

Improving Inverse Probability Weighting by Post-calibrating Its Propensity Scores

 Rom Gutman,^{a,b}  Ehud Karavani,^a and  Yishai Shimoni^a

Abstract: Theoretical guarantees for causal inference using propensity scores are partially based on the scores behaving like conditional probabilities. However, prediction scores between zero and one do not necessarily behave like probabilities, especially when output by flexible statistical estimators. We perform a simulation study to assess the error in estimating the average treatment effect before and after applying a simple and well-established postprocessing method to calibrate the propensity scores. We observe that postcalibration reduces the error in effect estimation and that larger improvements in calibration result in larger improvements in effect estimation. Specifically, we find that expressive tree-based estimators, which are often less calibrated than logistic regression-based models initially, tend to show larger improvements relative to logistic regression-based models. Given the improvement in effect estimation and that postcalibration is computationally cheap, we recommend its adoption when modeling propensity scores with expressive models.

Keywords: Average treatment effect; Causal inference; Calibration; Model validation; Propensity score

(*Epidemiology* 2024;35: 473–480)

The propensity score is defined as the conditional probability of being assigned to a treatment (exposure) given


Submitted March 15, 2023; accepted March 18, 2024

From the ^aIBM Research, University of Haifa Campus; ^bTechnion - Israel Institute of Technology, Haifa, Israel. Rom Gutman and Ehud Karavani contributed equally to this work.

Conceptualization: EK. Methodology: EK, RG. Software: RG. Formal analysis: RG. Investigation: RG, EK. Visualization: EK, RG. Writing - Original Draft: EK, RG. Writing - Review & Editing: EK, RG, YS. Supervision: EK, YS. Funding acquisition: YS.

The authors report no conflicts of interest.

Code for data generation, analysis, and producing the results are available on Github: https://github.com/RomGutman/propensity_calibration. Additional implementation examples for illustrative purposes in both Python (Listing 1) and R (Listing 2) programming languages are available in the eAppendix <http://links.lww.com/EDE/C136>.

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

Correspondence: Ehud Karavani, IBM R&D Labs in Israel, University of Haifa Campus, Mount Carmel, Haifa 3498825, Israel. ehudk@ibm.com;

Copyright © 2024 The Author(s). Published by Wolters Kluwer Health, Inc. This is an open-access article distributed under the terms of the Creative Commons Attribution-Non Commercial-No Derivatives License 4.0 (CCBY-NC-ND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.

ISSN: 1531-5487/24/354-473480

DOI: 10.1097/EDE.0000000000001733

one's observed confounding variables. It is commonly used in methods for estimating causal effects from observational data, such as inverse probability weighting (IPW),¹ propensity matching,^{2,3} propensity stratification,⁴ and doubly robust methods.^{5–8}

Rosenbaum and Rubin² identified theoretical guarantees that ensure that adjusting for the propensity score, instead of the covariates themselves, is sufficient for achieving the conditional exchangeability required for estimating a causal effect. Namely, they ensure that conditioning on the propensity scores obtained from a set of confounding variables is as good as conditioning on the confounding variables themselves for removing confounding bias. Specifically, these theoretical guarantees require the propensity scores to be the true conditional probabilities (see also eAppendix; <http://links.lww.com/EDE/C136>). In practice, however, not every model that inputs data and outputs a number between zero and one correctly estimates true probabilities. Therefore, the obtained propensity scores might not represent true probabilities reliably.

A prediction model that outputs probabilities accurately is referred to as calibrated (note that this is unrelated to a previous notion of “propensity score calibration” by Stürmer et al.⁹) Specifically, in this article we focus on the notion of moderate calibration,¹⁰ where the observed rate of events should equal the predicted rate among samples with a similar predicted score. Formally, an estimated model \hat{f} that regresses a binary variable A on covariates X and produces probabilistic predictions $\hat{f}(X) = \hat{\pi}$ is considered well-calibrated if it satisfies $E[A = 1 | \hat{\pi}] = \hat{\pi}$. To illustrate this notion, if we take all the observations for which the model predicted a score of 0.8, about 80% of them should have a positive label. Calibration can be empirically evaluated with a calibration curve (reliability diagram), comparing the predicted scores with their corresponding label rates for a range of scores.¹¹ However, because an entire curve is not always actionable, there are multiple metrics that try to capture this notion with a single numeric value.^{12–14}

While studies using propensity-based methods might check for overlapping propensity distributions or covariate balancing,^{15–17} it is somewhat uncommon for them to evaluate their propensity scores for calibration. Fortunately, because the majority of classical statistical literature estimates propensity scores using logistic regression models, those studies may have inadvertently overcome the need. Logistic

regression models are fitted by optimizing the log-loss objective function (binary cross-entropy, negative Shannon entropy),^{18,19} which is, by itself, a metric for calibration.²⁰ Additionally, its logit link function is the canonical link function for its log-loss and thus results in the balance property^[21 Chapter 5]. Therefore, a well-specified logistic regression is normally well-calibrated.

However, not all statistical estimators are inherently calibrated. Models of higher complexity, such as tree-based or neural-network-based models, may not be calibrated out of the box, like logistic regression. Even regularizing logistic regression (least absolute shrinkage and selection operator [LASSO], ridge regression, or elastic-net models) could harm calibration, as the penalty added moves the objective function away from the “pure,” calibrating log-loss.^{22,23} As these models become more popular for propensity estimation,²⁴ it is worth investigating if calibration of propensity models is beneficial for effect estimation.

Furthermore, it is of interest to explore if we can break down the trade-off between model expression and calibration by postcalibrating estimators. Therefore, postcalibration may hold promise for using complex high-dimensional data for propensity score estimation when performing causal inference from observational data.

In this article, we use simulations to quantify the downstream effect of poorly calibrated conditional probabilities on the estimation of causal effects. We hypothesize that well-calibrated propensities are indeed imperative for estimating causal effects properly and that in cases where calibration is poor, effect estimation will be improved by postcalibrating the propensity models.

METHODS

To assess the importance of calibration, we use simulations so we have access to individual-level propensity scores and counterfactual outcomes. Below, we describe the data-generating processes used, the effect estimation methods, and the measurements obtained from the various estimations.

Causal Inference Framework

We denote the binary treatment assignment for each individual i as A_i , the covariates (ideally confounding variables) as X_i , and the true propensity to be treated as $\pi_i = Pr[A_i = 1 | X = x_i]$. Using Rubin’s potential outcomes framework,²⁵ we denote Y_i^1 as the hypothetical outcome that would have been observed had individual i been treated, and similarly, Y_i^0 as the hypothetical outcome that would have been experienced had they not been treated. Then, assuming causal consistency, the observed outcome is defined as $Y_i = A_i Y_i^1 + (1 - A_i) Y_i^0$. Finally, we define the average treatment effect (ATE) as $E[Y_i^1 - Y_i^0]$.

Estimation

To estimate the causal effect from the observed data, we first estimate the propensity score, denoted as $\hat{\pi}$, by

regressing the treatment assignment on the covariates. We fit various estimators that are common in the literature: logistic regression, regularized logistic regression—LASSO²⁶ and ridge,²⁷ random forest,²⁸ and gradient boosting trees (additive trees).²⁹ Briefly, LASSO and ridge penalize the logistic regression coefficients using L1 and L2 norms, respectively; a random forest aggregates multiple decision trees fitted on different bootstrap samples of the dataset; and a gradient boosting trees algorithm adds up a sequence of decision trees.

Because these estimators require hyperparameters, we perform a hyperparameter search using cross-validation. For the tree-based models, we search over maximal tree depth and number of trees in the ensemble. For regularized logistic regression models we search over regularization (penalty) strength. To elicit calibration, we select the configuration that minimizes the Brier score.³⁰ Brier score is defined as the mean squared error between the predicted probabilities and the binary outcome variable. It can be used to assess the accuracy of probabilistic predictions and, if optimized, can be used as a calibration-inducing metric.

To avoid overfitting, the propensity scores are predicted on unseen data points using cross-validation. Hence, a nested cross-validation approach is taken: inner cross-validation is used for hyperparameter optimization, and outer cross-validation is used for propensity score prediction and subsequent effect estimation.

Once propensity scores, $\hat{\pi}_i$, are obtained, we plug them into an IPW estimator³¹ to estimate the ATE in a sample of size n , defined as $\widehat{ATE} = \frac{1}{n} \sum_{i=1}^n y_i \left(\frac{a_i}{\hat{\pi}_i} - \frac{1-a_i}{1-\hat{\pi}_i} \right)$. The eAppendix; <http://links.lww.com/EDE/C136> contains additional results obtained by estimating the effect using propensity score matching and propensity score stratification.

Postcalibration

Optionally, we can postprocess the predicted propensity scores by calibrating them before using them with IPW. There are multiple methods for performing this postcalibration that should result in scores functioning more like probabilities. In this study, we focus on Platt’s correction,³² as it is most appropriate for our data-generating process. The method takes advantage of the well-calibrated properties of log-loss by fitting a logistic regression over the scores outputted from an estimator against the treatment assignment labels. Formally, it models $\text{logit}(Pr[A_i = 1 | \hat{\pi}_i]) = \eta_0 + \eta_1 \hat{\pi}_i$, where A denotes the binary treatment assignment and $\hat{\pi}$ the (uncalibrated) estimated propensity scores.

Measurements

For each set of propensity scores and postcalibrated propensity scores, we take three main measurements.

First, we measure the calibration error. This is done graphically with calibration curves and numerically with the integrated calibrated index (ICI).¹⁴ Calibration curves present the notion of

empirical calibration. It bins the predicted scores and counts the positive and negative labels in each bin. A well-calibrated model will have the same rate of observed labels as the average score of a bin, thus resulting in a diagonal line along the $x=y$ curve. ICI is a way to extract a numeric value from the notion of the diagonal calibration curve. It fits a locally estimated scatterplot smoothing regression between the binary classes and the predicted scores, calculates the difference between the resulting predicted line and the optimal $x=y$ diagonal, and takes the mean of the absolute of those differences.

Second, we measure the effect estimation error. The propensity scores are transformed into inverse propensity weights and the ATE is estimated (\widehat{ATE}). We further define the ground truth ATE as the difference between the means of the simulated potential outcomes $\frac{1}{n} \sum_{i=1}^n y_i^1 - y_i^0$ in a sample of size n . The estimation error is then defined as the absolute difference between ATE and \widehat{ATE} .

Third, we measure the covariate balance between the treatment groups. We calculate the absolute standardized mean difference after inverse propensity weighting for each covariate and then select the maximum value over all covariates.

DATA

Simulation Data

To estimate the effect of miscalibration on effect estimation, we first use an estimation-free propensity score simulation. We apply the following simple data-generating process:

$$\begin{aligned} Y &= 5A + 1.2X_1 + 3.6X_2 + 1.2X_3 + 1.2X_4 + \varepsilon \\ A &\sim \text{Bernoulli}(\pi) \\ \text{logit}(\pi) &= \gamma(-0.1X_1 + 0.05X_2 + 0.2X_3 - 0.05X_4 + \varepsilon') \\ X_1, X_2, X_3, X_4 &\sim \text{Normal}(0, 3^2) \\ \varepsilon, \varepsilon' &\sim \text{Normal}(0, 0.5^2) \\ \gamma &\in [0, \infty] \end{aligned}$$

We set $\gamma = 1$ while generating the data. We imitate uncalibrated models by purposefully decalibrating the true propensity scores. We do so by multiplying $\text{logit}(\pi)$ by different γ values in the range of $[0.125, 3]$, which changes the shape of the propensity score distribution. For each γ value, we generate 10 repetitions, each of $n = 10,000$ observations.

We also use data from the $\gamma = 1$ setting when fitting the statistical estimators whose results appear in the eAppendix; <http://links.lww.com/EDE/C136>.

ACIC 2016 data

To estimate the impact of postcalibration on different statistical estimators in a more realistic scenario, we use a more complex data generation process from the 2016 Atlantic Causal Inference Conference (ACIC) Data Challenge. The data is semi-synthetic and is based on real covariates from the Collaborative Perinatal Project longitudinal study that are used to synthetically simulate both treatment assignments and potential outcomes. There are multiple generating processes

and multiple realizations for each process, but each dataset has the same 4802 observations of the same 58 covariates.

Both the treatment assignment mechanism and the outcome response surface were generated using two separate “generalized additive functions,” which are the functional generating process corresponding to generalized additive models (GAMs).³³ For example, two covariates X_1, X_2 can be combined as $f(X_1, X_2) = f_1(X_1) + f_2(X_2) + f_3(X_1)f_4(X_2)$, with f_k being either an indicator function, a sum over a polynomial of a random degree, or a step function of a randomly chosen threshold. For the treatment mechanism, an additional inverse logit link function was applied to bound the propensity scores between zero and one, from which a binary treatment assignment was drawn using a Bernoulli distribution. Last, only a random subset of the 58 covariates was used; thus, all confounding variables were observed, but not all covariates were confounding variables (i.e., affecting both the exposure and the outcome). Full details on the data-generating process can be found in Dorie et al.³⁴

Using generalized additive functions creates highly flexible nonlinear response surfaces for the treatment mechanism and the potential outcomes, which can make estimation challenging. For this study, we selected five instances of a single data-generating process (numbered 42) that has a polynomial-based treatment model and a step-based potential outcomes model. This creates heterogeneity in the treatment effect and increases the chances that the models applied to the data will be ill-specified.

As mentioned above, in each of the five instances we obtain the propensity scores for each statistical model, use them to estimate the average causal effect with IPW, and take multiple measurements (e.g., calibration error and estimation error). We then postcalibrate the propensity scores and repeat the process.

RESULTS

Estimation-free Propensity Scores

In the first experiment, we check the impact of calibration on the downstream effect estimation by simulating synthetic treatment assignment and potential outcomes. We take the true propensities to treat and gradually decalibrate them by scaling them in logit space. We first use the decalibrated propensities to estimate the effect of IPW. Then we recalibrate the decalibrated propensities using Platt’s correction³² and reestimate the treatment effect. This allows us to examine the downstream effect of decalibration and postcalibration in a tightly controlled setting with dose-response-like intervention while holding many other “real-world” degrees of freedom constant.

Figure 1A shows that the error in effect estimation increases with the magnitude of the decalibration of the true propensity to treat. Specifically, it shows that before calibration (square markers) stronger deformation (darker colors) leads to larger errors (increase in the y-axis). It further shows

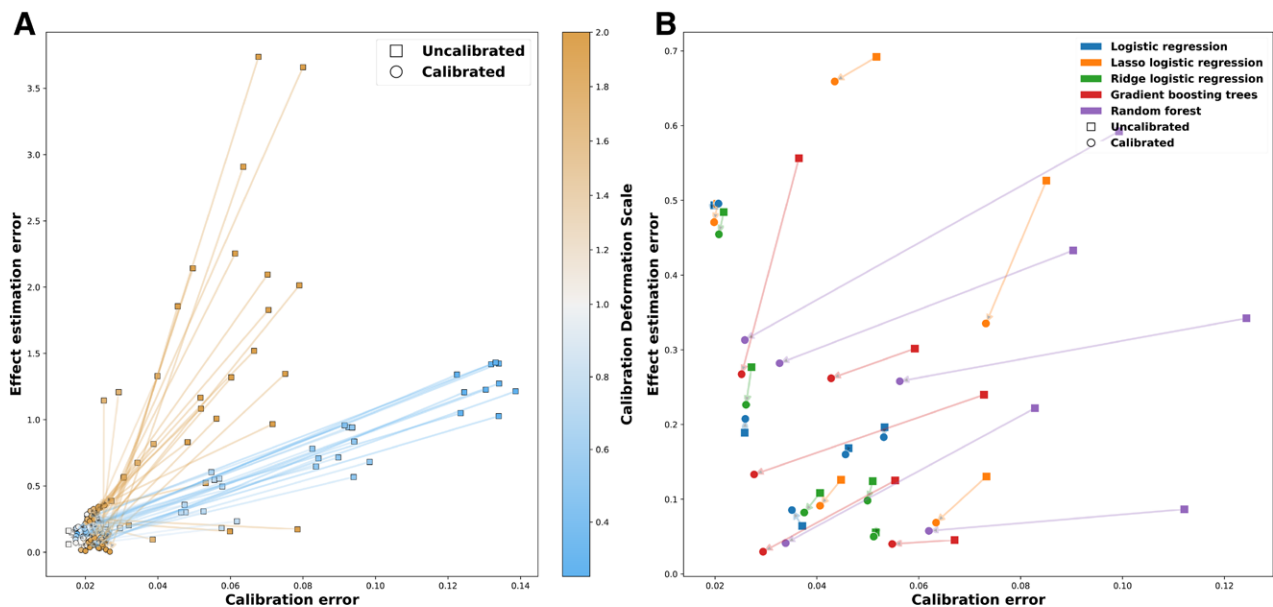


FIGURE 1. Postcalibration improves effect estimation. Using synthetic data with known effect sizes and propensities-to-treat, we plot the errors in effect estimation (y-axis) and calibration (x-axis) before (rectangular markers) and after (circles) postcalibration with arrows connecting corresponding pairs. In the left panel, propensity scores are deformed to decalibrate them and are then recalibrated using Platt's correction. The magnitude of the deformation—evaluated at $\gamma = 0.25, 0.5, 0.75, 1, 1.5, 1.75$, and 2 —is coded by color (the farther the value from 1 , the greater the deformation is and the darker its color). Before postcalibration (rectangles), the greater the propensity deformation (darker color), the greater the calibration and the effect estimation errors are. After postcalibration (circles), both the calibration error (mean integrated calibrated index) and the average treatment effect estimation error (the absolute difference between the true and the estimated effect) decrease. In the right panel, propensity scores are estimated using different statistical estimators (color-coded) on a dataset from the 2016 Atlantic Causal Inference Conference Data Challenge. Again, we see that postcalibrated propensity scores reduce calibration error and achieve smaller effect estimation error overall.

that recalibration (circle markers), reduces effect estimation error, meaning that the ATE is more accurate. eFigures 1, 2, and 3; <http://links.lww.com/EDE/C136> present additional points of view and gradual construction of this plot.

We can further quantify this improvement by calculating the average slopes between the decalibrated and the recalibrated points on the plane stretched by the calibration error and the effect estimation error. eTable 1; <http://links.lww.com/EDE/C136> shows the following for 1000 repetitions per deformation scale: First, the magnitude of the slope increases with the strength of the decalibration. Second, the sign of the slope is consistently negative, meaning postcalibration consistently reduces both the calibration error and the effect estimation error.

To better understand the decalibration and recalibration process, we can take a deeper look into the actual calibration curves of one instance for each deformation scale. In calibration curves, we bin the predicted probabilities into 10 bins and compare the rate of observed events in each bin. For well-calibrated models, we expect to see $\hat{\pi}\%$ of events among samples with a predicted propensity of $\hat{\pi}\%$, aligning with the $x=y$ diagonal. Figure 2 shows the effect of the deformation on the distribution of propensities and calibration curves. The

deformation magnitude controls the kurtosis of the propensity distribution, with small values leading to high kurtosis and large values leading to uniformly-looking, low kurtosis propensity distributions. It also shows that as deformation magnitude increases, that is, moves further away from 1 , calibration indeed deteriorates (orange line), moving further away from the $x=y$ diagonal (dashed line). However, correcting it with postcalibration (green line) improves calibration, getting it closer to the diagonal and the curve determined by the true propensity scores (blue line) and the optimal diagonal (dashed line).

The distribution of propensity scores under different γ values can explain the estimation bias in the uncalibrated results. As γ gets very small, the propensity scores cluster towards 0.5 . As a result, the IPW estimator collapses to a difference in the means estimator, which is biased because of the large covariance between the propensity scores and the potential outcomes. As γ gets very large, the propensity scores are pushed out to be very close to 0 and 1 , which means the IPW estimator overweighs a handful of treated points with propensities near 0 and control points with propensities near 1 , and relatively ignores the other points in between. The resulting estimator has a very high variance due to the extreme weights,

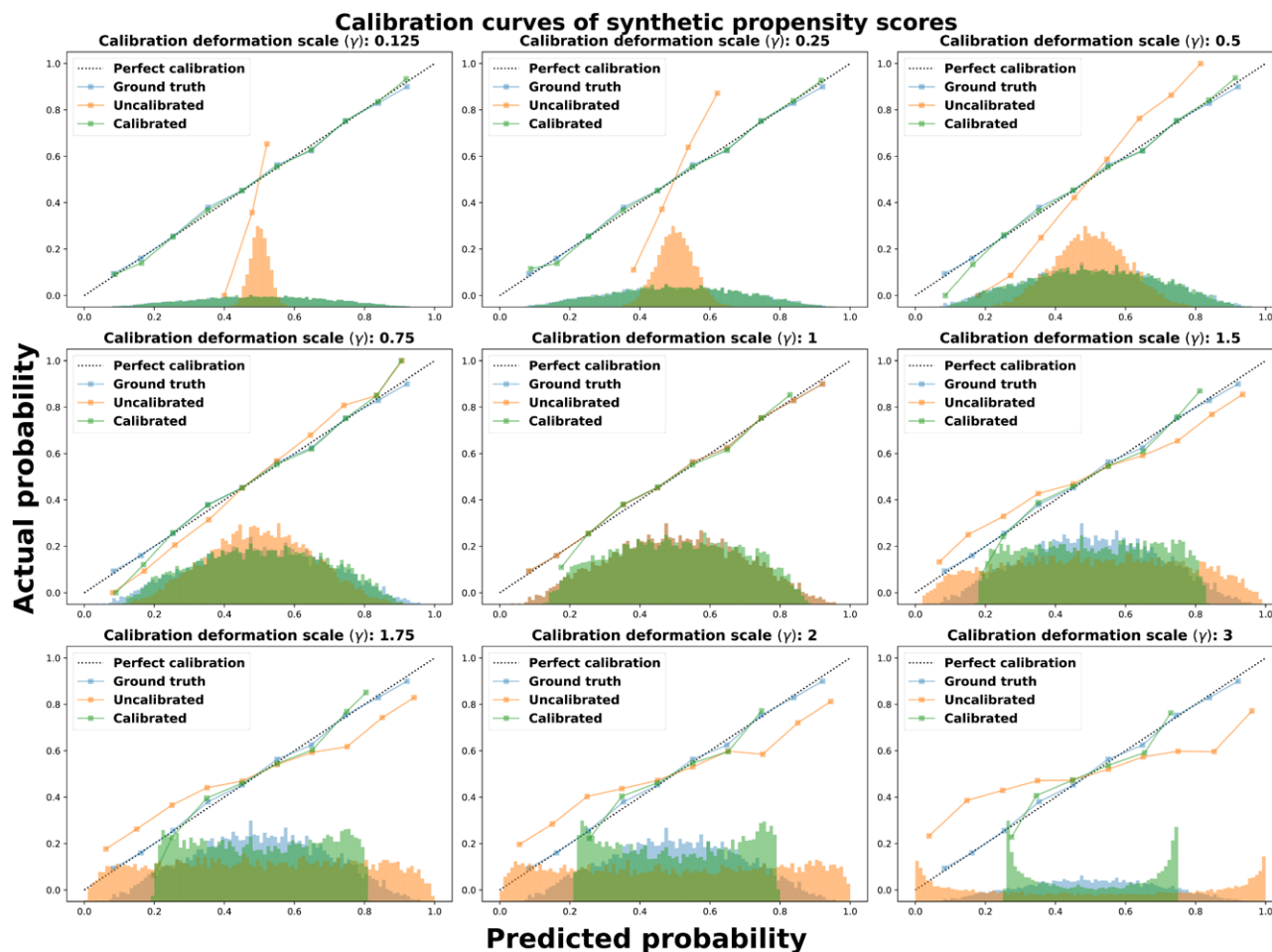


FIGURE 2. The greater the decalibration, the greater the deviation from the optimal $x=y$ diagonal of the calibration curves. This figure shows the distribution of the propensity scores and the calibration curves for different deformation scales. The greater the deformation (the farther the scale value γ is from 1), the greater the deviation from the diagonal (orange line). Postcalibrating the scores results in a calibration curve (green) closer to the curve defined by the true (before deformation) curve (blue), which is very close to the optimal diagonal curve (dashed line).

which explains the wide spread of the effect estimation error values for the boxes with large calibration deformation values.

Model-estimated Propensity Scores

In the second experiment, we examine the effect of postcalibrating different propensity estimators on the effect estimation, using more complex semi-synthetic data. We fit different propensity models using random forests, gradient boosting trees (additive trees), logistic regression, and regularized logistic regressions (LASSO and ridge), and use IPW to reweigh the outcomes to obtain an ATE estimation. We measure the calibration error and the effect estimation error, then postcalibrate the propensity scores and measure again. Figure 1B shows that postcalibration consistently reduces the error between the true and the estimated ATE.

The calibration curves from one instance of the ACIC data-generating process, shown in Figure 3, tell a similar yet less decisive story than the model-free propensity simulation

in Figure 2 (and the model-based results applied on the simpler data in eFigure 5; <http://links.lww.com/EDE/C136>). We first note that the distribution of the true propensity scores (blue lines) is skewed and has sparse tails. This increases the variance in the extreme bins, resulting in more off-the-diagonal behavior at the tails relative to the center of mass. Consequently, out of the box, the models (orange lines in each panel) are not very calibrated and diverge from the $x=y$ diagonal (dashed line). However, postcalibrating the propensity scores does improve calibration (green line), especially for the tree-based models, and moves the curve closer to the diagonal. The only small exception is the left tail of the tree-based models, which stretches further to the left and is therefore more prone to variance due to sparsity.

DISCUSSION

We performed a simulation study to assess the effect of post-calibrating propensity scores on the downstream

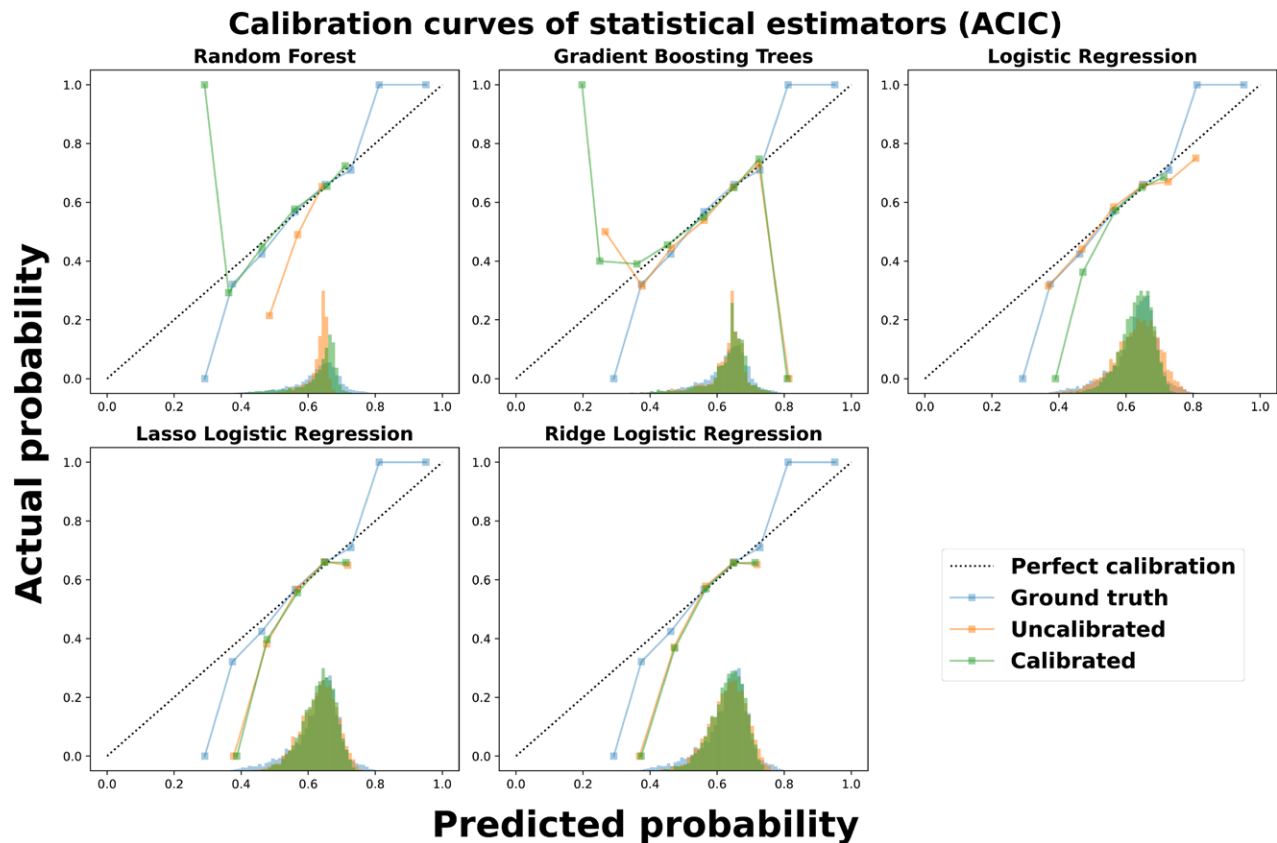


FIGURE 3. Calibration curves of different models on a single ACIC dataset. The true simulated propensity scores (blue) do not lie on the optimal $x=y$ diagonal since the tails of the distribution are thin and the variance is high. Original scores from the estimators (orange) lie farther from the diagonal because they are estimated by a binary instantiation of the true propensities. Post-calibrated scores (green) are slightly closer to the diagonal, relative to the original scores, as expected. This effect is more prominent for the tree-based models and less prominent for the logistic regression-based models.

estimation of causal effects. We show that miscalibration of propensity score models results in poor effect estimation and that postcalibrating models improve the treatment effect estimated via IPW.

We rely on simulated data because we need access to two variables that are normally unobserved. The first variable is the individual's true propensity to be treated. The second variable is the counterfactual outcomes, which is necessary for calculating the true ATE and comparing it to the estimated effect.

Postcalibration is the process of transforming the scores provided by an estimator so that they behave more like probabilities. Overall, when used for inverse weighting, postcalibrating propensity scores have consistently improved effect estimation by reducing estimation bias. Figure 1A shows this in a theoretical scenario, where the true propensity scores are gradually deformed and then recalibrated. This process allows us to perform a tightly controlled dose-response-like analysis, which demonstrates two key points. First, the stronger we deform and decalibrate the propensity scores, the more biased the effect estimated by using them is. Second, postcalibrating

these decalibrated propensity scores leads to a more accurate effect estimation.

Figure 1B (and, to a lesser extent, eFigure 4; <http://links.lww.com/EDE/C136>) shows the same conclusion in a more realistic scenario where actual statistical estimators are used. Mainly, we see that not all estimators are properly calibrated out-of-the-box and that postcalibrating them improves downstream effect estimation. Additionally, because the functional form of the treatment assignment and potential outcomes in the ACIC data is a complex set of generalized additive functions, the models used for estimation are practically misspecified. Yet, although this misspecification can introduce information bias, postcalibration still proves beneficial, and consistently so, in terms of effect estimation. This benefit, however, does not extend to more extreme cases where unconfoundedness does not hold at all (see eFigure 8a; <http://links.lww.com/EDE/C136>). But in more relaxed settings, where instead of removing a variable we just add additional noise to it—adjusting for a proxy variable—we regain that added benefit from calibration (eFigures 8b and 8c; <http://links.lww.com/EDE/C136>).

We hypothesize that the general improvement in effect estimation is mediated by improvement in balancing. The mathematical argument in the eAppendix; <http://links.lww.com/EDE/C136> demonstrates that the true conditional probabilities to be treated are the best choice for inverse probability weights. Thus, improving calibration improves weighting and subsequently improves effect estimation. Optimizing for calibration may therefore be a more sound approach than optimizing directly for balancing, because some metrics may be an imperfect proxy for balancing (see the eAppendix; <http://links.lww.com/EDE/C136> for further discussion on the common standardized mean difference) and optimizing for them directly may lead to suboptimal effect estimation.

Not all models benefit the same from postcalibration. Logistic regression and its regularized variants tend to improve less than their tree-based counterparts. This phenomenon is in line with regression-based models being more calibrated to begin with by minimizing the log-loss or a biased (i.e., penalized) log-loss. However, tree-based models seem to benefit substantially, with some instances achieving a smaller error after calibration compared with the logistic regression-based models. This seems to break the trade-off between model expressiveness and model calibration, allowing both to exist simultaneously. This also supports the claim that if more expressive models are required for modeling the exposure, they are more likely to benefit from postcalibration.

Furthermore, estimators—especially complex ones—are sensitive to underfitting and overfitting. Underfitted propensity models fail to capture the signal of the treatment assignment mechanism, resulting in their propensity scores being less informative, to begin with, and therefore might not benefit from postcalibration compared with models that are specified properly. Conversely, overfitted propensity models predict treatment assignment so well that the distribution of their scores differs across treatment groups, making overfitted models indistinguishable from positivity (overlap) violations, hence preventing us from converting our statistical estimations into causal claims. However, these issues are not unique to propensity score models but apply to general prediction models as well, and they can be partially automated through hyperparameter search and cross-validation.

Beyond the underlying estimators used to obtain propensity scores, there are also various ways to use propensity scores for adjustment. Three such main ways are through matching, stratification, and weighting. In this study, we focused on IPW because it provides a smoother, more continuous transformation of the propensities. We observed that this sensitivity to the continuous values of the propensity is what drives the change in estimates. eFigures 6 and 7 in the eAppendix; <http://links.lww.com/EDE/C136> show the results using propensity score matching and stratification that lead to no change in estimation error. The

main reason is that calibration is a monotonic function of the propensity scores, which scales and shifts all scores—of both treatment and control units—similarly. In matching, this results in the same nearest neighbors before and after calibration. Similarly, in quantile-based stratification, because the transformation is monotonic, it preserves the order (ranking) of the propensity scores, and therefore the units fall to the same quantiles before and after calibration. For fixed-interval (non-quantile) binning, there are small random changes for those units whose scores are closer to the bins' edges, which can make the calibration tip them onto the neighboring bin. More details on this result can be found in the eAppendix; <http://links.lww.com/EDE/C136>.

The Fundamental Problem of Causal Inference³⁵—the fact we cannot observe counterfactual outcomes and therefore can never have ground truth causal effects—requires us to make use of simulations for this study. Therefore, the extent to which our conclusions generalize is also dependent on how much the properties of these simulated data do exist in real data and is thus the reason why we show results for both a simple case and a complex case. In the former, while admittedly overly simplistic, we can more easily obtain insights into the effect of postcalibration on balancing, as the absolute standardized mean difference is a proper metric in such scenarios of no covariance between covariates (see eAppendix and eFigure 9; <http://links.lww.com/EDE/C136>). In the latter, the more complex ACIC data shows how postcalibration improves effect estimation in more realistic scenarios where the true underlying functional form is a complex nonlinear function that is not necessarily captured by the statistical estimator used to obtain the propensity scores.

Limitations notwithstanding, postcalibrating estimators allow us to reconcile theory—guarantees that rely on the true conditional probability of exposure—and practice—as in applied numerical modeling from data using statistical software. Postcalibration is a simple postprocessing procedure available in common statistical software, and that can be done on any statistical estimator. Normally, it is not computation-intensive, allowing us to utilize more complex models at a relatively small additional cost. Therefore, we conclude that postcalibrating propensity score models can be beneficial for effect estimation.

ACKNOWLEDGMENTS

We would like to thank our reviewers for their help in improving this manuscript. We specifically thank our Anonymous Reviewer 1 for their mathematical interpretation of the distributions under different deformation values and their effect on estimation bias in the deformation-recalibration results, which we incorporated almost verbatim to this work. We would additionally like to thank our colleagues from IBM Research and the Technion for proof-reading this article.

REFERENCES

- Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11:550–560.
- Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70:41–55.
- Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am Stat*. 1985;39:33–38.
- Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc*. 1984;79:516–524.
- Kang JDY, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Stat Sci*. 2007;22:523–539.
- Schuler MS, Rose S. Targeted maximum likelihood estimation for causal inference in observational studies. *Am J Epidemiol*. 2017;185:65–73.
- Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics*. 2005;61:962–973.
- Glynn AN, Quinn KM. An introduction to the augmented inverse propensity weighted estimator. *Political Anal*. 2010;18:36–56.
- Stürmer T, Schneeweiss S, Rothman KJ, Avorn J, Glynn RJ. Performance of propensity score calibration—a simulation study. *Am J Epidemiol*. 2007;165:1110–1118.
- Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol*. 2016;74:167–176.
- Zadrozny B and Elkan C. Transforming classifier scores into accurate multiclass probability estimates. In Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2002:694–699.
- Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. *J Am Med Inform Assoc*. 2020;27:621–633.
- Harrell FE. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer; 2001: 608.
- Austin PC, Steyerberg EW. The integrated calibration index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Stat Med*. 2019;38:4051–4065.
- Tazare J, Wyss R, Franklin JM, et al. Transparency of high-dimensional propensity score analyses: guidance for diagnostics and reporting. *Pharmacoepidemiol Drug Saf*. 2022;31:411–423.
- Granger E, Watkins T, Sergeant JC, Lunt M. A review of the use of propensity score diagnostics in papers published in high-ranking medical journals. *BMC Med Res Methodol*. 2020;20:1–9.
- Shimoni Y, Karavani E, Ravid S, et al. An evaluation toolkit to guide model selection and cohort definition in causal inference. *arXiv*. 2019.
- Jordan MI. *Why the logistic function? A tutorial discussion on probabilities and neural networks*; 1995.
- Friedman J, Hastie T, and Tibshirani R. *The Elements of Statistical Learning*. New York: Springer series in statistics; 2001.
- Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. *J Am Stat Assoc*. 2007;102:359–378.
- Wüthrich MV and Merz M. *Generalized Linear Models chapter 5, pages 111–205*. Cham: Springer International Publishing; 2023.
- Van Calster B, van Smeden M, De Cock B, Steyerberg EW. Regression shrinkage methods for clinical prediction models do not guarantee improved performance: simulation study. *Stat Methods Med Res*. 2020;29:3166–3178.
- Šinkovec H, Heinze G, Blagus R, Geroldinger A. To tune or not to tune, a case study of ridge logistic regression in small or sparse datasets. *BMC Med Res Methodol*. 2021;21:199–213.
- Westreich D, Lessler J, Funk MJ. Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. *J Clin Epidemiol*. 2010;63:826–833.
- Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66:688–701.
- Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol*. 1996;73:273–282.
- Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*. 1970;42:80–67.
- Breiman L. Random forests. *Mach Learn*. 2001;45:5–32.
- Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001;1189–1232.
- Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev*. 1950;78:1–3.
- Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *J Am Stat Assoc*. 1952;47:663–685.
- Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*. 1999;10:61–74.
- Hastie T, Tibshirani R. Generalized additive models: some applications. *J Am Stat Assoc*. 1987;82:371–386.
- Dorie V, Hill J, Shalit U, Scott M, Cervone D. Automated versus do-it-yourself methods for causal inference: lessons learned from a data analysis competition. *Stat Sci*. 2019;34:43–68.
- Holland PW. Statistics and causal inference. *J Am Stat Assoc*. 1986;81:945–960.