Improving Inverse Probability Weighting by post-calibrating its propensity scores: **online appendix**

Rom Gutman, Ehud Karavani, and Yishai Shimoni

Contents

eAppendix

A Code for computing calibrated propensity scores

This is a minimal example code for estimating causal treatment effects with IPW using calibrated propensity scores. It is provided for illustrative purposes and comes with absolutely no warranty.

```python
# Let `X` denote a matrix of covariates (ideally confounding variables)
# Let `a` denote a vector of binary treatment assignment
# Let `y` denote a vector of observed outcomes
# Let `model` denote some statistical estimator, e.g., `sklearn.ensemble.
    GradientBoostingClassifier()`

model.fit(X, a)
ps = model.predict_proba(X)[:, 1]
# Optionally, calculate calibration error and/or balancing error
effect = calculate_effect(ps, a, y)

cal_ps = calibrate(y, ps)
# Optionally, calculate calibration error and/or balancing error
cal_effect = calculate_effect(cal_ps, a, y)


def calibrate(treatment, propensity):
    from sklearn.calibration import _sigmoid_calibration
    b, a = _sigmoid_calibration(propensity, treatment)
    calibrated_logits = a + b * propensity
    calibrated_propensity = 1 / (1 + np.exp(-calibrated_logits)
    return calibrated_propensity
    # Equivalent to:
    # from sklearn.linear_model import LogisticRegression
    # calibrator = LogisticRegression(penalty="none")
    # calibrator.fit(propensity.reshape(-1, 1), treatment)
    # calibrated_propensity = calibrator.predict_proba(propensity.reshape(-1, 1))

def calculate_effect(propensity, treatment, outcome):
   ip_weights = (treatment / propensity) + ((1 - treatment) / (1 - propensity))
   ate = np.average(outcome, weights=ip_weights)
   return ate
```

Listing 1: An illustration of IPW with calibrated propensity scores using Python.

```r
install.packages("devtools")
library("devtools")
install_github("etlundquist/eRic")
library(etlundquist/eRic)  # for Platt's Calibration

# Let `X` denote a dataframe of covariates (ideally confounding variables)
# Let `a` denote a vector of binary treatment assignment
# Let `y` denote a vector of observed outcomes

model <- glm(a ~ ., family = "binomial", data = cbind(X, a))
ps <- predict(model, type = "response")
# Optionally, calculate calibration error and/or balancing error
effect <- calculate_effect(ps, a, y)
```

```
cal_ps <- calibrate(a, ps)
# Optionally, calculate calibration error and/or balancing error
cal_effect <- calculate_effect(cal_ps, a, y)

calibrate <- function(treatment, propensity){
   calibrated_propensity = prCalibrate(treatment, propensity)
   # Equivalent to:
   # calibrator <- glm(treatment ~ 1 + propensity, family = "binomial")
   # calibrated_propensity = predict(calibrator, type = "response")
   return(calibrated_propensity)
}

calculate_effect <- function(propensity, treatment, outcome){
   ip_weights = (treatment / propensity) + ((1 - treatment) / (1 - propensity))
   ate = weighted.mean(outcome, ip_weights)
   return(ate)
}
```

Listing 2: An illustration of IPW with calibrated propensity scores using R. Not tested.

B Estimation-free propensity scores

In Figure 1a, we show the results for ten different random datasets per deformation scale. To obtain more rigorous statistics, we repeat this process a thousand times per deformation scale, rather than ten. For each such instance, we calculate the slope between the pre- and post-calibration point on the calibration-error effect estimation error plane. The summary of these 1000 slopes per deformation scale is presented in eTable 1. Here we can see two things: First, the greater the deformation on each side of the deformation scale (below and above 1), the greater the magnitude of the slope. Second, the negative sign suggests that post-calibration consistently reduces the calibration error and the effect estimation error. For the deformation scale of 1.0, meaning, no deformation at all, we would expect a zero slope. However, we do see a small median slope of minus one, but with a large interquantile range below and above 0, hinting the slope is practically zero with some variation between instances.

Deformation scale	1st quartile	Median	3rd quartile
2.0	-54.10	-39.40	-25.06
1.75	-52.43	-36.47	-21.51
1.5	-96.84	-56.91	-18.41
1.0	-20.07	-1.03	23.10
0.75	-9.31	-7.41	-0.74
0.5	-10.08	-8.91	-5.89
0.25	-10.83	-9.95	-8.04

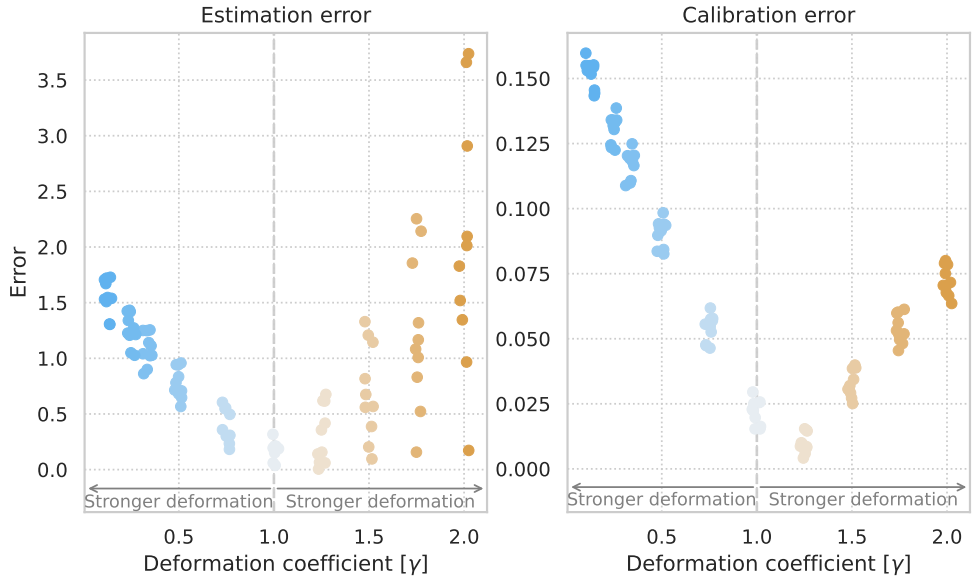
eTable 1: Slopes between pre- and post-calibration points on the calibration-error over effect-estimation error plane. Statistics taken for 1000 datasets per deformation scale show us the following: First, the greater the deformation, the greater the magnitude of the slope. Second, the negative sign suggests that post-calibration constantly reduces both the calibration error and effect estimation error.

C Extended view of Figure 1a

Figure 1a in the main text packs multiple dimensions into a single plot. Here, we provide additional views of the same data that gradually build towards the presentation in Figure 1a.

We focus on the model-free propensity scores, those obtained by deforming the true propensity score and then recalibrating it back. As a reminder, we have a deformation parameter (γ) depicting the deformation strength that affects the propensity score on the multiplicative scale. Therefore, the further γ is from 1, the stronger the deformation it applies to the propensity score. Smaller γ concentrate the distribution more closely around 0.5 (treatment prevalence), and larger γ pushes the distribution to the extreme values close to 0 and 1.

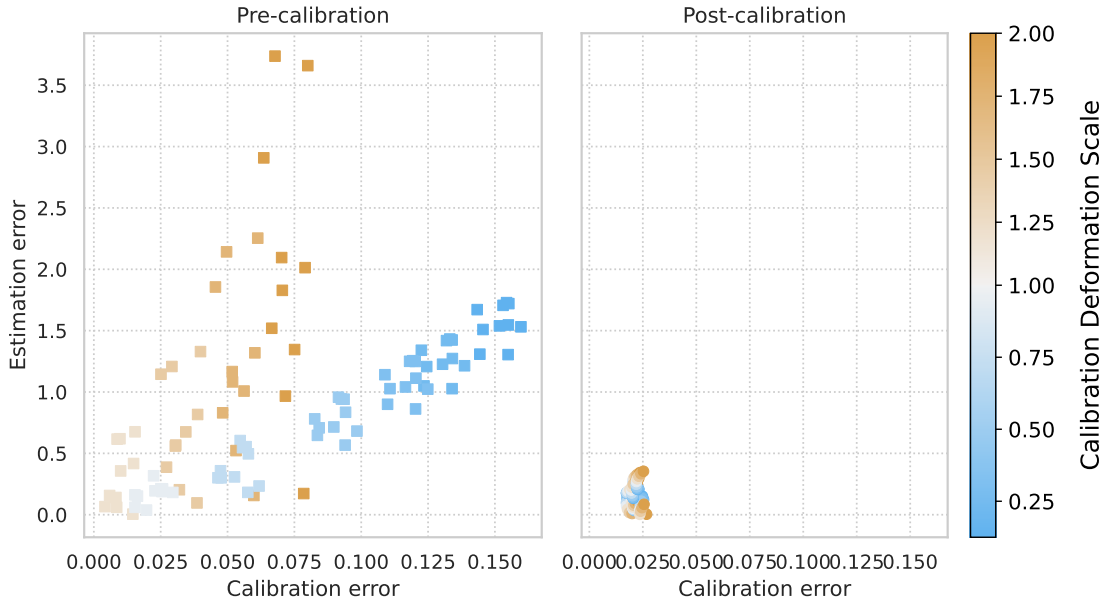
First, eFigure 1 shows the estimation error and calibration error as a function of the deformation strength (γ). We explicitly place the deformation parameter on the x-axis (with slight random jitter so the points less overlap) to clearly show the relationship. However, we keep the coloring – that also correspond to the x-axis location – although redundant in this case, because this is how code the deformation strength in Figure 1a and subsequent explanatory figures. The coloring is a diverging palette around the inflection point of 1, with darker blues corresponding to smaller γs, darker oranges to larger γs, and in-between values closer to 1 being lighter and more transparent. We see the further away γ is from 1 the larger the estimation error (left) and the calibration error (right).



eFigure 1: Estimation error (left) and calibration error (right) as a function of the deformation strength γ (x-axis, jittered). We color the dots according to their γ value to correspond with its coding in Figure 1a and in subsequent explanatory figures of this section.

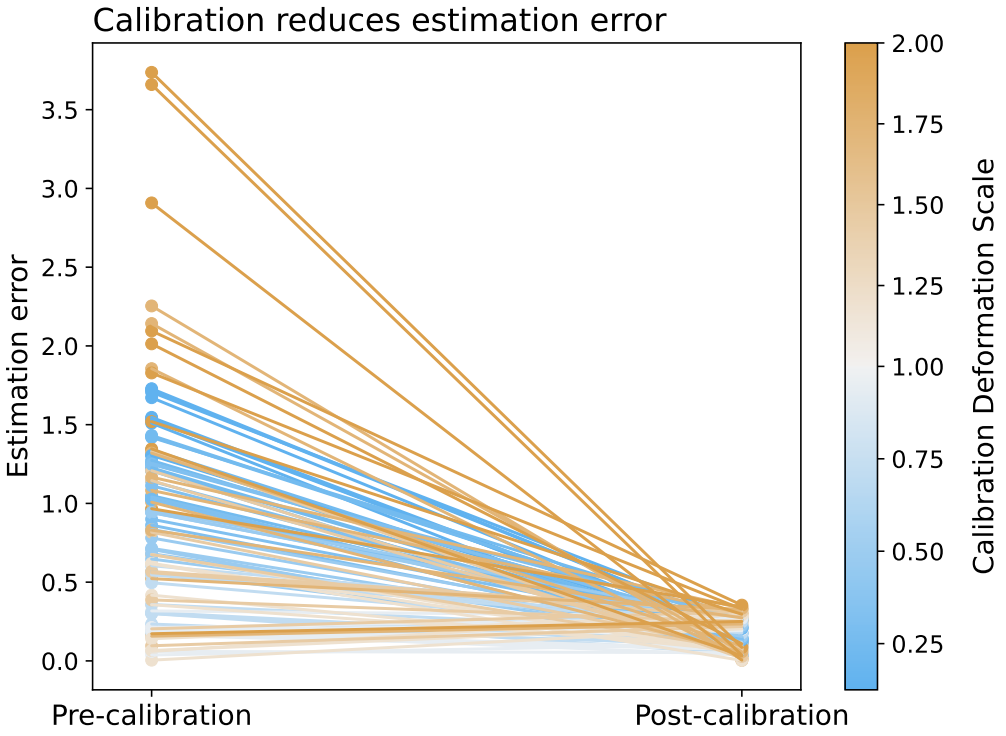
Second, we present the same error-plane as depicted in Figure 1. In eFigure 2 we have the error in calibration (average of Integrated Calibration Index) on the x-axis and the error in estimation (absolute difference between estimated ATE and true ATE) on the y-axis. However, unlike Figure

1, we separate the plot into two panels: before calibration (left) and after calibration (right). Additionally, although redundant here, we keep the same marker coding as in Figure 1a with pre-calibration points coded as squares and post-calibration points coded as circles. We can again see that the stronger the deformation (darker color) the larger the calibration error (rightward) is and the larger the estimation error (upward). But mainly, we can see that after calibration (right) the error is reduced in both calibration and estimation. We note that unlike Figure 1, this presentation only shows that calibration reduces error *on average*. We cannot track a single instance across the panels the same way Figure 1 does when connecting pre- to post-calibration instances in a line. However, the removal of lines does relieve the clutter and enable seeing the relationship on the left panel more clearly



eFigure 2: The average effect of calibration in reducing calibration error and estimation error. Comparing the errors before calibration (left) and after calibration (right).

Third, in order to show reduction *per-instance*, rather than on average, we need to connect the pre- and post-calibration of the instance in a line. This allows us to follow the path of estimation-error reduction for every individual instance and see how the action of calibration affects it. However, in order to reduce clutter, the slope graph in eFigure 3 does not show the calibration *error* on the x-axis (like Figure 1 and eFigure 2, but only whether post-calibration was done. Although this removes the relationship between the *size* of the calibration error and the estimation error and deformation strength, it allows us to better see the effect of the *action* of performing calibration for any single instance.

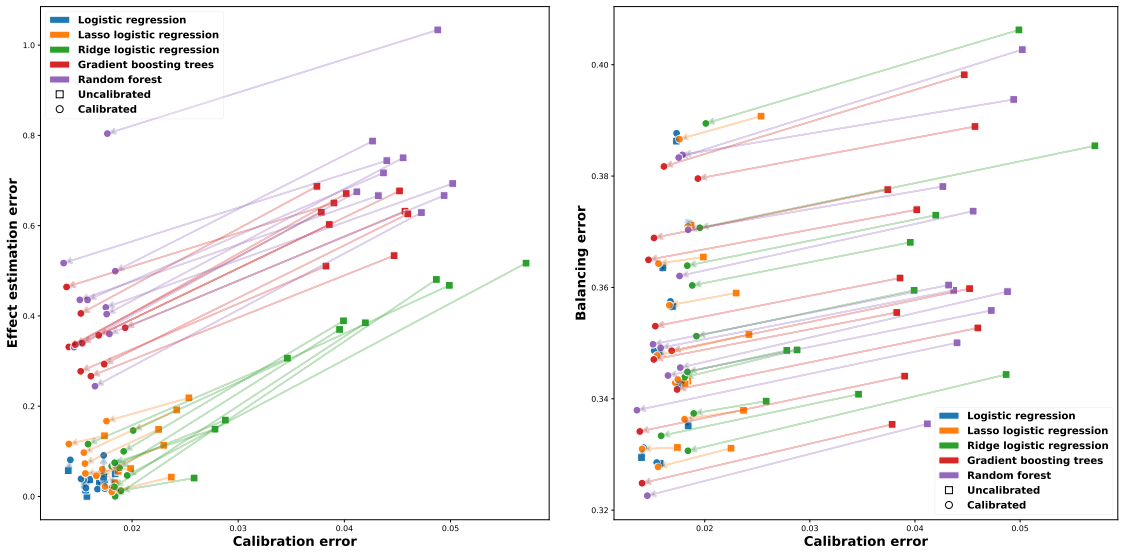


eFigure 3: The individual effect of calibration in reducing estimation error. Here we ignore the reduction in calibration error and focus on estimation error alone, showing how the act of calibration (before on the left, and after calibration on the right) reduces the error.

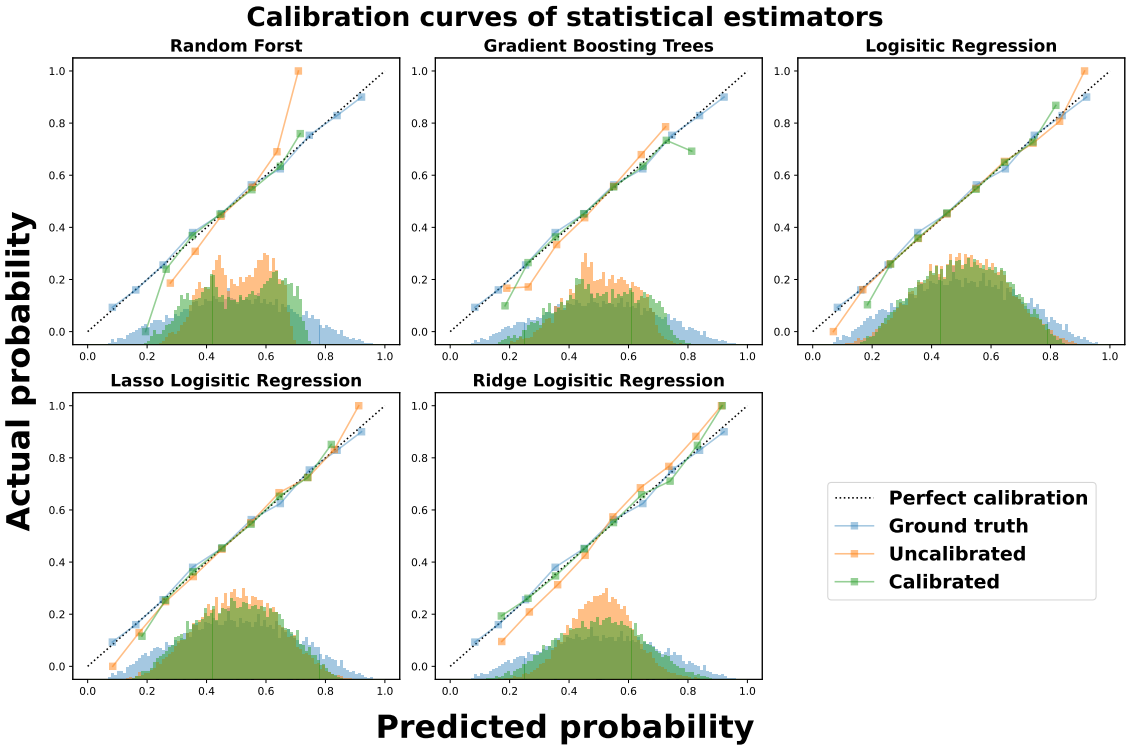
D Model-estimated propensity scores on simple simulation

We also applied statistical estimators to the data generating process described in the Methods section, setting $\gamma = 1$. The results in eFigure 4 show that the improvement caused by post-calibration is consistent, similar to what was shown on the 2016 Atlantic Causal Inference Conference data (Figure 1b). Here, we also see that the logistic regression model has a small and inconsistent benefit. This might be due to the model already being perfectly specified to the data and well-calibrated, having no additional benefit from an additional logistic regression-based calibration. We also see that tree-based models perform worse than the logistic regression-based models. This is probably due to the data being generated via regression, and trees being less adequate to fit linear trends (given finite complexity).

To further see the effect of post-calibration, we choose an arbitrary simulation instance and plot its corresponding calibration curves. In eFigure 5, we see that the post-calibrated curves (green) are closer to the optimal x=y diagonal than the original curves. We also see that the post-calibrated score distributions are more dispersed (relative to the uncalibrated scores distribution) to match the ground truth distribution. However, the distribution of scores from the logistic regression-based methods do not differ substantially from the uncalibrated ones. This might be because the simple simulation uses a logit function of a linear combination, making the logistic regression perfectly-specified, and leaving no place for improvement by the post-calibration procedures.



eFigure 4: Calibration improves effect estimation in a simple simulation scenario.



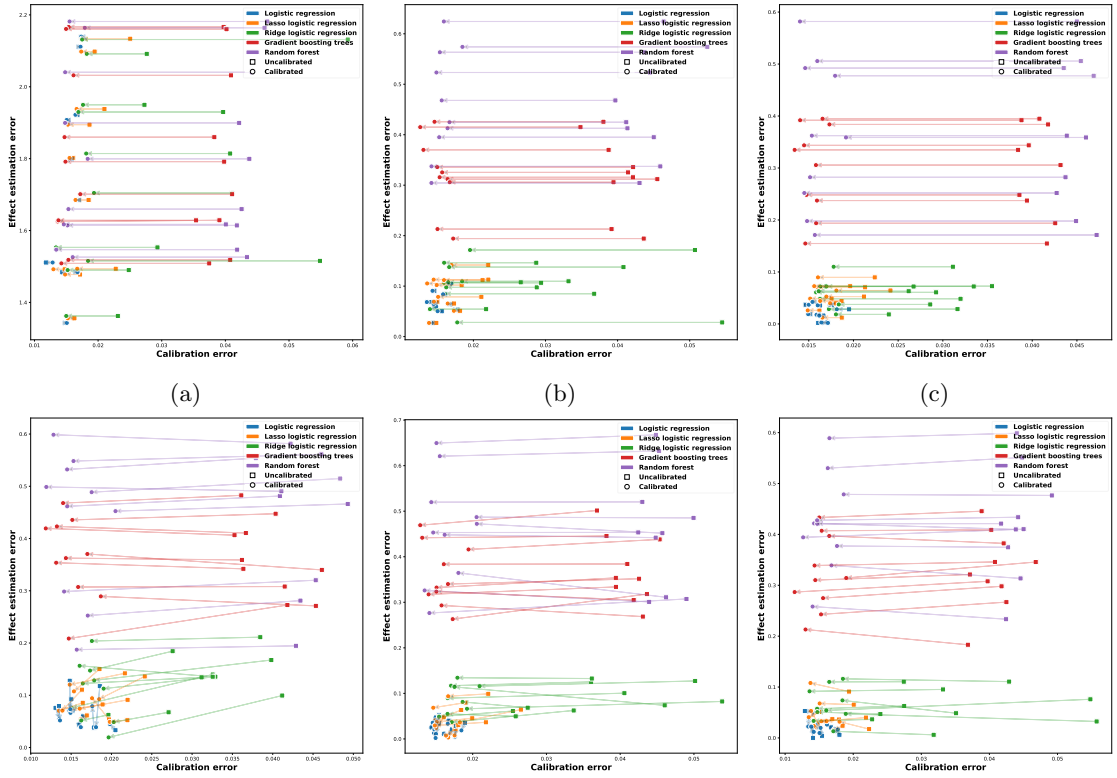
eFigure 5: Post-calibration improves calibration curves when different statistical estimators are applied to simple simulation data.

E Post-calibration with matching and stratification

The main text focuses on estimation using Inverse Propensity Weighting (IPW). However, we also ran the exact same experiment design using nearest neighbor matching with euclidean distance, quantile-based stratification and fixed-interval stratification for 10, 20, and 30 strata [1]. In these experiments, shown in eFigure 6 for the statistical estimators applied on the simple simulated data described in the Methods and in eFigure 7 for the deformation-recalibration experiment, we do not observe any change in average treatment effect estimation.

This can be explained by the combination of the calibration being a monotonic transformation and the coarsed treatment of the propensity scores. The calibration shifts propensity scores similarly, and therefore, for matching the same nearest neighbor before calibration is also the nearest neighbor after calibration. For quantile-stratification the same units are kept in the same bins before and after calibration, since it is a monotonic transformation and thus preserves the order (ranking) of the propensity scores. Since the order is invariant to calibration, the qunatiles (e.g., the lowest decile or the highest decile) are also invariant to it. For fixed-interval stratification, we do see some minor changes, but these seem to improve or worsen estimation error at random. This can be explained by the calibration tipping units on the edges of the bins into the neighboring bins. Since the values of the bins edges are an arbitrary choice of the researcher, units closer to the edges are closer to it due to chance, and thus tipping them to the neighboring bin results in a non-consistent effect on the estimation error. We can further see that the magnitude of these changes, while still small, is larger for 10 strata, as the difference across two neighboring strata is larger (each strata has more units and is less homogeneous); and becomes even smaller in 30 strata where the discretization of the propensity scores is fine enough that two neighboring bins are still very similar and the average effect in the two bins is more similar. Lastly, the added benefit for IPW seems to be due its continuous handling of the propensity scores, allowing it to extract the information added by the small changes imposed by the calibration process.

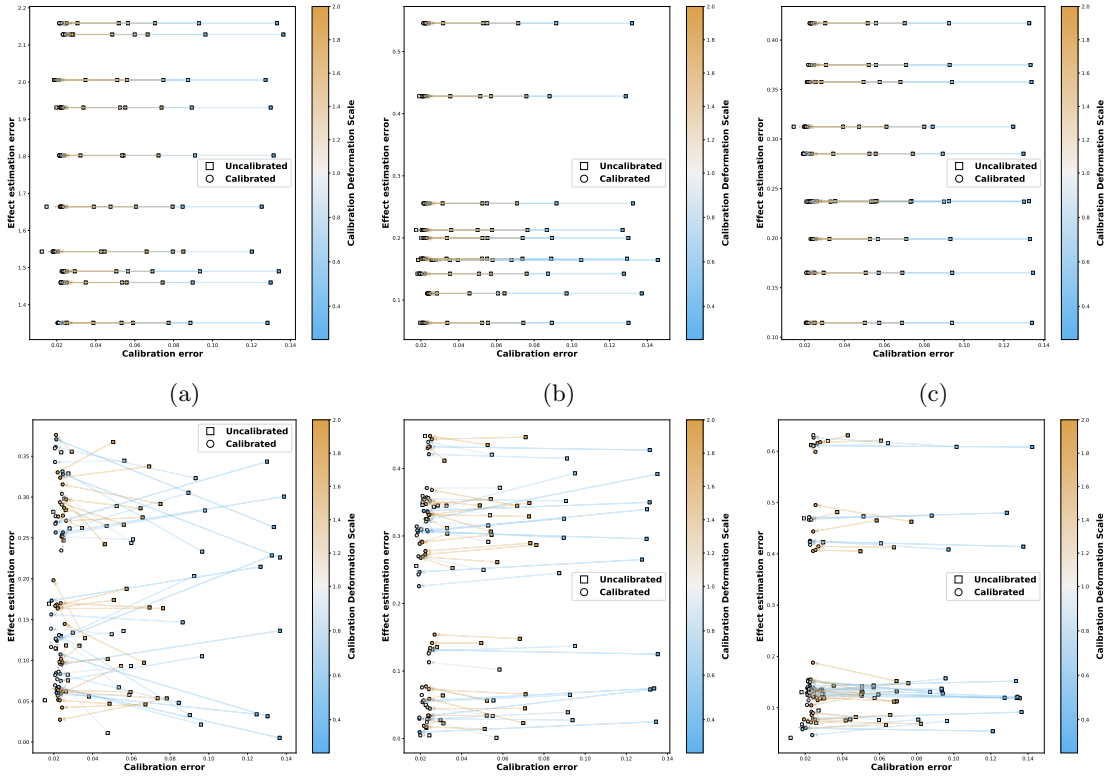
In addition to IPW's sensitivity to continuous values, there is also its dependence on the scale of these continuous values. IPW is the only method that cares for the absolute value of the propensity score. Matching and stratification care more about the propensity scores relative to other units' propensity scores; they only need to be closer to the "correct" units from the other group, but they can be scale-invariant. As an example, in logistic regression models, one could match or stratify on the linear predictions of the model, before the inverse-logit transformation, and obtain the same results. Such an approach will not work for IPW, which requires the single-scalar summary of the covariate space to strictly be in the zero-one probability space.



(a) (b) (c)

(d) (e) (f)

eFigure 6: No (substantial) changes in effect estimation are observed when using (a) nearest-neighbor matching, (b) quantile-stratification with 10 strata, (c) quantile-stratification with 30 strata, (d) fixed-interval stratification with 10 strata, (e) fixed-interval stratification with 20 strata, (f) fixed-interval stratification with 30 strata,



(a) (b) (c)

(d) (e) (f)

eFigure 7: matching and stratification

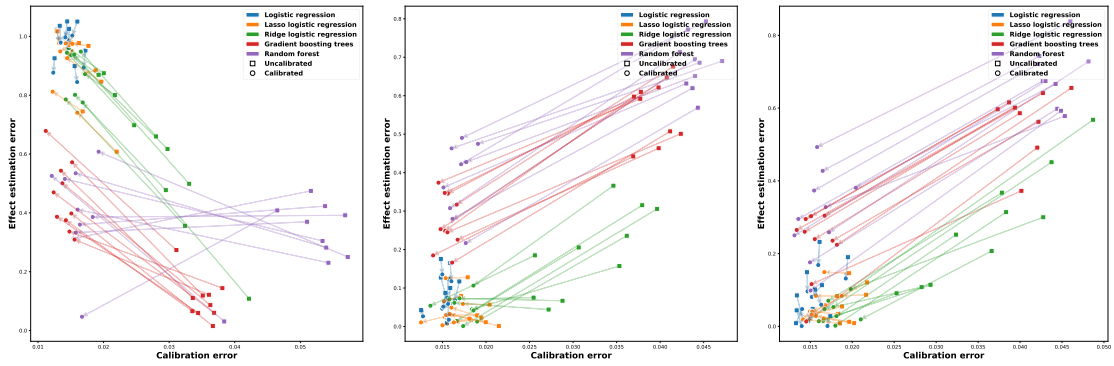
F Post-calibration under confoundedness and noisy proxies

We also tested the effect of post-calibration in cases of strong confoundedness and under noisy proxies of the variables. We used the simple data generating process described in the Methods section as a basis for these experiments.

For the confoundedness experiment, we restricted the models' access to one of the variables (X_1). Namely, the data generating process used it, but it was dropped before the statistical estimators could use it. eFigure 8a shows that post-calibration had generally worsen effect estimation. Especially the highly-adaptive model of gradient boosting trees, which was able to accomplish a relatively low estimation error in its original fit, has post-calibration worsening it. On the other hand, unregularized logistic regression – although quite calibrated to begin with – has some marginal improvement in effect estimation while having no improvement calibration error. Random forests, while being quite unclibrated to begin with and improving their calibration error substantially, had no consistent improvement in effect estimation.

We further relaxed the confoundedness setting, and instead of removing a single variable (X_1), we added noise to it. Namely, the data generating process used the original variable, but it was changed before the statistical estimators could use it. Formally, we created a variable $X_1' = X_1 + \xi$, once with unbiased noise $\xi \sim \text{Normal}(0, 2)$ (eFigure 8b) and once with biased noise $\xi \sim \text{Normal}(2, 2)$ (eFigure 8c). We see that in both cases calibration consistently improves the effect estimation.

These results extend the ones obtained on the ACIC data (Figure 1b). In the ACIC setting, all confounding variables are observed, but the model used misspecify the functional form with high probability, resulting in information bias due to model misspecification. In that setting, there is still consistent benefit for calibrating the propensity scores.



(a) (b) (c)

eFigure 8: Calibration under confoundedness and noisy proxies. In a setting when a confounding variable is entirely unobserved (a) we see calibration worsens the accuracy of effect estimation, even though it still improves calibration. In slightly more relaxed setting where the statistical estimators are exposed to an unbiased noisy proxy (b) or a biased noisy proxy (c) of subset of the true confounding variables, we see calibration improves the effect estimation in a consistent way.

G Calibration and balancing

Mathematical motivation

We motivate the desiderata for propensity score to be calibrated by referring theoretical guarantees requiring the true conditional probability to be treated. We now lay this argument more mathematically. In general, a weighting method, provided with confounding variable X and binary treatment assignment A, tries to find weights $w(X, A)$ such that $p(X) = w(X, A) \cdot p(X|A)$. Namely, find weights so that the weighted confounding variables distribution in either treatment groups (conditional) is equal to the overall (marginal) confounding variables distribution. Following transitivity, achieving that also guarantees the confounding variables distribution in the treatment group matches that of the controls: $p(X|A = 1) \cdot w(X, A) = p(X) = w(X, A) \cdot p(X|A = 0)$. This weighted distribution is what commonly referred to as the *pseudo-population* [2]. When $w(X, A)$ achieve this equality, we say that these weights *balanced* the confounding variable distribution of the treatment groups.

Since $p(X|A)$ is a multivariate distribution, we can simplify it by using Bayes' theorem to get:

$$\begin{aligned} p(X) &= w(X, A) \cdot p(X|A) \\ p(X) &= w(X, A) \cdot \frac{p(A|X)p(X)}{p(A)} \\ 1 &= w(X, A) \cdot \frac{p(A|X)}{p(A)} \end{aligned}$$

which provides

$$w(X, A) = \frac{p(A)}{p(A|X)}$$

which are the (stabilized) inverse propensity score[1] based on the true conditional probability of being treated. Therefore, weighting using a good propensity model is equivalent to full balancing.

Standardized mean difference

The (Absolute) Standardized Mean Difference (ASMD) is the most common method to evaluate imbalance in propensity score-based studies [3]. It evaluates the difference between the treatment and control groups using a Cohen's d-based metric, subtracting the standardized mean of a covariate between groups:

$$d_j = \frac{\bar{x}_{j,t} - \bar{x}_{j,c}}{\sqrt{\frac{\sigma^2_{j,t} + \sigma^2_{j,c}}{2}}}$$

Where $\bar{x}_{j,t}, \bar{x}_{j,c}$ are the average of covariate x_j in the treatment and control groups, respectively, and, likewise, $\sigma^2_{j,t}, \sigma^2_{j,c}$ are the variance of covariate x_j in the treated and control groups.

Often we only care about the magnitude of the difference, not the direction, and therefore the absolute value $|d_j|$ is the metric of interest. When there are multiple covariates and still want a single scalar summary, we often take the maximum value over all d_js. The maximum is a worst-case measurement, ensuring the absolute differences between *all* covariates are below a certain

[1] Often, Bayes theorem is written proportionally and without the evidence term, $p(A)$, in the denominator. In this case the weights can be some arbitrary scale of unstabilized $1/p(A|X)$ weights. And, namely, the unstabilized weights themselves.

acceptable threshold (with 0.1 being regarded as an indication of good balancing). Hence, this often may assure no single covariate can introduce too large of a bias.

Unfortunately, the ASMD can sometimes be a poor metric to evaluate multivariate distributional differences between two groups. Primarily, it checks each covariate marginally, disregarding any information of the joint distribution. However, it is possible for two covariates to be marginally balanced while their joint distribution is imbalanced [4, Figure 1].

Therefore, the results in this study examining the relationship between calibration error and balancing error before and after calibration, can sometimes be inconsistent. First, we can examine the relationship between the calibration error and the maximal ASMD (balancing error) in a relatively simple setting where all covariates are independent. eFigure 9a shows a consistent reduction in max ASMD after calibration. This behavior concur with our mathematical formulation that better calibration leads to better balancing. However, the reduction is not as consistent in the more complicated ACIC data (eFigure 9b, which may be an example for when ASMD serves as a poor measure for balancing. Namely, the models might balance the joint distribution which is poorly capture by examining each covariate separately. This is further suggested by noting that although improvement in the max ASMD is inconsistent, the improvement in the effect estimation error is more consistent (Figure 1b. Additionally, there is the confounded case as described before in the eAppendix, where covariates don't share any covariance. In it (eFigure 9c, we still see a constant reduction in the ASMD of the *observed* covariates. However, eFigure 8a shows consistent *increase* in estimation bias, reinforcing the prevailing reliance on the no hidden confounding variables assumption. Lastly, we also present the improvement in max ASMD when using matching (eFigure 9d), stratification (9e), and quantile-stratification (9f) –instead of IPW– which resulted in no significant change in effect estimation although we see consistent improvement in max ASMD reduction (eFigure 6), further demonstrating these models are invariant to post calibration.

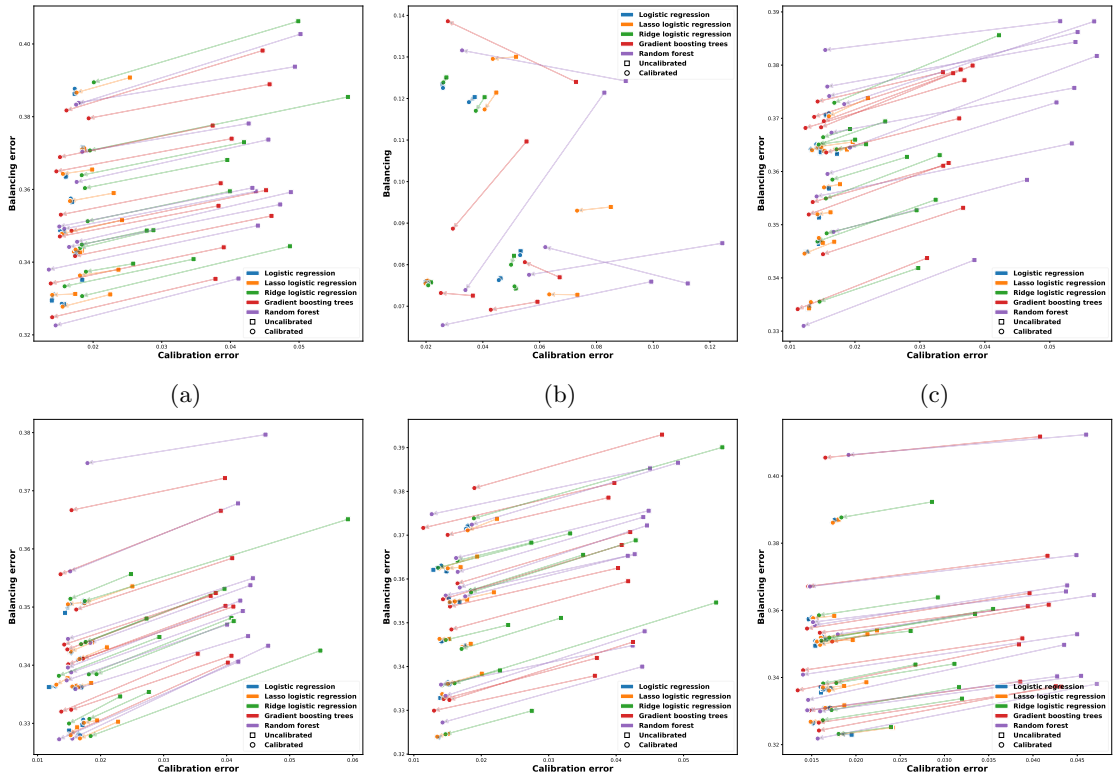
References

[1] Markus Neuhäuser, Matthias Thielmann, and Graeme D Ruxton. The number of strata in propensity score stratification for a binary outcome. *Archives of Medical Science*, 14(3):695–700, 2018.

[2] James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

[3] Emily Granger, Tim Watkins, Jamie C Sergeant, and Mark Lunt. A review of the use of propensity score diagnostics in papers published in high-ranking medical journals. *BMC Medical Research Methodology*, 20(1):1–9, 2020.

[4] Ehud Karavani, Peter Bak, and Yishai Shimoni. A discriminative approach for finding and characterizing positivity violations using decision trees. *arXiv preprint arXiv:1907.08127*, 2019.



(a) (b) (c)

(d) (e) (f)

eFigure 9: The relationship between calibration error and imbalance error, measured as the maximal absolute standardized mean difference across covariates. In (a) we present a simple setting of statistically independent covariates, where max ASMD is a good proxy for imbalance and we see a reduction in imbalance after calibration. In (b) we show a more complex data generating process with structural covariance among covariates where the max ASMD can be a poor proxy for imbalance, and we see inconsistent effect of calibration on balancing. Furthermore, (c) shows a setting where unconfoundedness does not hold, and while the ASMD on the *observed* covariates is reduced after calibration, it does not guarantee reduction in effect estimation. Lastly, we also exhibit the improvement in max ASMD when the causal model is matching (d), stratification (e), and quantile-based stratification (e), which resulted in no change in effect estimation.