## MAIN TAKEAWAY

Complementary to Standardized Mean Difference, which evaluates bias correction by features, our metric measures how well these methods retrieve the unbiased, cutting effect estimation errors by up to 50%.

# Improving Bias Correction Standards by Quantifying its Effects on Treatment Outcomes

Alexandre Abraham and Andrés Hoyos Idrobo

implicity Rakuten

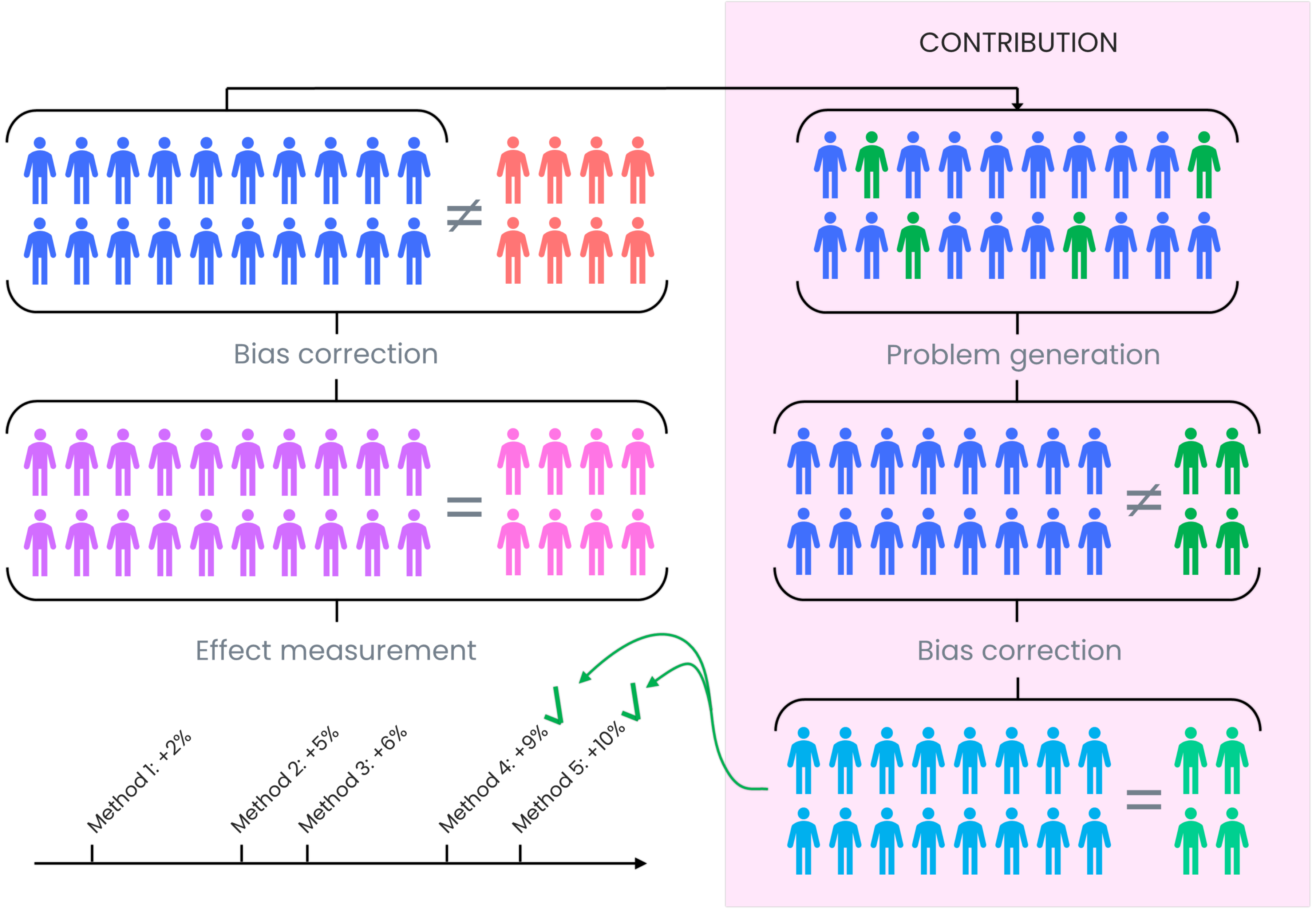
This poster and more !

**Problem.** Bias correction methods considered valid by today's standards can yield significantly different results depending on the method used.

**Contribution.** We propose a new metric that generates numerous artificial bias correction tasks from one of the populations. We claim that the ability of bias correction methods to retrieve the true effect in these artificial tasks correlates with their ability to do so in the real task.

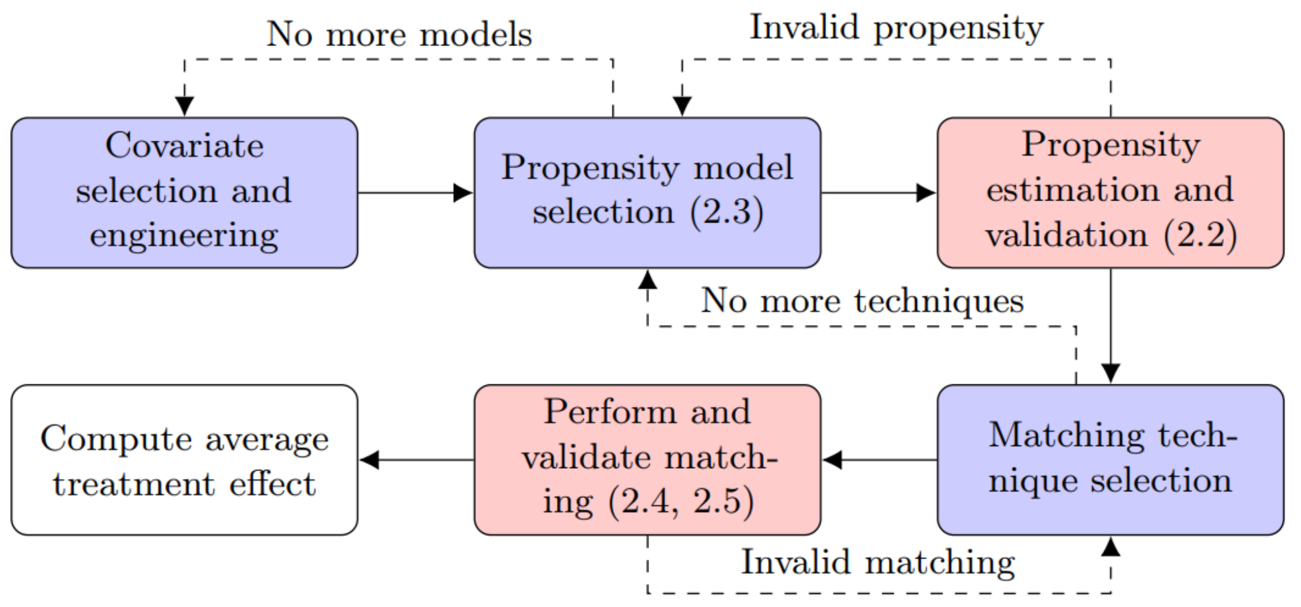
**Corollary contribution.** Because our metric requires running hundreds of bias corrections, we automate the entire process, including validation, to replace manual steps.

**Results.** In simulations, methods that perform best on artificial tasks exhibit a lower average estimation error of the true effect. On real tasks from public datasets, we demonstrate a reduction of up to 90% in the variability of estimated corrected effects.



## FIGURES AT A GLANCE: ASK ME FOR MORE DETAILS

The typical propensity score matching pipeline involves two key decisions: selecting the propensity model and choosing the matching technique.


No more models
Invalid propensity
Covariate selection and engineering
Propensity model selection (2.3)
Propensity estimation and validation (2.2)
No more techniques
Compute average treatment effect
Perform and validate matching (2.4, 2.5)
Matching technique selection
Invalid matching

There are many different propensity estimation models available. In some cases, only one model is considered valid, while in others, multiple models may be applicable. Here, we present six models, though many more exist. The scores reflect an overlapping metric between population propensities, with a propensity estimation considered valid if it exceeds 50%.

| Transform | None | | | Logit | | |
|---|---|---|---|---|---|---|
| Model | LR | RF | CLR | LR | RF | CLR |
| Groupon | **0.75** | 0.70 | 0.67 | 0.25 | 0.27 | 0.26 |
| Horse | 0.51 | 0.29 | **0.84** | 0.18 | 0.28 | 0.19 |
| NHANES | 0.47 | 0.46 | **0.83** | 0.12 | 0.00 | 0.19 |
| Synth. data | | | | | | |
| # confounders 0 | **1.00** | 0.86 | 0.97 | 0.00 | 0.62 | 0.45 |
| 1 | **1.00** | 0.81 | 0.97 | 0.00 | 0.61 | 0.43 |
| 2 | 0.85 | 0.60 | **0.97** | 0.36 | 0.41 | 0.44 |
| 3 | **0.99** | 0.73 | 0.96 | 0.20 | 0.53 | 0.43 |
| 4 | **1.00** | 0.87 | 0.97 | 0.17 | 0.52 | 0.43 |
| 5 | 0.88 | 0.91 | **0.95** | 0.38 | 0.57 | 0.44 |
| 6 | **1.00** | 0.86 | 0.97 | 0.00 | 0.27 | 0.40 |
| 7 | **1.00** | 0.89 | 0.97 | 0.00 | 0.59 | 0.44 |
| 8 | 0.88 | 0.83 | **0.97** | 0.35 | 0.40 | 0.43 |
| 9 | 0.91 | 0.96 | **0.98** | 0.38 | 0.76 | 0.43 |
| 10 | 0.91 | 0.90 | **0.96** | 0.39 | 0.04 | 0.44 |

A2A is computed using an algorithm that generates artificial matching problems similar to the problem being addressed. This involves selecting two subpopulations from the control group that exhibit the same differences as those between the control and treated groups, but halved to avoid creating overly challenging problems with numerous invalid corrections. Since there is no actual difference between the two subpopulations drawn from the same control population, we can evaluate how effectively methods can debias this problem. Creating the artificial problem boils down to minimizing the following loss:

$$\mathcal{L}\left(X, Y, \left\{X_0^{(0)}, X_0^{(1)}\right\}\right) = \left(\frac{1}{2}\underbrace{\mathsf{ATE}(Y_0, Y_1)}_{\text{reference task}} - \underbrace{\mathsf{ATE}\left(Y_0^{(0)}, Y_0^{(1)}\right)}_{\text{artificial task}}\right)^2 + \left(\frac{1}{2}\underbrace{\mathsf{SMD}(X_0, X_1)}_{\text{reference task}} - \underbrace{\mathsf{SMD}\left(X_0^{(0)}, X_0^{(1)}\right)}_{\text{artificial task}}\right)^2$$

We now show SMD and A2A values for various tasks and propensity models. We observe that the results of a given matching method can vary significantly depending on the propensity estimation used. Generally, A2A and SMD are complementary: when one is low, the other tends to be high.

| | SMD | | | A2A | | |
|---|---|---|---|---|---|---|
| Matching | Groupon | Horse | NHANES | Groupon | Horse | NHANES |
| *Optimal* | | | | | | |
| ElasticNet | 0.118 | 0.128 | 0.046 | 84.327 | 0.036 | 0.002 |
| Rpart | 0.164 | 0.151 | 0.070 | 88.880 | 0.038 | 0.003 |
| CBPS | 0.052 | 0.117 | 0.016 | 118.526 | 0.036 | 0.006 |
| Bart | 0.034 | 0.106 | 0.024 | 145.374 | 0.053 | 0.003 |
| GAM | 0.026 | 0.115 | 0.015 | 220.050 | 0.035 | 0.003 |
| GLM | 0.026 | 0.115 | 0.015 | 220.050 | 0.035 | 0.003 |
| *Bipartify* | | | | | | |
| RF | 0.051 | 0.077 | 0.044 | 588.262 | 0.043 | 0.002 |
| LR | 0.039 | 0.086 | 0.032 | 624.086 | 0.048 | 0.001 |
| CLR | 0.054 | 0.076 | 0.056 | 834.433 | 0.047 | 0.002 |
| *PsmPy* | | | | | | |
| CLR | 0.026 | 0.107 | 0.028 | 292.465 | 0.037 | 0.004 |
| RF | 0.051 | 0.179 | 0.166 | 1491.992 | 0.039 | 0.001 |
| LR | 0.043 | 0.158 | 0.165 | 1597.057 | 0.041 | 0.003 |

This figure illustrates the range of estimated corrected effects among valid methods, showing the difference between the maximum and minimum values for valid models. Large ranges typically correspond to a high mean average error. By reducing the number of valid models using A2A, we achieve smaller ranges and lower errors, indicating that the selected models are more accurate. The optimal strategy for combining the two metrics depends on the number of confounders: when there are many confounders, SMD is more relevant, and methods that prioritize it perform better. Conversely, when there are few confounders, methods that emphasize A2A are more effective.

| | Range of ATE values | | | ATE estimation error | | | | |
|---|---|---|---|---|---|---|---|---|
| Dataset | SMD | SMD ×A2A | Pareto | SMD | SMD ×A2A | Pareto | Min A2A | Min SMD |
| Groupon | 1417 | 1176 | **659** | *Real ATE unknown* | | | | |
| Horse | 0.014 | **0.000** | 0.014 | | | | | |
| NHANES | 0.022 | **0.000** | 0.005 | | | | | |
| Synth. data | | | | | | | | |
| # confounders 0 | 0.067 | 0.032 | **0.031** | 0.054 | **0.022** | 0.035 | 0.063 | **0.006** |
| 1 | 0.099 | **0.003** | **0.003** | 0.070 | **0.043** | 0.045 | **0.039** | 0.052 |
| 2 | 0.099 | **0.013** | 0.013 | 0.063 | **0.043** | 0.046 | **0.030** | 0.058 |
| 3 | 0.122 | **0.079** | 0.094 | 0.069 | **0.062** | 0.066 | **0.028** | 0.103 |
| 4 | 0.096 | 0.046 | **0.012** | 0.059 | 0.060 | **0.046** | 0.041 | **0.039** |
| 5 | 0.087 | 0.064 | **0.018** | 0.057 | 0.057 | **0.026** | **0.013** | 0.039 |
| 6 | 0.095 | **0.000** | 0.028 | 0.066 | 0.094 | **0.063** | 0.094 | **0.049** |
| 7 | 0.067 | 0.043 | **0.033** | 0.041 | 0.046 | **0.037** | 0.074 | **0.030** |
| 8 | 0.046 | 0.029 | **0.007** | 0.034 | 0.034 | **0.030** | 0.035 | **0.025** |
| 9 | 0.079 | 0.070 | **0.038** | 0.049 | 0.053 | **0.043** | 0.093 | **0.019** |
| 10 | 0.078 | 0.038 | **0.032** | 0.052 | 0.051 | **0.030** | **0.023** | 0.044 |

**Remaining work.** We still need a way to determine the best method among SMDxA2A and Pareto on the problems at hand (ie low or high number of confounders). Additionally, A2A currently requires computing all possible matchings, which can be prohibitive in some cases. We need to find a way to translate A2A into a more absolute metric, similar to SMD.