Stephen R. Hooper
Department of Allied Health Sciences
School of Medicine
University of North Carolina-Chapel Hill
Chapel Hill, NC

REFERENCES

1. Blouin B, Casapia M, Kaufman JS, Joseph L, Larson C, Gyorkos TW. Bayesian methods for exposure misclassification adjustment in a mediation analysis: hemoglobin and malnutrition in the association between Ascaris and IQ. *Epidemiology.* 2019;30:659–668.
2. Wechsler D. *Wechsler Preschool and Primary Scale of Intelligence*. 3rd ed. San Antonio, TX: The Psychological Corporation; 2002.
3. Freeman S. Wechsler preschool and primary scale of intelligence. In: Volkmar FR, ed. *Encyclopedia of Autism Spectrum Disorders*. New York, NY: Springer New York; 2013:3351–3360.
4. Baker FB, Kim S-H. The information function. In: Baker FB, Kim S-H, eds. *The Basics of Item Response Theory Using R. Statistics for Social and Behavioral Sciences*. Cham, Switzerland: Springer International Publishing; 2017:89–104.
5. Blouin B, Casapía M, Joseph L, Kaufman JS, Larson C, Gyorkos TW. The effect of cumulative soil-transmitted helminth infections over time on child development: a 4-year longitudinal cohort study in preschool children using Bayesian methods to adjust for exposure misclassification. *Int J Epidemiol.* 2018;47:1180–1194.

The Impact of Undersampling on the Predictive Performance of Logistic Regression and Machine Learning Algorithms

A Simulation Study

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).

Code to reproduce the data and analyses is available on GitHub at https://github.com/ainaimi/Undersampling-sim.

ISSN: 1044-3983/20/3105-0e42
DOI: 10.1097/EDE.0000000000001198

To the Editor:

Machine learning techniques may improve risk prediction and disease screening. Class imbalance (ratio of noncases to cases > 1) routinely occurs in epidemiologic data and may degrade the predictive performance of machine learning algorithms.[1–4] Of the many techniques developed to address class imbalance,[5,6] here, we investigated simple undersampling. This method is straightforward and accessible, but evidence on its performance is mixed and practical guidance is needed. Using simulated data, we assessed the predictive performance of the ensemble machine learning algorithm SuperLearner and logistic regression in imbalanced and undersampled data to investigate whether undersampling alters predictive accuracy.

DATA-GENERATING MECHANISM

We used Monte Carlo simulation with four groups of 1,000 Monte Carlo samples each. We simulated each Monte Carlo sample to have a sample size of 1,000, 10 independent standard normal covariates generated from a random uniform distribution, and 10 independent dichotomous covariates generated from a binomial distribution. A dichotomous outcome was simulated from a logistic regression model conditional on all 20 covariates. Parameters were chosen to lie between −1 and 1 for the continuous variables, and the outcome prevalence was set to lie between 0.15 and 0.50.

STUDY DESIGN

In two of the four groups of Monte Carlo samples, we left all samples unbalanced. In the remaining two groups, we performed undersampling to balance each sample by randomly selecting a number of noncases equal to the number of cases. To avoid overfitting, we split each Monte Carlo sample into training (70%) and testing (30%) sets with similar outcome prevalences.[7] We generated predicted probabilities on 1,000 undersampled and 1,000 unbalanced samples parametrically via logistic regression and nonparametrically via stacking (SuperLearner).[4] We implemented SuperLearner with 10-fold cross-validation and a library of five algorithms with default tuning parameters: extreme gradient boosting, random forests, kernel k-nearest neighbors, kernel support vector machines, and penalized regression (LASSO). Logistic regression was implemented as a generalized linear model with binomial variance and a logit link function. We evaluated average performance metrics (sensitivity, specificity, positive and negative predictive value, and overall accuracy) across all 1,000 Monte Carlo samples in each group using a classification threshold close to the outcome prevalence of 0.2 for unbalanced groups and 0.5 for undersampled groups. We generated areas under the receiver operating curve for each sample using the roc() function in the “pROC” package.[8] We conducted all analyses using R version 3.6.1.

The Figure shows the receiver operating characteristic (ROC) curves for all 1,000 Monte Carlo samples in each group and average predictive performance metrics. Areas under the curve across all Monte Carlo samples were similar for all groups. Performance metrics were higher for logistic regression than SuperLearner regardless of data preprocessing method except sensitivity and positive predictive value, which were higher for SuperLearner than logistic regression. Undersampling did not have a substantial impact on logistic regression performance; however, undersampling improved SuperLearner accuracy, specificity, and positive predictive value and worsened SuperLearner sensitivity and negative predictive value. Repeating the analysis with a lower outcome prevalence (2%–10%) did not substantially affect the results.

We observed generally more accurate predictive performance with logistic regression than with SuperLearner regardless of data preprocessing method. This is expected because we simulated our data from a logistic model. However, SuperLearner performed nearly as well on average as the true data-generating

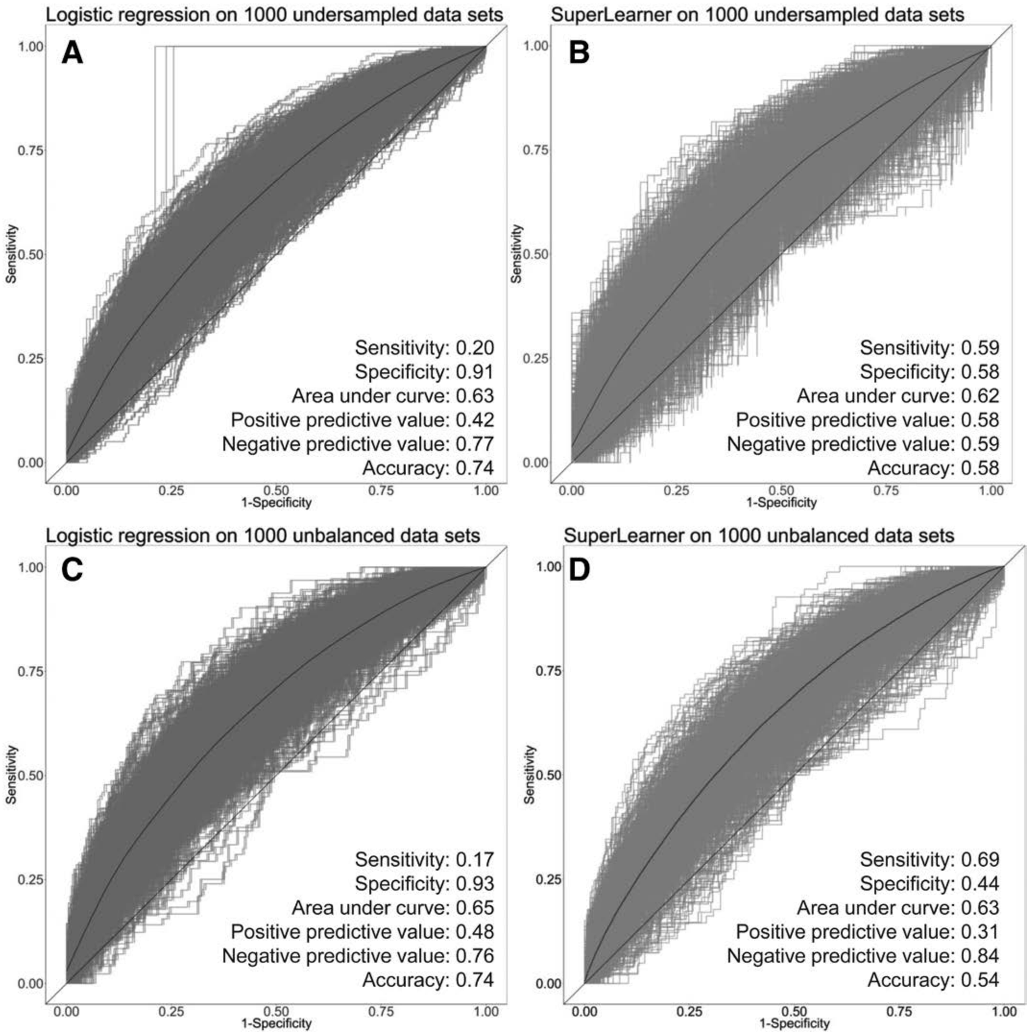

FIGURE. Receiver operating characteristic curves of each Monte Carlo sample by data preprocessing method and prediction technique. The figure displays individual ROC curves (gray lines) for each of the 1,000 Monte Carlo sample within each of the four groups listed below, as well as the average ROC curve (black line) across all 1,000 Monte Carlo samples within each group. Average performance metrics for all 1,000 Monte Carlo samples are also displayed in each panel. A, Logistic regression with undersampling; (B) SuperLearner with undersampling; (C) logistic regression with undersampling; (D) SuperLearner with undersampling.

mechanism although logistic regression was intentionally excluded from the SuperLearner library. In our simulations, undersampling did not dramatically improve predictive performance, suggesting that ensemble machine learning can achieve adequate performance in similar settings with moderate class imbalance. These results provide some insight on the optimal use of machine learning for predicting imbalanced outcomes. Example code to reproduce these analyses is available in the eSupplement; http://links.lww.com/EDE/B675.

Abigail R. Cartus, MPH
Department of Epidemiology
Graduate School of Public Health
University of Pittsburgh
Pittsburgh, PA

Lisa M. Bodnar, RD, PhD
Department of Epidemiology
Graduate School of Public Health
University of Pittsburgh
Pittsburgh, PA
Department of Obstetrics, Gynecology,
and Reproductive Sciences
University of Pittsburgh School of Medicine
Pittsburgh, PA
Magee-Womens Research Institute
University of Pittsburgh
Pittsburgh, PA

Ashley I. Naimi, PhD
Department of Epidemiology
Graduate School of Public Health
University of Pittsburgh
Pittsburgh, PA
ashley.naimi@pitt.edu

REFERENCES

1. Sun Y, Wong, A, Kamel M. Classification of imbalanced data: a review. *Int J Pattern Recogn Artificial Intelligence*. 2011;23:687–719.
2. Chawla NV, Japkowicz N, Kotcz A. Special issue on learning from imbalanced data sets.

ACM SIGKDD Explorations Newsletter. 2004;6:1–6.

3. Naimi AI, Balzer LB. Stacked generalization: an introduction to super learning. *Eur J Epidemiol*. 2018;33:459–464.
4. *SuperLearner: Super Learner Prediction*. [computer program]. Version R package version 2.0-2.4. Available at: https://CRAN.R-project.org/package=SuperLearner2018. Accessed 1 October 2019.
5. Batista G, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*. 2004;6:20–29.
6. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artifi Intell Res*. 2002;16:321–357.
7. Kuhn M. Building predictive models in R using the caret package. *J Stat Soft*. 2008;28:1–26.
8. *pROC: an open-source package for R and S+ to analyze and compare ROC curves* [computer program]. *BMC Bioinformatics*. 2011;12:77.

Simple Sensitivity Analysis for Control Selection Bias

To the Editor:

Case–control studies allow for efficient sampling schemes but are subject to bias when controls fail to represent the exposure distribution in the population from which the cases were sampled. Identifying this population, known as the study base, is often a challenge, and controls may be chosen out of convenience or to avoid other types of bias, such as exposure misclassification.[1] On the other hand, it may be straightforward to completely ascertain or randomly sample cases, as they may be enumerated in registries, hospital records, or other sampling frames.

When inappropriate control selection is suspected to have occurred, it can be informative to conduct a sensitivity analysis to investigate the possible extent of the resulting bias. In this letter, we show that a recently developed framework for simple sensitivity analysis[2–4] can be extended to this situation. We demonstrate with an example, and we provide a more detailed derivation in the eAppendix; http://links.lww.com/EDE/B670.

MacMahon et al[5] conducted a case–control study of pancreatic cancer patients whom they compared with controls who were patients of the same physicians as the cases but who had different illnesses. After adjusting for age, cigarette smoking, and sex, they found an odds ratio of 2.7 (95% confidence interval = 1.6, 4.7) comparing drinkers of at least 3 cups per day to non-coffee drinkers.

However, soon after the study was published, multiple possible sources of bias were described.[6] In particular, many of the control patients had gastrointestinal disorders, which the investigators failed to account for. If the controls drank less coffee than the general source population due to their illnesses, selection bias would result, exaggerating the association between coffee and pancreatic cancer.

To quantify the possible size of this bias, consider the ratio of the observable odds ratio from case–control data (OR_{obs}) to the odds ratio that would have been estimated had the entire study base been sampled (OR_{true}). For simplicity, assume that any bias from the case–control study is due to poor control selection.

It is possible to derive a bound similar to that in Smith and VanderWeele[4] but with different definitions for the parameters resulting from the different causal structure (Figure) and estimand of interest. Specifically, if we assume that selection ($S = 1$) of cases ($Y = 1$) is independent of exposure status ($A \in \{0,1\}$) (possibly conditional on measured covariates C), but that control ($Y = 0$) selection is not independent of exposure without additionally conditioning on unmeasured factor(s) U, then:

$$OR_{\text{obs}} / OR_{\text{true}} \le \left\{ \frac{RR_{UA_1} \times RR_{S_0U}}{RR_{UA_1} + RR_{S_0U} - 1} \right\} \times \left\{ \frac{RR_{UA_0} \times RR_{S_1U}}{RR_{UA_0} + RR_{S_1U} - 1} \right\}$$

where

$$RR_{UA_1} = \frac{\max_u \Pr(A = 1 \mid Y = 0, u, c)}{\min_u \Pr(A = 1 \mid Y = 0, u, c)}$$

$$RR_{UA_0} = \frac{\max_u \Pr(A = 0 \mid Y = 0, u, c)}{\min_u \Pr(A = 0 \mid Y = 0, u, c)}$$

$$RR_{S_1U} = \max_u \frac{\Pr(U = u \mid Y = 0, S = 1, c)}{\Pr(U = u \mid Y = 0, S = 0, c)}$$

$$RR_{S_0U} = \max_u \frac{\Pr(U = u \mid Y = 0, S = 0, c)}{\Pr(U = u \mid Y = 0, S = 1, c)}.$$

To understand these parameters, suppose that U represents a binary indicator of gastrointestinal illness that affects coffee drinking and also makes hospital visits (and therefore selection as a control) more likely. With respect to the example, RR_{UA_1} describes the increased probability of drinking ≥3 cups of coffee per day in eligible controls without gastrointestinal disorders compared with those with gastrointestinal disorders, RR_{UA_0} is the increased probability of no coffee drinking in eligible controls with gastrointestinal disorders compared with those without gastrointestinal disorders, RR_{S_1U} is the increased probability of gastrointestinal disorders in controls who were selected for the study compared with those who were not, and RR_{S_0U} is the increased probability of a healthy GI system in controls who were *not* selected for the study compared with those who *were* selected.

We could propose various values for these parameters to "correct" for, or bound, selection bias. For example, suppose that among eligible controls with gastrointestinal disorders, only 5% drink at least 3 cups of coffee daily. However, among those with healthy gastrointestinal tracts, 30% drink that amount. Then $RR_{UA_1} = 0.3/0.05 = 6$

This work was supported by grant R01CA222147 from the National Institutes of Health (T.V.W.).

The authors report no conflicts of interest.

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).

The analysis can be replicated at the website linked in the manuscript, http://selection-bias.louisahsmith.com.

ISSN: 1044-3983/2020/3105-e44
DOI: 10.1097/EDE.0000000000001207