Can machine learning effectively reduce potential nonresponse bias?

YES

David Haziza
University of Ottawa, Canada, dhaziza@uottawa.ca

Short answer: yes, it can — but with important caveats.

To start with, reducing non-response bias fundamentally relies on having auxiliary variables that are observed for all sampled units and that are related both to the outcome of interest and to the response mechanism. Under a missing-at-random type assumption, conditioning on such variables allows us, at least in principle, to remove non-response bias. Additional variables that are related to the outcome can also help improve efficiency; see, e.g., Little and Vartivarian (2005). From that perspective, machine learning methods can be very useful. Whether we are dealing with unit non-response or item non-response, the idea is similar. In the unit non-response case, we estimate response probabilities and construct adjusted weights. In the item non-response case, we impute missing values using predicted outcomes. In both settings, machine learning can help by providing flexible estimators of either the response model or the outcome model.

The main advantage of machine learning is its ability to capture complex relationships without requiring a fully specified parametric model. In practice, parametric approaches require us to get both the functional form and the set of covariates right, including interactions and nonlinearities. This is rarely realistic. Machine learning methods are much more robust to these kinds of misspecifications, which means they can, in principle, lead to better bias reduction — and sometimes improved efficiency as well.

However, there is no free lunch.

NO

Roderick J. Little
University of Michigan, Ann Arbor, USA, rlittle@umich.edu

Machine learning methods use training data to develop algorithms for prediction. Statistical methods also apply algorithms fitted to data, and can provide predictions, as when applied to imputation for nonresponse. I view machine learning as a form of statistical modeling, whereas computer scientists might consider statistical modeling as a form of machine learning. These semantic squabbles do little to advance science – although machine learners might pay more attention to, and reference, the extensive statistical literature on missing data.

Breiman (2001) advocated "black box" algorithmic approaches over more classical statistical models like linear or logistic regression (see also Little 2026, chapter 12). Since that paper, machine learning has come to be associated with algorithmic methods such as regression trees and forests, neural networks and gradient boosting. A more specific question is whether these methods can be used to reduce nonresponse bias. In their favor, these methods tend to be more flexible than classical models, and have the ability to allow for interactions that might otherwise be missed. There is some evidence that they produce better predictions than standard statistical models. I believe that mixing a variety of methods can be better than relying on one single approach, as in random forests rather than regression trees; I favor Bayesian forms of mixing as in BART (Chipman et al. 2010), because they give more weight to better-fitting models. On the other hand, machine learning algorithms are not easy to interpret, and underlying assumptions tend to be buried. Because of this, the methods tend to be viewed as "assumption-free", and hence always superior to classical statistical models. This is not true – the only universal approach is to avoid missing data! I think that

First, machine learning methods typically converge more slowly than parametric estimators, and therefore may require larger sample sizes to perform well. Second, and more importantly in the context of official statistics, point estimation is not enough. We also need valid variance estimators and, sometimes, confidence intervals. This is where things become much more challenging.
While recent developments such as double/debiased machine learning (Chernozhukov et al., 2018) provide promising solutions in some settings (for instance in the context of imputation for the treatment of item nonresponse), they often require rethinking the estimator itself in order to recover valid statistical inference.

The situation is even more complex in the case of unit non-response. Here, we construct adjusted weights that are meant to be used across all variables in the survey. Ensuring valid inference in that context, including variance estimation and asymptotic normality, remains a difficult problem, and in many cases an open research question; see Dagdoug and Haziza (2026) for a discussion.

Another practical issue is that machine learning methods require the user to choose an architecture and tuning parameters. In practice, default settings are often used, but for highly flexible methods this can be risky and may lead to unstable or biased estimators.

Finally, it is important to emphasize that machine learning does not solve the fundamental identification problem. If key variables related to both response and the outcome are not observed, no method — machine learning or otherwise — can fully eliminate non-response bias.

In summary, machine learning offers powerful tools to reduce non-response bias by better approximating complex relationships in the data. But it also introduces new challenges, particularly for statistical inference and practical implementation. Addressing these challenges remains an active area of research.

attention to the setting and underlying causal mechanisms of nonresponse can be important. To give two examples: (a) one application is to predict probabilities of response, which can be used in design to target units where nonresponse is high, or in analysis for nonresponse weighting. The "probability of response" depends on the set of predictors (Little 2021), and including auxiliary variables that are highly predictive of response but unrelated to survey outcomes is counterproductive, increasing variance without reducing bias (Little & Vartivarian 2005). Identifying such variables is context-dependent, and varies across the survey variable being imputed. Throwing the kitchen sink into a prediction algorithm ignores this issue. (b) More directly, machine-learning methods can be used to predict the missing values of survey variables. Here the fact that the algorithms are trained on existing data implies some form of missing at random (MAR, Rubin 1976) or very specific missing not at random (MNAR) mechanisms if response indicators are included as predictors (Little 2020; Fischer et al. 2026). MNAR models are chronically unidentified, and often best treated using a sensitivity analysis (Little & Rubin 2019, chapter 15).

Machine-learning algorithms focus on best predictions of the missing values; surveys concern statistical inferences for model parameters or finite population quantities. When missing values are replaced by best estimates, standard errors based on the filled-in data are underestimated. Also estimates of nonlinear functions of the data are biased. Multiple imputation (MI, Rubin 1987) imputes draws of the missing values from a predictive distribution. Estimates are obtained for each imputed data set and then averaged, rather than averaging the draws, the simulation analog of imputing best predictions. The MI approach is ingenious, less subject to bias for nonlinear statistics, and yields valid standard errors as well as point estimates. However, MI requires a statistical model to generate a predictive distribution for each missing value, which machine learning algorithms generally avoid. Machine learning approaches could be incorpo-

References

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., & Newey, W. (2017). Double/debiased/Neyman machine learning of treatment effects. *American Economic Review*, 107(5), 261-265.

Dagdoug, M., & Haziza, D. (2026). Machine learning methods for finite population parameter estimation in survey sampling. *arXiv preprint arXiv:2604.01160*

Little, R. J., & Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means?. *Survey Methodology*, 31(2), 161.

rated into fully-specified statistical models, and MI applied to the predictive distributions of the missing values.

Finally, a comment about the question: it focuses on bias, assuming that a sizeable sample size make variance irrelevant; but the number of ultimate clusters in multistage sampling is often small, and for questions like subgroup analysis and small area estimation variance really matters. Methods that trade some bias for lower mean squared error, such hierarchical models, ridge regression, or the machine-learning methods discussed here, have become increasingly important. For me, these “biased” approaches constitute one of the seminal ideas in modern statistics (Little 2026, Chapter 7).

References

Breiman, L. (2001). Statistical modeling: two cultures. *Statist. Sci.* 16, 3, 199-231.

Fischer, M., Little, R.J. & West, B. T. (2026). Multiple imputation under missing not at random: incorporating response indicators into sequential imputation. In press, *Journal of Statistical Computation and Simulation*.

Chipman, H.A., George, E.I. & McCulloch, R.E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Statist.*, 4, 1, 266-298.

Little, R.J. (2021). A Note About the Definition of Propensity Weights. *Journal of Survey Statistics and Methodology.* 10 (4), 1098-1106.

Little, R.J. (2026). *Seminal Ideas and Controversies in Statistics*, Chapter 7. Baseball averages, foreign cars, and shrinkage estimation, and Chapter 12: Exploratory data analysis and data science. Chapman & Hall/CRC Press. Little, R.J. and Rubin, D.B. (2019). *Statistical Analysis with Missing Data*, 3rd edition, Wiley Press.

Little, R.J., and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Meth.*, 31, 161-168. Rubin, D.B. (1976). Inference and missing data (with discussion), *Biometrika*, 63, 581-592.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*, Wiley Press.