Published in final edited form as:
IEEE J Biomed Health Inform. 2022 November ; 26(11): 5728–5737. doi:10.1109/JBHI.2022.3201752.

Robust Methods for Quantifying the Effect of A Continuous Exposure from Observational Data

Roshan Tourani,
University of Minnesota, Institute for Health Informatics

Sisi Ma,
Institute for Health Informatics and Department of Medicine at University of Minnesota

Michael Usher,
Department of Medicine at University of Minnesota

Gyorgy J. Simon
Institute for Health Informatics and Department of Medicine at University of Minnesota

Abstract

A cornerstone of clinical medicine is intervening on a continuous exposure, such as titrating the dosage of a pharmaceutical or controlling a laboratory result. In clinical trials, continuous exposures are dichotomized into narrow ranges, excluding large portions of the realistic treatment scenarios. The existing computational methods for estimating the effect of continuous exposure rely on a set of strict assumptions. We introduce new methods that are more robust towards violations of these assumptions. Our methods are based on the key observation that changes of exposure in the clinical setting are often achieved gradually, so effect estimates must be "locally" robust in narrower exposure ranges. We compared our methods with several existing methods on three simulated studies with increasing complexity. We also applied the methods to data from 14k sepsis patients at M Health Fairview to estimate the effect of antibiotic administration latency on prolonged hospital stay. The proposed methods achieve good performance in all simulation studies. When the assumptions were violated, the proposed methods had estimation errors of one half to one fifth of the state-of-the-art methods. Applying our methods to the sepsis cohort resulted in effect estimates consistent with clinical knowledge.

Index Terms—

Causal effect estimation; Causal inference; Continuous exposure; Sepsis; Time to treatment

I. Introduction

Estimating the effect of exposures and interventions on outcomes is one of the cornerstones of biomedical research. When the exposure is binary, a multitude of effective methods exist. However, the exposures of interest in biomedicine are often continuous. For example,

Corresponding author: roshan@umn.edu.

the effect of A1c value on cardiovascular risk [14], [38], the effect of time to antibiotics on sepsis-related adverse outcomes [27], [32], [41], and the effect of potassium level on mortality due to acute myocardial infarction [13] are all examples of continuous exposures. Quantifying the effect of continuous exposures remains a challenge.

Randomized clinical trials (RCT) are considered the gold standard for estimating the effect of interventions. For some questions, however, the use of RCTs would be deemed unethical, and even when RCTs are applicable, they have difficulty handling continuous exposures. Typically, the exposure is dichotomized into two clinically meaningful narrow ranges to compare. In a diabetes example [38], patients with A1c of 6.0% (intensive control) are compared to patients with A1c between 7.0 and 7.9% (standard control) in terms of cardiovascular outcomes. The desired interpretation is that more intensive A1c control (being in the 6.0% range as opposed to the 7.0–7.9% range) leads to lower cardiovascular risk. Unfortunately, this leaves gaps in our knowledge about the potential benefit of the intervention in patients who are outside the range of exposure. Patients with A1c levels far in excess of 8% could potentially also benefit from more intensive control, however reducing their A1c to 6.0% may not be safe or even possible. To experimentally determine the effect of reducing A1c along the entire range of A1c measurements would require multiple RCTs.

In lieu of running multiple RCTs, effects can be estimated from observational studies. These studies, in turn, suffer from their own problems. Key among these problems is confounding, when the contribution of observed (or unobserved) variables is incorrectly attributed to the exposure. To overcome this problem, a vast array of *causal modeling* techniques have been developed. These techniques are used for better prediction [7], [12], [30], [39], learning the causal mechanisms underlying the data [11], [31], [33], [36], and estimating the effect of exposures or interventions on outcomes [8], [23], [32] in a broad range of contexts including time-series data (e.g. Granger causality in [12]), longitudinal data [30], pre-post analyses [29], retrospective subpopulation analysis [43], and propensity score (PS) analysis using aggregated Electronic Health Records (EHR) data [8].

In this paper, our focus is on estimating the effect size of a continuous exposure variable E on an outcome variable Y. We consider the most common study design in observational EHR data, where patients' longitudinal data is aggregated into a single cross section. Commonly applied effect estimation methods include the basic Direct Adjustment (DA) method (our implementation is equivalent to G-computation method in [3]) and methods of the Generalized Propensity Score (GPS) class, which were first introduced as extensions of the binary exposure problem [35] to the continuous exposure [19], [22].

These methods critically depend on three assumptions, which are often violated in practice, yielding potentially misleading results or results that hold no clinical value.

Common support.

Common support assumes that patients who are identical, except for their exposure level, exist along the entirety of the exposure range. Returning to the A1c example, the common support assumption implies that there are patients along the entire A1c range, who are comparable in all respects relevant to A1c and the outcome, and only differ in their A1c levels. The common support assumption is often violated at the extreme values of the exposure range: e.g. patients with normal A1c and those with excessively high A1c are typically not comparable; disease in the latter patients has progressed beyond comparability.

Validity of the outcome model.

A popular and straightforward method to estimate the effect of a continuous exposure on an outcome is to regress the outcome on the continuous exposure and potential confounders [13]. After acquiring this model, the average effect size between two levels of exposure can be estimated as the average of the difference in model predictions calculated at these two levels. Correct effect estimation is dependent on the correct specification of the outcome model, which could be a difficult task for real world problems where complex and non-linear relationships exist.

Validity of the propensity model.

Many of the methods that adjust for confounding utilize propensity scores to this end. The use of propensity score methods can allow us to sidestep issues regarding the validity of the outcome model as long as the propensity model itself is correctly specified. However, the correct specification of the propensity model can be challenging when explicit knowledge about the causal mechanism is absent.

Many recent methods that are technically in the GPS class [6], [10], [18], [24], [44], [45], either try to estimate only a single global propensity model and/or a single outcome model, and assume global common support. The term *global* refers to the requirement that common support exists throughout the entire range of possible exposure values, while the term *local* means that common support is required only in a narrower range of exposure values. In this paper, we relax the global common support assumption to the local common support assumption and suggest a framework where the effect gradient is estimated locally and some of the effect gradients are then integrated into an effect estimate. Using local samples relaxes the assumption of validity of propensity (or outcome) model, because we focus on a more uniform sample with common support in the distribution of confounders (no extrapolation). We also introduce a common support diagnostic and show how to interpret the result of integrating the effect gradient based on the result of this diagnostic.

There are also some very recent novel techniques for estimating the effect of continuous exposures. These methods include the entropy balancing method [16], which has been extended from the binary to the continuous exposure problem very recently [40], [42], and the generative adversarial deconfounding method [25]. Both of these methods follow similar assumptions to GPS relying on global common support or global positivity. These methods rely on weighting the samples so that in different exposure levels the distribution of confounders become more similar in a certain sense.

Contributions.—The main contributions of this paper are (i) to point out that in biomedical studies, the common support assumption is often too strong and when the common support assumption is violated, the effect estimates by GPS can suffer considerably; (ii) to establish a general framework that can use any existing GPS or binary

PS methods locally to estimate the effect gradient when only local common support is present (relaxing global common support assumption); and (iii) to estimate the extent of common support for the integrated effect curve and discuss the causal interpretation of this curve. Specifically, in (ii), we introduce two example methods using the proposed framework and compare them to several existing methods on simulated datasets, where the true effect size is known. We also apply these methods to a real-world clinical problem, where we estimate the effect of the latency of the first antibiotic administration on prolonged hospital stay for sepsis patients.

II. Materials and Methods

A. Problem Definition

For a continuous exposure E and two of its levels e_1 and e_2, we wish to estimate the average effect of changing E from e_1 to e_2 with respect to an outcome Y. This is analogous to ATT for binary treatment, where control group is $E = e_1$ and treated (or exposed) group is $E = e_2$. We denote this average effect by $A(e_1, e_2)$.

B. Proposed Methods

Fig 1 shows an illustration of the proposed method. To estimate the effect of changing the exposure from e_1 to e_2, we create a sequence of propensity matched populations at a sequence of exposure levels ϵ between e_1 and e_2. Then, in each propensity matched population, we estimate the local effect $a(\epsilon)$ of the exposure in the small neighborhood of ϵ and finally, we sum up the local effects to obtain $A(e_1, e_2)$. Local estimation allows for more relaxed assumptions: we only need common support locally, in the neighborhood of ϵ, as opposed to globally (along the entire exposure range) and similarly, the propensity and outcome models only need to provide a good approximation of the true relationship between exposure, confounders, and the outcome locally, in the small neighborhood around ϵ.

Specifically, we propose two methods for estimating $A(e_1, e_2)$, the Local Gradient Propensity Score Matching (grad-PSM) and the Greater Than Less Than Propensity Score Matching (gtlt-PSM).

```
**Data:** $\{X_i, E_i, Y_i\}$ for patinets $i = 1, \ldots, N_{\text{tot}}$
**Result:** Effect curve $A(e_1, e_2)$
**Step 1.** Initialization:
Generate $\epsilon_k$ sequence along exposure, $k = 1, \ldots, n$
Choose $S_\epsilon$ local sample, a minumum sample size $N$ selected by
 power sample analysis of local models;
**Step 2.** Local PSM:
Fit binary propensity score model for each $S_\epsilon$(above vs below $\epsilon$)
 and compute matched sample at $\epsilon$, $S_\epsilon^{(m)}$;
(Use $\delta$ parameter as minimum difference between matched samples
 to numerically stabilize gradient estimation. See Appendix A for
 details .)
**Step 3.** Fit local outcome model and compute
 $a(\epsilon) = \mathbb{E}_{S_\epsilon^{(m)}}[\partial\hat{y}/\partial\epsilon]$;
**Step 4.** Use trapezoid rule numerical integration to compute,
 $A(e_1, e_2) = \int_{e_1}^{e_2} \mathrm{d}\,\epsilon a(\epsilon)$;
**Step 5.** Estimate the extent of common support: Use matching
 efficiency (under the exact conditions of Step 2 for balance criteria
 etc .) from $e_1$ to further and further target exposures $e_2$. Compute
 extent of common support for each $e_1$ as exposure $e_2$ on each side
 of $e_1$ where matching efficiency drops to a cut-off value (e.g. 90%).
**Algorithm 1:** grad-PSM algorithm.
```

```
**Data:** $\{X_i, E_i, Y_i\}$ for patinets $i = 1, \ldots, N_{\text{tot}}$
**Result:** Effect curve $A(e_1, e_2)$
**Step 1.** Initialization:
Generate $\epsilon_k$ sequence along exposure, $k = 1, \ldots, n$;
**Step 2.** Local PSM:
Fit binary propensity score model for each $\epsilon$ (greater than vs less
 than $\epsilon$ using all data) and compute matched sample at $\epsilon$, $S_\epsilon^{(m)}$;
(Use $\delta$ parameter as minimum difference between matched samples
 to numerically stabilize gradient estimation . See Appendix A for
 details .)
**Step 3.** Use mean difference in outcome divided to change in
 exposure for matched pairs to estimate the effect gradient,
 $a(\epsilon) = \mathbb{E}_{(i,j) \in S_\epsilon^{(m)}}\left[\left(Y_i - Y_j\right)/\left(E_i - E_j\right)\right]$;
**Step 4.** Same as grad-PSM;
**Step 5.** Same as grad-PSM.
**Algorithm 2:** gtlt-PSM algorithm.
```

1) Local Gradient Propensity Score Matching (grad-PSM): (Step 1) Consider a sample of patients i with covariates X_i, exposure E_i and outcome Y_i. A sequence ϵ_1, ϵ_2, …, ϵ_n of exposure values is generated and the exposure range is divided into a sequence of (possibly overlapping) neighborhoods centered around $\epsilon_k (k = 1, 2, \ldots, n)$. For brevity, below we use $\epsilon \in \{\epsilon_1, \epsilon_2, \ldots, \epsilon_n\}$, dropping the index.

(Step 2) In the neighborhood around each ϵ, propensity score matching is carried out resulting in a matched set of patients $S_\epsilon^{(m)}$,

$$S_\epsilon^{(m)} = \Big\{(i, j) \mid i, j \in S_\epsilon, \quad E_i \le \epsilon, \quad \epsilon < E_j, \\ p_i \approx p_j, \quad E_j - E_i > \delta\Big\}, \tag{1}$$

where S_ϵ is a set of patients around ϵ, i (and j) are patients with exposure levels E_i (and E_j, respectively) and propensity scores p_i and p_j that fall within the required calipers ($p_i \approx p_j$), and their exposure levels differ at least by δ. The minimum difference δ in exposure levels numerically stabilizes the gradient estimate by avoiding division by very small numbers. More details about selecting S_ϵ and matching procedure are covered in Appendix A.

(Step 3) In the propensity matched population $S_\epsilon^{(m)}$, $a(\epsilon)$ is computed. To this end, we apply an outcome model, which is a logistic regression model for binary outcome Y, $Y \sim E + X$, constructed on the matched samples $S_\epsilon^{(m)}$, and estimate,

$$a(\epsilon) = \mathbb{E}_{S_\epsilon^{(m)}}\left[\frac{\partial \hat{y}}{\partial \epsilon}\right] = \mathbb{E}_{S_\epsilon^{(m)}}\left[\frac{\beta_E}{4\cosh^2([1, X, E]\beta/2)}\right], \tag{2}$$

where $\hat{y}$ is a model-based estimate of the outcome, $\beta = [\beta_0, \beta_X, \beta_E]$ is denoting the intercept, coefficients for confounders, and coefficient for exposure, respectively, and a reminder that $\cosh(x) = (e^x + e^{-x})/2$. Notice that this model needs to approximate the true relationships between X, E, and Y only in the narrow neighborhood of $E = \epsilon$. The quantity $a(\epsilon)$ can be interpreted as the gradient of the average treatment effect in the treated (ATT) when exposure level is reduced approximately from $\epsilon + \delta/2$ to $\epsilon - \delta/2$.

(Step 4) We integrate $a(\epsilon)$ to obtain $A(e_1, e_2)$, the effect of changing exposure from e_1 to e_2 using the trapezoid rule,

$$A(e_1, e_2) = \int_{e_1}^{e_2} \mathrm{d}\epsilon \; a(\epsilon), \tag{3}$$

keeping in mind that $A(e_1, e_2)$ in only meaningful up to the *extent of common support* from e_1 to e_2.

(Step 5) To estimate the extent of common support, we use the same matching method as above, and the proportion of samples at e_1 that could be matched to samples at e_2 or beyond can be used (e.g. using a 90% cut-off) as the extent of common support.

2) Greater than vs Less than Propensity Score Matching (gtlt-PSM):

Inspired by [32], [41], gtlt-PSM is a simplified version of grad-PSM. The gtlt-PSM is identical to the grad-PSM except for steps 2 and 3. In gtlt-PSM, we use the entire population when matching,

$$S_{\epsilon}^{(m)} = \left\{(i, j) \mid \quad E_i \le \epsilon, \quad \epsilon < E_j, \right. \\ \left. p_i \approx p_j, \quad E_j - E_i > \delta \right\}, \tag{4}$$

and the gradient $a(\epsilon)$ is computed with mean difference in outcome (no outcome model),

$$a(\epsilon) = \mathbb{E}_{(i, j) \in S_{\epsilon}^{(m)}}\left[\frac{Y_i - Y_j}{E_i - E_j}\right]. \tag{5}$$

The details regarding the selection of the sequence of ϵ, propensity matching, and the choice of δ are identical to grad-PSM. These implementation details are covered in the Appendix A.

C. Comparison methods

Unadjusted is the most naive calculation of effect size, without adjusting for the confounders,

$$A(e_1, e_2) = \mathbb{E}[Y(E_i \approx e_2)] - \mathbb{E}[Y(E_i \approx e_1)], \tag{6}$$

where “$E_i \approx e$” is implemented as the 200 nearest samples to e (i denotes the samples).

Direct Adjustment is the most commonly used method to estimate the effect of continuous exposures, where a model is constructed for the outcome using the covariates. The effect curve in this case is calculated as,

$$A(e_1, e_2) = \mathbb{E}[\hat{y}(E_i = e_2, X_i) - \hat{y}(E_i = e_1, X_i)], \tag{7}$$

where $\hat{y}$ denotes the estimated outcome from the model. Specifically we consider two different models. GLM-DA utilizes a generalized linear model and GBM-DA uses gradient boosted trees as implemented in the gbm R package.

Covariate Balancing Generalized Propensity Score (CB-GPS) As discussed earlier, the GPS methods follow very similar assumptions, especially regarding the global common support. Here we include the CB-GPS method [10], a representative method from this class, that was recently found to stands out as one of the top-performing methods [3].

D. Simulation Studies

Real data with complex confounding is often observational data and the true effect of the interventions are not known. To know the true effect sizes and to be able to judge whether a causal method is successful in removing the confounding bias, the use of simulated data is necessary and common practice. We use a sequence of increasingly challenging studies designed to highlight the limitations of the different methods. Fig. 2 shows the causal graph used for data generation in all three studies. In Study 1, none of the assumptions are violated and our aim is to see whether the proposed method can perform as well as the theoretically best method. In Study 2, although both the propensity and the outcome models for the GPS methods are specified correctly, the global common support assumption is violated. This study demonstrates the importance of the global common support assumption. In Study 3, neither the propensity nor the outcome model is specified correctly and the global common support assumption is also partially violated.

Study 1 is simple, patients are comparable across the entire range of the exposure. The exposure model is linear in the confounders and the outcome is linear in all variables; the outcome model we use is correctly specified,

$$E = \sum_j C_j + 2.25\ \xi, \tag{8}$$

$$O = -1 + 0.2E + 0.5\sum_j C_j + 0.5\sum_j X_j + \varepsilon\,. \tag{9}$$

where $\xi \sim \mathcal{N}(0,1)$, $\varepsilon \sim \mathcal{N}(0,1)$, and the available outcome data is binary, $Y = \mathrm{binom}(\mathrm{logit}(O))$. Notice that about half the variation in exposure is due to noise (the choice of SD=2.25).

This study meets all the theoretical assumptions for GLM-DA and CB-GPS methods.

Study 2 is exactly the same as Study 1, except,

$$E = 1.4\sum_j C_j + \xi, \tag{10}$$

and $\xi \sim \mathcal{N}(0,1)$. We kept the total variation of E almost the same as Study 1, but now most of this variation is due to confounders. This way the global common support assumption is violated.

In **Study 3** we vary the effect size across exposure (non-linear U-shaped relation) and make patients only locally comparable across the exposure range. Also we use piecewise linear curves both in the exposure and the outcome models (still monotonic in each variable). This reflects the common behavior of labs and vitals in real biomedical studies, where a lab value mostly affects the exposure or outcome above or below a threshold.

$$\begin{aligned} E = 2.5[&r(C_1, a) + l(C_2, b) + r(C_3, a) + l(C_4, b) + C_5] \\ &+ \xi \end{aligned} \tag{11}$$

$$\begin{aligned} O = &-2 \\ &+ 1.5[l(C_1, -c) + l(C_2, b) + r(C_3, c) + r(C_4, a) + C_5] \\ &+ r(X_1, a) + l(X_2, b) + r(X_3, a) + l(X_4, b) + X_5 \\ &+ 0.1\ u(E, -0.5, 1) + \varepsilon, \end{aligned} \tag{12}$$

where $\xi \sim \mathcal{N}(0,1)$, $\varepsilon \sim \mathcal{N}(0,1)$, $a = 0.5$, $b = -0.25$, and $c = 1$ are used. $l(X, a)$ is a left-hinged curve, where the curve is linearly decreasing on the left of 'a' and is constant on the right; $r(X, a)$ is a right-hinged curve that is linearly increasing on the right of 'a' and is constant on the left; finally, $u(X, a, b)$ is a u-shaped curve that quadratically decreases on the left of 'a', quadratically increases on the right of 'b', and zero in between ($a < b$),

$$l(X,a) = \begin{cases} X & X \le a, \\ a & X > a\,. \end{cases} \tag{13}$$

$$r(X,a) = \begin{cases} a & X < a, \\ X & X \ge a\,. \end{cases} \tag{14}$$

$$u(X,a,b) = \begin{cases} (X-a)^2 & X \le a, \\ 0 & a < X < b, \\ (X-b)^2 & X \ge b\,. \end{cases} \tag{15}$$

Again available outcome data is binary $Y = \text{binom}(\text{logit}(O))$. This example is an attempt to mimic the considered real data set and a bit more challenging test for methods introduced.

Usually in biomedical studies one knows the direction of effect size (it being positive or negative). Given that we defined the effect size to be U-shaped, for GLM-DA and CB-GPS, we estimate the effect size for $E < 0$ and $E > 0$ separately and combine the results. These methods have no built-in facility to handle U-shaped effect curves.

Evaluation: Since the data generation method assumes path/direction independence for $A(e_1, e_2)$ function, for any e_0 we have,

$$A(e_1, e_2) = A(e_1, e_0) + A(e_0, e_2) = A(e_0, e_2) - A(e_0, e_1)\,. \tag{16}$$

It is thus sufficient to reconstruct the $A(e_0, e)$ curve, a single-valued function of e, for all exposure levels e relative to a "reference" initial exposure e_0. In other words, evaluating $A(e_1, e_2)$ in two dimensions will not add any new information. The path independence assumption can be interpreted as averaging over a distribution of paths that patients take from e_1 to e_2; or more formally, $\langle A(\gamma(e_1, e_2))\rangle =: A(e_1, e_2)$ where γ is the paths patients take from e_1 to e_2 in the population that is comparable between e_1 and e_2 (confounders or propensity of exposure are changed by a very small amount).

We chose $e_0 = 0$ as the reference level, because it is the approximate median of the exposure data. Changing e_0 to e_0' only shifts all the curve by a constant $A(e_0, e_0')$ and does not affect the evaluation results in any other way.

Similar to simulations in previous studies, e.g. [10], the population in simulation is taken to be homogeneous (a single outcome model is used) and the effect size does not depend on the direction, meaning,

$$A(e_1, e_2) = -A(e_2, e_1), \tag{17}$$

therefore different treatment effects are equal, i.e. ATE=ATT, which helps simplify and focus our discussion in this paper.

For improved visibility, smoothing is applied to the curve for all the methods. We use Gaussian kernel smoother with the same smoothing widths (bandwith) across all methods. See Appendix B.

We generated 10k samples for Studies 1 and 2, and 15k for Study 3. We applied the algorithms to estimate the effect curves and plotted them. We computed the RMSEs for three exposure ranges of decreasing data density: 35–65th, 25–75th and 5–95th percentiles. To quantify the variability of the RMSE, we generated 400 additional data sets and computed the RMSE on each. We report the mean and 95% CI of the 400 RMSEs for each exposure range. Also we use t-test for all method pairs to check whether the mean differences are significant, using p-value < 0.001 as the significance level with Holm-Bonferroni method to adjust for multiple comparisons.

Finally, to evaluate the extent of common support we use bins with 400 sample size at initial exposure, plot the heat-map plot of matching proportions for different target exposures and mark the 90% contours.

E. Antibiotic Latency Effect on Prolonged Hospital Stay for Patients with Sepsis on Admission

Study design and setting.—This is a retrospective cohort study of 14k hospitalized septic patients from M Health Fairview (FV) from Sep 2010 to Dec 2016. Septic patients are identified as those who are Systemic Inflammatory Response Syndrome (SIRS) positive or have acute organ dysfunction within two days of admission and follow a full course of antibiotics [34]. For patients with multiple hospitalizations, only one, a randomly chosen hospitalization is included.

Outcome.—The primary outcome is prolonged hospital stay, defined as length of stay of 10 days or longer.

Exposure.—We define the continuous exposure variable antibiotic latency as the elapsed time from first the time the patient was SIRS positive (which we call onset time) until the first antibiotic administration, measured in hours.

Variables of interest.—We consider the following variables as potential confounders: chronic comorbidities (e.g. tumor, chronic kidney disease, liver disease, chronic obstructive pulmonary disease), baseline labs, the variables related to acuity of infection at the onset time: vital signs, WBC, lactate, and PCO2, plus other labs which can be related to organ dysfunction, again measured up to the onset time.

Statistical Analysis.—The proposed methods grad-PSM and gtlt-PSM were applied. As part of the method, propensity models are constructed. We use causal feature discovery [28] to discover the actual confounders and these confounders were used in the propensity model. As discussed, grad-PSM uses an outcome model for further adjustment after matching.

Evaluation.: Given that exposure is time, the path independence assumption holds (similar to the simulated studies), meaning that we can study a single valued curve rather than two-dimensional $A(e_1, e_2)$. Given that the latency distribution is concentrated around 2.5h, we focus on estimating the ATT of moving a patient from latency e to 2.5h denoted by $-A(e,$ 2.5h), where the negative sign is used to denote the benefit (outcome is adverse).

Estimating variability with propensity score methods is debated [4], [17], [37]. Here we simply report the estimated variability using 95% CI of bootstrap sampling (with replacement) following [17], [32]. Using all sample bootstrap (with replacement) potentially overestimates sample variation (especially for grad-PSM), but arguably it is the safer choice when reporting the existence or significance of effect sizes.

III. Results

A. Simulation Studies

First we applied the methods to the data generated for Study 1. Fig 3a shows the estimated effect size for each method from a single run. The vertical axis depicts the effect size of $A(0, e)$, changing the exposure from 0 to the target exposure e, e ranging from −5.3 to 5.2 (5%−95% percentile of data) depicted on the horizontal axis. The effect size is ATE for GLM-DA and GBM-DA and ATT for the other methods. The solid black line represents the true effect size, for which ATE=ATT, and the dashed line is the raw (unadjusted) effect size.

Fig 3b visualizes the extent of common support. The horizontal axis corresponds to the original level e_1 of exposure and the vertical axis corresponds to the target level e_2. The greenness of the pixel represents the percentage of samples within gray box that could be matched. The contour lines corresponds to the range within which 90% of the samples was matched. For example, when the source exposure is 1.5 (1.5 on the horizontal axis), at least 90% of the samples can be matched for target exposures between almost −1.5 and positive 3.5 (this is where the two contour lines are).

Table Ia displays the errors (RMSE) from the various methods in different ranges of exposure (35%−65%, 25%−75%, and 5%−95%) (scaled up by 10^2 for easier reading). It shows the mean RMSE and its empirical 95% CI over the 400 runs. ‘Unadjusted’ has the highest error, followed by GBM-DA. The other methods perform very comparably and their RMSE is about 10% of the RMS of actual effect size. The mean differences can be read from this table and except the difference between CB-GPS and grad-PSM, all other differences were significant (using t-test for all the method pairs with significance level of p-value < 0.001 with Holm-Bonferroni adjustment for multiple comparisons).

Next, we performed the analysis on Study 2. Results are shown in Fig 3 second row. Recall that Study 2 violates the (global) common support assumption, but allows for correct specification of the outcome and exposure models. The results are similar to those from Study 1 but unlike in Study 1, CB-GPS fails to reliably estimate the effect size. Table Ib contains the RMSE for the various methods. All the mean differences were significant other than the difference between grad-PSM and gtlt-PSM (using t-test for all method pairs and significance level of p-value < 0.001).

Finally, the analyses were carried out on Study 3 and are shown in Fig 3 third row. Given that all three assumptions are violated, all methods performed worse than in the previous studies, but the two proposed methods performed better than the others. GBM-DA was slightly more biased towards the Unadjusted curve. RMSE values are reported in Table Ic. In the middle (25%−75%) grad-PSM performs exceptionally well, but at the edges it becomes similar to gtlt-PSM as the sample density decreases. All the mean differences in Table Ic were significant (using t-test for all method pairs and significance level of p-value < 0.001).

B. Antibiotic Latency Effect on Prolonged Hospital Stay for Sepsis Cohort

The effect curves for grad-PSM and gtlt-PSM are shown in Fig 4(a). The vertical axis is $-A(e, 2.5\text{h})$ which denotes the effect size of reducing the outcome (benefit) by changing the latency from e to $e_0 = 2.5\text{h}$.

Fig 4(b) shows that we have a rather large unconfounded variation in exposure, which allows patients to be matched down to about 1h (blue dashed line). This means we can get reliable effect estimates for reducing the exposure (latency) from 6h not only to 2.5h but all the way down to 1h and estimate the effect using equation (16): $A(e, 1\text{h}) = A(e, 2.5\text{h}) + A(2.5\text{h}, 1\text{h})$ for any $e \gtrsim 2\text{h}$. Using the average curves, reducing the delay in antibiotic administration from 6h to 1h can reduce prolonged hospital stay by about 3%.

IV. Discussion

We consider the problem of estimating the effect of changing a continuous exposure from an original value e_1 to another e_2. Currently existing methods make three critical assumptions. These are (global) common support, validity of the outcome model, and validity of the propensity model. We designed a new method that reduces reliance on these assumptions and we demonstrated that the proposed methods can achieve good performance in face of violations of these three assumptions.

In a series of three simulation studies of escalating difficulty, the first one was very simple, where none of the assumptions were violated, and accordingly, most methods performed well, except for GBM-DA which overestimated the effect of exposure. The other methods (with their assumptions met) performed well with minimal differences (RMSE range from 0.0046 to 0.012, which is 5.7% to 15% of the RMS of estimated effect.). GBM-DA could fit the outcome data almost perfectly, yet its effect estimates were biased because the confounders had similar coefficients in the exposure and the outcome models, allowing GBM to mistakenly attribute the confounders’ effect to the exposure. This is an example where a flexible (and less interpretable) outcome model, that fits the outcome data well, does not help with better causal effect estimation.

The second study primarily differs from the first one in that it violates the (global) common support assumption, only having local common support. With the outcome and propensity models being correctly specified, most models performed well, except for CB-GPS and GBM-DA. CB-CPS relies on the global common support assumption heavily and even failed to determine the effect direction correctly. Similarly to the first study, the flexibility of GBM-DA hurt its performance.

The second study highlights one of the key differences between the proposed method and existing methods, which is the departure from the (global) common support assumption. Interpretation for Study 2 is driven by the local common support: $A(e_1, e_2) = A(0, e_2) - A(0, e_1)$ is the ATT of changing the exposure level from e_1 to e_2, provided that e_1 and e_2 has sufficient common support (but neither e_1 nor e_2 has to have common support with $E = 0$). In practice this means that e_1 and e_2 cannot be arbitrarily far from each other. This reflects clinical practice. For example in patients with very high A1c levels, the target A1c level can be raised (to the more attainable 8% in the Minnesota Community Measures [1]) instead of the normal target of 7% [2].

The third study highlights that under more realistic conditions, where none of the assumptions are met, the proposed method still performs adequately. Study 3 simulates a situation where the exact functional form of the propensity and outcome models are unknown. This favors flexible methods such as GBM-DA and the proposed methods. Even though the models were incorrectly specified, using many local point estimates helped the proposed methods achieve a reasonably good approximation.

For the antibiotics treatment latency example, there is no gold standard thus we did not perform a thorough comparison and only included the results for the proposed methods. The findings are consistent with prior studies [27], [32], [41], and provide evidence that starting antibiotics earlier may also benefit “healthier” patients (“healthier” in the sense that we excluded in-hospital mortality). This is not to say that patients with about 6h latency can necessarily be compared to patients with about 2h latency and will benefit according to the resultant effect curve. The diagnostic ambiguities that cause large latency are very challenging to address, if at all possible. But looking at this result and the methods used, physicians will know how much benefit there is in reducing everyone’s latency by about an hour or two, which is an attainable amount. The interpretability and the focus on measuring attainable changes in exposure, better helps physicians make necessary changes in treatment planning.

Contrasting grad-PSM and gtlt-PSM.

In terms of the global support assumption, an assumption critical to the causal interpretability of the results, the various methods lie on a continuum. On one extreme are the existing methods that require global support. On the other extreme is grad-PSM that only requires local support. In between these two is gtlt-PSM, which is local due to binarization, but can use more data than grad-PSM. Also fusing grad-PSM and gtlt-PSM is possible. Rather than using a sharp cutoff at N samples in S_ϵ for grad-PSM or using all possible samples in gtlt-PSM, S_ϵ can be constructed by weighting the samples. More generally, many method variaties can be created from our proposed framework, for example, by replacing the PSM estimator with inverse propensity weighting or the doubly robust estimator in grad-PSM.

Limitations.

(1) The methods used in the study, the proposed and comparison methods alike, rely on having sufficient local sample size for estimation. This is why in all studies above, the proposed methods performed better around central regions of exposure with higher sample density compared to the edge regions where the density was much lower. (2) The proposed methods use binary propensity matching as an operation and support virtually any such method [5], [9], [15], [20], [21], [26], [37]. Therefore, all the limitations of binary propensity score matching apply to the framework discussed in this paper. Though, due to the pointwise estimation, as we have shown, the provided framework can be more robust compared to single PS effect analyses (e.g. common binary PS or GPS methods).

V. Conclusion

When the three critical assumptions are met (having global common support and correctly specifying outcome and propensity models), the effect analysis problem is simple and most methods perform well. When they, especially the common support assumption, are violated, the proposed methods (grad-PSM and gtlt-PSM) can estimate the effects better than the existing methods. By focusing on estimating the effect of attainable or clinically actionable changes in exposure, the proposed methods can make more realistic assumptions and obtain more reliable effect estimates, thus they can better inform clinical decision making.

Acknowledgment

This work was supported in part by NIGMS award R01 GM 120079, AHRQ award R01 HS024532, and the NCATS University of Minnesota CTSA UL1 TR002494. The views expressed in this manuscript are those of the authors and do not necessarily reflect the views of the funding agencies.

Appendix

A. Proposed Methods

(Step 1) A sequence $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ of exposure values is generated and the exposure range is divided into a sequence of (possibly overlapping) neighborhoods centered around ϵ_k, $k = 1, \ldots, n$.

Recall that we perform three simulation studies and an application to a real-world clinical problem of estimating the effect of delay in antibiotic administration for sepsis cohort. In the simulation studies we used $n = 100$ points across the exposure range, and in the sepsis study, the sequence of ϵ is equally spaced at 5 min increments of delayed antibiotic administration, total of $n = 73$.

(Step 2) In the neighborhood around each ϵ, propensity score matching is carried out resulting in a matched set of patients $S_\epsilon^{(m)}$.

Propensity scores are estimated by binarizing the exposure around $E = \epsilon$ and fitting a logistic regression model ($E < \epsilon$ vs $E \geq \epsilon$). For matching, we are using nearest neighbor matching with caliper 0.25. Balance diagnosis utilizes the standardized mean differences (SMD) of each variable: whenever SMD exceeds 0.1 (a commonly used SMD threshold), we re-match the population using univariate subclassification matching on that variable [37], which refers to exact matching for binary and categorical variables, and to matching within quintiles for continuous variables. Additionally, matched pairs of patients are required to have a minimum exposure difference of δ. If sufficient balance could not be achieved following this matching procedure, we discard this ϵ from the gradient estimation.

The required number of patients in S_ϵ (for propensity model) is determined based on power sample analysis and selecting a least sample size of N below and above $E = \epsilon$. Here we use $N = 500$ for simulated studies and $N = 1000$ for the sepsis study.

We require that the matched pair's exposure levels differ at least by δ. The minimum difference in exposure levels numerically stabilizes the gradient estimate by avoiding division by very small numbers.

The selection of δ can be domain driven: differences in antibiotic administration delays of less than 1 hour are not expected to impact outcomes measurably so we set $\delta = 1$h; or can be chosen arbitrarily: we chose $\delta = 0.25$ for all simulation studies. Sensitivity analysis is carried to verify that the choice of N and δ did not unduly influence the results.

(Step 3) In the propensity matched population $S_\epsilon^{(m)}$, the effect gradient $a(\epsilon)$ is computed. As explained in the main text, for grad-PSM method we use the mean of derivative of outcome modelh preidiction with respect to exposure variable, $a(\epsilon) = \mathbb{E}_{S_\epsilon^{(m)}}\left[\frac{\partial \hat{y}}{\partial \epsilon}\right]$, and for gtlt-PSM we use the mean of outcome differences divided by exposure differences, $a(\epsilon) = \mathbb{E}_{(i,\, j) \in S_\epsilon^{(m)}}\left[\frac{Y_i - Y_j}{E_i - E_j}\right]$. Other estimation methods are also possible, e.g. using inverse propensity weighting or doubly robust (DR) estimators.

(Step 4) We integrate $a(\epsilon)$ to obtain the $A(e_1, e_2)$, the effect of changing exposure from $E = e_1$ to $E = e_2$. The way we use PSM makes $A(e_1, e_2)$ directional, meaning that we are estimating ATT going from $E = e_1$ (control) to $E = e_2$ (treated). Using other estimators one can estimate ATE instead.

The numerical integration is done using the trapezoid rule. For the simulation studies, 100 point estimates were used for the gradient $a(\epsilon)$. About 10 points got dropped from distribution tails with insufficient sample size for S_ϵ or $S_\epsilon^{(m)}$. In the sepsis study every 5 minutes a point estimate is acquired (73 points from 60min to 420min, and again, about 10 of them dropped from tails for not having enough samples).

(Step 5) The exact same criteria used in Step 2 for accepting matched groups is applied for matching from $E \approx e_1$ to further exposure value. $E \approx e_1$ means 400 samples around $E = e_1$. The target exposure samples are selected from samples at $E = e_2$ or further than $E = e_2$ (measured from $E = e_1$). And the extent of common support is defined as 90% matching ratio between 400 samples at $E \approx e_1$ and samples that are at $E = e_2$ or further (at least 360 matched pairs which must satisfy the same matching/balancing criteria given in Step 2).

Computational complexity.

The computational complexity is linear in the number of ϵ steps n, multiplied to PS matching complexity (PSM). For PSM we used nearest neighbor greedy algorithm in our code, which is linear in sample size. So computational complexities are $O(nN)$ for grad-PSM and $O(nN_{\mathrm{tot}})$ for gtlt-PSM. For evaluating the extent of common support (or when path/direction independence assumption does not hold) one has to estimate the matching efficiency (or causal effect) on the surface of (e_1, e_2). This makes the complexity quadratic in the number of steps. We should mention that the algorithm is straightforward to fully parallelize in step iterations (independent for-loop parallelization). So with access to a computational grid, one can parallelize away the linear or quadratic in steps part and only deal with the underlying PSM complexity (or any propensity score method used in each local region).

B. Simulated Studies

Smoothing.

A final smoothing is applied on the effect curves. The integration step in proposed methods result in a smooth curve, however, to make all the results comparable, we use Gaussian kernel smoother on the final results for all the methods studied, with the same smoothing widths: 0.1 for simulated data and 20min for the real study.

U-shaped effect in GPS or GLM models.

Usually in biomedical studies one knows the direction of effect size (it being positive or negative). In Study 3, we defined the effect size to be U-shaped, so for GLM-DA and CB-GPS, we estimated the effect size for $E < 0$ and $E > 0$ separately and combined the results. These methods have no built-in facility to handle U-shaped effect curves.

References

[1]. D5 treatment goals for diabetes, https://mncm.org/measurement-resources/, 2014.

[2]. Ada guidelines, ada standards of care, https://care.diabetesjournals.org/content/44/Supplement_1, 2021.

[3]. Austin Peter C. Assessing the performance of the generalized propensity score for estimating the effect of quantitative or continuous exposures on binary outcomes. Statistics in medicine, 37(11):1874–1894, 2018. [PubMed: 29508424]

[4]. Austin Peter C and Small Dylan S. The use of bootstrapping when using propensity-score matching without replacement: a simulation study. Statistics in medicine, 33(24):4306–4319, 2014. [PubMed: 25087884]

[5]. Austin Peter C and Stuart Elizabeth A. Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. Statistics in medicine, 34(28):3661–3679, 2015. [PubMed: 26238958]

[6]. Brown Derek W, Greene Thomas J, Swartz Michael D, Wilkinson Anna V, and DeSantis Stacia M. Propensity score stratification methods for continuous treatments. Statistics in Medicine, 40(5):1189–1203, 2021. [PubMed: 33305367]

[7]. Dandu Sriram Raju, Engelhard Matthew M, Qureshi Asma, Gong Jiaqi, Lach John C, Brandt-Pearce Maite, and Goldman Myla D. Understanding the physiological significance of four

inertial gait features in multiple sclerosis. IEEE journal of biomedical and health informatics, 22(1):40–46, 2017.

[8]. Datta Arghya, Matlock Matthew K, Dang Na Le, Moulin Thiago, Woeltje Keith F, Yanik Elizabeth L, and Swamidass Sanjay Joshua. ‘black box’to ‘conversational’machine learning: Ondansetron reduces risk of hospital-acquired venous thromboembolism. IEEE Journal of Biomedical and Health Informatics, 25(6):2204–2214, 2020.

[9]. Deb Saswata, Austin Peter C, Tu Jack V, Ko Dennis T, David Mazer C, Kiss Alex, and Fremes Stephen E. A review of propensity-score methods and their use in cardiovascular research. Canadian Journal of Cardiology, 32(2):259–265, 2016. [PubMed: 26315351]

[10]. Fong Christian, Hazlett Chad, Imai Kosuke, et al. Covariate balancing propensity score for a continuous treatment: application to the efficacy of political advertisements. The Annals of Applied Statistics, 12(1):156–177, 2018.

[11]. Glymour Clark, Zhang Kun, and Spirtes Peter. Review of causal discovery methods based on graphical models. Frontiers in genetics, 10:524, 2019. [PubMed: 31214249]

[12]. Gong Jiaqi, Qi Yanjun, Goldman Myla D, and Lach John. Causality analysis of inertial body sensors for multiple sclerosis diagnostic enhancement. IEEE journal of biomedical and health informatics, 20(5):1273–1280, 2016. [PubMed: 27411232]

[13]. Goyal Abhinav, Spertus John A, Gosch Kensey, Venkitachalam Lakshmi, Jones Philip G, Van den Berghe Greet, and Kosiborod Mikhail. Serum potassium levels and mortality in acute myocardial infarction. Jama, 307(2):157–164, 2012. [PubMed: 22235086]

[14]. ADVANCE Collaborative Group. Intensive blood glucose control and vascular outcomes in patients with type 2 diabetes. New England journal of medicine, 358(24):2560–2572, 2008. [PubMed: 18539916]

[15]. Guo Shenyang and Fraser Mark W. Propensity score analysis: Statistical methods and applications, volume 11. SAGE publications, 2014.

[16]. Hainmueller Jens. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. Political analysis, 20(1):25–46, 2012.

[17]. Hill Jennifer and Reiter Jerome P. Interval estimation for treatment effects using propensity score matching. Statistics in medicine, 25(13):2230–2256, 2006. [PubMed: 16220488]

[18]. Hines Oliver, Diaz-Ordaz Karla, and Vansteelandt Stijn. Parameterising the effect of a continuous exposure using average derivative effects. arXiv preprint arXiv:2109.13124, 2021.

[19]. Hirano Keisuke and Imbens Guido W. The propensity score with continuous treatments. Applied Bayesian modeling and causal inference from incomplete-data perspectives, 226164:73–84, 2004.

[20]. Ho Daniel E, Imai Kosuke, King Gary, and Stuart Elizabeth A. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political analysis, 15(3):199–236, 2007.

[21]. Imai Kosuke and Ratkovic Marc. Covariate balancing propensity score. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(1):243–263, 2014.

[22]. Imai Kosuke and Van Dyk David A. Causal inference with general treatment regimes: Generalizing the propensity score. Journal of the American Statistical Association, 99(467):854–866, 2004.

[23]. Imbens Guido W and Rubin Donald B. Causal inference in statistics, social, and biomedical sciences Cambridge University Press, 2015.

[24]. Kennedy Edward H, Ma Zongming, McHugh Matthew D, and Small Dylan S. Non-parametric methods for doubly robust estimation of continuous treatment effects. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79(4):1229–1245, 2017. [PubMed: 28989320]

[25]. Kuang Kun, Li Yunzhe, Li Bo, Cui Peng, Yang Hongxia, Tao Jianrong, and Wu Fei. Continuous treatment effect estimation via generative adversarial de-confounding. Data Mining and Knowledge Discovery, 35(6):2467–2497, 2021.

[26]. Lee Brian K, Lessler Justin, and Stuart Elizabeth A. Improving propensity score weighting using machine learning. Statistics in medicine, 29(3):337–346, 2010. [PubMed: 19960510]

[27]. Liu Vincent X, Fielding-Singh Vikram, Greene John D, Baker Jennifer M, Iwashyna Theodore J, Bhattacharya Jay, and Escobar Gabriel J. The timing of early antibiotics and hospital mortality in sepsis. American journal of respiratory and critical care medicine, 196(7):856–863, 2017. [PubMed: 28345952]

[28]. Ma Sisi and Statnikov Alexander. Methods for computational causal discovery in biomedicine. Behaviormetrika, 44(1):165–191, 2017.

[29]. Ning Yilin, Støer Nathalie C, Ho Peh Joo, Kao Shih Ling, Ngiam Kee Yuan, Khoo Eric Yin Hao, Lee Soo Chin, Tai E, Hartman Mikael, Reilly Marie, et al. Robust estimation of the effect of an exposure on the change in a continuous outcome. BMC medical research methodology, 20(1):1–11, 2020.

[30]. Orphanou Kalia, Stassopoulou Athena, and Keravnou Elpida. Dbn-extended: a dynamic bayesian network model extended with temporal abstractions for coronary heart disease prognosis. IEEE journal of biomedical and health informatics, 20(3):944–952, 2015. [PubMed: 25861090]

[31]. Pearl Judea. Causality Cambridge university press, 2009.

[32]. Pruinelli Lisiane, Westra Bonnie L, Yadav Pranjul, Hoff Alexander, Steinbach Michael, Kumar Vipin, Delaney Connie W, and Simon Gyorgy. Delay within the 3-hour surviving sepsis campaign guideline on mortality for patients with severe sepsis and septic shock. Critical care medicine, 46(4):500–505, 2018. [PubMed: 29298189]

[33]. Ramsey Joseph D, Hanson Stephen José, Hanson Catherine, Halchenko Yaroslav O, Poldrack Russell A, and Glymour Clark. Six problems for causal inference from fmri. neuroimage, 49(2):1545–1558, 2010. [PubMed: 19747552]

[34]. Rhee Chanu, Dantes Raymund, Epstein Lauren, Murphy David J, Seymour Christopher W, Iwashyna Theodore J, Kadri Sameer S, Angus Derek C, Danner Robert L, Fiore Anthony E, et al. Incidence and trends of sepsis in us hospitals using clinical vs claims data, 2009–2014. Jama, 318(13):1241–1249, 2017. [PubMed: 28903154]

[35]. Rosenbaum Paul R and Rubin Donald B. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55, 1983.

[36]. Spirtes Peter, Glymour Clark N, Scheines Richard, and Heckerman David. Causation, prediction, and search MIT press, 2000.

[37]. Stuart Elizabeth A. Matching methods for causal inference: A review and a look forward. Statistical science: a review journal of the Institute of Mathematical Statistics, 25(1):1, 2010. [PubMed: 20871802]

[38]. Action to Control Cardiovascular Risk in Diabetes Study Group. Effects of intensive glucose lowering in type 2 diabetes. New England journal of medicine, 358(24):2545–2559, 2008. [PubMed: 18539917]

[39]. Tourani Roshan, Murphree Dennis H, Melton-Meaux Genevieve, Wick Elizabeth, Kor Daryl J, and Simon Gyorgy J. The value of aggregated high-resolution intraoperative data for predicting post-surgical infectious complications at two independent sites In 17th World Congress on Medical and Health Informatics, MEDINFO 2019, pages 398–402. IOS Press, 2019.

[40]. Tübbicke Stefan. Entropy balancing for continuous treatments. Journal of Econometric Methods, 11(1):71–89, 2022.

[41]. Usher Michael G, Tourani Roshan, Webber Ben, Tignanelli Christopher J, Ma Sisi, Pruinelli Lisiane, Rhodes Michael, Sahni Nishant, Olson Andrew PJ, Melton Genevieve B, et al. Patient heterogeneity and the j-curve relationship between time-to-antibiotics and the outcomes of patients admitted with bacterial infection. Critical care medicine, 50(5):799–809, 2021.

[42]. Vegetabile Brian G, Griffin Beth Ann, Coffman Donna L, Cefalu Matthew, Robbins Michael W, and McCaffrey Daniel F. Nonparametric estimation of population average dose-response curves using entropy balancing weights for continuous exposures. Health Services and Outcomes Research Methodology, 21(1):69–110, 2021. [PubMed: 34483714]

[43]. Winterhoff Boris, Kommoss Stefan, Heitz Florian, Konecny Gottfried E, Dowdy Sean C, Mullany Sally A, Park-Simon Tjoung-Won, Baumann Klaus, Hilpert Felix, Brucker Sara, et al. Developing a clinico-molecular test for individualized treatment of ovarian cancer: the interplay of precision medicine informatics with clinical and health economics dimensions In AMIA

Annual Symposium Proceedings, volume 2018, page 1093. American Medical Informatics Association, 2018.

[44]. Zhang Zhiwei, Zhou Jie, Cao Weihua, and Zhang Jun. Causal inference with a quantitative exposure. Statistical methods in medical research, 25(1):315–335, 2016. [PubMed: 22729475]

[45]. Zhao Shandong, van Dyk David A, and Imai Kosuke. Propensity score-based methods for causal inference in observational studies with non-binary treatments. Statistical methods in medical research, 29(3):709–727, 2020. [PubMed: 32186266]

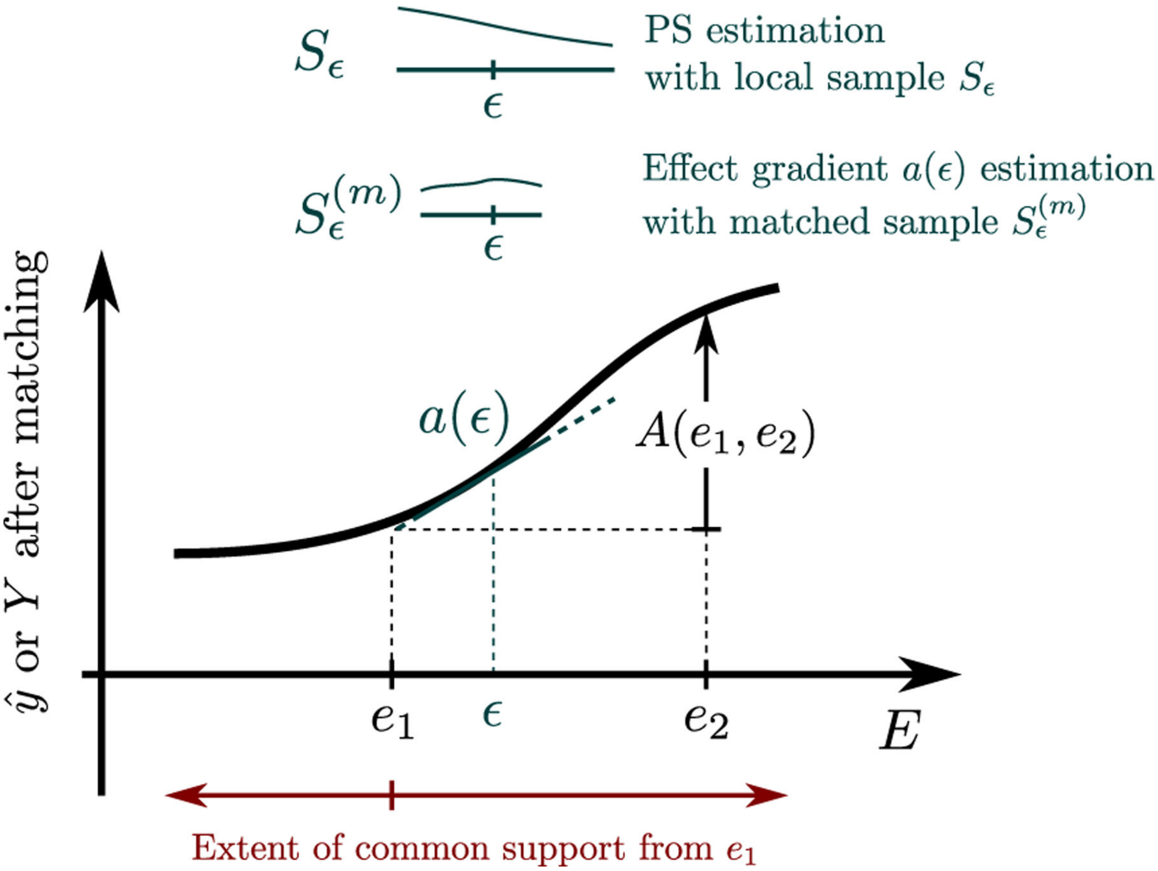
S_ϵ
ϵ
PS estimation with local sample S_ϵ
$S_\epsilon^{(m)}$
ϵ
Effect gradient $a(\epsilon)$ estimation with matched sample $S_\epsilon^{(m)}$
$\hat{y}$ or Y after matching
$a(\epsilon)$
$A(e_1, e_2)$
e_1
ϵ
e_2
E
Extent of common support from e_1

Fig. 1:

After selecting a sequence of samples S_ϵ, the local samples are matched $S_\epsilon^{(m)}$ and the local effect $a(\epsilon)$ is computed. $A(e_1, e_2)$ is estimated by numerically integrating over local effects $a(\epsilon)$. Finally the extent of common support is estimated e.g. by checking the matching ratio from e_1 to further exposures.

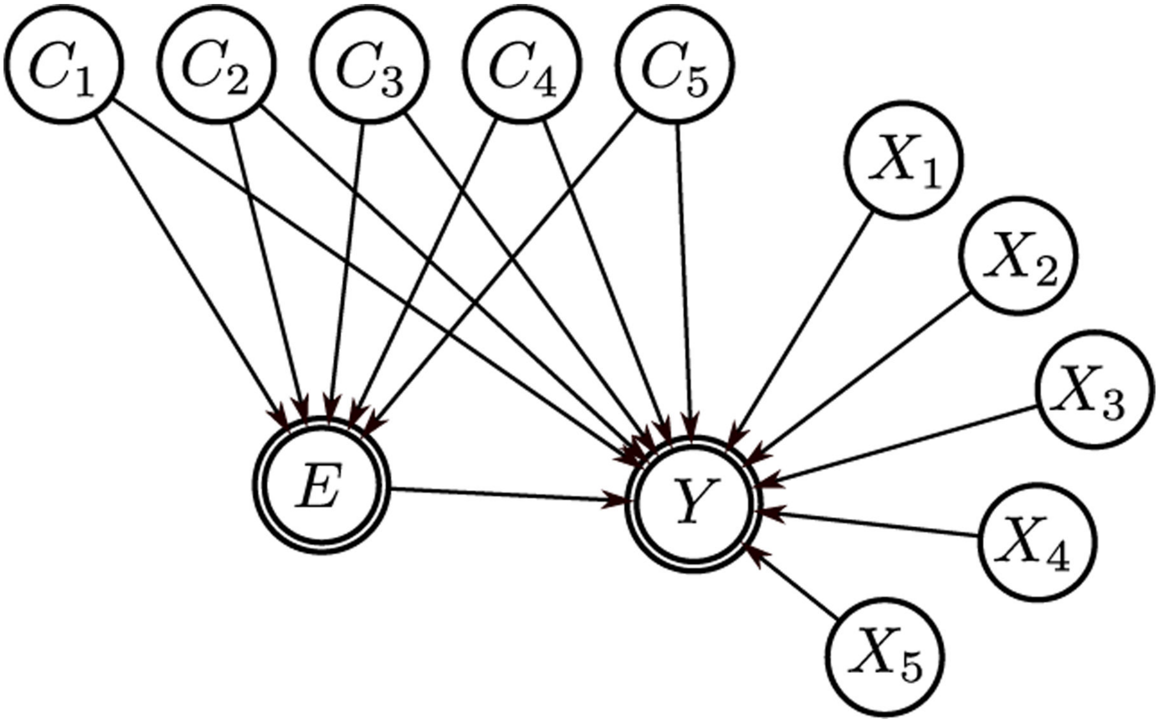

Fig. 2:

The causal graph used for data generation in simulation studies. For clarity the true confounder variables X_5, …, X_{10} are called C_1, …, C_5, respectively.

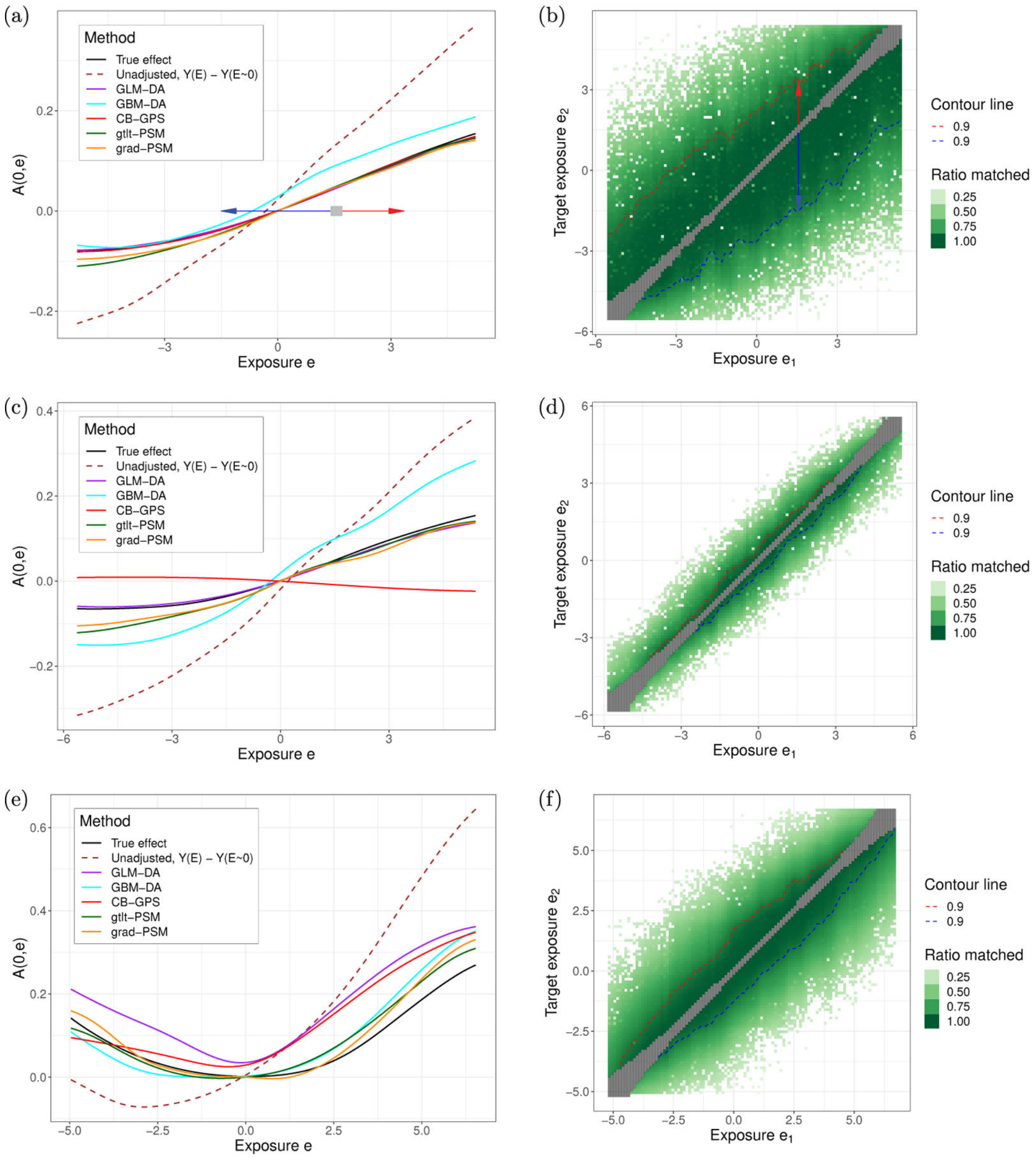

Fig. 3:

(a) The effect curve $A(0, e)$ for Study 1, a single run. (b) Extent of common support. Matching a bin of 400 samples around a initial exposure (horizontal axis) with a matching proportion of 90% (360/400) was possible to the exposure range (vertical axis) shown by red and blue dashed lines. For example, a sample from $e = 1.5$ is likely to have a match from $e = -1.5$ (blue arrow) to $e = 3.5$ (red arrow). In this example, the extent of common support is large. (c) $A(0, e)$ results for Study 2, a single run. (d) Extent of common support in Study 2. It is much smaller than Study 1 and to achieve 90% matching, one can only change exposure by about $\lesssim 0.5$ below or above. In other words, on $A(0, e)$ plot the estimated effects from $e = e_1$ to $e = e_2$ are only meaningful for $|e_2 - e_1| \lesssim 0.5$. (e) The effect curve for Study 3, a single run. (f) The extent of common support for Study 3.

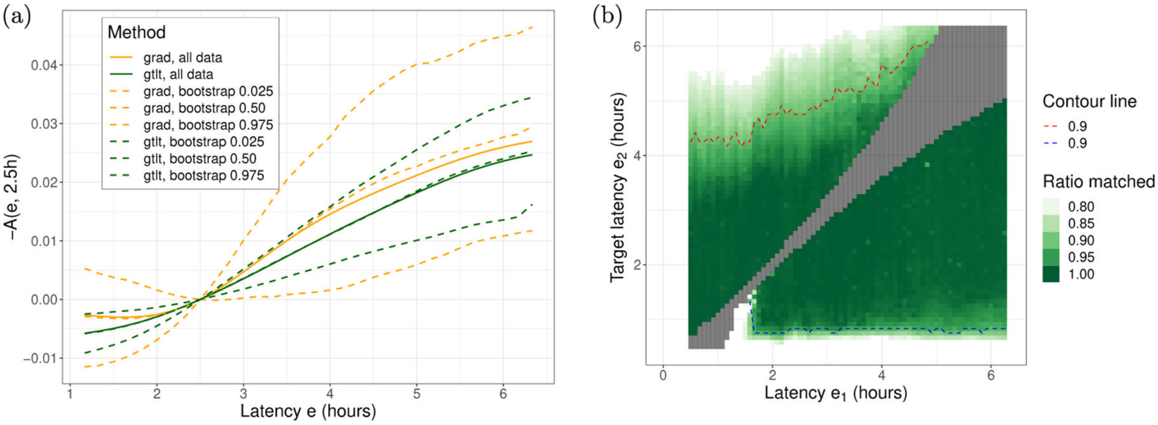

Fig. 4:

(a) Effect of antibiotic latency on prolonged (10+ days) hospital stay (excluding in-hospital mortality). $-A(e, 2.5\text{h})$ means how much patients will benefit (lowering the adverse outcome) if their latency is changed from e to 2.5h. (b) An estimation of the extent of common support. Patients with antibiotic latency of 2h or above can be matched down to 1h. So using the effect size curve in (a) it is reasonable to start from exposure 2h or above and reduce latency all the way down to 1h.

TABLE I:

RMSE ×100 for different methods over three intervals of exposure for (a) Study 1, (b) Study 2, and (c) Study 3. Reported values are the mean (95% CI) of the 400 RMSEs for each exposure range, comparing methods in regions with high density (35%−65%) going towards extreme exposure values with less densities (25%−75% and 5%−95%), pushing methods to rely more and more on extrapolating. t-test is used for all method pairs with significance level of p-value <0.001 with Holm-Bonferroni adjustment for multiple comparisons.

	35%−65%	**25%−75%**	**5%−95%**
GLM-DA	0.10 (0.01,0.3)	0.17 (0.03,0.6)	0.38 (0.08,1.2)
gtlt-PSM	0.18 (0.05,0.4)	0.32 (0.08,0.7)	1.11 (0.43,1.9)
grad-PSM	0.29 (0.07,0.7)	0.49 (0.16,1.1)	1.22 (0.47,2.2)
CB-GPS	0.26 (0.02,1.0)	0.44 (0.04,1.7)	0.96 (0.12,3.7)
GBM-DA	1.16 (0.35,3.0)	1.47 (0.57,3.5)	1.90 (0.90,3.8)
Unadjusted	3.42 (2.10,7.2)	5.07 (3.79,8.0)	10.90 (9.85,13.0)

(a) RMSE ×100 results for Study 1. t-tests are conducted for all method pairs. The only non-significant mean differences are between grad-PSM and CB-GPS.

	35%−65%	**25%−75%**	**5%−95%**
GLM-DA	0.24 (0.02,0.8)	0.39 (0.04,1.3)	0.84 (0.13,2.9)
gtlt-PSM	0.51 (0.11,1.2)	0.83 (0.26,2.0)	2.08 (0.75,4.0)
grad-PSM	0.50 (0.12,1.2)	0.84 (0.22,1.9)	2.16 (0.82,4.3)
CB-GPS	0.77 (0.05,3.0)	1.29 (0.10,5.0)	2.80 (0.30,11.2)
GBM-DA	2.35 (0.88,3.9)	3.74 (1.46,5.7)	7.89 (3.21,12.6)
Unadjusted	4.22 (2.87,7.0)	6.66 (5.47,8.8)	14.85 (13.68,16.4)

(b) RMSE ×100 results for Study 2. t-tests are conducted for all method pairs. The only non-significant mean differences are between grad-PSM and gtlt-PSM.

	35%−65%	**25%−75%**	**5%−95%**
grad-PSM	0.5 (0.2,1.6)	0.9 (0.3,2.0)	3.3 (1.9,4.9)
gtlt-PSM	1.4 (0.7,2.1)	2.0 (1.1,2.9)	3.1 (1.7,4.6)
GBM-DA	1.8 (0.3,3.4)	2.5 (0.6,4.1)	5.4 (2.4,7.7)
GLM-DA	6.8 (6.1,7.6)	8.3 (7.4,9.2)	10.0 (8.6,11.4)
CB-GPS	6.3 (4.3,9.0)	7.6 (5.0,11.2)	8.5 (4.9,14.4)
Unadjusted	6.5 (4.1,9.7)	9.0 (7.0,11.3)	18.0 (16.4,19.9)

(c) RMSE ×100 results for Study 3. t-tests are conducted for all method pairs. All the mean differences are significant.