# A Principled Approach for Designing Observational Studies from Within RCTs*

Luke Keele† David Harrison‡ Richard Grieve§

April 8, 2016

**Abstract**

Do randomized Controlled Trials (RCTs) offer the opportunity to estimate the effectiveness of treatments, other than those randomized? We develop a principled approach for estimating treatment effects for non-randomized interventions within the RCT setting. Our approach is motivated by a multicentre RCT in critical care (the PROMISE trial). The primary analysis reported that randomization to alternative clinical protocols for sepsis did not have a statistically significant effect on the primary outcome (all-cause mortality). However, as often happens in pragmatic RCTs, there was variation in the care received by patients randomised to the control arm. This raised an important clinical question: did receipt of this concomitant intervention have a causal effect on outcome? Motivated by this example, we develop a general approach for drawing causal inferences from observational contrasts in RCT data by harnessing aspects of the randomised design. We define and assess the requisite causal assumptions for identifying the effect of the new intervention. We re-define the intervention and control groups according to the receipt of the concomitant intervention of interest. We use matching based on integer programming to balance baseline covariates between these groups in a way congruent with the original random assignment. We assess the causal assumptions for identifying the effect of the concomitant intervention by conducting placebo tests made possible by the random assignment.

*We thank ... for comments and discussion.

†Associate Professor, Department of Political Science, 211 Pond Lab, Penn State University, University Park, PA 16802 Email: ljk20@psu.edu, corresponding author.

‡

§Professor of Health Economics Methodology, Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, 15-17 Tavistock Place, London, WC1H 9SH, Email: richard.grieve@lshtm.ac.uk

# 1 Introduction

## 1.1 Pragmatic Trials as Opportunities

Randomized studies have become the standard for the assessment of the effectiveness of medical interventions. In a randomized controlled experiment (RCT), subjects are randomly assigned to either a treatment group which receives a new treatment or to a control group which does not receive the intervention and maintains usual care. Randomization ensures that the treated and control groups have the same distribution on observed and unobserved pretreatment characteristics. In a randomized pragmatic trial, the intervention is designed to occur in a wide range of usual care settings, and the study seeks to enroll the kinds of participants for whom the intervention will be applied once effectiveness is established.

Given the emphasis on real world settings in pragmatic trials, there tends to be natural variation in the delivery of treatment. In these trials, patients assigned to usual care often receive elements of the new treatment, and those assigned to treatment may not fully receive the treatment. We argue that treatment variation in many pragmatic trials offers additional opportunities for gaining clinical insights. In this paper, we propose a framework for the analysis of treatments that are non-randomly assigned in pragmatic trials.

We fully acknowledge that studies of this type will require stronger assumptions since interest will focus on treatment variation that is self rather than randomly selected. However, the framework of a pragmatic trial offers natural strengths. For example, confining an observational study to the population enrolled in the trial ensures that treatment comparisons are confined to a well defined group of patients that have met a careful set of inclusion criteria. Moreover, part of the trial design is identification of the key prognostic variables that have to be balanced. These covariates are likely to be important predictors of treatment variation and may allow us to render such variation as-if randomly administered.

Here, we propose a framework for conducting observational studies of treatment variation in pragmatic trials. We clearly articulate the assumptions needed for studies of this type.

We discuss under what conditions these assumptions are more likely to be valid. We propose a testing plan that seeks to reveal if key assumptions are violated. Throughout we illustrate concepts and ideas using the ProMISe RCT a multicenter pragmatic trial conducted in the United Kingdom that focused on evaluating a sepsis management protocol. We conclude by analyzing whether treatment variation in the ProMISe RCT contributed to higher survival rates for some patients. Next, we provide more details on the ProMISe trial.

## 1.2 The ProMISe Trial

Protocolised Management In Sepsis (ProMISe) was an open, parallel group, multicentre, randomised controlled trial of the clinical and cost-effectiveness of early goal-directed protocolised resuscitation (EGDT) for emerging septic shock. EGDT as a complex intervention with many components. EGDT aims to optimize tissue oxygen transport using continuous monitoring of pre-specified physiological targets e.g. central venous pressure (CVP), mean arterial pressure (MAP) and central venous oxygen saturation (ScvO2) – to guide the delivery of IV fluids, vasoactive drugs and red cell transfusions. One potentially important aspect of EGDT is the insertion of a central venous catheter (CVC) that allows CVP monitoring which allows SCVO2 to be measured and may form the basis for subsequent changes to patient care. Prior to the ProMISe trial, resuscitation guidance was based on Rivers et al. (2001) single-centre study that found that protocolised delivery of six hours of EGDT to patients presenting at the emergency department (ED) with early septic shock reduced, hospital mortality and hospital stay. Despite the findings of the Rivers trial, the uptake of EGDT protocols was not universal amid skepticism about the generalizability of the findings. To address these concerns, multicenter trials of EGDT in America (ProCESS), Australasia (ARISE), and England (ProMISe) were conducted.

In the ProMISe trial 1260 patients admitted to an ED in 56 hospitals in England were randomly assigned to either EGDT (a six-hour resuscitation protocol) (n=630) or usual resuscitation (n=630). The study had a pragmatic design with broad inclusion criteria, and

patients randomised to the usual resuscitation (control) arm continued to receive monitoring, investigations and treatment determined by the treating clinician(s), while the EGDT (treatment) arm commenced the resuscitation protocol. The EGDT protocol required that during the first hour a central venous catheter (CVC) capable of continuous ScvO2 monitoring was inserted. Beyond the EGDT protocol, other treatment, during the intervention period and after, was at the discretion of the treating clinician(s). The ProMISe trial reported that the primary clinical outcome, all-cause mortality at 90 days, was similar between the randomised arms; in the EGDT arm 184 (29.5%) patients died versus 181 (29.2%) in the usual resuscitation group (P=0.90; absolute risk reduction -0.3%, 95% confidence interval -5.4 to 4.7; relative risk 1.01, 0.85-1.20), and average costs were higher in the EDGT versus the usual resuscitation arm. Similarly, the ARISE and ProCESS RCTs did not report a reduction in 90-day mortality for EGDT compared with usual resuscitation.

As clinicians in the 'control arm' were left to manage patients according to local discretion there was natural variation in the care received within that randomised arm, which reflected not just heterogeneity in the patient mix, but in clinical opinion clinicians as to which aspects of "usual resuscitation" should be provided according to the individual patient's prognosis. In particular, while in the intervention arm, around 95% of patients had a CVC inserted as per the EDGT protocol, in the control group around 50% of patients had a CVC inserted. CVC insertion is a controversial intervention, with some clinicians judging the insertion of this device a vital part of patient management, since it allows clinicians to closely monitor SCVO2 levels and to change fluid levels quickly. However, other clinicians point to the increased risk of infection (Parienti et al. 2015). While some clinicians believe CVC insertion is helpful, it is unclear whether it is sufficiently effective to justify the additional costs for target populations as broad as those eligible for inclusion in the PROMISE study. We explore this clinical question in our study. Specifically, the aim of our re-analysis is to estimate the causal effect of CVC insertion versus not for patients admitted to ED with emerging septic severe septic shock. We estimate this effect by contrasting outcomes for patients who received

a CVC versus those who did not in the usual resuscitation (control) arm of ProMISe. While the PROMISE study found that there was no overall difference between EGDT versus usual care, it is conceivable that there were aspects of the intervention (such as CVC insertion that were received in both randomized arms that were effective). We also exploit the availability of outcome data for patients in the EGDT arm who had a CVC inserted to assess the requisite identification assumptions. More generally, we use this clinical question in the ProMISe RCT to illustrate how one might redesigning an RCT as an observational study to address new casual questions. We next outline a general framework to guide observational studies based on randomized trials.

## 2 Framework for Analysis

Next, we outline a framework for conducting observational studies using data from RCTs, which we apply to the PROMISE trial. We define an observational study as an empirical analysis where the objective is to elucidate cause-and-effect relationships in contexts where subjects select their own treatment status (Cochran and Chambers 1965). The difficulty, of course, is that when subjects select into treatments, outcomes may reflect pretreatment differences in treated and control groups rather than treatment effects (Cochran and Chambers 1965; Rubin 1974). For the patients in the control arm that had a CVC insertion, their outcomes may result from the effects of CVC or because these patients were healthier or sicker than patients in the control group that did not have a CVC inserted. These differences in treated and control groups may be measurable and thus form overt biases or these difference may be unmeasured which are hidden biases. In an observational study, analysts use pretreatment covariates and a statistical adjustment strategy to remove overt biases in the hopes of consistently estimating treatment effects. We also attempt to assess whether our conclusions are sensitive to bias from unobserved confounders. We follow this basic pattern in our observational study. However, we first outline notation and use the potential outcomes

framework to clearly define the causal contrast and the assumptions needed to identify the causal effects of interest. We then outline a set of principles that should aid in the analysis of observational data from an RCT.

## 2.1 Notation

In our application, there are $n$ subjects, $i = 1, \ldots, n$, and subject $i$ has two potential outcomes: the outcome $Y_i(1)$ that would be observed if $i$ were assigned to ProMISe and the outcome, $Y_i(0)$ that would be observed if $i$ were assigned to usual care (Neyman 1923; Rubin 1974). Of the $n$ subjects, $m$ are randomly assigned to receive treatment, and the remaining $n - m$, control. Each of the possible $\binom{n}{m}$ treatment assignments has the same probability $\binom{n}{m}^{-1}$. For each RCT participant $i$, the indicator $Z_i = z \in \{0, 1\}$, records the randomly assigned treatment status. Combining potential outcomes with the treatment indicator, we can define the observed outcome for each unit $i$: $Y_i = Z_i Y_i(1) + (1 - Z_i)Y_i(0)$. Next, $D_i(z)$ is the potential CVC status for subject $i$ if assigned to treatment $z$, $D_i(z) = 1$ means subject $i$ had a CVC inserted if assigned to $z$, and $D_i(z) = 0$, means subject $i$ did not have CVC insertion. We also define the compound potential outcome $Y_i(z, d)$ as the outcome subject $i$ would have if he or she were assigned level $z$ of the treatment and has $d$ CVC status.

## 2.2 Estimand, Identification, Assignment Mechanism

In the application, we are interested in the following counterfactual comparison: $Y_i(0, 1) - Y_i(0, 0)|Z_i = 0$, which is the counterfactual contrast in mortality for those who received CVC versus those that did not among those assigned to usual care. As such, our primary interest will be in the following empirical contrast $E(Y_i|D_i = 1, Z_i = 0) - E(Y_i|D_i = 0, Z_i = 0)$. Our estimand is a form of an as-treated estimand, though a more typical as-treated contrast would be $E(Y_i|D_i = 1) - E(Y_i|D_i = 0)$ (Imbens and Rubin 2015). Note that our estimand differs from a instrumental variable estimand, which focuses on the causal effect in a different subpopulation. To identify as-treated estimands, it must be true that selection on $D_i$ is as-if

randomized. For this to be true, we would have to first randomize units to treatment and then randomly assign the units to compliance status. The difficultly is that units select into compliance status and as such it is typically non-ignorable. For example, in ProMISe, we have no reason to think that clinicians were randomly selecting patients in the usual care arm for CVC insertion. More likely, these patients were sicker and thus selected for CVC.

To identify this causal estimand, we must invoke a set of assumptions often referred to as identification assumptions. First, we assume that the SUTVA holds (Rubin 1986) which has the two following components: 1) there are no hidden forms of treatment, which implies that for unit $i$ under $Z_i = z$ and $D_i = d$, we assume that $Y_i(z, d) = Y_i$ and 2) a subject's potential outcome is not affected by other subjects' exposures. The first component of SUTVA is often referred to as the consistency assumption in the epidemiological literature. Next, we assume that conditional ignorability holds (Rosenbaum and Rubin 1983). First, we define $\mathbf{x}_i$ as a set of pretreatment covariates, and $u_i$ is an unobserved binary covariate. Formally, the conditional ignorability assumption is comprised of two parts. First, it must be the case that

$$\pi_i = Pr(D_i = 1|Y_i(z, d), \mathbf{x}_i, u_i) == Pr(D_i = 1|\mathbf{x}_i).$$

Second, we must assume there is overlap: $0 < \pi_i < 1$. The first assumption is the critical departure from the randomized trial. We are assuming that within the usual care arm receipt of CVC is effectively as if random once we condition on covariates. In short, we must assert that there is some set of covariates such that treatment assignment is random conditional on these covariates (Barnow et al. 1980). Critically, the selection on observables assumption is nonrefutable, insofar as it cannot be verified with observed data (Manski 2007). Next, we outline our central methodological contribution in that we outline the key elements needed to plausibly identify causal effects from an observational data based on RCTs.

# 3 Observational Studies from RCTs: Critical Elements

Given that we are assuming that treatment decisions are observable, we need to probe these assumptions as much as possible. There are some strategies that we think should generally be adopted in designs of this type given that we must adopt a strong assumption for identification of the causal effects.

## 3.1 Eligibility Criteria

First, we note one built advantage of building an observational study from an RCT. Hernán and Robins (2016) note that observational studies seeking to mirror randomized trials must use the same eligibility criteria as the trial. Otherwise, study populations may be too heterogenous to yield valid causal comparisons. One general advantage of using data from a pragmatic RCT to define the target population is that all patients in our study were subject to the same eligbility criteria of the PROMISE study. Since all subjects are in the same trial, they are all from the same well defined population defined in the RCT analysis plan. As such, for the new causal question focusing on the effects of CVC insertation all the ProMISe RCT participants are potentially in the target population. However, even in the EGDT arm, 50 patients did not receive the CVC insertion hence these patients and their unknown counterparts randomised to the control arm, are outside the target population of interest. Within the remaining RCT participants, to address the causal question of interest (CVC versus not) it will be important to design the study to account for this feature.

## 3.2 Assignment Mechanism

Under the adopted identification strategy, we are assuming that selection into the CVC insertion treatment is based on observable covariates. In essence, we seek to model the treatment assignment mechanism for a subpopulation of the ProMISe population. As (Rubin 2008) notes investigators of any observational study must ask themselves who are the decision

makers in charge of selection and what criteria did they use in the selection process. We argue that observational studies based on RCTs, will be more likely to remove overt biases when treatment decisions are centralized with clinicians rather than with patients. That is, we are more likely to identify the causal estimand if we can model a decision that doctors make with clinical data, rather than situations where patients select a treatment. We would argue that the latter case is more likely to be a decision that is based on unobservables. This works in our favor in ProMISe, since the decision for CVC insertion was centralized with clinicians. However, it will also be important to have a thick description of the clinical decision-making leading to the insertion of a CVC (Rosenbaum 2001). This qualitative information will be critical for both the selection of covariates and the prioritization of covariates when we adjust for overt bias.

### 3.3 Use An Analysis Plan

One approach to reducing analysis and publication bias is for investigators to articulate a specific study design and analysis plan before the study begins. The hope is that there will be less room for investigators to selectively report results or focus on only on significant findings (John et al. 2012; Casey et al. 2012). The analysis plan may include a pre-specification of covariates to be used in adjustment, the contrasts of interest, and perhaps whether the analysis will be blinded to outcomes. See Rubin (2008) on importance of not looking at outcomes. In general, we can use various aspects of the RCT to formulate the analysis plan. We can make our statistical adjustments for covariates consistent with the level of balance in the RCT, follow the same subgroup analyses, and rely on equivalent outcomes for placebo test. Whether this analysis plan is one that is registered in publicly or simply a part of the analysis outline is something we leave to the analyst. Many RCTs follow an analysis plan, and it may serve as a template for the observational study.

### 3.4 Testing Plan

Following the logic in Rosenbaum (2010), we develop a plan for our statistical tests. As he notes, a planned analysis outlines a set of statistical tests that mirror the clinical expectations of the study, which increases transparency. Here, we develop a plan for the analysis, based on a design outlined in Rosenbaum (2010, ch. 19). We argue that this testing plan is critical for bolstering the plausibility of our conclusions. Our testing plan is based on two devices: pattern specificity and placebo tests.

First, a pattern of specific confirmatory tests provides better evidence than a single test for a causal hypothesis. As Cook and Shadish (1994, pg. 95) write: "Successful prediction of a complex pattern of multivariate results often leaves few plausible alternative explanations." Under pattern specificity, part of the design is the generation and testing of multiple hypotheses based on the causal theory. Thus we seek to test a pattern of results from the ProMISe data. Second, we use placebo or falsification tests. Causal theories often both predict the presence of a causal effect in some settings, but also may predict the absence of causal effects. When we find causal effects where they should not be, this is often a sign of hidden confounders and a failure of the identification strategy. Based on clinical knowledge, we expect that administration of a CVC to patients in the control arm offered some protective benefit. As such, control patients with CVC should have better outcomes than control patients without CVC, and should have outcomes more like those in the treated arm. The ProMISE trial found that the full EGDT intervention did not have an effect compared to usual care, so as the other aspects of the randomised intervention did not appear to have an effect (unless they were negative and removed the effect of CVC), as such, it is reasonable to assume that they do not have an effect. Therefore, if CVC (EGDT) versus CVC (usual care) produces an effect then this is likely to be unobserved confounding rather than residual differences in the treatment. Based on this expectation, we expect that in the comparison within the control arm outcomes should differ, and in the comparison across treated arms outcomes should not differ. We now formalize this expectation in a testing plan. While we

have tailored this testing plan to the ProMISe data, we think it serves a a general template for any study with a similar form of treatment variation.

First, we define the three relevant groups for comparison. The first group is where $D_i = Z_i = 1$, these are treated patients that received CVC; in the second group $D_i = 1$ and $Z_i = 0$, these are control patients that had a CVC insertion; and for the final group $D_i = Z_i = 0$, these are control patients that did not receive a CVC. We denote these groups with $g_{dz}$: $g_{00}$, $g_{10}$, and $g_{11}$. Next, we define $\mu_{11}$ to be the mean of $g_{11}$, $\mu_{10}$ is the mean of $g_{10}$ and so forth. Based on clinical knowledge, we expect that $\mu_{11} = \mu_{10}$ but those two groups should not have means that are equal to $\mu_{00}$ since these patients most closely adhered to the standard of usual care and did have a CVC inserted.

Therefore, we want to test the following proposition: that the mean of $g_{00}$ exceeds the means of groups $g_{11}$ and $g_{10}$ by more than the means of groups $g_{11}$ and $g_{10}$ differ from each other. That is to say, the usual care group is more different from either the treated group or the control group that received CVC than the two groups that received CVC are from each other. This set of propositions implies a pattern of effects that should be observable in series of tests that include a placebo test.

The following planned analysis allows us to fully test these propositions. First, we are interested in being able to reject the following hypothesis: $H_0$: $\mu_{00} \leq \mu_{11}$. This hypothesis implies the following second hypothesis: $\tilde{H}_0$: $\mu_{00} \leq \mu_{10}$. We could test these hypothesis separately or we could also test that the expected outcome in $g_{00}$ is no higher than the average outcome in groups $g_{10}$ and $g_{11}$. We represent this test with the following hypothesis: $\bar{H}_0^{(\tau)} : (\mu_{00} - \tau) \leq (\mu_{11} + \mu_{10})/2$, where we set $\tau = 0$. We can test this hypothesis by estimating the following contrast $1 \times \mu_{00} - (1/2) \times \mu_{10} - (1/2) \times \mu_{11}$. However, as Rosenbaum (2010) notes, $H_0$ and $\tilde{H}_0$ do not preclude one another, but if $\bar{H}_0^{(\tau)}$ is false they do preclude each other. As such, he suggests testing $\bar{H}_0^{(\tau)}$ before testing $H_0$ and $\tilde{H}_0$.

Next, we wish to test that the groups $g_{10}$ and $g_{11}$ are nearly equivalent in their outcomes. That is, we need to test whether the placebo test holds. To assert that two groups are close

in their outcomes is to assert that $|\mu_{11} - \mu_{10}| < \delta$ is small for some $\delta > 0$. Thus to pass the placebo test, we must reject the following hypothesis $H_{\neq}^{(\delta)} : |\mu_{11} - \mu_{10}| < \delta$. Rejecting $H_{\neq}^{(\delta)}$ provides a basis for asserting with confidence that $|\mu_{11} - \mu_{10}| < \delta$. $H_{\neq}^{(\delta)}$ is the union of two exclusive hypotheses: $\overleftarrow{H}_0^{(\delta)} : \mu_{11} - \mu_{10} \leq -\delta$ and $\overrightarrow{H}_0^{(\delta)} : \mu_{11} - \mu_{10} \geq \delta$, and $H_{\neq}^{(\delta)}$ is rejected if both $\overleftarrow{H}_0^{(\delta)}$ and $\overrightarrow{H}_0^{(\delta)}$ are rejected (Rosenbaum and Silber 2009). We can apply the two tests without correction for multiple testing since we test two mutually exclusive hypotheses. How should we select the magnitude for $\delta$? There are two possible routes. We could select a plausible value for $\delta$ based on the trial. In ProMISe, we might set $\delta$ equal to the difference produced across the EDGT and usual care arms. Alternatively, we could be more agnostic and set $\delta = \infty$ and continue testing smaller and smaller values of $\delta$ until a value of $\delta$ is encountered such that either $\overleftarrow{H}_0^{(\delta)}$ or $\overrightarrow{H}_0^{(\delta)}$ is not rejected. We then assert with 95% confidence that $|\mu_{11} - \mu 10| < \delta$.

However, while this test of equivalence allows for a straightforward placebo test, it has a specific weakness. In short, we may reject $H_{\neq}^{(\delta)}$, for a given value of $\delta$ larger than zero. However, we may also wish to judge whether $\delta$ is small when it is larger than zero. Thus we might reject the placebo test by using too narrow of a test. As formulated, we might reject that $H_0$: $\mu_{00} < \mu_{11}$ and $H_0$: $\mu_{00} < \mu_{10}$ even though group $g_{00}$ is close to one of the other two groups, but the difference between the two groups $g_{11}$ and $g_{10}$ is larger than zero. Rosenbaum (2010) argues that a difference between groups $g_{11}$ and $g_{10}$ is problematic unless that difference is small compared to the difference between $g_{00}$ and $g_{11}$ and $g_{10}$. We are not necessarily interested in whether the difference between $g_{11}$ and $g_{10}$ is significantly different from zero. Instead, we care more about whether the magnitude of bias needed to explain the difference between $g_{11}$ and $g_{10}$ is smaller than the magnitude of the bias need to explain the difference between $g_{00}$ and $g_{11}$ and $g_{10}$.

Rosenbaum (2010) formalizes this idea in the following way: for a fixed number $\delta$, we wish to assert $(\mu_{00} - \delta) - \max(\mu_{11}, \mu_{10}) > \max(\mu_{11}, \mu_{10}) - \min(\mu_{11}, \mu_{10})$ We could assert this

is true if we rejected the following hypothesis:

$$\bar{H}_{\triangle}^{(\delta)} : (\mu_{00} - \delta) - \max(\mu_{11}, \mu_{10}) \leq \max(\mu_{11}, \mu_{10}) - \min(\mu_{11}, \mu_{10})$$

For example, if we reject $\bar{H}_{\triangle}^{(\tau)}$ at level 0.05 when $\tau = 0$, we have 95% confidence that the mean of group $g_{00}$ exceeds the means of groups $g_{00}$ and $g_{11}$ by more than the groups $g_{00}$ and $g_{11}$ differ from each other. In the context of our application, we satisfy the placebo test if the difference between those without CVC in the control arm and those with CVC in both the treated and control arms is statistically larger than the difference between those with CVC in the control and treatment arms. This leads to a testing plan based on the following set of tests:

1. Test $\bar{H}_0^{(\tau)} : (\mu_{00} - \delta) \leq (\mu_{11} + \mu_{10})/2$. If the $p$-value from this test is greater than 0.05 stop, we are unable to reject $\bar{H}_{\triangle}^{(\delta)}$, if not proceed to step 2.

2. Test both $H_0^{(\delta)} : (\mu_{00} - \delta) \leq \mu_{11}$ and $\tilde{H}_0^{(\delta)} : (\mu_{00} - \delta) \leq \mu_{10}$. If either $p$-value is above 0.05 stop, we are unable to reject $\bar{H}_{\triangle}^{(\delta)}$, if not proceed to step 3.

3. Test both $H_{\circ}^{(\delta)} : (\mu_{00} - \delta) - \mu_{11} \leq \mu_{11} - \mu_{10}$ and $H_{\bullet}^{(\delta)} : (\mu_{00} - \delta) - \mu_{10} \leq \mu_{10} - \mu_{11}$. If both $p$-values are less than 0.05 we can reject $\bar{H}_{\triangle}^{(\delta)}$. If either $p$-value is larger than 0.05 we are unable to reject this hypothesis.

If we able to pass all three steps, we can say with 95% confidence that the mean of $g_{00}$ exceeds both of the means of groups $g_{00}$ and $g_{11}$ by more than these two means exceed each other. For example, this would allow us to assert with 95% confidence that the mortality rate for the usual care group that did not receive CVC exceeds the mortality rate of both the groups that did receive CVC by more than the mortality rate differs across the two CVC groups. In essence, this builds outcome testing and the placebo test into a single testing procedure.

### 3.5 Sensitivity Analysis

Finally, we argue that a sensitivity analysis is necessary. Formally a sensitivity analysis is designed to *quantify* the degree to which a key identification assumption must be violated in order for a researcher's original conclusion to be reversed. A sensitivity analysis provides a quantifiable statement about the plausibility of an identification strategy. If a causal inference is sensitive, a slight violation of the assumption may lead to substantively different conclusions. In our case, the key identification assumption is that selection into CVC insertion is only a function of observable covariates. As such, we should apply a sensitivity analysis to probe the plausibility of this assumption. We detail the exact form of the sensitivity analysis below, but we use it to state the level of hidden bias that would need to be present to reverse the study conclusions.

## 4 Analysis of ProMISe

We now implement these ideas within the context of our application. We start by explicating our analysis plan. We then describe the data and the statistical methods we used to both adjust for overt bias and to analyze outcomes. Finally, we implement our testing plan.

### 4.1 Analysis Plan

We begin by outlining the elements of our analysis plan. For our study of the data from the ProMISe RCT, our analysis plan includes three key components. First, we define the relevant target population of interest. Second, we define our adjustment strategy to remove overt biases. Critically, we use the RCT itself to guide choices about covariates and balance in the adjustment strategy. Third, we trim the study sample to the subset of patients that are comparable. That is, we remove patients from the usual care arm that were unlikely to receive CVC based on covariates. Fourth, we report results from a sensitivity analysis for matched samples. Finally, we implemented our testing plan to make judgements about the

plausibility of causal effects

### 4.1.1 Eligbility Criteria

While the elgibility criteria in our study are largely based on those for inclusion into ProMISe, we added one additional element in the form of an exclusion criterion. In the treated arm of the study, 50 patients (8% of those assigned to treatment) did not receive a CVC. We exclude these patients from the population of treated patients that we will use as the placebo test. We also matched these 50 patients to 50 patients from the the control arm that did not have a CVC inserted. We then excluded these 50 matches controls from further use in the study.

### 4.1.2 Adjustment Strategy

Under our identification strategy, we use a statistical adjustment method, to make patients more alike in terms of observable covariates. We use multivariate matching as our method of statistical adjustment. Other methods could be applied as well, however, matching will facilitate trimming the sample to those patients most likely to receive CVC. We use matching to perform two statistical adjustments. First, we use matching to find a set of patients in the usual care arm that are received CVC that are comparable to patients in the usual care arm that did not. Second, we use matching to find a set of treated patients that are comparable to those who received CVC in the first match. Thus, we use it to bolster the placebo test.

However, we use a newer form of matching that allows us to use the ProMISe trial balance directly in the matching process. Under most forms of matching, a distance metric is minimized perhaps subjects to constraints, but users cannot target balance specifically. That is, the level balance returned by the matching algorithm is a by-product of minimizing distance within matched pairs. We use an alternate form of matching based on integer programming (Zubizarreta 2012). Matching based on integer programming achieves covariate balance directly by minimizing the total sum of distances while constraining the measures

of imbalance to be less than or equal to certain tolerances. This form of matching also allows us to impose constraints for exact and near-exact matching, and near and near-fine balance for more than one nominal covariate. Under this form of matching, the analyst sets a tolerance for the imbalance for each covariate. For example, we might set the tolerance on the difference in means on a covariate to be less than 2 points across treated and control groups. The algorithm then attempts to achieve that level of balance. If this balance constraint is infeasible the algorithm stops and reports an error. We implemented the matches using the `R` package `designmatch` (Zubizarreta and Kilcioglu 2016) .

This form of matching has two strengths for our purposes. First, we use an integer programming based match to target the same level of balance produced by the ProMISe trial. For example, in the ProMISe trial the difference in the proportion of females across arms of the study was 2 points. When we match, we set the balance constraint in the match for the proportion of females to be two points. In fact, for every covariate on which we match, we set a balance constraint based on the level of balance for the trial. For binary covariates, we set balance constraints using the difference in proportions. For continuous covariates, we use the Kolmogorov-Smirov (KS) test, which is the maximum discrepancy in the empirical cumulative distribution functions. Tables 7 and 8 in the appendix contain the balance statistics produced by randomization.

Second, this form of matching allows us to prioritize balance among the covariates. From the baseline variables measured at randomisation clinicians defined two sets according to their prognostic importance (Ramsahai et al. 2011). Based on clinical input, we designated one set of covariates that were of high priority and a second set that were of lower priority. Hence in each stage of the subsequent matching, most attention was given balancing the high priority variables and the balance on these variables are reported in the main results tables, with the corresponding results for the low priority variables given in the appendixes. The logic behind this decision is as follows. We are in the fortunate position that patients were not themselves selecting to receive CVC. Instead this was a clinical decision made

by doctors using available evidence. Thus we seek to mimic this assignment process by prioritizing covariates that would have been given the most weight in the clinical decision making process.

### 4.1.3 Outcome and Sensitivity Analysis

We below we perform outcome analyses to compare between-group differences in 90 day mortality and in the mean 90 day cost. For outcome analysis, we use methods based on randomization inference. The notation and methods under randomization inference depend on whether we are working with an unmatched or matched sample. We first review methods for the unmatched data. Under the sharp null hypothesis, we test whether the treatment effect is zero for all units. In potential outcomes notation, if the sharp null hypothesis holds then $Y_i(1) = Y_i(0)$ for every $i$. To test the sharp null, we define a test statistic as a function of the data. The sharp null hypothesis is tested by calculating the observed value of a test-statistic and comparing it to the randomization distribution. For a binary treatment and outcome, Fisher's exact test is a randomization test (Rosenbaum 2002b, ch. 5). The test statistic is the total number of units in the treatment group with a response equal to 1. Under the null hypothesis of no treatment effect, the randomization distribution of this test statistic follows a hypergeometric distribution. When $n$ is large, the $\chi^2$ test is an approximation to the exact $p$-value. For continuous outcomes, we use Wilcoxon's sum rank test and the associated Hodges-Lehmann estimate of the treatment effect.

After matching, there are $I$ matched pairs, $i = 1, \ldots, I$, with two subjects, $j = 1, 2$, one treated and one control for $2I$ total subjects. Treatment assignment, potential outcomes, and observed outcomes are respectively $Z_{ij}$, $Y_{ij}(1)$, $Y_{ij}(0)$, and $Y_{ij}$. We test Fisher's sharp null hypothesis using McNemar's test, which is based on the number of discordant pairs in matched outcomes. In the case of matched pairs with binary responses, pair $i$ is discordant if it contains exactly one person who voted, $Y_{i1} + Y_{i2} = 1$. McNemar's statistic is the number of deaths, $T$, among treated subjects in discordant pairs, $T = \sum_{i \in K} \sum_{j=1}^{2} Z_{ij} Y_{ij}$, where $K$ is

a set of indices for the $I^* \leq I$ discordant pairs. Some of the deaths recorded in $T$ may have been caused by a CVC insertion and others might have occurred whether there was a CVC inserted or not. The unobservable quantity $T_c = \sum_{i,j} Z_{ij} y_{Cij}$ is the number of deaths that would have occurred without a CVC inserted. Fisher's sharp null hypothesis, $H_0 : \boldsymbol{\delta} = 0$, says that no deaths were caused or prevented by the ballot initiative, implying that $T = T_c$. Therefore, this hypothesis may be tested by comparing $T$ with the randomization distribution of $T_c$, which follows a binomial distribution with sample size $I^*$ and probability of success $1/2$. For continuous outcomes, we use the sign rank test.

In an observational study, we can base a test of the sharp null hypothesis on the randomization distribution of $T_c$ (Rosenbaum 2002a). In particular, if every unit $j$ in pair $i$ has the same probability of receiving treatment, $\Pr(Z_{ij} = 1) = 1/2$. This mode of treatment assignment would be true by construction in a pair randomized experiment since we would choose one unit at random from each pair to receive treatment. In our analysis, we assume this model of treatment assignment holds after conditioning on $\mathbf{x}_j$. One model for a sensitivity analysis of this assumption stipulates that $1/(1+\Gamma) \leq \Pr(Z_{ij} = 1|\mathbf{x}_j) \leq \Gamma/(1+\Gamma)$ for a specified value of $\Gamma$ greater than one, such that randomization inference corresponds to $\Gamma = 1$; see Rosenbaum (2002b, §4) for a discussion. We use values of $\Gamma > 1$ to compute a range of possible inferences, which indicates the magnitude of bias due to an unobserved covariate that would need to be present to alter the conclusions reached when we assume that random assignment of the treatment holds given the observed covariates.

## 4.2 Data

In the match, we included the following covariates: age in years, female, APACHE II score[1], MEDS score[2], SOFA score[3]. We also matched on indicators for whether a patient had a severe liver, renal, respiratory cardiovascular condition in his or her past medical history or had ever been immuno-compromised in the past. Finally, we also matched on blood lactate concentration (mmol/L), systolic blood pressure (SBP – mmHg), mean arterial pressure (MAP – mmHg), median total intravenous fluids pre hospital and in the hospital up to randomization, supplemental oxygen, median time from emergency department presentation to randomization, amount of blood products administered (ml), an indicator for whether a patient would have been admitted direct to ICU from ED it not enrolled in ProMISe.

As we noted above, we chose to prioritize balance on some covariate over others. In consultation with the physicians that actively participated in the ProMISe trial, we identified that we should prioritize balance on the APACHE II score, the MEDS score, the SOFA score, age, blood lactate concentration, systolic blood pressure, and mean arterial pressure. Clinical standards suggest that patients should be comparable on these covariates in particular.

## 4.3 Matching

We performed three different matches. As we noted in our analysis plan, clinical information suggested that we should remove from the usual care arm the set of patients most like the patients in the ProMISe arm that did not received CVC. To that end, we matched the 50 patients who did not receive CVC in the treatment arm to 50 patients in the usual care arm

[1]Scores on the Acute Physiology and Chronic Health Evaluation (APACHE) II range from 0 to 71, with higher scores indicating greater severity of illness. The APACHE II score was calculated on the basis of the last recorded data before randomization.

[2]Scores on the Mortality in Emergency Department Sepsis (MEDS) scale range from 0 to 27, with higher scores indicating greater severity of illness. The MEDS score was calculated on the basis of the last recorded data before randomization.

[3]Scores on the Sequential Organ Failure Assessment (SOFA) range from 0 to 24, with higher scores indicating a greater degree of organ failure. The SOFA score was calculated on the basis of the last recorded data before randomization. The SOFA renal score was based on the plasma creatinine level only and did not include urine output.

that did not receive CVC. Given the high ratio of treated to control in this match, we found it quite easy to find a comparable subset in the usual care arm. After this exclusion, we are left with 318 usual care patients that received CVC and 258 that did not. We next sought to match the 258 non-CVC patients to the 318 CVC patients.

When we completed the matching process within the usual care arm, we found that there was no set of matches that had a balance that was equivalent to the balance produced by randomization in the trial. This not entirely surprising. The pool of units available for matching is not large, which often makes balance difficult to achieve. To better balance the data we used optimal subset matching, which can be applied to find the largest set of treated units for which balance constraints are met (Rosenbaum 2012). We implemented this through the use of cardinality matching (Zubizarreta et al. 2014) Using integer programming, a cardinality match seeks to preserve the largest sample that meets a set of balance constraints. Thus, we set balance constraints for each covariate consistent with the balance achieved by randomization, while prioritizing the key clinical variables. The algorithm trims the sample so that we have the largest set of matched pairs that meet the balance constraints.

Trimming units to achieve a specific level of balance changes the causal estimand such that it only applies to the population of units for which the effect of treatment is marginal: units that may or may not receive the treatment. Generally changing the estimand in this way is viewed as unproblematic if the goal is to estimate the effect of a treatment when the data do not represent a well-defined population (Rosenbaum 2012). Here, we are removing from the study patients that did not receive CVC in the usual care arm. We would argue that this is not a well-defined population, and, we are in fact only interested in the patients where there was some probability that they could have received a CVC. Undoubtedly, for some patients there was little chance they would have been administered CVC. These patients do not directly interest us.

After cardinality matching, we are left with 204 matched pairs where balance is close to or slightly better than that in the trial. The balance results are in Tables 1 and 2. Due to

Table 1: Balance Comparison Between Patients in the Usual Care Arm that Received CVC and those that did not After Matching

| | Treated Mean | Control Mean | Std. Diff | p-value |
|---|---|---|---|---|
| Age | 64.29 | 65.19 | -0.06 | 0.56 |
| Apache Score | 16.67 | 17.36 | -0.10 | 0.26 |
| SOFA Score | 3.96 | 4.01 | -0.02 | 0.78 |
| MEDS Score | 7.81 | 7.88 | -0.02 | 0.83 |
| Blood Lactate | 4.55 | 4.63 | -0.02 | 0.78 |
| Systolic Blood Pressure (SBP) | 97.43 | 95.09 | 0.10 | 0.26 |
| Mean Arterial Pressure (MAP) | 66.41 | 65.71 | 0.09 | 0.28 |
| Male | 0.55 | 0.57 | -0.04 | 0.69 |
| Liver Severe Cond | 0.01 | 0.01 | 0.00 | 1.00 |
| Renal Severe Cond | 0.00 | 0.01 | -0.07 | 0.56 |
| Immuno Severe Cond | 0.13 | 0.11 | 0.06 | 0.55 |
| Resp Severe Cond | 0.13 | 0.11 | 0.06 | 0.54 |
| Cardio Severe Cond | 0.03 | 0.03 | -0.03 | 0.78 |
| Fluids Pre Hosp | 107.18 | 122.67 | -0.06 | 0.58 |
| Fluids Pre Rand | 1751.56 | 1794.35 | -0.04 | 0.66 |
| Supp Oxygen | 0.50 | 0.51 | -0.03 | 0.77 |
| Time from ED to Rand | 2.84 | 2.89 | -0.04 | 0.72 |
| Direct ICU to ED | 0.48 | 0.91 | -1.10 | 0.00 |
| Blood Products | 4.95 | 8.13 | -0.03 | 0.65 |
| Blood Lactate Miss | 0.07 | 0.07 | 0.02 | 0.85 |
| SBP Miss | 0.07 | 0.09 | -0.08 | 0.46 |
| MAP Miss | 0.77 | 0.80 | -0.06 | 0.55 |
| Fluid PH Miss | 0.01 | 0.01 | 0.00 | 1.00 |
| Fluid PR Miss | 0.00 | 0.02 | -0.12 | 0.18 |
| Oxygen Miss | 0.12 | 0.15 | -0.09 | 0.39 |

the prioritization of certain variables, the balance is not identical to that in the trial but is very close. We then conducted one additional match. Following the analysis plan, we matched the 204 patients that received CVC in the usual care arm to the 575 patients that received CVC in the ProMISe arm. Balance results can be found in the appendix. We had little trouble finding a highly comparable set of patients.

Table 2: KS Test Balance Comparison Btw Those in Usual Care Arm that got CVC and those that did not.

| | KS Test Stat | KS p-value |
|---|---|---|
| Age | 0.05 | 0.93 |
| Apache Score | 0.08 | 0.56 |
| SOFA Score | 0.05 | 0.97 |
| MEDS Score | 0.04 | 1.00 |
| Blood Lactate | 0.08 | 0.48 |
| Systolic Blood Pressure | 0.09 | 0.34 |
| Mean Arterial Pressure | 0.03 | 1.00 |
| Fluids Pre Hosp | 0.06 | 0.80 |
| Fluids Pre Rand | 0.05 | 0.97 |
| Supp Oxygen | 0.08 | 0.56 |
| Blood Products | 0.03 | 1.00 |

## 4.4 Outcome Analysis

We present the outcome analysis in two parts. First, we report unadjusted estimates. We report these in addition to the estimates after matching to understand whether the results change after our analytic adjustments to the data. For review, the mortality rate in the ProMISe arm of the trial was 0.294 and was 0.289 in the usual care arm for a risk differences of 0.01 (95% CI -0.04–0.055). Next, we compare risks within the usual care arm. Table 3 contains the results for those in the usual care arm that received CVC compared to those in the usual care arm that did not received CVC. We observe that usual care patients who received had higher mortality rates and higher costs.

Table 3: Unadjusted Associations between CVC and non-CVC Cohorts

| | CVC | non-CVC | Risk Difference | 95% CI | $p$-value |
|---|---|---|---|---|---|
| 90 Day Mortality | 0.346 | 0.236 | 0.109 | [0.039, 0.179] | 0.003 |
| 90 Day Costs | 8993 | 4622 | 3741 | [4775, 2755] | 0.000 |

Note: Entries for length of stay and costs are medians. Tests for risk difference based on Fisher's exact test and Wilcoxon summed rank test.

Table 4 contains the same results comparing patients that received CVC in the usual care arm to treated patients that received the full ProMISe protocol including CVC. While

the differences are not has large, CVC patients in the usual care arm had higher rates of mortality and higher costs.

Table 4: Unadjusted Associations between CVC in Usual Care Arm and Treated Arm

| | CVC | Treated | Risk Difference | 95% CI | $p$-value |
|---|---|---|---|---|---|
| 90 Day Mortality | 0.346 | 0.300 | -0.045 | [-0.108, 0.021] | 0.175 |
| 90 Day Cost | 8993 | 7744 | 883 | [-61.6, 1844] | 0.067 |

Note: Entries for length of stay and costs are medians. Tests for risk difference based on Fisher's exact test and Wilcoxon summed rank test.

Next, we examine the outcomes after matching. Table 5 matching, the mortality rate, among those who received CVC in the usual care arm was 0.262, and was 0.269 among those in the usual care arm that did not received CVC for a risk differences of 0.007 (95% CI -0.094–0.095). Based on the outcome analysis, it would appear that patients in the usual care arm that received a CVC were particularly sick, which caused clinicians to administered additional care via the insertion of a CVC. However, once, we find a comparable set of patients, these differences in mortality are removed.

Next, we examine the secondary outcome of costs. As we noted above, patients that had CVC inserted in the usual care arm tended to have higher costs. Even once we examine comparable patients, we find those patients that received CVC had higher costs. We find that the median risk difference between the two groups is 4770 95% CI (95% 3285–6365). If there is no hidden bias, $\Gamma = 1$, then the sharp null hypothesis of no treatment effect is implausible as the $p$-value from the test is 0.000. The upper bound on the $p$-value is 0.048 for $\Gamma = 2.09$ and 0.050 for $\Gamma = 2.10$, which indicates that a hidden confounder that doubles the odds of CVC insertion would explain the associate we observe.

Next, we seek to check our results using our testing plan. As we noted above, we matched the patients with CVC in the control arm to patients with a CVC in the treatment arm. We expect that if use of a CVC has key protective benefits, these patients should have outcomes similar to those in the treatment arm. Table 6 contains the results after matching for the

Table 5: Associations between CVC and non-CVC in Matched Cohort

| | CVC | non-CVC | Risk Difference | 95% CI | $p$-value |
|---|---|---|---|---|---|
| 90 Day Mortality | 0.269 | 0.262 | 0.007 | [-0.095, 0.095] | 0.914 |
| 90 Day Cost | 8998 | 4791 | 4770 | [3285, 6365] | 0.000 |

Note: Entries for length of stay and costs are medians. Tests for risk difference based on McNemar's test and Wilcoxon signed rank test.

placebo analysis. We find that in terms of survival outcomes are nearly identical. The risk differences is 0.001 (95% -0.085–0.104). Moreover, the usual care patients that had a CVC inserted still had higher costs. The median difference in costs is 2170 (95% 251–4265). While these differences in costs are smaller, they remain statistically significant, these results emphasize the need for our testing plan. That is, we find that there are differences between the two groups that received CVC, but we would like to tests whether those differences are larger than the difference between the non-CVC patients and those that had a CVC inserted.

We first use the testing plan with the survival outcome. Here, we are unable to reject $\bar{H}_0^{(\tau)}$ ($p = 0.496$), which implies that we need not proceed with the rest of the testing plan. Next, we proceed to the costs outcome. Here we can reject $\bar{H}_0^{(\tau)}$ ($p < 0.001$). We are also able to reject both $H_0$ and $\tilde{H}_0$ ($p < 0.001$ and $p = 0.032$). This implies that patients that did not have CVC inserted has lower costs that either group that had a CVC insertion. Next, we proceed to Step 3 in the testing plan. Using th modified test of equivalence, we can say with 95% confidence that the mean cost in group $g_{00}$ exceeds the means in both $g_{11}$ and $g_{00}$ by at least 23 pounds more than the two control differ from each other. Thus while the difference in means is between $g_{11}$ and $g_{00}$ is statistically significant, we can say with 95% confidence that the difference between $g_{00}$ and both $g_{11}$ and $g_{00}$ is substantially larger than the difference between $g_{11}$ and $g_{00}$. In words, this implies that the placebo test holds for the cost outcome.

Table 6: Placebo Test Using Treated Arm Matched to CVC in Usual Care Arm

| | CVC | Treated | Risk Difference | 95% CI | $p$-value |
|---|---|---|---|---|---|
| 90 Day Mortality | 0.269 | 0.261 | 0.009 | [-0.085, 0.104] | 0.914 |
| 90 Day Cost | 8998 | 7558 | 2170 | [4265, 251] | 0.028 |

# 5 Discussion

No protective benefit, CVC increased costs in both other arms by more than cost in usual care arm without CVC.

# References


Barnow, B., Cain, G., and Goldberger, A. (1980), "Issues in the Analysis of Selectivity Bias," in *Evaluation Studies*, eds. Stromsdorfer, E. and Farkas, G., San Francisco, CA: Sage, vol. 5, pp. 43–59.

Casey, K., Glennerster, R., and Miguel, E. (2012), "Reshaping institutions: Evidence on aid impacts using a pre-analysis plan," *Quarterly Journal of Economics*, 127, 1755–1812.

Cochran, W. G. and Chambers, S. P. (1965), "The Planning of Observational Studies of Human Populations," *Journal of Royal Statistical Society, Series A*, 128, 234–265.

Cook, T. and Shadish, W. (1994), "Social Experiments: Some Developments Over the Past Fifteen Years," *Annual Review of Psychology*, 45, 545–580.

Hernán, M. A. and Robins, J. M. (2016), "Usig Big Data to Emulate a Target Trial When a Randomized Trial is Not Available," *American Journal of Epidemiology*, Forthcoming.

Imbens, G. W. and Rubin, D. B. (2015), *Causal Inference For Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge, UK: Cambridge University Press.

John, L. K., Loewenstein, G., and Prelec, D. (2012), "Measuring the prevalence of questionable research practices with incentives for truth telling," *Psychological science*, 0956797611430953.

Manski, C. F. (2007), *Identification For Prediction And Decision*, Cambridge, Mass: Harvard University Press.

Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science*, 5, 465–472. Trans. Dorota M. Dabrowska and Terence P. Speed (1990).

Parienti, J.-J., Mongardon, N., Mégarbane, B., Mira, J.-P., Kalfon, P., Gros, A., Marqué, S., Thuong, M., Pottier, V., Ramakers, M., et al. (2015), "Intravascular Complications of Central Venous Catheterization by Insertion Site," *New England Journal of Medicine*, 373, 1220–1229.

Ramsahai, R. R., Grieve, R., and Sekhon, J. S. (2011), "Extending iterative matching methods: an approach to improving covariate balance that allows prioritisation," *Health Services and Outcomes Research Methodology*, 11, 95–114.

Rivers, E., Nguyen, B., Havstad, S., Ressler, J., Muzzin, A., Knoblich, B., Peterson, E., and Tomlanovich, M. (2001), "Early Goal-Directed Therapy in the Treatment of Severe Sepsis and Septic Shock," *New England Journal of Medicine*, 345, 1368–1377, pMID: 11794169.

Rosenbaum, P. R. (2001), "Effects Attributable To Treatment: Inference In Experiments And Observational Studies With A Discrete Pivot," *Biometrika*, 88, 219–231.

— (2002a), "Attributing Effects to Treatment in Matched Observational Studies," *Journal of the American Statistical Association*, 97, 1–10.

— (2002b), *Observational Studies*, New York, NY: Springer, 2nd ed.

— (2010), *Design of Observational Studies*, New York: Springer-Verlag.

— (2012), "Optimal Matching of an Optimally Chosen Subset in Observational Studies," *Journal of Computational and Graphical Statistics*, 21, 57–71.

Rosenbaum, P. R. and Rubin, D. B. (1983), "The Central Role of Propensity Scores in Observational Studies for Causal Effects," *Biometrika*, 76, 41–55.

Rosenbaum, P. R. and Silber, J. H. (2009), "Sensitivity Analysis for Equivalence and Difference in an Observational Study of Neonatal Intensive Care Units," *Journal of the American Statistical Association*, 104, 501–511.

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 6, 688–701.

— (1986), "Which Ifs Have Causal Answers," *Journal of the American Statistical Association*, 81, 961–962.

— (2008), "For Objective Causal Inference, Design Trumps Analysis," *The Annals of Applied Statistics*, 2, 808–840.

Zubizarreta, J. R. (2012), "Using Mixed Integer Programming for Matching in an Observational Study of Kidney Failure after Surgery," *Journal of the American Statistical Association*, 107, 1360–1371.

Zubizarreta, J. R. and Kilcioglu, C. (2016), "`designmatch`: Construction of Optimally Matched Samples for Randomized Experiments and Observational Studies that are Balanced by Design," `R` package version 0.1.1.

Zubizarreta, J. R., Paredes, R. D., and Rosenbaum, P. R. (2014), "Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile," *The Annals of Applied Statistics*, 8, 204–231.

# Appendices

## A.1 Balance Produced By Randomization in the ProMISe RCT

Table 7: Balance produced by ProMISe trial, select covariates

| | Treated Mean | Control Mean | Std. Diff | p-value |
|---|---|---|---|---|
| Age | 66.42 | 0.14 | 0.01 | |
| Apache Score | 18.70 | 0.10 | 0.08 | |
| SOFA Score | 4.21 | -0.02 | 0.66 | |
| MEDS Score | 7.99 | 0.03 | 0.63 | |
| Blood Lactate | 5.16 | 0.02 | 0.67 | |
| Systolic Blood Pressure (SBP) | 99.84 | 0.10 | 0.09 | |
| Mean Arterial Pressure (MAP) | 66.72 | 0.11 | 0.05 | |
| Male | 0.57 | -0.03 | 0.55 | |
| Liver Severe Cond | 0.02 | 0.00 | 1.00 | |
| Renal Severe Cond | 0.01 | 0.02 | 0.70 | |
| Immuno Severe Cond | 0.13 | 0.07 | 0.22 | |
| Resp Severe Cond | 0.15 | 0.06 | 0.32 | |
| Cardio Severe Cond | 0.04 | 0.05 | 0.41 | |
| Fluids Pre Hosp | 95.87 | -0.06 | 0.25 | |
| Fluids Pre Rand | 1790.50 | -0.04 | 0.48 | |
| Supp Oxygen | 0.52 | -0.01 | 0.93 | |
| Time from ED to Rand | 2.70 | -0.07 | 0.24 | |
| Direct ICU to ED | 0.67 | -0.03 | 0.66 | |
| Blood Products | 5.14 | -0.08 | 0.15 | |
| Blood Lactate Miss | 0.07 | 0.04 | 0.43 | |
| SBP Miss | 0.06 | -0.03 | 0.57 | |
| MAP Miss | 0.83 | 0.07 | 0.20 | |
| Fluid PH Miss | 0.01 | 0.00 | 1.00 | |
| Fluid PR Miss | 0.02 | 0.01 | 0.82 | |
| Oxygen Miss | 0.14 | 0.01 | 0.86 | |

## A.2 Additional Balance Results

Balance results for match between those who received CVC in the usual care arm and those that received CVC in the treatment arm.

Table 8: KS Test Balance produced by ProMISe trial, select covariates

| | KS Test Stat | KS p-value |
|---|---|---|
| Age | 0.08 | 0.03 |
| Apache Score | 0.07 | 0.06 |
| SOFA Score | 0.06 | 0.21 |
| MEDS Score | 0.04 | 0.84 |
| Blood Lactate | 0.04 | 0.81 |
| Systolic Blood Pressure | 0.07 | 0.10 |
| Mean Arterial Pressure | 0.04 | 0.75 |
| Fluids Pre Hosp | 0.03 | 0.98 |
| Fluids Pre Rand | 0.02 | 1.00 |
| Supp Oxygen | 0.04 | 0.85 |
| Blood Products | 0.01 | 1.00 |

Table 9: Mean Balance Comparison Btw Those in Usual Care Arm that got CVC and those in Promise arm that received CVC after matching

| | Treated Mean | Control Mean | Std. Diff | p-value |
|---|---|---|---|---|
| Age | 65.19 | 64.92 | 0.02 | 0.86 |
| Apache Score | 17.36 | 16.73 | 0.10 | 0.31 |
| SOFA Score | 4.01 | 3.96 | 0.02 | 0.80 |
| MEDS Score | 7.88 | 7.84 | 0.01 | 0.91 |
| Blood Lactate | 4.63 | 4.68 | -0.02 | 0.86 |
| Systolic Blood Pressure (SBP) | 95.09 | 97.27 | -0.09 | 0.33 |
| Mean Arterial Pressure (MAP) | 65.71 | 66.20 | -0.07 | 0.42 |
| Male | 0.57 | 0.55 | 0.04 | 0.69 |
| Liver Severe Cond | 0.01 | 0.01 | 0.00 | 1.00 |
| Renal Severe Cond | 0.01 | 0.00 | 0.13 | 0.16 |
| Immuno Severe Cond | 0.11 | 0.13 | -0.06 | 0.55 |
| Resp Severe Cond | 0.11 | 0.13 | -0.06 | 0.54 |
| Cardio Severe Cond | 0.03 | 0.02 | 0.05 | 0.56 |
| Fluids Pre Hosp | 122.67 | 108.06 | 0.06 | 0.58 |
| Fluids Pre Rand | 1794.35 | 1792.67 | 0.00 | 0.99 |
| Supp Oxygen | 0.51 | 0.50 | 0.03 | 0.73 |
| Time from ED to Rand | 2.89 | 2.82 | 0.06 | 0.59 |
| Direct ICU to ED | 0.91 | 0.63 | 0.73 | 0.00 |
| Blood Products | 8.13 | 1.52 | 0.10 | 0.20 |
| Blood Lactate Miss | 0.07 | 0.08 | -0.04 | 0.71 |
| SBP Miss | 0.09 | 0.06 | 0.11 | 0.26 |
| MAP Miss | 0.80 | 0.83 | -0.08 | 0.45 |
| Fluid PH Miss | 0.01 | 0.00 | 0.05 | 0.56 |
| Fluid PR Miss | 0.02 | 0.02 | 0.00 | 1.00 |
| Oxygen Miss | 0.15 | 0.16 | -0.01 | 0.89 |

Table 10: KS Test Balance Comparison Btw Those in Usual Care Arm that got CVC and those in Promise arm that received CVC

| | KS Test Stat | KS p-value |
|---|---|---|
| Apache Score | 0.07 | 0.64 |
| SOFA Score | 0.09 | 0.41 |
| MEDS Score | 0.12 | 0.09 |
| Blood Lactate | 0.05 | 0.97 |
| Systolic Blood Pressure | 0.06 | 0.80 |
| Mean Arterial Pressure | 0.03 | 1.00 |
| Fluids Pre Hosp | 0.04 | 1.00 |
| Fluids Pre Rand | 0.05 | 0.97 |
| Supp Oxygen | 0.07 | 0.64 |
| Blood Products | 0.01 | 1.00 |