ for the interaction term in this model. Here, $T_{\text{imp},i} - T_{\text{pred},i}$ is less than 0 for 56 of the 500 cases, so that the estimated $P = 0.224$. Thus, for this interaction coefficient, we do not have strong evidence that the imputations seriously distort the relations in the observed data.

After screening many models, we do not find substantial evidence that the sequential tree imputations are implausible. Figure 3 displays the posterior predictive P values for 99 regression coefficients for variables other than the standard controls; none of these P values is below 0.10. We exclude the standard control variables from Figure 3, because each of these variables is missing in 4 or fewer ($<$0.4%) of the records, so that regression analyses are insensitive to any reasonable imputation model for these variables. The covariates related to the coefficients in Figure 3 are missing in 1.9%–24.5% of the records. Among the standard control variables, the P value for the indicator of mother's first pregnancy is consistently small; in a few regressions, we even estimate $P = 0$ from the 500 pairs of data sets. The small P value indicates that the CART imputation model did not accurately re-create the conditional distribution of first pregnancy for the entire data set. However, because previous pregnancy data are missing for only 3 mothers, we are not particularly concerned with a potential misspecification of the imputation model for the first pregnancy indicator.

CONCLUSION

Researchers often avoid tree-based regressions because they can be difficult to interpret unless the trees are relatively small. Interpretation also can be strained by the volatility of the fitting process: When small changes in the observed data would lead to different initial splits, the resulting trees could be very different from the original one. As an imputation engine, however, neither of these issues is particularly consequential. We are not interested in interpreting the trees or making inferences related to them. Their ability to provide sensible imputations and to preserve complexity is all that matters.

With that in mind, one might consider using more exotic nonparametric modeling techniques, such as random forests, neural networks, or Bayesian additive regression trees (14, 32). Such techniques generate results that can be even more difficult to interpret, but their predictive performance can be excellent. One drawback of these approaches compared with CART is the typically much slower speed of the fitting algorithms. This is especially important when using posterior predictive checks; for example, performing imputations along with the posterior predictive checks in the adverse birth outcome study conservatively requires half a million model fits. Nonetheless, we anticipate increased use of nonparametric methods to implement MICE as computing power continues to grow.

ACKNOWLEDGMENTS

Authors affiliation: Department of Statistical Science, Duke University, Durham, North Carolina.

This work was supported by Environmental Protection Agency grant R833293.

The authors thank Dr. Marie Lynn Miranda, Dr. Geeta Swamy, and Dr. Redford Williams for suggesting the models used to check the imputations.

Conflict of interest: none declared.

REFERENCES

1. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Hoboken, NJ: Wiley-IEEE; 1987.
2. Reiter JP, Raghunathan TE. The multiple adaptations of multiple imputation. *J Am Stat Assoc*. 2007;102(480):1462–1471.
3. Barnard J, Meng XL. Applications of multiple imputation in medical studies: from AIDS to NHANES. *Stat Methods Med Res*. 1999;8(1):17–36.
4. Harel O, Zhou XH. Multiple imputation: review of theory, implementation and software. *Stat Med*. 2007;26(16):3057–3077.
5. Klebanoff MA, Cole SR. Use of multiple imputation in the epidemiologic literature. *Am J Epidemiol*. 2008; 168(4):355–357.
6. Rubin DB. Multiple imputation after 18+ years. *J Am Stat Assoc*. 1996;91(434):473–489.
7. Schafer JL. Multiple imputation: a primer. *Stat Methods Med Res*. 1999;8(1):3–15.
8. Stuart EA, Azur M, Frangakis C, et al. Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative. *Am J Epidemiol*. 2009;169(9):1133–1139.
9. Raghunathan T, Solenberger P, Van Hoewyk J. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv Methodol*. 2002;27(1): 85–96.

10. Su YS, Gelman A, Hill J, et al. Multiple imputation with diagnostics (mi) in R: opening windows into the black box. *J Stat Softw*. 2009;20(1):1–27.
11. Van Buuren S, Oudshoorn K. *Flexible Multivariate Imputation by MICE*. Leiden, the Netherlands: TNO Prevention Center; 1999.
12. Gelman A, Speed TP. Characterizing a joint probability distribution by conditionals. *J R Statist Soc B*. 1993;55(1):185–188.
13. Breiman L, Friedman JH, Olshen RA, et al. *Classification and Regression Trees*. Boca Raton, FL: Chapman and Hall/CRC; 1984.
14. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer; 2009.
15. Ripley B. Tree: classification and regression trees. Oxford, United Kingdom: University of Oxford, 2009. (http://cran.r-project.org).
16. He Y, Zaslavsky A, Landrum M. Multiple imputation in a large-scale complex survey: a practical guide. In:*Statistical Methods in Medical Research*. Thousand Oaks, CA: SAGE; 2009:1–18.
17. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Stat Sci.* 1992;7(4):457–472.
18. Raghunathan T, Solenberger P, Van Hoewyk J. *IVEware: Imputation and Variance Estimation Software*. Ann Arbor, MI: Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan; 2002.
19. Friedman JH. Multivariate adaptive regression splines. *Ann Stat*. 1991;19(1):1–67.
20. Reiter JP. Using CART to generate partially synthetic public use microdata. *J Off Stat*. 2005;21(3):7–30.
21. Rubin DB. The Bayesian bootstrap. *Ann Stat*. 1981;9(1): 130–134.
22. Dai JY, Ruczinski I, LeBlanc M, et al. Imputation methods to improve inference in SNP association studies. *Genet Epidemiol*. 2006;30(8):690–702.
23. Conversano C, Cappelli C. Missing data incremental imputation through tree based methods. In: Härdle W, Rönz B, eds. *COMPSTAT 2002—Proceedings in Computational Statistics: 15th Symposium Held in Berlin, Germany*. Heidelberg, Germany: Physica-Verlag; 2002:455–460.
24. Gelman A, Hill J, Yajima M, et al. mi: missing data imputation. Vienna, Austria: The R Foundation, 2009. (http://cran.r-project.org).
25. Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prev Sci.* 2007;8(3):206–213.
26. Little RJA. Regression with missing Xs—a review. *J Am Stat Assoc.* 1992;87(420):1227–1237.
27. Moons KG, Donders RA, Stijnen T, et al. Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol*. 2006;59(10):1092–1101.
28. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol.* 2007;6(1):1–21.
29. Cohen S, Mermelstein R, Kamarck T, et al. Measuring the functional components of social support. In: Sarason IF, Sarason BR, eds. *Social Support: Theory, Research and Applications*. The Hague, the Netherlands: Martinus Nijhoff; 1985:74–94.
30. Abayomi K, Gelman A, Levy M. Diagnostics for multivariate imputations. *J R Stat Soc Ser C Appl Stat*. 2008;57(3):273–291.
31. Meng XL. Posterior predictive *P*-values. *Ann Stat*. 1994;22(3): 1142–1160.
32. Chipman HA, George EI, McCulloch RE. BART: Bayesian additive regression trees. *Ann Appl Stat*. 2010;4(1):266–298.