From SATE to PATT: Combining Experimental with Observational Studies to Estimate Population Treatment Effects

Erin Hartman

University of California at Berkeley, Berkeley, USA.

Richard Grieve

London School of Hygiene and Tropical Medicine, London, UK.

Roland Ramsahai

Statistical Laboratory, Cambridge, UK.

Jasjeet S. Sekhon

University of California at Berkeley, Berkeley, USA.

Summary. Randomized controlled trials (RCTs) can provide unbiased estimates of sample average treatment effects. However, a common concern is that RCTs may fail to provide unbiased estimates of population average treatment effects. We derive assumptions sufficient for identifying population average treatment effects from RCTs. We advocate relying on stronger identification assumptions than required because the stronger assumptions allow for falsification tests. We offer new research designs for estimating population effects that use non-randomized studies (NRSs) to adjust the RCT data. This approach is considered in a cost-effectiveness analysis of a clinical intervention, Pulmonary Artery Catheterization (PAC).
Keywords: causal inference, external validity, placebo tests, randomized controlled trials, observational studies, cost-effectiveness studies

1. Introduction

Randomized controlled trials (RCTs) can provide unbiased estimates of the relative effectiveness of alternative interventions within the study sample. Much attention has been given to improving the design and analysis of RCTs to maximise internal validity. However, policy-makers require evidence on the relative effectiveness and cost-effectiveness of interventions for target populations that usually differ to those represented by RCT participants (Hoch et al., 2002; Mitra and Indurkhya, 2005; Mojtabai and Zivin, 2003; Nixon and Thompson, 2005; Willan et al., 2004; Willan and Briggs, 2006). A key concern is that estimates from RCTs and meta-analyses may lack external validity (Allcott and Mullainathan, 2012; Deaton, 2009; Heckman and Urzua, 2009; Heckman and Vytlacil, 2005; Hotz et al., 2005; Imbens, 2009). In RCTs, treatment protocols and interventions differ to those administered routinely,

and trial participants—for example individuals, hospitals, or schools—are generally unrepresentative of the target population (Gheorghe et al., 2013). These concerns pervade RCTs across different areas of public policy and are key objections to using RCTs for policy-making (Deaton, 2009). There is also growing interest in using big observational data sources that contain detailed information about the target population of interest (National Research Council, 2013). Our approach combines the benefits of RCTs with those of large observational data sources, and it maintains the advantages of both types of data. We establish sufficient conditions under which RCTs can identify population treatment effects in combination with observational data, and we develop methods to test if these conditions hold in a given application.

Previous research has proposed using non-randomized studies (NRSs) to assess whether RCT-based estimates apply to a target population (Cole and Stuart, 2010; Greenhouse et al., 2008; Kline and Tamer, 2011; Imai et al., 2008; Shadish et al., 2002; Stuart et al., 2011). A common concern is that there may be many baseline covariates, including continuous measures, which differ between the RCT and target population, and modify the treatment effect. In these situations simple post-stratification approaches for reweighting the treatment effects from the RCT to the target population may not fully adjust for observed differences between the settings (Stuart et al., 2011). There may also be unobserved differences between the RCT and target population participants, providers, or settings. And the form of treatment or control may vary. For example, the dose of a drug or the rigor of a protocol may differ between the settings (Cole and Frangakis, 2009). Hence, the RCT may provide biased estimates of the effectiveness and cost-effectiveness of the routine delivery of the treatment in the target population.

Heckman et al. (1998) and Imai et al. (2008) introduced frameworks for decomposing the biases that arise when estimating population treatment effects. Stuart et al. (2011) proposed the use of propensity scores to assess the generalizability of RCTs. We extend this literature by defining the assumptions that are sufficient to identify population treatment effects from RCTs, and providing accompanying placebo tests to assess whether the assumptions hold. These tests can use observational studies to establish when treatment effects for the target population can be inferred from a given RCT. Such tests have challenging requirements: they have to follow directly from the identifying assumptions, be sensitive to key design issues, and have sufficient power to test the assumptions– not just for overall treatment effects, but also for subgroups of prime interest. The formal derivations and the placebo tests allow for a number of research designs for estimating population treatment effects. These research designs can be used with a variety of different estimation techniques, and the best estimation approach for a given problem will depend on the application in question.

We illustrate our approach in an evaluation of the effectiveness and cost-effectiveness of Pulmonary Artery Catheterization (PAC), an invasive and controversial cardiac monitoring device used in critical care. While the evidence

from RCTs and meta-analyses suggests that PAC is not effective or cost-effective (Harvey et al., 2005), concerns have been raised about the external validity of these findings (Sakr et al., 2005). For this empirical application, we employ an automated matching approach, Genetic Matching (GenMatch) (Diamond and Sekhon, 2013; Sekhon and Grieve, 2012), to create matched strata within the RCT. We use maximum entropy (MaxEnt) weighting to reweight the individual RCT strata according to the observed characteristics in the target population.

The paper proceeds as follows. Section 2 introduces the motivating example and the problem to be addressed. Section 3 derives the assumptions required for identifying population average treatment effects. Section 4 describes the placebo tests for checking the underlying assumptions, while section 5 outlines estimation strategies. In Section 6, we illustrate the approach with the PAC case study. Section 7 proposes an alternative design identified by the main theorem, Section 8 discusses related work, and Section 9 concludes.

2. Motivating Example

Pulmonary Artery Catheterization (PAC) is a cardiac monitoring device used in the management of critically ill patients (Dalen, 2001; Finfer and Delaney, 2006). The controversy over whether PAC should be used was fuelled by NRSs that found PAC was associated with increased costs and mortality (Chittock et al., 2004; Connors et al., 1996). These observational studies encouraged RCTs and subsequent meta-analyses, all of which found no statistically significant difference in mortality between the randomized groups (Harvey et al., 2005). The largest of these RCTs was the UK publicly funded PAC-Man Study, which randomized individual patients to either monitoring with a PAC, or no PAC monitoring (no PAC) (Harvey et al., 2005). This RCT had a pragmatic design, with broad inclusion criteria and an unrestrictive treatment protocol, which allowed clinicians to manage patients as they would in routine clinical practice. The study randomized 1,014 subjects recruited from 65 UK hospitals during 2000-2004, and reported that overall, PAC did not have a significant effect on mortality (Harvey et al., 2005), but that there was some heterogeneity in the effect of PAC according to patient subgroup (Harvey et al., 2008). An accompanying CEA used mortality and resource use data directly from the RCT, and reported that PAC was not cost-effective (Stevens et al., 2005). However, despite the pragmatic nature of the RCT, commentators suggested that the patients and centres differed from those where PAC was used in routine clinical practice (Sakr et al., 2005). The major concern was that subgroups for which PAC might be relatively effective (e.g. elective surgical patients), were underrepresented in the RCT, and the unadjusted estimates of effectiveness and cost-effectiveness from the RCT, might not apply to the target population.

To consider the costs and outcomes following PAC use in routine clinical practice, a prospective NRS was undertaken using data from the Intensive Care National

Audit Research Centre (ICNARC) Case Mix Program (CMP) database. The ICNARC CMP database contains information on case-mix, patient outcomes, and resource use for about 1.5 million admissions to 250 critical care units in the United Kingdom (Harrison et al., 2004). A total of 57 units from the CMP collected additional prospective data on PAC use for consecutive admissions between May 2003 and December 2004.[1] The NRS applied the same inclusion and exclusion criteria for individual patients as the corresponding PAC-Man Study, which resulted in a sample of 1,052 PAC cases and 31,447 potential controls. The overall control group is not exchangeable with those who received PAC in practice (Sakr et al., 2005; Sekhon and Grieve, 2012). Hence we only use information from the 1,052 patients who received PAC in routine clinical practice, and from 1,013 RCT participants.

We assume throughout that the patients who received treatment in the NRS represent the target population of interest, as these are the patients who receive PAC in routine clinical practice. Therefore, as is common, the estimand of policy interest is the population average treatment effect on the treated (PATT)—i.e. the average treatment effect of PAC on those individuals in the target population who received it. Information is available on baseline prognostic covariates common to both the RCT and NRS settings, and includes those covariates anticipated to modify the effect of PAC. For a center to participate in the PAC-Man Study required that local clinicians were in equipoise about the potential benefits of the intervention (Harvey et al., 2005), and the patients randomized had to meet the inclusion criteria. The net effect is that the baseline characteristics of the RCT participants differed somewhat from those who received PAC in routine clinical practice (Table 2). The baseline prognosis of the RCT patients was more severe, with a higher mean age, a higher proportion of patients admitted following emergency surgery and a higher proportion having mechanical ventilation. The RCT patients were less likely to be admitted to teaching hospitals than those who received PAC in the target population. For both studies the main outcome measure was hospital mortality, which was higher in the RCT, than for the PAC patients in the NRS. The studies reported similar hospital costs. The effect of PAC on costs and mortality can be incorporated into a measure of cost-effectiveness such as the incremental net

[1] Over this time period, 10 units recorded no PAC use and were excluded from this analysis, as were units participating in the RCT (PAC-Man Study). The RCT data used, excludes one participant for whom no endpoint data were available.

monetary benefit (INB) (Willan et al., 2003; Willan and Lin, 2001).[2]

Table 1. Baseline characteristics and endpoints for the PAC-Man Study, and for patients in the NRS who received PAC. Numbers are N (%) unless stated otherwise

	RCT		NRS
	No PAC n=507	PAC n=506	PAC n=1052
Baseline Covariates			
Admitted for elective surgery	32 (6.3)	32(6.3)	98 (9.3)
Admitted for emergency surgery	136 (26.8)	142 (28.1)	243 (23.1)
Admitted to teaching hospital	108 (21.3)	110 (21.7)	447 (42.5)
Mean (SD) Baseline probability of death	0.55 (0.23)	0.53 (0.24)	0.52 (0.26)
Mean (SD) Age	64.8 (13.0)	64.2 (14.3)	61.9 (15.8)
Female	204 (40.2)	219 (43.3)	410 (39.0)
Mechanical Ventilation	464 (91.5)	450 (88.9)	906 (86.2)
ICU size (beds)			
5 or less	57 (11.2)	59 (11.7)	79 (7.5)
6 to 10	276 (54.4)	272 (53.8)	433 (41.2)
11 to 15	171 (33.7)	171 (33.8)	303 (28.8)
Endpoints			
Deaths in Hospital	333 (65.9)	346 (68.4)	623 (59.3)
Mean Hospital Cost (£)	19,078	18,612	19,577
SD Hospital Cost (£)	28,949	23,751	24,378

This study is an example of where estimates of effectiveness and cost-effectiveness from an RCT may not be directly externally valid for a target population, but there is information from an NRS on the baseline characteristics and outcomes that can inform the estimation of population treatment effects. The next section defines the assumptions required for estimating PATT in this context.

[2] Net monetary benefits can be calculated by weighting each life year using a quality adjustment anchored on a scale from 0 (death) to 1 (perfect health), in order to report quality-adjusted life years (QALYs) for each treatment. Then net monetary benefits for each treatment group can be calculated by multiplying the QALY by an appropriate threshold willingness to pay for a QALY gain (e.g. the threshold recommended by NICE in England and Wales is £20,000 to £30,000 to gain a QALY), and subtracting the cost. Finally, the INB of the new treatment can be estimated by contrasting the mean net monetary benefits for each alternative.

3. Identifying PATT from an RCT

For simplicity we consider those circumstances where data come from a single RCT and a single NRS. It is assumed that the treatment subjects in the NRS represent those in the target population of interest. This section outlines sufficient assumptions for identification of PATT.

A random sample is taken from an infinite population. Let Y_{ist} represent potential outcomes for a unit i assigned to study sample s and treatment t, where $s = 1$ indicates membership of the RCT and $s = 0$ the target population. For simplicity, we assume that in either setting a unit is assigned to treatment ($t = 1$) or control ($t = 0$), and that, as in the motivating example, there is compliance with treatment assignment and no missing outcome data.[3] We define S_i as a sample indicator, taking on value s, and T_i as a treatment indicator taking on value t. For subjects receiving the treatment, we define W_i^T as a set of observable covariates related to the sample selection mechanism for membership in the RCT versus the target population. Similarly $W_i^{C_T}$ is a set of observable covariates related to the sample assignment for inclusion of controls in the RCT, versus the target population.

The sample average treatment effect (SATE) is defined as:

$$\tau_{SATE} = \mathbb{E}(Y_{11} - Y_{10}|S = 1),$$

where the expectation is over the (random) units in $S = 1$ (the RCT sample). Within the RCT, randomization ensures that the difference in the mean outcomes between the treatment versus control units is an unbiased estimate of the SATE.

Other estimands include the average treatment effect on the treated in the sample (SATT), and the average treatment effect on the controls in the sample (SATC):

$$\tau_{SAT*} = \mathbb{E}(Y_{11}|S = 1, T = t) - \mathbb{E}(Y_{10}|S = 1, T = t),$$

where $t = 0$ for τ_{SATC} and $t = 1$ for τ_{SATT}. SATT estimates the average treatment effect conditional on the distribution of potential outcomes under treatment, and SATC estimates the average treatment effect conditional on the distribution of potential outcomes under control. Randomization implies that the potential outcomes in the treatment and control groups are exchangeable ($(Y_{11}, Y_{10}) \perp\!\!\!\perp T|S = 1$), and that the alternative estimands are asymptotically equivalent.[4]

The Population Average Treatment Effect (PATE) is defined as the effect of treatment in the target population, the Population Average Treatment Effect on Controls (PATC) as the treatment effect conditional on the distribution of potential outcomes under control, and the Population Average Treatment Effect on Treated

[3] When there is non-random attrition, causal effects even for the experimental sample cannot be estimated without additional assumptions.

[4] The treatment effects discussed here refer to infinite populations and samples, whereas Imai et al. (2008) refers to treatment effects in infinite populations as super population effects.

(PATT) as the treatment effect conditional on the distribution of potential outcomes under treatment:

$$\begin{aligned}\tau_{PATE} &= \mathbb{E}(Y_{01} - Y_{00}|S = 0)\\ \tau_{PATC} &= \mathbb{E}(Y_{01} - Y_{00}|S = 0, T = 0)\\ \tau_{PATT} &= \mathbb{E}(Y_{01} - Y_{00}|S = 0, T = 1). \end{aligned} \tag{1}$$

Our main quantity of interest is (1). Because treatment in the target population is not randomly assigned, these three population estimands differ even asymptotically, and they may be difficult to estimate without bias.

The following proof outlines the conditions under which population treatment effects can be identified from RCT data. The following assumptions are sufficient to derive the identifiable expression for τ_{PATT} in Theorem 1. Figure 1 represents the assumptions, and demonstrates the result of Theorem 1.

Assumption 1: Consistency under Parallel Studies

$$Y_{i01} = Y_{i11} \tag{2}$$

$$Y_{i00} = Y_{i10} \tag{3}$$

For either the treatment or control group, assumption 1 restricts an individual's potential outcomes for the RCT and the target population. Intuitively, it is assumed that if units in the target population were assigned their observed treatment randomly, then their outcome would be the same as if they were assigned that particular treatment in the RCT. This essentially ensures that any differences in the treatment between the RCT and the NRS, for example, in a clinical protocol, do not affect the outcome. Assumption 1 is similar to the assumption of consistency under the parallel experiment design in Imai et al. (2013). Assumption 1 may be violated if, for example, the clinical protocol for insertion of the PAC differs between the RCT and the NRS. The pragmatic design of the PAC-Man Study helped ensure that this assumption was met. Further examples of violation of the consistency assumption are given in Cole and Frangakis (2009).

Assumption 2: Strong Ignorability of Sample Assignment for Treated

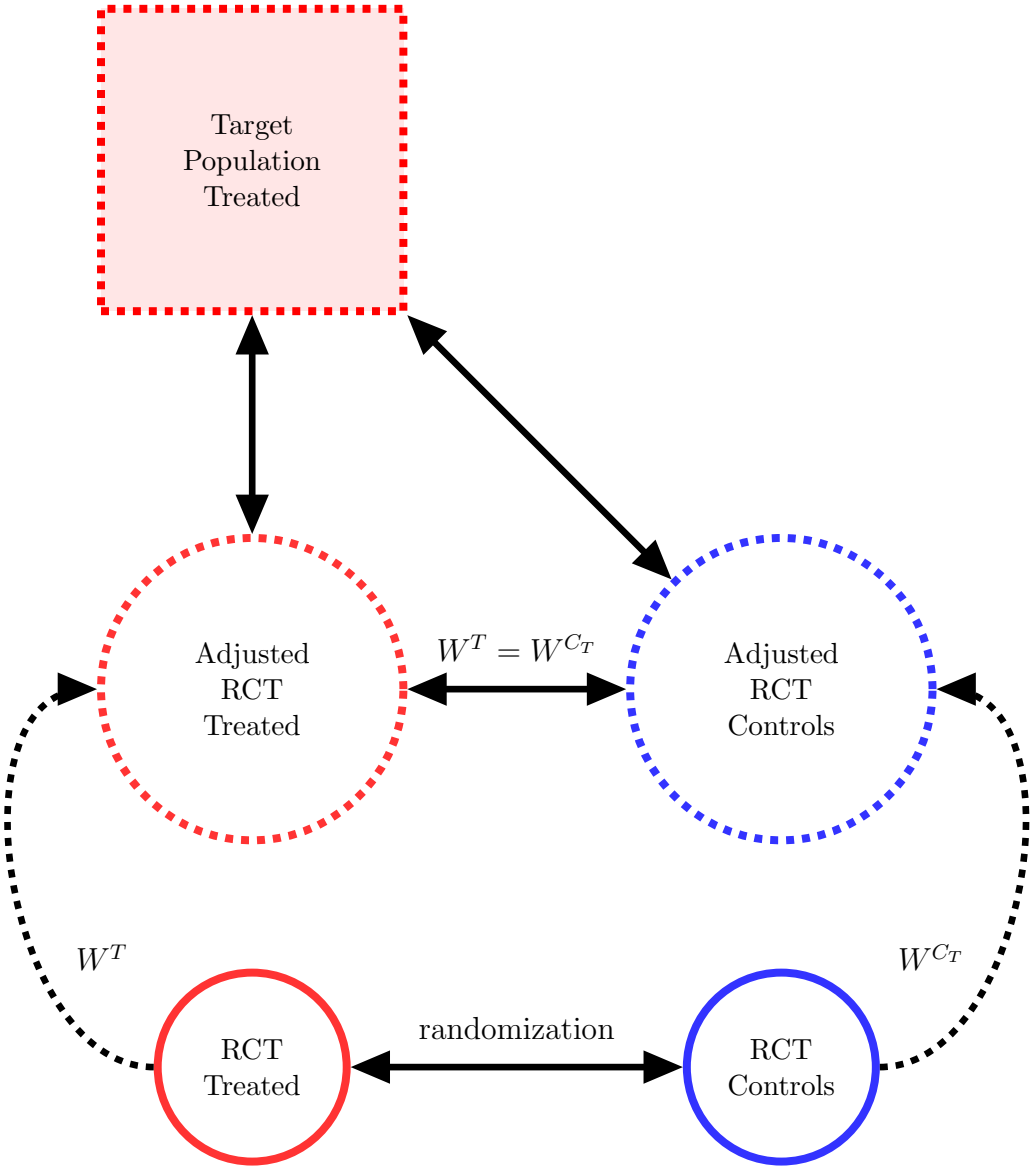
$$(Y_{01}, Y_{11}) \perp\!\!\!\perp S|(W^T, T = 1), \qquad 0 < Pr(S = 1|W^T, T = 1) < 1.$$

Assumption 2 states that the potential outcomes for treatment are independent of sample assignment, for treated units with the same W^T. Assumption 2 implies that

$$\mathbb{E}(Y_{s1}|S = 0, T = 1) = \mathbb{E}_{01}\{\mathbb{E}(Y_{s1}|W^T, S = 1, T = 1)\}, \tag{4}$$

for $s = 0, 1$. The expectation $\mathbb{E}_{01}\{\cdot\}$ is a weighted mean of the W^T specific means, $\mathbb{E}(Y_{s1}|W^T, S = 1, T = 1)$, with weights according to the distribution of W^T in the

Fig. 1. Schematic showing adjustment of sample effect to identify population effect. Double arrows indicate exchangeability of potential outcomes, and dashed arrows indicate adjustment of the covariate distribution.



treated target population, $Pr(W^T|S=0,T=1)$. Essentially, on the right side of Equation (4), the characteristics of the treated units in the RCT, W^T, are adjusted to match those of the treatment group in the target population. Figure 1 illustrates this process with the single arrow from the RCT treated in the solid red circle, to the adjusted group in the dashed red circle. The adjustment can be performed with the weighting methods discussed in Section 5.

The right side of Equation (4) is the expectation in the adjusted RCT treated group, depicted as the dashed red circle in Figure 1. The left side of Equation (4) is the expectation in the treatment group in the target population, depicted as the dashed red square in Figure 1. Thus by Equation (4) the adjusted treatment group in the RCT replicates the Y_{s1} potential outcomes of the treatment group in the target population. In Figure 1, the double arrow between the dashed red circle and square represents the assumed exchangeability of potential outcomes between settings for the treated units.

Assumption 3: Strong Ignorability of Sample Assignment for Controls

$$(Y_{00}, Y_{10}) \perp\!\!\!\perp S|(W^{C_T}, T=1), \qquad 0 < Pr(S=1|W^{C_T}, T=1) < 1.$$

Assumption 3 states that the potential outcomes for control are independent of sample assignment, for treated units with the same W^{C_T}.

Assumption 3 implies that

$$\mathbb{E}(Y_{s0}|S=0,T=1) = \mathbb{E}_{01}\{\mathbb{E}(Y_{s0}|W^{C_T}, S=1, T=0)\}, \tag{5}$$

for $s=0,1$, since treatment assignment is random in the RCT, i.e. $Y_{s0} \perp\!\!\!\perp T|(W^{C_T}, S=1)$. The characteristics of the units in the control group in the RCT, W^{C_T}, are adjusted to match those of the treatment group in the target population. This process is depicted in Figure 1 as the single arrow from the RCT control in the solid blue circle to the adjusted group in the dashed blue circle.

The right side of Equation (5) is the expectation in the adjusted RCT control group, which is depicted as the dashed blue circle in Figure 1. The left side of Equation (5) is the expectation in the treated group in the target population, which is depicted as the dashed red square in Figure 1. Thus it follows by Equation (5) that the adjusted control group in the RCT replicates the expected Y_{s0} potential outcomes of the treated group in the target population.

Assumption 4: Stable Unit Treatment Value Assumption (SUTVA)

$$Y_{ist}^{L_i} = Y_{ist}^{L_j} \qquad \forall i \neq j,$$

where L_j is the treatment and sample assignment vector for unit j. This is a stable unit treatment value assumption (SUTVA), which states that the potential outcomes of unit i are constant regardless of the treatment or sample assignment of any other unit.

Theorem 1 follows from Assumptions 1-4, with the proof given in Appendix A.

Theorem 1. *Assuming consistency and SUTVA hold, if*

$$\begin{aligned} &\mathbb{E}_{01}\{\mathbb{E}(Y_{s1}|W^T, S=0, T=1)\} - \mathbb{E}_{01}\{\mathbb{E}(Y_{s0}|W^{C_T}, S=0, T=1)\} \\ &= \mathbb{E}_{01}\{\mathbb{E}(Y_{s1}|W^T, S=1, T=1)\} - \mathbb{E}_{01}\{\mathbb{E}(Y_{s0}|W^{C_T}, S=1, T=1)\}, \end{aligned} \tag{6}$$

or sample assignment for treated units is strongly ignorable given W^T, and sample assignment for controls is strongly ignorable given W^{C_T}, then

$$\tau_{PATT} = \mathbb{E}_{01}\{\mathbb{E}(Y|W^T, S=1, T=1)\} - \mathbb{E}_{01}\{\mathbb{E}(Y|W^{C_T}, S=1, T=0)\},$$

where $\mathbb{E}_{01}\{\mathbb{E}(\cdot|W^T, \ldots)\}$ denotes $\mathbb{E}_{W^T|S=0,T=1}\{\mathbb{E}(\cdot|W^T, \ldots)\}$

and $\mathbb{E}_{01}\{\mathbb{E}(\cdot|W^{C_T}, \ldots)\}$ denotes $\mathbb{E}_{W^{C_T}|S=0,T=1}\{\mathbb{E}(\cdot|W^{C_T}, \ldots)\}$.

From Theorem 1, it is possible to identify τ_{PATT} from the adjusted RCT data alone. In Figure 1, the adjusted experimental controls and treated units are only exchangeable if $W_i^T = W_i^{C_T}$. As Figure 1 makes plain, in identifying τ_{PATT}, the adjusted RCT controls are being used in place of the subset of population controls who have the same distribution of observable characteristics as the treated units in the target population. The adjusted RCT controls are not a substitute for all population controls, since the controls and treated in the target population are not assumed to be exchangeable.

As the randomized arms within the RCT are exchangeable, adjusting both groups by the same observable characteristics will yield (asymptotically) exchangeable groups. This implies that if $W_i^T = W_i^{C_T}$, then the adjusted RCT treated and controls are exchangeable with each other, and they can replace their counterparts in the target population. To gain precision, matching or stratifying between the treated and control units within the RCT can be undertaken, before adjustment to the target population (Miratrix et al., 2013). Hence τ_{PATT} can be estimated by reporting the treatment effect for each matched pair from the RCT, and then adjusting these unit level treatment effects to the characteristics of those treated in the target population. The corresponding estimate of the SATT is given by the average of the unadjusted unit level effects from the RCT.

4. Placebo Tests for Checking Assumptions

Placebo tests are generally used to assess the plausibility of a model or identification strategy when the treatment effect is known, from theory or design (Sekhon, 2009).

This section describes placebo tests for checking the identifiability assumptions of Theorem 1, regardless of the estimation strategy subsequently chosen. From Section 3, if Equation (2) in assumption 1, assumptions 2, and 4 all hold, then the Y_{s1} potential outcomes of the adjusted RCT treated group, and the target population are exchangeable, i.e. Equation (11) holds. Since the potential outcomes Y_{01} are observed in the treated group of the target population, then $\mathbb{E}(Y_{01}|S = 0, T = 1)$ is equal to $\mathbb{E}(Y|S = 0, T = 1)$ and

$$\mathbb{E}(Y|S = 0, T = 1) - \mathbb{E}_{01}\{\mathbb{E}(Y|W^T, S = 1, T = 1)\} = 0, \tag{7}$$

from Equation (11) in Appendix A. Hence, if these assumptions hold, then the expected outcomes will be the same for the treatment group in the RCT after adjustment and the target population. A placebo test can be used to check whether the average outcomes differ between the adjusted RCT treatment group and the treatment group in the target population. If the placebo test detects a significant difference in these outcomes, then either Equation (2) in assumption 1, assumption 2 or assumption 4 is violated.[5] If Equation (3) in assumption 1, and assumptions 3 and 4 hold, then the Y_{s0} potential outcomes of the adjusted RCT treated group and the target population are exchangeable, i.e. Equation (12) in Appendix A holds. However, since Y_{00} is not observed in those treated in the target population, then $\mathbb{E}(Y_{00}|S = 0, T = 1)$ is not necessarily equal to $\mathbb{E}(Y|S = 0, T = 0)$. Therefore the mean outcome in the adjusted RCT control group is not necessarily the same as the mean outcome in the target treated population. This implies that a placebo test cannot be used to check whether Equation (3) in assumption 1, assumption 3 or assumption 4 fails.

Placebo can be used to highlight the failure of several underlying assumptions, but they cannot delineate the bias from the failure of each individual assumption. Also, the tests cannot exclude the possibility that each assumption is violated but the ensuing biases cancel one another out. Traditional, placebo tests have a null hypothesis, that there is no difference in the average outcome between groups, and the null hypothesis is rejected if the test statistic is significant. If the null hypothesis is not rejected then a standard conclusion is that there is evidence to support the identification strategy. However, the failure to reject the null hypothesis may be because of insufficient power to detect a true difference between the groups, particularly if treatment effects by subgroup are of interest, or if there are endpoints, such as cost, that have a high variance. CEA typically have both these features.

To address this concern, Hartman and Hidalgo (2011) introduce equivalence based placebo tests, with the null hypothesis that "the data are *inconsistent* with a valid research design." In this context, the null hypothesis can be stated as: the adjusted endpoints for the treatment group in the RCT, are not equivalent to

[5] If assumption 1 is violated and there is a constant difference between the potential outcomes in the target population and the RCT, then the PATT can still be identified by Theorem 1. See Section 7.1.

those for the treatment group in the target population. This null hypothesis of non-equivalence is only rejected if there is sufficient power.[6] Hence, a low p-value would offer support for the identification strategy. The advantage of the proposed test is that it only supports the identification strategy when the test reports that the two groups are equivalent, *and* when the test has sufficient power. Specifying an alternative null hypothesis has implications for the test statistic and, just as in a sample size calculation, requires that the threshold for a meaningful difference in outcomes is pre-defined. Appendix B and Hartman and Hidalgo (2011) give further details.

5. Estimating PATT

Estimation strategies for predicting population-level treatment effects from RCT data fall into two broad classes. One class of strategies use weighting methods, such as Inverse Propensity Score Weighting (IPSW) (Stuart et al., 2011) and Maximum Entropy (MaxEnt) weighting (Kullback, 1997; Jaynes, 1957), which rely on ancillary information, for example from a NRS, to reweight the RCT data. The other prominent approach is to estimate the response surface using the RCT data and extrapolate this response surface to the target population, and includes methods such as Bayesian Additive Regression Trees (BART) (Chipman et al., 2010), Classification and Regression Trees (CART) (Breiman, 2001; Liaw and Wiener, 2002; Stuart et al., 2011), and linear regression. The result in Theorem 1 is agnostic to the estimation strategy; the adjustment of the RCT data by W^T and W_T^C can either use weights from the first class of estimators, or predicted values from a response surface model. Either way, in order to identify the population estimand of interest, the estimation strategy must pass the proposed placebo tests.

While Theorem 1 does not require a specific estimation strategy, we do provide a new research design that employs a weighting method. Our proposed strategy firstly matches treated and control units within the RCT to create matched pairs or strata (Diamond and Sekhon, 2013; Sekhon, 2011), from which we estimate the SATT overall and by pre-specified subgroup. We then reweight the matched pairs according to the characteristics of the target population to report PATT, both overall and for subgroups.

5.1. Matching Treated and Control Units within the RCT

We create matched pairs within the RCT data, by matching controls to treated units within the RCT using Genetic Matching (GenMatch) to maximise the balance between the randomized groups (Diamond and Sekhon, 2013; Sekhon, 2011). We recommend including in the matching algorithm, those covariates anticipated to

[6] This alleviates the issues of confounding the notion of statistical equivalence with a tests relationship to sample size discussed in Imai et al. (2008).

influence not only the endpoints, but also the selection of patients into the RCT. Covariates related to the selection into the RCT are part of the conditioning sets W^T and W_T^C, and therefore care should be taken to ensure these covariates are balanced.

5.2. Weighting Methods

We focus on an reweighting approach, MaxEnt, that can be applied when either summary or individual data are available for the target population.[7] MaxEnt, goes back to at least Jaynes (1957); Kullback (1997); Ireland and Kullback (1968), and has much in common with method of moments estimators (e.g. Hansen, 1982; Hellerstein and Imbens, 1999). In brief, this approach does not assume the propensity score is correctly specified, nor does it make additional assumptions about the distribution of weights. Under MaxEnt the cell weights, marginal distributions, or other population moments for the conditioning covariates, W^T, are used as constraints. MaxEnt ensures that the weights chosen for the matched pairs sum to one, but simultaneously satisfy the MaxEnt constraints given by the population characteristics. See Appendix D for more details on MaxEnt.

6. Empirical Example: PAC

We illustrate our new strategy for extrapolating from an RCT to a target population using the PAC example. Here, we estimate PATT overall and for pre-specified subgroups; patients' surgical status (elective, emergency, non-surgical) and type of admission hospital (teaching or not).

6.1. Matching and Weighting in the PAC Example

We used GenMatch to create matched pairs within the RCT data, by matching a control unit to each treated unit. The matching algorithm included those covariates anticipated to influence the selection of patients into the RCT and the endpoints (See Appendix E Table E.1). The GenMatch loss function was specified to require that balance, according to t-tests and KS tests, was not made worse on those covariates anticipated to be of high prognostic importance after matching. GenMatch matched 1-1 with replacement using a population size of 5,000. Matching was repeated within each subgroup to report SATT at the subgroup level.[8] Variance estimates were calculated conditional on the matched data (Imai et al., 2008). The matching identified a control for each treated observation, resulting in 507 matched pairs for the overall estimate. Each baseline covariate

[7] IPSW is considered in Appendix H.

[8] The aggregated subgroup estimates may not be equivalent to the overall estimate because different matches are used for the overall and subgroup estimates.

was well balanced after matching according to both t-tests and KS tests, as shown in Figure F.1, Appendix F.

The SATT results, both overall and at subgroup level, were similar to the SATE estimates from the RCT.

We use MaxEnt weighting to adjust the distribution of observable baseline covariates in the matched RCT data to the distribution of the PAC patients in the NRS. We constructed the weights for the covariates and interactions listed in Appendix F (Table E.2). For each covariate used to construct the weights, the mean for the PAC patients after reweighting was balanced with the observed means for the PAC patients in the NRS. The t-tests for difference in means all have a p-value of 1.

We then apply these weights to adjust the individual matched pairs from the RCT according to the observed characteristics of the PAC patients in the NRS. To recognize the uncertainty in the estimation of the weights, standard errors for both SATT and PATT were estimated using subsampling (Politis and Romano, 1994). Abadie and Imbens (2008) show that the bootstrap is not valid for estimating the standard error of a matching estimator, but note that subsampling (Politis and Romano, 1994) is valid. We used the algorithm described in Bickel and Sakov (2008) to select the subsampling size, m, and found that the optimal subsample was the sample size, n, in the RCT. We used 1000 bootstrap replicates.[9]

MaxEnt provided weights that were reasonably stable (see Appendix G.1), with a mean weight of 1 and a maxima of 8; no individual stratum was given an extreme weight.

6.2. Results of the Placebo Tests

We now report placebo tests that test the underlying assumptions for identifying PATT by comparing the mean endpoints for the PAC patients in the NRS with the adjusted means for the PAC patients in the RCT. The results are reported in Figure 6.2, for all three endpoints: survival rates (black), cost (red), and INB (blue). We present the equivalence based placebo test p-values for the overall estimate, and each subgroup, and allow for multiple comparisons, by presenting p-values with a false discovery rate correction (FDR) using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). Following MaxEnt, for the overall stratum all the placebo tests are passed; mean differences between the settings are small, and there is sufficient power to assess whether such differences are statistically significant.

[9] One of the arguments against using the bootstrap for matching estimators is that individual matches can be no better than in the full sample, and typically are worse. However, in the RCT, where the true propensity of each individual to be assigned to treatment is constant, there are many potential matches for each unit. Therefore, in each bootstrap sample, the probability of a close match for each unit is high. Therefore, it may not be surprising then that the Bickel and Sakov (2008) algorithm selects $m = n$.

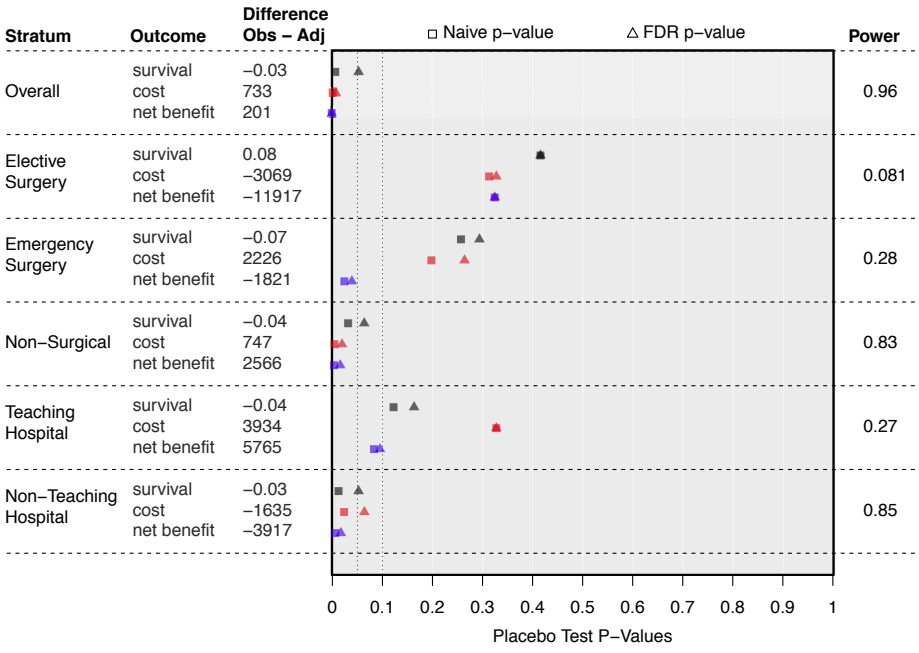


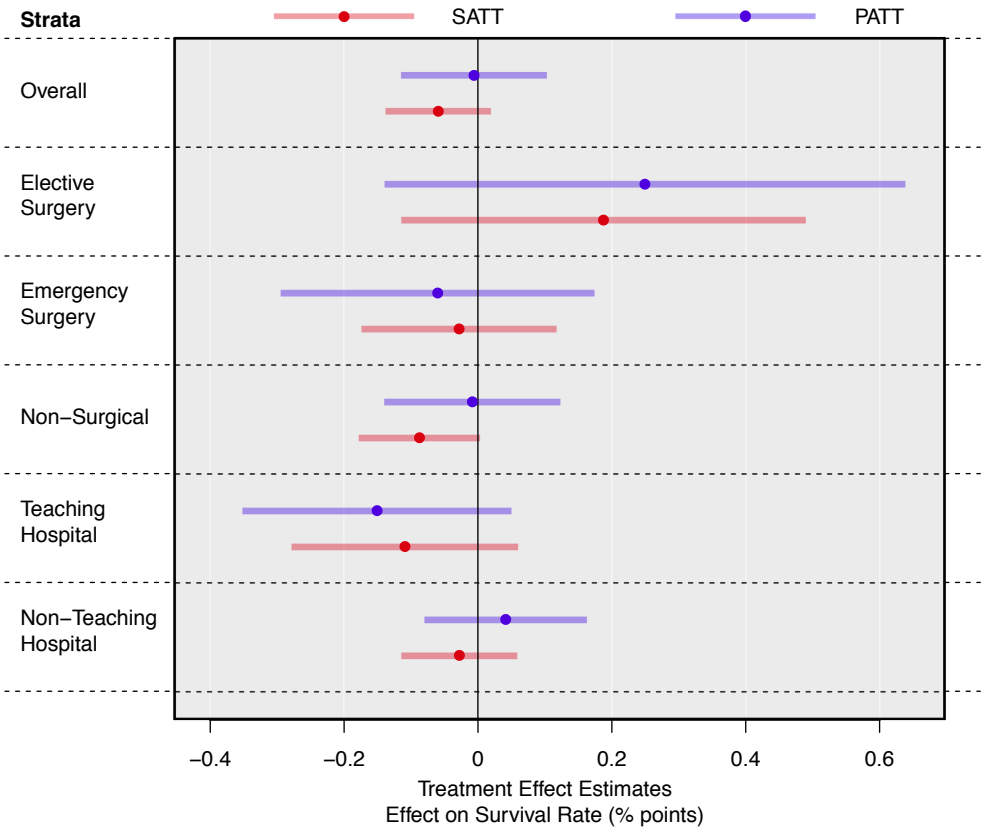
Fig. 2. MaxEnt Placebo Tests

Results of the equivalence placebo tests comparing the mean outcome of NRS treated to the reweighed mean of the RCT treated. The column labelled "Difference" presents the difference between the observed outcomes for the PAC group in the NRS and the PAC group in the RCT after reweighting. The p-values presented are before (squares) and after (triangles) FDR adjustment. The column labelled "Power" presents the power of the equivalence t-test for each stratum.

For some subgroups (teaching hospitals, elective and emergency surgery) there is insufficient power to detect differences between the settings, and the placebo test fails; for other subgroups (non-surgical and non-teaching hospital), the mean differences are small after reweighting, and as there is also sufficient power, the placebo tests are passed.

Details on applying IPSW to reweight RCT data to the target population and for the PAC example are given in Appendix H.

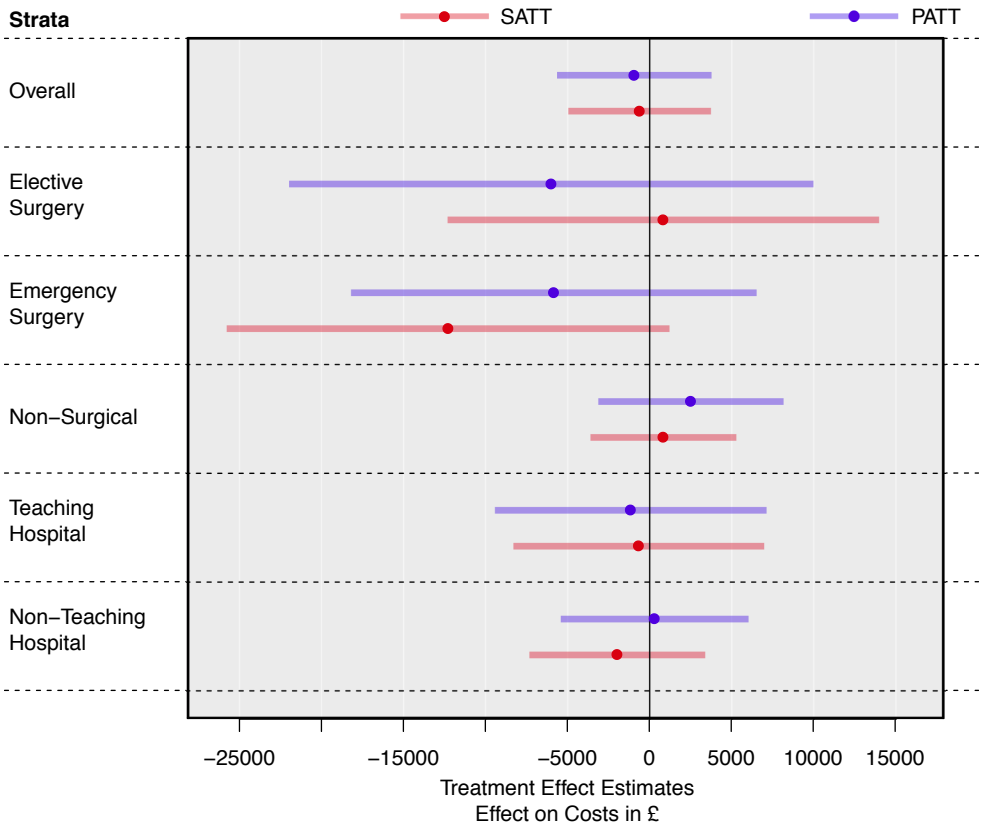
Fig. 3. Population Treatment Effects on Hospital Survival Rates



6.3. Population Estimates in the PAC Example

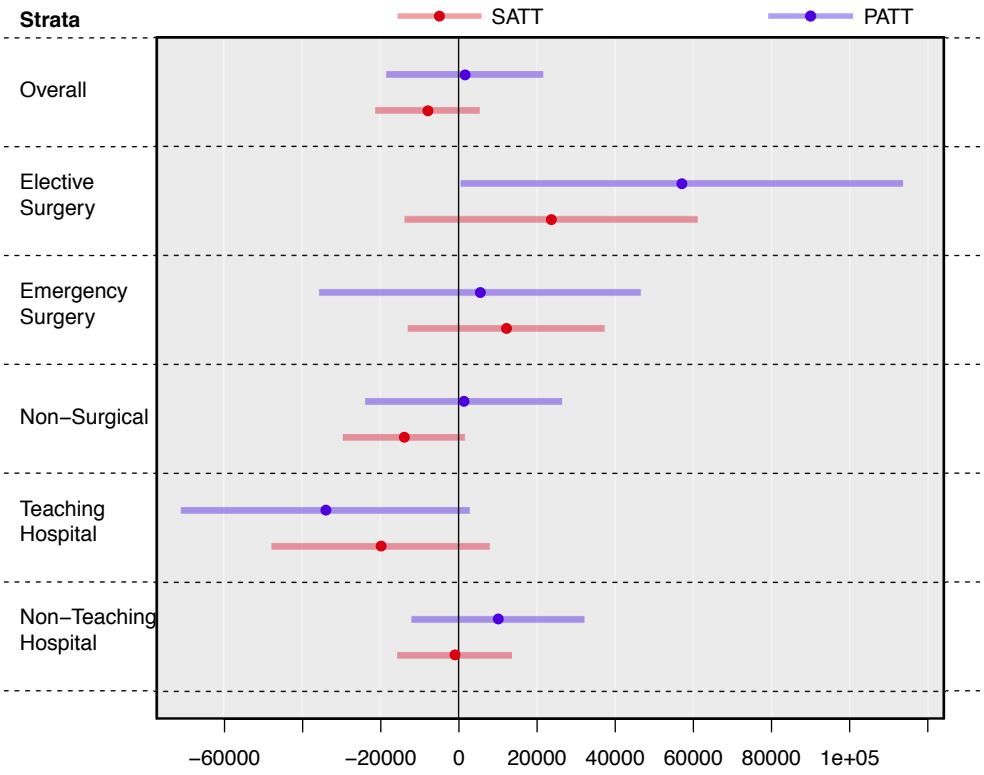
We report SATT estimated from the matched RCT data, and PATT after using the MaxEnt weights to adjust the SATT estimates. The 95% confidence intervals

Fig. 4. Population Treatment Effects on Costs (GBP £)



(CIs) are obtaining using subsampling (Figures 3–5). For the overall group, the PATT and SATT estimates are similar for each endpoint. For the non-teaching hospital subgroup, which passed the placebo tests, the positive point estimate for PATT suggested a somewhat more beneficial effect for PAC on survival, than the corresponding SATT. The accompanying cost-effectiveness estimates were a negative INB for the SATT, but for the PATT, the estimated INB was positive. This finding suggests that for non-teaching hospitals in the target population, PAC was relatively cost-effective. However, the CIs for each estimate overlapped zero, and in general, the CIs for the PATT estimates were wider than those for the corresponding SATT estimates.

Fig. 5. Population Treatment Effects on Incremental Net Monetary Benefits



Incremental Net Monetary Benefits calculated by valuing each quality-adjusted life year (QALY) gain at a £20,000 per QALY threshold

6.4. Response Surface Models

An alternative estimation strategy is to use response surface models to estimate covariate-endpoint relationships in the RCT, and use these estimates to predict population treatment effects in the target population. For example, in the case of OLS regression, the response surface can be estimated from the RCT data, the βs held fixed, and the population treatment effects predicted from the covariate distribution of the NRS treated. This approach may achieve efficiency gains relative to weighting approaches, especially if not all the covariates included in the adjustment are predictive of potential outcomes.

The proposed placebo tests can be used following the response surface approach, by comparing the average outcomes predicted by the model with the average of the

observed outcomes for the treated group. Again given sufficient power, a failure to find equivalence between the predicted and observed outcomes indicates a failure of at least one assumption underlying Theorem 1 and bias in the estimated population treatment effects.

We implement response surface modeling with a statistical machine learning algorithm for classification that uses a non-parametric Bayesian regression approach–Bayesian Additive Regression Trees (BART) (Chipman et al., 2010). BART is a "sum-of-trees" model where each tree is constrained by a regularization prior to be a weak learner. It is a nonparametric method that uses dimensionally adaptive random basis elements. The flexibility of BART confers potential advantages in that it does not require the analyst to specify particular parametric relationships between the covariates, the sample assignment, or the endpoints, and it can incorporate a large number of predictors.

We apply BART in the PAC example, by estimating a response surface model on the patients randomized to receive PAC. We estimate a model for the relationship between the baseline characteristics and each endpoint (mortality, cost and net monetary benefit). We then predict the outcomes the target population would have had, if they had been included in the RCT. We predict these outcomes by combining the coefficients from the response models, with the baseline characteristics of each of the PAC patients in the NRS. The equivalence based placebo tests were then applied by contrasting the means of the predicted versus the observed outcomes.

As can be seen in Figure I.5 in Appendix I, this response surface modeling approach provided estimates that did not pass the requisite placebo tests for estimating the overall treatment effects, and so in this example, this approach was not applied to the estimation of population treatment effects.

7. Alternative Designs Identified under Theorem 1

7.1. Using the Population Treated

A main assumption in the derivation of Theorem 1 is that selection on observables assumptions are sufficient to recognize the selection of the RCT participants. However, if a placebo test rejects the null hypothesis given by Equation (7) then Equation (2) in assumption 1, assumption 2 or assumption 4 is violated. In such a case the results of Theorem 1 are no longer valid. However, if assumption 4 is not violated and if assumption 3 and Equation (3) in assumption 1 are valid, PATT can still be identified by

$$\tau_{PATT} = \mathbb{E}(Y|S=0,T=1) - \mathbb{E}_{01}\{\mathbb{E}(Y|W^{C_T},S=1,T=0)\}, \tag{8}$$

from (12) in Appendix A. This estimator makes direct use of the population treated, and it is valid if there is a constant difference in the potential outcomes between

the population and the RCT. One can see this by rewriting (8) as:

$$\tau_{PATT_{DID}} = \mathbb{E}_{01}\{\mathbb{E}(Y|W^T, S=1, T=1) - \mathbb{E}(Y|W^T, S=1, T=0)\} \quad (9)$$
$$-[\mathbb{E}_{01}\{\mathbb{E}(Y|W^T, S=1, T=1)\} - \mathbb{E}(Y|S=0, T=1)], \quad (10)$$

assuming $W^T = W^{C_T}$. The first difference (9) is the adjusted experimental estimand and is intuitively a measure of the adjusted average effect. The second difference (10) is defined as the difference between the outcomes of the treatment groups in the RCT and the NRS.

The major concern with this estimator is that there is no longer a placebo test available to check if the identifying assumptions hold. Hence, while the main approach proposed makes a somewhat stronger identifying assumption, a key advantage is that this design allows the implications of the assumptions to be tested.

8. Other Related Literature

Heckman and Vytlacil (2005) show that all of the estimands we consider (e.g., PATE, PATT, PATC) are weighted averages of Marginal Treatment Effects (MTEs). The MTE is the treatment effect for a fixed value of the observed covariates for units who are equally indifferent between treatment and control. The indifference is conceptualized as an unobserved random variable that measures utility. Heckman et al. (2006) show that if, conditional on observed covariates, selection into treatment is a function of the gain from treatment (i.e., there is *essential heterogeneity*), the usual estimators do not in general estimate a policy relevant estimand. This is why the MTE conditions on unobserved utility. In our case, essential heterogeneity cannot occur in the RCT because there is full compliance, but the issue can arrise between selection into the RCT versus the NRS. If there is essential heterogeneity in that selection process, it would violate Assumptions 2 and 3, and the placebo tests we offer would be sensitive to this problem if it were present. In short, our approach is for the case where the MTE is just a function of observed covariates, and we offer specification tests to help assess if this indeed is the case.

Hotz et al. (2005) examine how the efficacy of worker training programs differ from one location to another. They offer a formalization that is similar to ours, but there are key differences because of the setup they examine: they formalize the comparison of two randomized trials undertaken in different locations. The setup allows the treatments to differ between the two locations, but by design, the control conditions are assumed the same. To evaluate whether there is unconfoundedness across locations, they conduct placebo tests that contrast outcomes for controls across settings. They then conduct a placebo test contrasting endpoints for treatment versus treatment to assess whether treatment was homogenous (conditional on passing the control-control placebo). This interpretation of the placebos and the setup pertains to the setting with RCTs

undertaken in different locations. Our theorem, placebo tests, and estimators differ because we have a RCT and observational data on the target population.

Allcott (2014, 2011), and Allcott and Mullainathan (2012) find that treatment effects vary greatly across the different experimental locations they consider, and that this variance cannot be explained by observed variables. Therefore, the external validity of the experimental estimate from one location to another is limited. Allcott (2011) show that in their setting non-experimental estimates have poor external validity, and that is worse than when non-experimental effects are predicted using experimental results from another location. This is consistent with our setup where we reweight the experimental estimand, and we do not resort to the observational estimator which is available to us.

While the main population estimand that we consider in this paper is PATT, policy-makers may also be interested in PATC or PATE. In Appendix C we outline identification strategies for these alternative population estimands of interest, and we link to previous work by Stuart et al. (2011) for estimating PATE.

9. Discussion

This paper derives conditions under which treatment effects can be identified from RCTs for the target population of policy relevance. We provide placebo tests, which follow directly from the conceptual framework, that can assess whether the requisite assumptions are satisfied. These placebo tests contrast the reweighted RCT endpoints with those of the target population provided, for example, by a NRS. The general framework is illustrated with estimation strategies that reweight the matched RCT data, but we could also exploit alternative estimation strategies such as double-robust estimators. Whichever estimation strategy is taken, the placebo tests presented can assess whether or not the assumptions required for identification are met. The paper builds on previous approaches for considering external validity (Heckman and Vytlacil, 2005; Hotz et al., 2005; Imai et al., 2008; Stuart et al., 2011), by defining the assumptions required for estimating population treatment effects, and providing a general strategy for assessing their plausibility.

We illustrate the framework for estimating population treatment effects in a context where the treatment, in this case a medical device, has been defused to the target population without adequate evaluation, and the parameter of interest is the PATT. The framework can be applied to other situations: for example, in evaluations of new pharmaceuticals, where the only individuals who receive the treatment are those included in the phase III RCT. Then, the target population is defined by those who would meet the criteria for treatment in routine practice but receive usual care, and the estimand of interest is the PATC. In these settings, the proposed framework can assess the identification strategies with placebo tests that compare the weighted outcomes from the RCT control group versus those receiving usual care in the target population (Stuart et al., 2011). Failure of these

placebo tests would indicate that either participants' unobserved characteristics, or "usual care," differs between the RCT and target population settings. Hence the underlying assumptions are violated leading to biased estimates of the effectiveness and cost-effectiveness of treatment in the target population.

Our framework complements the move to RCTs with pragmatic designs which require that the participants and treatments included represent those in the target population (Tunis et al., 2003). As the case study illustrates, pragmatic RCTs can help ensure that the treatments delivered in RCTs are similar to routine practice, and that there is reasonable overlap in baseline characteristics between the settings. The PAC-Man RCT had broad inclusion criteria, many prognostic baseline covariates common to the RCT and NRS settings, good overlap in the distribution of the baseline covariates between the settings, and the RCT used the same treatment and usual care protocols as for routine practice. These design features were an important reason why the placebo test findings following MaxEnt reweighting, supported the underlying assumptions required for estimating the PATT, overall and for some subgroups. For those subgroups, for example teaching hospitals where the placebo test results showed the underlying assumptions were violated, this may reflect unobserved differences between the RCT and target population. RCTs generally apply restrictive exclusion criteria, or treat according to more rigid treatment protocols than would be applied in routine practice (Rothwell, 2005). Such study designs mean that assumptions pertaining to both the consistency of treatment, and strong ignorability will be violated; the placebo test would indicate the likely bias in the estimates of the population treatment effects.

The proposed approach encourages future studies to fully recognize the uncertainty in estimating population treatment effects, which comprises not just the random error in the sample estimates, the systematic differences between the RCT and the target population (Greenland, 2005), but also the uncertainty in estimating the requisite weights. It is anticipated that when the treatment effects are estimated for the population rather than the sample, there will be increased uncertainty. Future studies should anticipate the additional uncertainty at the design stage when developing the sampling strategy.

The paper motivates the following areas for further investigation. First, research is required to consider the proposed framework in evidence synthesis and meta-analyses of individual participant data from several RCTs. Here, rather than weighting the data from each setting according to their relative sample size or variance, weights should partly reflect each study's relative relevance, according for example to elicited opinion (Turner et al., 2009). Our approach can be extended to recognize systematic differences in the populations and the treatments in each study versus those in the target population. Second, we illustrate an approach for reweighting evidence from head-to-head RCTs, but the framework extends to settings which require comparisons across several interventions where there is a common comparator, as typically happens in network meta-analyses. In this setting,

the placebo tests can assess whether the underlying assumptions for estimating population treatment effects are met, by contrasting the reweighted endpoints for the common comparator (e.g. usual care) from each RCT with those of the target population. Lastly, the framework presented is for settings where there is full compliance with the treatment. For settings with non-compliance, further research is required to define and test the assumptions required to identify the complier-average causal effect for the target population.

Acknowledgments

We gratefully acknowledge the Associate Editor, two anonymous reviewers, Sheila Harvey (LSHTM), David Harrison (ICNARC) and Kathy Rowan (ICNARC) for access to data from the PAC-Man CEA and the ICNARC CMP database; and Chris McCabe and Katherine Stevens (SCHARR) for access to the cost data. We thank Daniel Hidalgo, Adrienne Hosek, Adam Glynn, Holger Kern, Dan Polsky, and the Statistics Department at the University of Pennsylvania for comments. The authors are responsible for all errors.

References

Abadie, A. and G. W. Imbens (2008). On the failure of the bootstrap for matching estimators. *Econometrica 76*(6), 1537–1557.

Allcott, H. (2011). Social norms and energy conservation. *Journal of Public Economics 95*(9), 1082–1095.

Allcott, H. (2014). Site selection bias in program evaluation. Working Paper.

Allcott, H. and S. Mullainathan (2012). External validity and partner selection bias. Technical report, National Bureau of Economic Research.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B 57*(1), 289–300.

Bickel, P. J. and A. Sakov (2008). On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica 18*, 967–985.

Breiman, L. (2001). Random forests. *Machine Learning 45*(1), 5–32.

Chipman, H. A., E. I. George, R. E. McCulloch, et al. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics 4*(1), 266–298.

Chittock, D. R., V. K. Dhingra, J. J. Ronco, J. A. Russell, D. M. Forrest, M. Tweeddale, and J. C. Fenwick (2004). Severity of illness and risk of death

associated with pulmonary artery catheter use. *Critical Care Medicine 32*, 911–915.

Cole, S. R. and C. E. Frangakis (2009). The consistency statement in causal inference: a definition or an assumption? *Epidemiology 20*(1), 3–5.

Cole, S. R. and E. A. Stuart (2010). Generalizing evidence from randomized clinical trials to target populations the actg 320 trial. *American journal of epidemiology 172*(1), 107–115.

Connors, A. F., T. S. Speroff, N. V. Dawson, C. Thomas, F. E. Harrell, D. Wagner, N. Desbiens, L. Goldman, A. W. Wu, R. M. Califf, W. J. Fulkerson, H. Vidaillet, S. Broste, P. Bellamy, J. Lynn, and W. A. Knaus (1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. *Journal of the American Medical Association 276*, 889–897.

Dalen, J. E. (2001). The pulmonary artery catheter—friend, foe, or accomplice? *Journal of the American Medical Association 286*, 348–350.

Deaton, A. (2009). Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development. NBER Working Paper 14690.

Diamond, A. and J. S. Sekhon (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics 95*(3), 932–945.

Finfer, S. and A. Delaney (2006). Pulmonary artery catheters as currently used, do not benefit patients. *British Medical Journal 333*, 930–1.

Gheorghe, A., T. E. Roberts, J. C. Ives, B. R. Fletcher, and M. Calvert (2013, February). Centre selection for clinical trials and the generalisability of results: A mixed methods study. *PLOS ONE 8*(2).

Greenhouse, J. B., E. E. Kaizar, K. Kelleher, H. Seltman, and W. Gardner (2008). Generalizing from clinical trial data: a case study. the risk of suicidality among pediatric antidepressant users. *Statistics in medicine 27*(11), 1801–1813.

Greenland, S. (2005). Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 168*(2), 267–306.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica 50*(4), pp. 1029–1054.

Harrison, D. A., A. R. Brady, and K. Rowan (2004). Case mix, outcome and length of stay for admissions to adult, general critical care units in england, wales and

northern ireland: the intensive care national audit & research centre case mix programme database. *Critical Care 8*, R99–111.

Hartman, E. K. and F. D. Hidalgo (2011, March). What's the alternative?: An equivalence approach to placebo and balance tests. Working Paper.

Harvey, S., D. A. Harrison, M. Singer, J. Ashcroft, C. M. Jones, D. Elbourne, W. Brampton, D. Williams, D. Young, and K. Rowan (2005). An assessment of the clinical effectiveness of pulmonary artery catherters in patient management in intensive care (pac-man): a randomized controlled trial. *Lancet 366*, 472–77.

Harvey, S., C. Welch, D. Harrison, and M. Singer (2008). Post hoc insights from pac-man—the uk pulmonary artery catheter trial. *Critical Care 35*(6), 1714–1721.

Heckman, J. J., H. Ichimura, J. Smith, and P. Todd (1998). Characterizing selection bias using experimental data. *Econometrica 66*(5), 1017–1098.

Heckman, J. J. and S. Urzua (2009, January). Comparing iv with structural models: What simple iv can and cannot identify. IZA Discussion Papers 3980, Institute for the Study of Labor (IZA).

Heckman, J. J., S. Urzua, and E. Vytlacil (2006). Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics 88*(3), 389–432.

Heckman, J. J. and E. Vytlacil (2005, May). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica 73*(3), 669–738.

Hellerstein, J. K. and G. W. Imbens (1999). Imposing moment restrictions from auxiliary data by weighting. *Review of Economics and Statistics 81*(1), 1–14.

Hoch, J. S., A. H. Briggs, and A. R. Willan (2002). Something old, something new, something borrowed, something blue: A framework for the marriage of econometrics and cost-effectiveness analysis. *Health Economics 11*, 415–430.

Hotz, V. J., G. Imbens, and J. H. Mortimer (2005). Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics 125*(1-2), 241–70.

Imai, K., G. King, and E. A. Stuart (2008). Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A 171*(2), 481–502.

Imai, K., D. Tingley, and T. Yamamoto (2013, January). Experimental design for identifying causal mechanisms. *Journal of the Royal Statistical Society, Series A 176*(1), 5–51.

Imbens, G. (2009, April). Better late than nothing: Some comments on deaton (2009) and heckman and urzua (2009). NBER Working Paper 14896.

Ireland, C. T. and S. Kullback (1968). Contingency tables with given marginals. *Biometrika 55*(1), pp. 179–188.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Reviews 106*(4), 620–630.

Kang, J. D. Y. and J. L. Schafer (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from imcomplete data (with discussion). *Statistical Science 22*, 523–539.

Kish, L. (1992). Weighting for unequal P_i. *Journal of Official Statistics 8*, 183–200.

Kline, B. and E. Tamer (2011, April 14). Using observational vs. randomized controlled trial data to learn about treatment effects. Available at SSRN: http://ssrn.com/abstract=1810114 or http://dx.doi.org/10.2139/ssrn.1810114.

Kullback, S. (1997). *Information theory and statistics*. New York: John Wiley.

Liaw, A. and M. Wiener (2002). Classification and regression by randomforest. *R News 2*(3), 18–22.

Mattos, R. and A. Viega (2004). Entropy optimization: Computer implementation of the maxent and minxent principles. Working Paper.

Miratrix, L. W., J. S. Sekhon, and B. Yu (2013). Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society, Series B 75*(2), 369–396.

Mitra, N. and A. Indurkhya (2005). A propensity score approach to estimating the cost-effectiveness of medical therapies from observational data. *Health Economics 14*, 805–815.

Mojtabai, R. and J. G. Zivin (2003). Effectiveness and cost-effectiveness of four treatment modalities for substance disorders: a propensity score analysis. *Health Services Research 38*, 233–59.

National Research Council (2013). *Frontiers in Massive Data Analysis*. Washington, DC: The National Academies Press.

Nixon, R. M. and S. G. Thompson (2005). Incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. *Health Economics 14*, 1217—1229.

Politis, D. N. and J. P. Romano (1994). Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics 22*, 2031–2050.

Porter, K. E., S. Gruber, M. J. van der Laan, and J. S. Sekhon (2011). The relative performance of targeted maximum likelihood estimators. *The International Journal of Biostatistics 7*(1).

Rothwell, P. M. (2005). External validity of randomised controlled trials:to whom do the results of this trial apply?. *The Lancet 365*(9453), 82–93.

Sakr, Y., J.-L. Vincent, K. Reinhart, D. Payen, C. J. Wiedermann, D. F. Zandstra, and C. L. Sprung (2005). Sepsis occurrence in acutely ill patients investigators. use of the pulmonary artery catheter is not associated with worse outcome in the ICU. *Chest 128*(4), 2722–31.

Sekhon, J. S. (2009, June). Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science 12*, 487–508.

Sekhon, J. S. (2011). Matching: Multivariate and propensity score matching with automated balance search. *Journal of Statistical Software 42*(7), 1–52. Computer program available at *http://sekhon.berkeley.edu/matching/*.

Sekhon, J. S. and R. Grieve (2012). A nonparametric matching method for covariate adjustment with application to economic evaluation (genetic matching). *Health Economics 21*(6), 695–714.

Shadish, W. R., T. D. Cook, and D. T. Campbell (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Stevens, K., C. McCabe, C. Jones, J. Ashcroft, S. Harvey, and K. Rowan (2005). The incremental cost effectiveness of withdrawing pulmonary artery catheters from routine use in critical care. *Appl Health Econ Health Policy 4*(4), 257–264. On behalf the PAC-Man Study Collaboration.

Stuart, E. A., S. R. Cole, C. P. Bradshaw, and P. J. Leaf (2011, April). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society, Series A 174*(2), 369–386.

Tunis, S. R., D. B. Stryer, and C. M. Clancy (2003). Practical clinical trials. *JAMA: the journal of the American Medical Association 290*(12), 1624–1632.

Turner, R. M., D. J. Spiegelhalter, G. Smith, and S. G. Thompson (2009). Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 172*(1), 21–47.

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority*. CRC Press.

Willan, A. R. and A. H. Briggs (2006). *Statistical Analysis of Cost-Effectiveness Data*. Wiley.

Willan, A. R., A. H. Briggs, and J. S. Hoch (2004). Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data. *Health Economics 13*, 461–475.

Willan, A. R., E. B. Chen, R. J. Cook, and D. Y. Lin. (2003). Incremental net benefit in randomized clinical trials with quality-adjusted survival. *Statistics in Medicine 22*(3), 353–362.

Willan, A. R. and D. Y. Lin (2001). Incremental net benefit in randomized clinical trials. *Statistics in Medicine 20*(11), 1563–1574.

A. Proof of Theorem 1

PROOF. From (2) and (4)

$$\begin{aligned}\mathbb{E}(Y_{01}|S=0,T=1) &= \mathbb{E}(Y_{11}|S=0,T=1)\\ &= \mathbb{E}_{01}\{\mathbb{E}(Y_{11}|W^T,S=1,T=1)\}\\ &= \mathbb{E}_{01}\{\mathbb{E}(Y|W^T,S=1,T=1)\}.\end{aligned} \tag{11}$$

From (3) and (5)

$$\begin{aligned}\mathbb{E}(Y_{00}|S=0,T=1) &= \mathbb{E}(Y_{10}|S=0,T=1)\\ &= \mathbb{E}_{01}\{\mathbb{E}(Y_{10}|W^{C_T},S=1,T=0)\}\\ &= \mathbb{E}_{01}\{\mathbb{E}(Y|W^{C_T},S=1,T=0)\}.\end{aligned} \tag{12}$$

The result follows by substituting Eqs. (11) and (12) in the quantity of interest τ_{PATT} in Equation (1). With strong ignorability of sample assignment, from (6),

$$\begin{aligned}&\mathbb{E}(Y_{01}|S=0,T=1) - \mathbb{E}(Y_{00}|S=0,T=1)\\ &= \mathbb{E}_{01}\{\mathbb{E}(Y_{11}|W^T,S=0,T=1)\} - \mathbb{E}_{01}\{\mathbb{E}(Y_{10}|W^{C_T},S=0,T=1)\}\\ &= \mathbb{E}_{01}\{\mathbb{E}(Y_{11}|W^T,S=1,T=1)\} - \mathbb{E}_{01}\{\mathbb{E}(Y_{10}|W^{C_T},S=1,T=1)\},\end{aligned}$$

and the result follows from randomization.

B. Equivalence Tests

Equivalence tests begin with the null hypothesis:

$$H_0 : \frac{\mu_{\text{adj samp}}-\mu_{\text{pop}}}{\sigma} \geq \epsilon_U \qquad \text{or} \qquad \frac{\mu_{\text{adj samp}}-\mu_{\text{pop}}}{\sigma} \leq \epsilon_L$$
$$\text{versus}$$
$$H_1 : \epsilon_L < \frac{\mu_{\text{adj samp}}-\mu_{\text{pop}}}{\sigma} < \epsilon_U$$

where $\mu_{\text{adj samp}}$ is the true mean of the reweighted sample treated and μ_{pop} is the true mean of the populated treated, and σ is the pooled standard deviation of the two groups. We define $\epsilon_L = 0.2$ and $\epsilon_U = 0.2$, as discussed above. The test uses the test statistic

$$T = \frac{\sqrt{mn(N-2)/N}(\bar{X}_{\text{adj samp}}-\bar{X}_{\text{pop}})}{\left\{\sum_{i=1}^{m}(X_{\text{adj samp}i}-\bar{X}_{\text{adj samp}})^2+\sum_{j=1}^{n}(X_{\text{pop}j}-\bar{X}_{\text{pop}})^2\right\}^2}$$

where $\bar{X}_{\text{adj samp}}$ is the observed mean of the reweighted sample treated, $\bar{X}_{\text{pop}}$ is the observed mean of the population treated, standardized by the observed standard deviation. m refers to the number of observations in the reweighted sample, and n to the number of observations in the population treated, and $N = m + n$. The test rejects the null of non-equivalence if:

$$|T| < C_{\alpha;m,n}(\epsilon)$$
$$\text{with}$$
$$C_{\alpha;m,n}(\epsilon) = F^{-1}(\alpha; df_1 = 1, df_2 = N-2, \lambda_{nc}^2 = mn\epsilon^2/N)^{\frac{1}{2}}$$

where $C_{\alpha;m,n}(\epsilon)$ is the square root of the inverse F distribution with level α, degrees of freedom $1, N-2$, and non-centrality parameter $\lambda_{nc}^2 = mn\epsilon^2/N$. One important aspect of equivalence testing is that it requires the definition of a range over which observed differences are considered substantively inconsequential. We follow the recommendations of Hartman and Hidalgo (2011), and define equivalence as a mean difference between the reweighted sample treated and the true population treated of no more than 0.2 standardized differences and use the t-test for equivalence defined in Wellek (2010).

C. Identifiability of Alternative Causal Quantities

The main population treatment effect considered in this paper is PATT, however there are numerous population treatment effects that policy makers might be interested in. If PATC is of interest then assumptions 2 and 3 can be replaced by

$$(Y_{01}, Y_{11}) \perp\!\!\!\perp S|(W^T, T=0) \qquad \text{and} \qquad (Y_{00}, Y_{10}) \perp\!\!\!\perp S|(W^{C_T}, T=0), \tag{13}$$

respectively. Additionally, if Equation (3) in assumption 1, $(Y_{00}, Y_{10}) \perp\!\!\!\perp S|(W^{C_T}, T=0)$ and assumption 4 hold then $\mathbb{E}(Y|S=0, T=0) = \mathbb{E}_{00}\{\mathbb{E}(Y|W^{C_T}, S=1, T=0)\}$. Therefore the mean outcomes would be the same for the control group in the target population and the adjusted RCT, adjusted such that W^{C_T} follows its distribution in the target control group. A placebo test can then be used to check the validity of the required assumptions. However, this is not necessary to apply Theorem 1 because (13) is not assumed in the current analysis.

In circumstances where the estimand of interest is the PATE then the estimand of interest is the effect in the entire target population, where $\tau_{PATE} = \mathbb{E}(Y_{01} - Y_{00}|S=0)$. In such a case, assumptions 1–4, as well as $(Y_{01}, Y_{11}) \perp\!\!\!\perp S|(W^{C_T}, T=0)$ and $(Y_{00}, Y_{10}) \perp\!\!\!\perp S|W^{C_T}, T=0$, are sufficient for identification. Assuming $W_i^T = W_i^{C_T}$, these assumptions and randomization imply that $Y_{st} \perp\!\!\!\perp (S,T)|W^T$, which means that the potential outcomes for units with the same W^T are exchangeable, regardless of whether they are assigned to treatment or control and whether they are in the target population or RCT. Under these assumptions and randomization, it can be shown that

$$\mathbb{E}(Y_{st}|S=0) = \mathbb{E}_{W^{C_T}|S=0}\{\mathbb{E}(Y|W^{C_T}, S=1, T=t)\} \tag{14}$$

$$\tau_{PATE} = \mathbb{E}_{W^{C_T}|S=0}\{\mathbb{E}(Y|W^{C_T}, S=1, T=1) - \mathbb{E}(Y|W^{C_T}, S=1, T=0)\}, \tag{15}$$

for $t = 0, 1$. Equation 14 implies that the mean outcome in the target population is the same as in the adjusted RCT $T=t$ group, adjusted such that W^{C_T} follows its distribution in the target population. Equation 15 implies that the results from an adjusted RCT can be used to identify PATE for a target population. The analysis of Stuart et al. (2011) makes the stronger assumptions required to justify Equations 14

and 15. Stuart et al. (2011) verify the assumptions by confirming the validity of Equation 14, for $t = 0$, which then justifies the generalizability of the RCT results, from Equation 15. It is possible for Equation (14) to be violated and Equation (15) to still hold. This occurs if treatment consistency is violated but the potential outcomes in the target population and RCT differ by some constant.

The assumptions required for Equation 14 can be checked by a placebo test of the mean outcome in the target population and the adjusted RCT $T = t$ group, for $t = 0, 1$. However, since the assumptions used in the analysis here are weaker and do not imply Equation 14, such a test is not done.

D. Maximum Entropy Weighting

The principle of maximum entropy is defined as:

$$\max_{\mathbf{p}} S(\mathbf{p}) = -\sum_{i=1}^{n} p_i \ln p_i \tag{16}$$

$$s.t. \begin{cases} \sum_{i=1}^{n} p_i = 1 \\ \sum_{i=1}^{n} p_i g_r(x_i) = \sum_{i=1}^{n} p_i g_{ri} = a_r \qquad r = 1, \ldots, m \\ p_i \geq 0 \qquad i = 1, 2, \ldots, n \end{cases} \tag{17}$$

where equation (16) maximises Shanon's measure of entropy, which is a form of probabilistic uncertainty. The first constraint in equation (17) is referred to as the natural constraint, and it simply states that all the probabilities must sum to one. The m moment constraints are referred to as the consistency constraints. Each a_r represents an r-th order moment, or characteristic moment, of the probability distribution (i.e. $g_{ri} = (x_i - \mu)^r$ where μ is the distribution mean). The distribution chosen for $\mathbf{p}$ is that most similar to the uniform that still satisfies the constraints. Due to the fact that there are $m+n$ equations and $m+n$ unknowns, corresponding to m Lagrange multipliers and n probabilities, it is not possible to derive an analytical solution for p_i and λ_r simultaneously using only the known moments. A solution must be found using an iterative search algorithm (Mattos and Viega, 2004).

In this context this ensures that individuals in the treatment group in the RCT who have identical values for all of the covariates used in the constraints are given equal weights. Here, the matched pairs from the RCT are reweighted using constraints from the target population as, for example represented by the NRS. The consistency constraints are constructed using moments such as the covariate means from a NRS. Typically this is done using covariate data contained in both the NRS and the RCT, however information from several external sources (e.g. disease registries) about the target population can also be incorporated into the constraints. Once the consistency constraints have been created, a set of weights that simultaneously satisfies the constraints while maximizing the entropy measure

is calculated. PATT can then be reported by weighting the SATT for each of the individual matched pairs.

E. Covariates used in Matching and Reweighting with MaxEnt

Table E.1. Covariates used in GenMatch

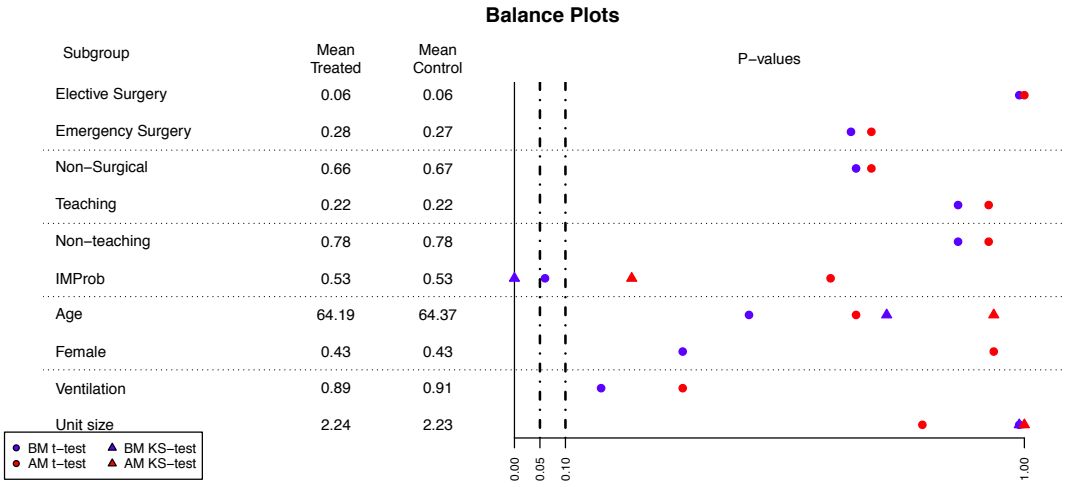
GenMatch Covariates	
Priority (balance enforced to be no worse than initial balance)	Age, baseline probability of death, elective surgery indicator, emergency surgery indicator, size of ICU, teaching hospital indicator, mechanical ventilator at admission, base excess
Additional Covariates used for Matching	physiology score, admission diagnosis, gender, past medical history variables on cardiac, respiratory, liver, and immune measures, heart rate and blood pressure physiology measures, temperature measures, respiratory measures
Additional covariates used for Balance	admission diagnosis, blood gas rate, Pf rate, Ph, Creatinine, Sodium, Urine output, white blood cell counts, Glasgow coma, cardiac and respiratory measures of organ failure, indicator for sedation or paralyzation, baseline PAC rate in unit, geographical region, APACHE II probability of death, indicators for missing values

Table E.2. Covariates used in MaxEnt

MaxEnt Margins
age, elective surgery indicator, emergency surgery indicator, teaching hospital indicator, gender, baseline probability of death, mechanical ventilator at admission, chemical measure of decline, past medical history variables on cardiac, respiratory, liver, and immune measures, categorical variables on blood pressure rates, categorical measures on temperature, geographical region, categorical variables on age (0-56, 57-66, 67+), categorical classification of diagnostic variable, categorical classification of base excess, base excess categories $\times$ age categories, unit size $\times$ teaching hospital indicator, teaching hospital indicator $\times$ base excess categories, mechanical ventilation $\times$ base excess categories, teaching hospital indicator $\times$ mechanical ventilator at admission, unit size $\times$ mechanical ventilator at admission, gender $\times$ teaching hospital, teaching hospital $\times$ age categories, gender $\times$ age categories, emergency surgery indicator $\times$ gender, elective surgery indicator $\times$ gender, teaching hospital indicator $\times$ past medical history variables on cardiac, respiratory, liver, and immune measures, age categories $\times$ base excess $\times$ gender, gender $\times$ past medical history variables on cardiac and respiratory measures, mechanical ventilation at admission $\times$ pasty medical history variables on cardiac, respiratory, and renal measures

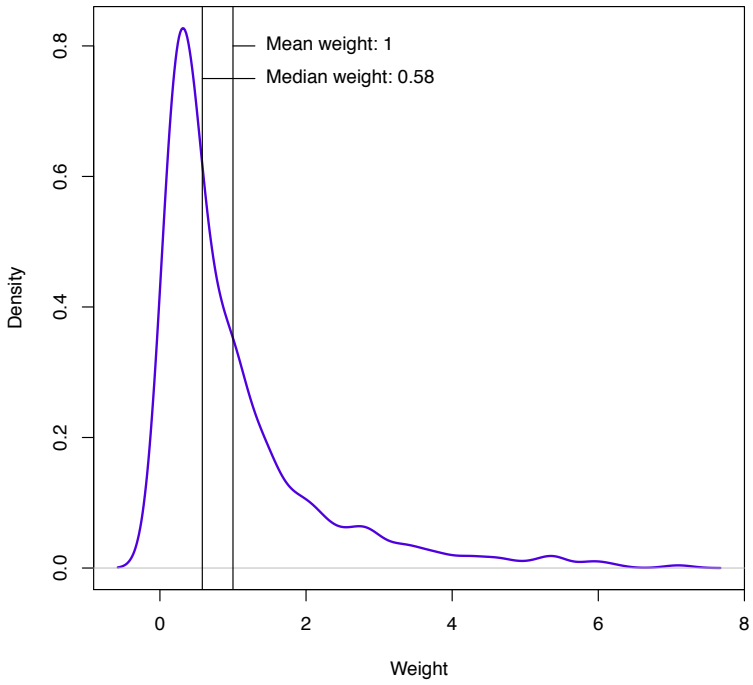
F. Covariate balance in the PAC-Man RCT before and after matching

Fig. F.1. Covariate balance in the PAC-Man RCT according to p-values from t-tests, before matching (BM) and after matching (AM) with GenMatch



G. Distribution of Maximum Entropy Weights

Fig. G.1. MaxEnt Weight Distribution



H. Inverse Propensity Score Weighting

Where individual unit data are available as in the PAC example, IPSW can also be used to reweight the RCT data. In this context the propensity score estimates the predicted probability of each individual unit being in the RCT, conditional on baseline characteristics observed in the RCT and the NRS. IPSW then gives each individual in both the RCT and NRS settings a weight, calculated as the inverse of the probability of being in the RCT according to baseline characteristics. However, IPSW weights can be extreme leading to unstable results and estimated treatment effects can be particularly sensitive to misspecification of the propensity score (Porter et al., 2011; Kang and Schafer, 2007; Kish, 1992). In recognition of the concern that the propensity score may be misspecified we used a machine learning algorithm for classification, random forests (Breiman, 2001), implemented in the `randomForest` package with the default parameters (Liaw and Wiener, 2002; Stuart et al., 2011).

We implemented the IPSW approach in the case study by estimating a propensity score using the covariates listed in the Table H.1. We calculated the IPSW weights separately for each subgroup to enforce interactions of variables with the subgroup classifications. Figure H.3 reports the covariate balance for the PAC patients in the NRS versus the RCT, after reweighting with IPSW.

We found that in our example the IPSW weights were stable (see Figure H.2), but there were differences between the observed covariate means for the PAC patients in the NRS versus those from the RCT after reweighting with IPSW.

After we applied the placebo tests following IPSW, we found there are still large differences in hospital mortality between the PAC patients in the adjusted RCT treated group and the NRS. Figure H.4 shows the equivalence based placebo results for the IPSW method for all three outcomes of interest. The placebo tests to do not consistently pass, we therefore do not consider IPSW for estimation of the PATT.

Table H.1. Covariates used in IPSW Estimation

Propensity Score Covariates
gender, age, categorical age variables, elective surgery indicator, emergency surgery indicator, past medical history variables on cardiac, respiratory, liver, and immune measures, categorical diagnostic variable, chemical decline variable, base excess categorical variables, heart rate categorical variables, blood pressure categorical variables, temperature categorical variables, blood gas rate categorical variables, Pf rate categorical variables, Ph categorical variables, Creatinine categorical variables, Sodium categorical variables, Urine output categorical variables, white blood cell counts categorical variables, Glasgow coma categorical variables, cardiac and respiratory measures of organ failure, mechanical ventilation at admission, unit size categorical variable, teaching hospital indicator

Fig. H.2. IPSW Weight Distribution

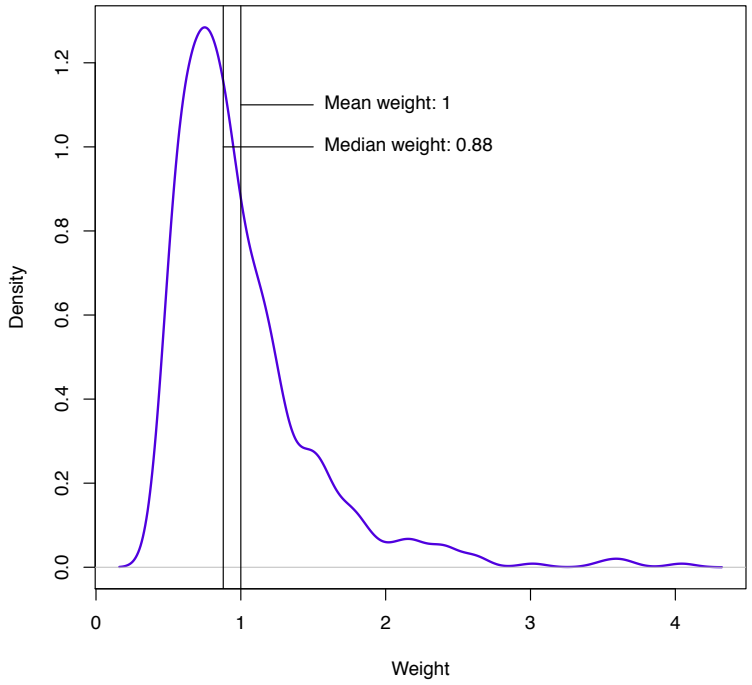
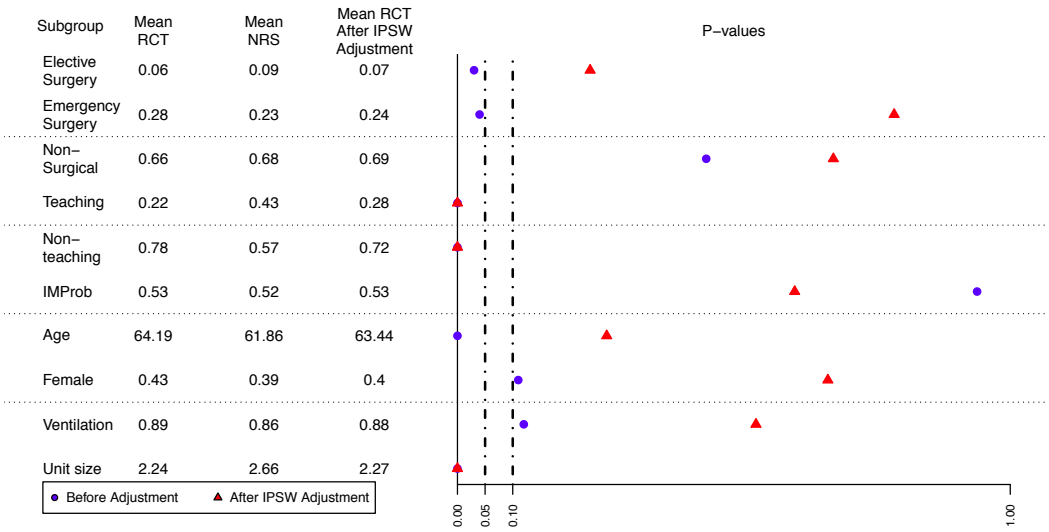


Fig. H.3. Balance on observable characteristics between the PAC patients in the NRS and the RCT, before and after adjustment of the RCT data with IPSW



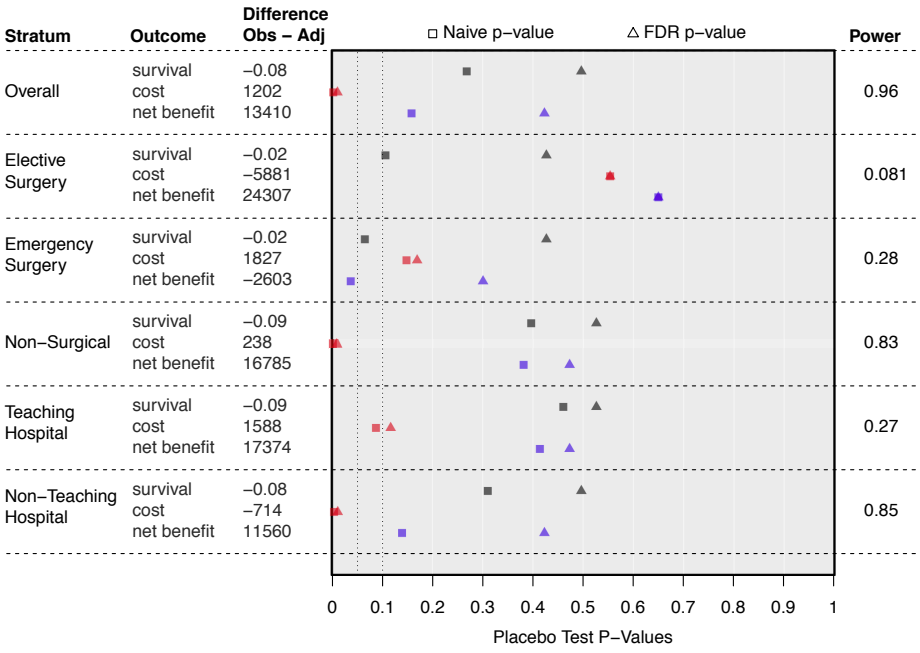


Fig. H.4. IPSW Placebo Tests

Above are the results of the equivalence placebo tests comparing the mean outcome of NRS treated to the reweighed mean of the RCT treated. The column labelled "Difference" presents the difference between the observed outcomes for the PAC group in the NRS and the PAC group in the RCT after reweighting. The column labelled "Power" presents the power of the equivalence t-test for each stratum.

I. Bayesian Additive Regression Trees: Placebo test results

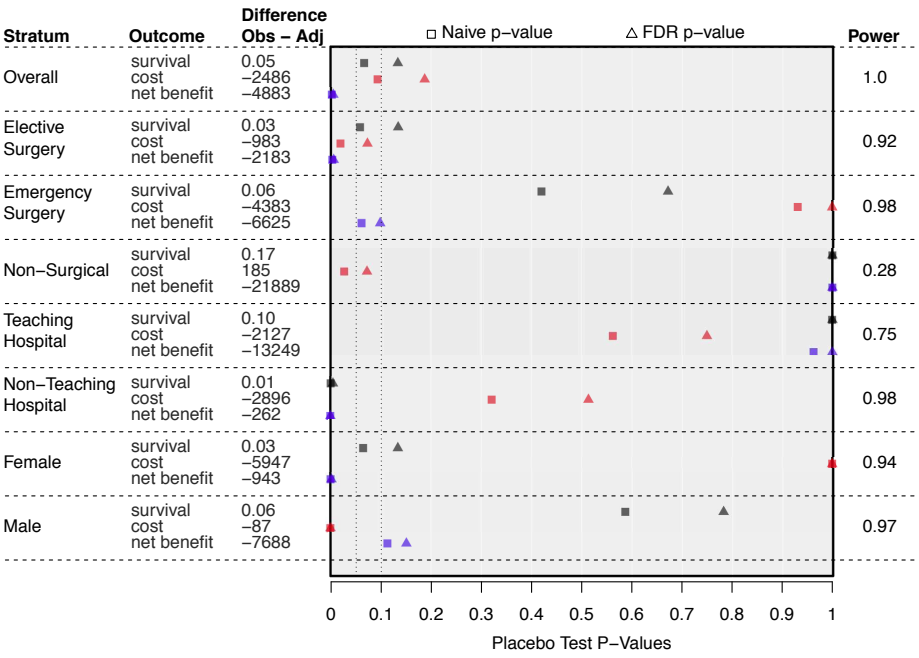


Fig. I.5. BART Placebo Tests

Results of the equivalence placebo tests comparing the mean outcome of NRS treated to the BART estimated population treated. The column labelled "Difference" presents the difference between the observed outcomes for the PAC group in the NRS and the PAC group in the BART estimated population treated. The column labelled "Power" presents the power of the equivalence t-test for each stratum.