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

Erin Hartman[†] Richard Grieve[‡] Roland Ramsahai[§] Jasjeet S. Sekhon[¶]

11/1/2013 (20:16)

Abstract

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 the assumptions required to identify population average treatment effects from RCTs. We provide placebo tests, which formally follow from the identifying assumptions and can assess whether they hold. 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).

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 (Willan and Briggs, 2006; Nixon and Thompson, 2005; Mitra and Indurkhya, 2005; Willan, Briggs and Hoch, 2004; Mojtabai and Zivin, 2003; Hoch, Briggs and Willan, 2002). A key concern is that estimates from RCTs and meta analyses may lack external validity (Imbens, 2009; Heckman and Urzua, 2009; Deaton, 2009; Heckman and Vytlacil, 2005). In RCTs treatment protocols and interventions differ to those administered routinely, and trial participants, for example, individuals, hospitals, or schools, tend to be atypical of those in the target population, which can limit the generalisability of results (Gheorghe et al., 2013). These concerns pervade RCTs across different areas of public policy, and are key objections to undertaking RCTs, or to using the results in policy-making (Deaton, 2009). It is therefore important to establish the conditions under which RCTs can identify population treatment effects, and to develop methods to test if these conditions hold in a given application.

[*] We gratefully acknowledge 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, the Statistics Department at the University of Pennsylvania for comments. The authors are responsible for all errors.

[†] Ph.D. Candidate, Department of Political Science, UC Berkeley.

[‡] Reader in Heath Economics, London School of Hygiene & Tropical Medicine.

[§] Statistical Laboratory, Cambridge.

[¶] Professor of Political Science and Statistics, UC Berkeley, http://sekhon.berkeley.edu.

Previous research has proposed using non-randomized studies (NRSs) to assess whether RCT-based estimates apply to a target population (Stuart et al., 2011; Green and Kern, 2012; Imai, King and Stuart, 2008; Kline and Tamer, 2011; Cole and Stuart, 2010; Shadish, Cook and Campbell, 2002; Greenhouse et al., 2008). 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). Furthermore, there may be unobserved differences between the RCT participants and the target population, and the form of treatment or control, for example the dose of a drug or the rigor of a protocol, may differ between the settings (SR 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.

Imai, King and Stuart (2008) introduced a framework 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 then consider two alternative techniques for weighting the individual RCT strata according to the observed characteristics in the target population: inverse propensity score weighting (IPSW) and maximum entropy (MaxEnt) weighting. Each technique aims to reweight the individual strata in the RCT so that they resemble the distribution of 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 the population average treatment effects. Section 4 describes the placebo tests for checking the underlying assumptions, while section 5 outlines estimation strategies for population treatment effects. In Section 6, we illustrate the approach with the PAC case study. Section 7 discusses an alternative design identified by the main theorem, and Section 8 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 (Connors et al., 1996; Chittock et al., 2004). 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 database contains information on case-mix, patient outcome and resource use for 200 critical care units in the United Kingdom (Harrison, Brady and Rowan, 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 receive PAC. 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 1). 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 monetary benefit (INB) (Willan et al.,

[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

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	PAC	PAC
	n=507	n=506	n=1051
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

2003; Willan and Lin, 2001).[2]

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.

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.

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, such as those in the NRS. For simplicity, we assume that in either study a unit is assigned to treatment $(t = 1)$ or

[2] Net 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 Walesis £20,000 to £30,000 to gain a QALY), and subtracting from this the cost. Finally, the INB of the new treatment can be estimated by contrasting the mean net benefits for each alternative.

control ($t = 0$), and that there is no crossover. 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 controls included in the RCT versus the target population.

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

$$\tau_{SATE} = \mathbb{E}(Y_{i11} - Y_{i10} | S_i = 1)$$

Randomization ensures that the difference in the mean outcome for the treated versus control units in the RCT is an unbiased estimate of 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), which in finite samples, they may differ from SATE even in RCTs. When treatment assignment is ignorable, they are:

$$\tau_{SAT*} = \mathbb{E}(Y_{i11} | S_i = 1, T_i = t) - \mathbb{E}(Y_{i10} | S_i = 1, T_i = t),$$

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 or $(Y_{i11}, Y_{i10}) \perp\!\!\!\perp T_i = 1 | S_i = 1$, and the alternative estimands are asymptotically equivalent.[3]

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

$$\begin{aligned}
\tau_{PATE} &= \mathbb{E}(Y_{i01} - Y_{i00} | S_i = 0) \\
\tau_{PATC} &= \mathbb{E}(Y_{i01} - Y_{i00} | S_i = 0, T_i = 0) \\
\tau_{PATT} &= \mathbb{E}(Y_{i01} - Y_{i00} | S_i = 0, T_i = 1). \qquad (1)
\end{aligned}$$

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 necessary to derive the identifiable expression for τ_{PATT} in Theorem 1. Figure 1 represents the assumptions, and demonstrates the result of Theorem 1.

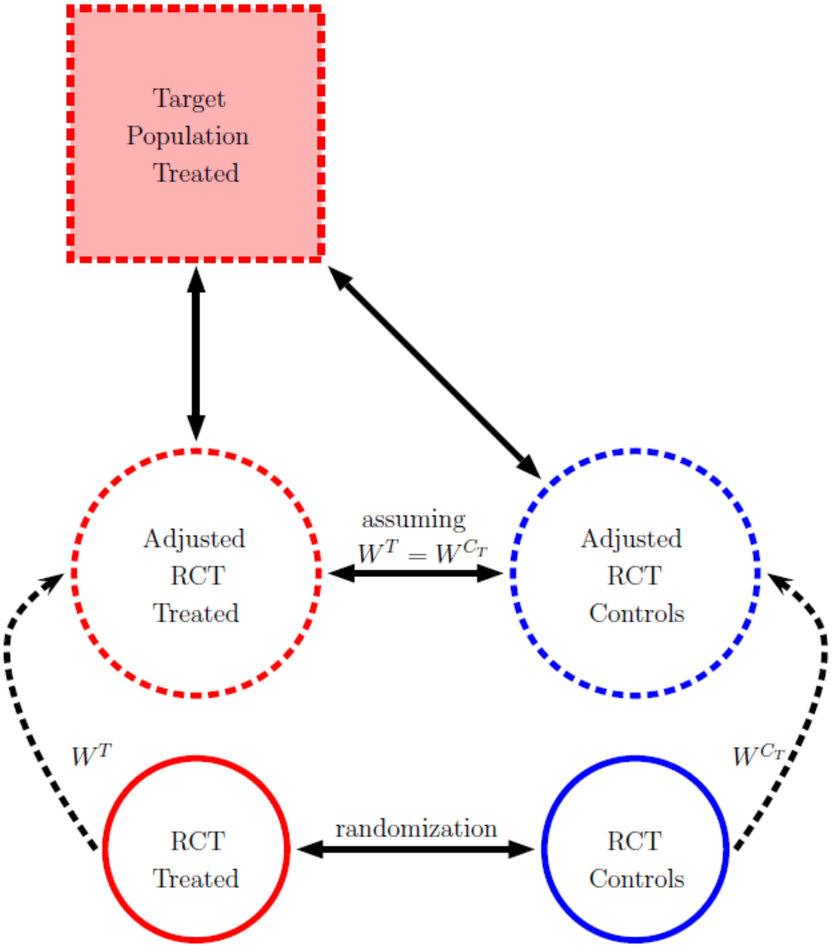
Assumption 1: Consistency under Parallel Studies

$$\begin{aligned}
Y_{i01} &= Y_{i11} \qquad (2) \\
Y_{i00} &= Y_{i10} \qquad (3)
\end{aligned}$$

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

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

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



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 effect the outcome. Assumption 1 is similar to the assumption of consistency under the parallel experiment design in Imai, Tingley and Yamamoto (2013). Assumption 1 may be violated for example if 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

$$(Y_{i01}, Y_{i11}) \perp\!\!\!\perp S_i | (W_i^T, T_i = 1), \qquad 0 < Pr(S_i = 1 | W_i^T, T_i = 1) < 1.$$

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

$$\mathbb{E}(Y_{is1} | S_i = 0, T_i = 1) = \mathbb{E}_{01}\{\mathbb{E}(Y_{is1} | W_i^T, S_i = 1, T_i = 1)\}, \tag{4}$$

for $s = 0, 1$. The expectation $\mathbb{E}_{01}\{\cdot\}$ is a weighted mean of the W_i^T specific means, $\mathbb{E}(Y_{is1} | W_i^T, S_i = 1, T_i = 1)$, with weights according to the distribution of W_i^T in the treated target population, $Pr(W_i^T | S_i = 0, T_i = 1)$. Essentially, on the right side of Equation (4), the characteristics of the treated units in the RCT, W_i^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_{is1} potential outcomes of the treatment group in the target population. This is depicted as the double arrow, representing equivalence between the dashed red circle and the dashed red square in Figure 1.

Assumption 3: Strong Ignorability of Sample Assignment for Controls

$$(Y_{i00}, Y_{i10}) \perp\!\!\!\perp S_i | (W_i^{C_T}, T_i = 1), \qquad 0 < Pr(S_i = 1 | W_i^{C_T}, T_i = 1) < 1.$$

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

Assumption 3 implies that

$$\mathbb{E}(Y_{is0} | S_i = 0, T_i = 1) = \mathbb{E}_{01}\{\mathbb{E}(Y_{is0} | W_i^{C_T}, S_i = 1, T_i = 0)\}, \tag{5}$$

for $s = 0, 1$, since treatment assignment is random in the RCT, i.e. $Y_{is0} \perp\!\!\!\perp T_i | (W_i^{C_T}, S_i = 1)$. Similarly to Equation (4), on the right side of Equation (5), the characteristics of the units in the control group in the RCT, $W_i^{C_T}$, 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. Again this adjustment can be performed by the various weighting methods discussed in Section 5.

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 Y_{is0} potential outcomes of the treated group in the target population. This is depicted as the double arrow representing equivalence between the dashed blue circle and the dashed red square in Figure 1.

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*

$$\begin{aligned}&\mathbb{E}_{01}\{\mathbb{E}(Y_{is1}|W_i^T, S_i = 0, T_i = 1)\} - \mathbb{E}_{01}\{\mathbb{E}(Y_{is0}|W_i^{C_T}, S_i = 0, T_i = 1)\}\\ &= \mathbb{E}_{01}\{\mathbb{E}(Y_{is1}|W_i^T, S_i = 1, T_i = 1)\} - \mathbb{E}_{01}\{\mathbb{E}(Y_{is0}|W_i^{C_T}, S_i = 1, T_i = 1)\},\end{aligned} \tag{6}$$

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

$$\tau_{PATT} = \mathbb{E}_{01}\{\mathbb{E}(Y_i|W_i^T, S_i = 1, T_i = 1)\} - \mathbb{E}_{01}\{\mathbb{E}(Y_i|W_i^{C_T}, S_i = 1, T_i = 0)\},$$

where $\mathbb{E}_{01}\{\mathbb{E}(\cdot|W_i^T, \ldots)\}$ denotes $\mathbb{E}_{W_i^T|S_i=0,T_i=1}\{\mathbb{E}(\cdot|W_i^T, \ldots)\}$.

From the expression for τ_{PATT} in Theorem 1, it is possible to identify τ_{PATT} from the adjusted RCT data alone, a schematic of which is provided in Figure 1. In Figure 1, the adjusted experimental controls and treated 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 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.

By randomization, the RCT treatment and control groups are exchangeable. Therefore adjusting both groups by the same observable characteristics will asymptotically yield 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 can replace their counterparts in the target population. In order to gain precision, matching or stratifying between the treated and control units within the RCT according to baseline characteristics, can be undertaken prior to adjustment to the target population (Miratrix, Sekhon and Yu, 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 according to the characteristics of the treatment group 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 and assumptions 2 and 4 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_{i01} are observed in the treated group of the target population, then $\mathbb{E}(Y_{i01}|S_i = 0, T_i = 1)$ is equal to $\mathbb{E}(Y_i|S_i = 0, T_i = 1)$ and

$$\mathbb{E}(Y_i|S_i = 0, T_i = 1) - \mathbb{E}_{01}\{\mathbb{E}(Y_i|W_i^T, S_i = 1, T_i = 1)\} = 0, \tag{7}$$

from Equation (11) in Appendix A. Therefore if Equation (2) in assumption 1 and assumptions 2 and 4 hold, the expected outcomes in the adjusted RCT treated group and the target population will be the same. A placebo test can be used to check whether there is a difference in average outcomes between the adjusted RCT treated group versus the target population. If the placebo test detects a significant difference in the above outcomes, then either Equation (2) in assumption 1, assumption 2 or assumption 4 is violated.[4] 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_{i00} is not observed in those treated in the target population, then $\mathbb{E}(Y_{i00}|S_i = 0, T_i = 1)$ is not necessarily equal to $\mathbb{E}(Y_i|S_i = 0, T_i = 0)$. 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 tests can therefore be used to validate the underlying identifying assumptions for estimating population-level causal estimates. They may highlight the failure of several underlying assumptions but 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, whose null hypothesis is 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 treated in the RCT, are not equivalent to those treated in the target population. This null hypothesis of non-equivalence is only rejected if there is sufficient power.[5] Therefore a low p-value would indicate the two groups are statistically equivalent, whereas high p-values in traditional tests 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 C and Hartman and Hidalgo (2011) give further details.

[4] 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.

[5] This alleviates the issues of confounding the notion of statistical equivalence with a tests relationship to sample size discussed in Imai, King and Stuart (2008).

5 Estimating PATT

There are a growing number of estimation strategies for predicting population-level treatment effects from RCT data, which can be assigned into two broad classes. One class of estimators estimates the response surface using the RCT data and extrapolates this response surface to the target population, and includes methods such as Bayesian Additive Regression Trees (BART) (Chipman, George and McCulloch, 2006; Green and Kern, 2012), Classification and Regression Trees (CART) (Breiman, 2001; Liaw and Wiener, 2002; Stuart et al., 2011), and linear regression. The other prominent approach is to use weighting methods, such as Inverse Propensity Score Weighting (IPSW) (Stuart et al., 2011) and Maximum Entropy (MaxEnt) weighting, which rely on ancillary information, for example from a NRS, to reweight the RCT data. 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 predicted values from a model of the response surface, or weights from the second class of estimators. 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 estimating 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 subgroup. We then reweight the matched pairs according to the characteristics of the target population to report PATT, both overall and for pre-specified 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 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

Stuart et al. (2011) consider IPSW, full matching and sub-classification for reweighting RCT estimates to the target population, and find that IPSW performs relatively well. Hence, we firstly consider IPSW, where in this context the propensity score estimates the predicted probability of a 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. In general IPSW methods can be particularly sensitive to misspecification of the propensity score; the weights can be extreme, leading to unstable results (Porter et al., 2011; Kang and Schafer, 2007; Kish, 1992).

We consider an alternative approach to reweighting—MaxEnt (Shannon, 1948; Mattos and Viega, 2004). 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.

For estimating PATT, the requisite placebo tests can be conducted by comparing the average of the observed outcomes for the treated in the target population, to the weighted outcomes for the treated in the RCT. Provided that the placebo test has sufficient power, failure to find equivalence between the above treated groups indicates that at least one assumption underlying Theorem 1 is not met, and that there is bias in the estimated PATT.

5.3 Response Surface Models

Response surface models can estimate the covariate endpoint relationships in the RCT, and use these estimates to predict population treatment effects the target population. A common concern is that such estimation methods use the coefficients estimated from the RCT data in predicting population estimates for the observed covariate distribution in the 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 assumes the response surfaces in the RCT and the target population are the same. This assumption may lead to efficiency gains relative to weighting approaches especially if some, but not all, covariates included in the adjustment are strongly predictive of potential outcomes. The placebo tests proposed can be used with the response surface approach by comparing average predicted values from the model estimated from the NRS treated covariate distribution, to the average of the observed outcomes for the NRS treated. Again given sufficient power, a failure to find equivalence between the above outcomes indicates a failure of at least one assumption underlying Theorem 1 and bias in the estimated population effects.

6 Empirical Example: PAC

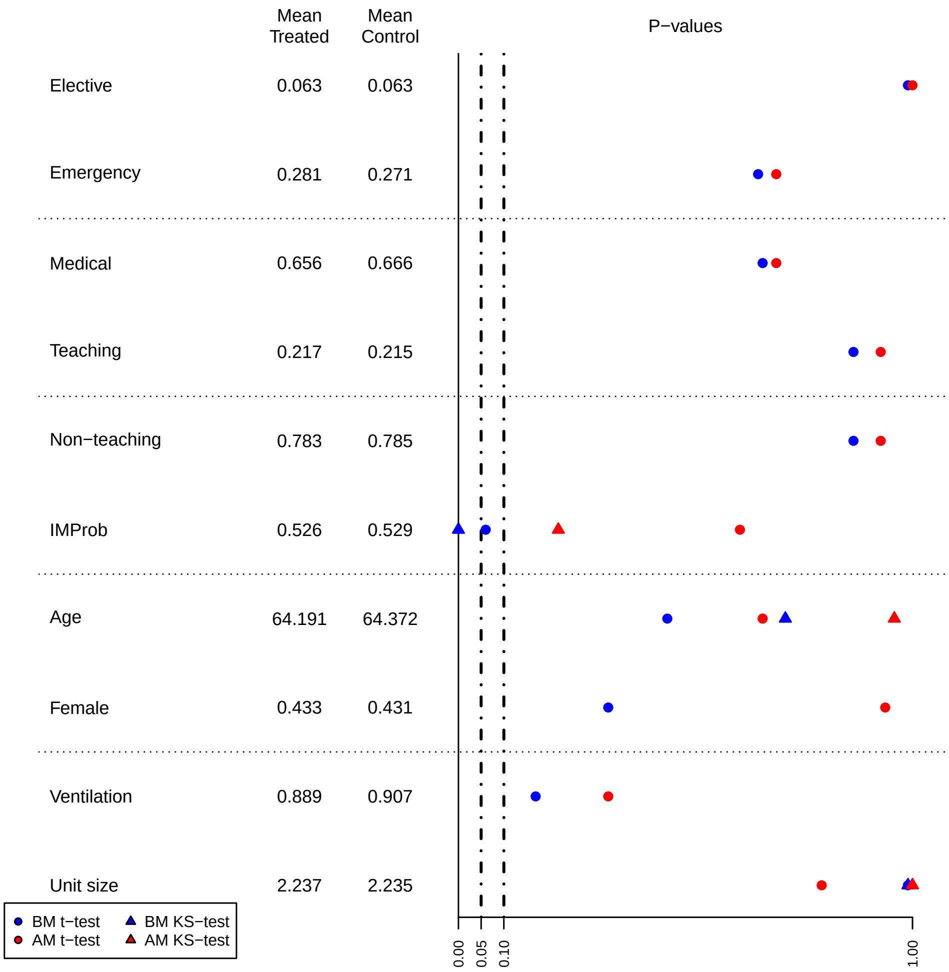
We illustrate our new estimating strategy for extrapolating from an RCT to a target population using the PAC example, where a major concern is that baseline covariates anticipated to modify the treatment effect, differ between the RCT and those who receive PAC in the target population Table 1. Here, 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, and included those covariates listed in the Appendix Table 2. 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.[6] Variance estimates were calculated conditional on the matched data (Imai, King and Stuart, 2008). The matching identified a control for each treated observation, resulting in 507 matched pairs for the overall estimate. Each baseline covariate was well balanced after matching according to both t-tests and KS tests. The SATT results, both overall and at subgroup level, were similar to the SATE estimates from the RCT.

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

Figure 2: Covariate Balance in the RCT



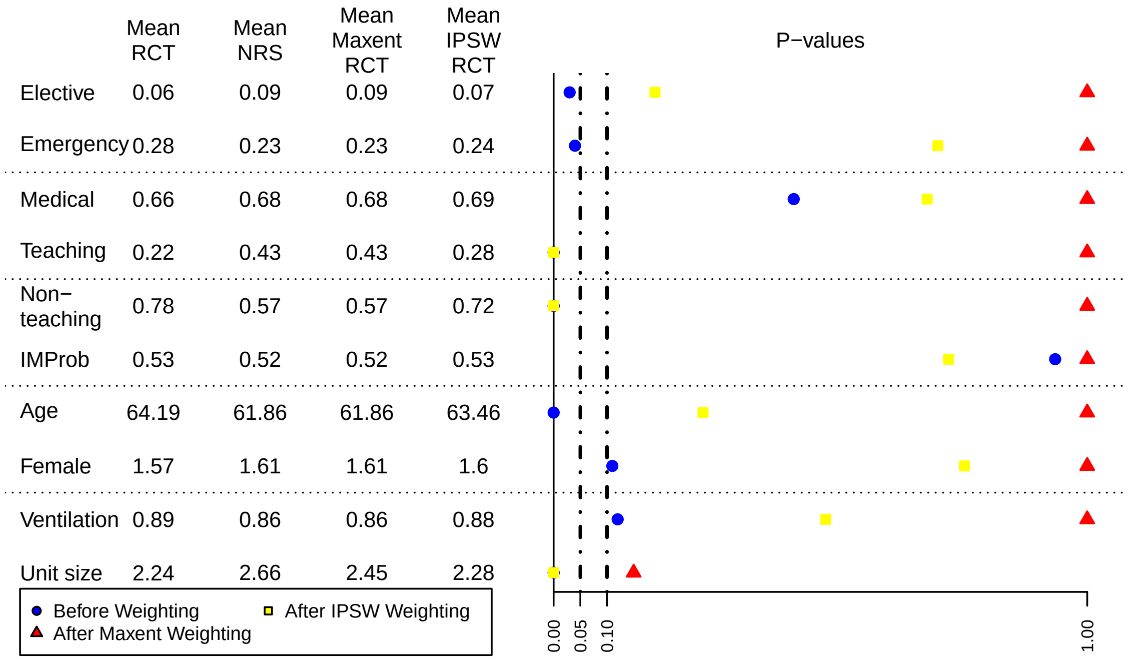
	Mean Treated	Mean Control
Elective	0.063	0.063
Emergency	0.281	0.271
Medical	0.656	0.666
Teaching	0.217	0.215
Non−teaching	0.783	0.785
IMProb	0.526	0.529
Age	64.191	64.372
Female	0.433	0.431
Ventilation	0.889	0.907
Unit size	2.237	2.235

We use two weighting methods to derive weights for the matched pairs. These weights adjusted the distribution of observable baseline covariates in the matched RCT data to the distribution of the PAC patients in the NRS. We firstly estimated a propensity score to predict participation in the RCT, according to the baseline covariates listed in the Appendix, Table 4. We recalculated the IPSW weights separately for each subgroup to enforce interactions of variables with the subgroup classifications. 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 also used MaxEnt, and constructed the weights for the covariates listed in Appendix, Table 3 using marginal interacted covariate distributions from the population. The consistency constraints were taken as the covariate means from the PAC patients in the NRS. The weights were selected to satisfy these constraints while maximizing the entropy measure. As Figure 3 shows, population moments used in constructing the weights matched exactly. The inclusion of interacted population moments helped ensure, that unlike IPSW, only one set of weights was required for the overall and subgroup estimates.

We applied both IPSW and MaxEnt to reweight the individual matched pairs from the RCT

Figure 3: Balance on observable characteristics between the PAC patients in the NRS and the RCT, before and after adjustment of the RCT data with MaxEnt and IPSW



	Mean RCT	Mean NRS	Mean Maxent RCT	Mean IPSW RCT
Elective	0.06	0.09	0.09	0.07
Emergency	0.28	0.23	0.23	0.24
Medical	0.66	0.68	0.68	0.69
Teaching	0.22	0.43	0.43	0.28
Non−teaching	0.78	0.57	0.57	0.72
IMProb	0.53	0.52	0.52	0.53
Age	64.19	61.86	61.86	63.46
Female	1.57	1.61	1.61	1.6
Ventilation	0.89	0.86	0.86	0.88
Unit size	2.24	2.66	2.45	2.28

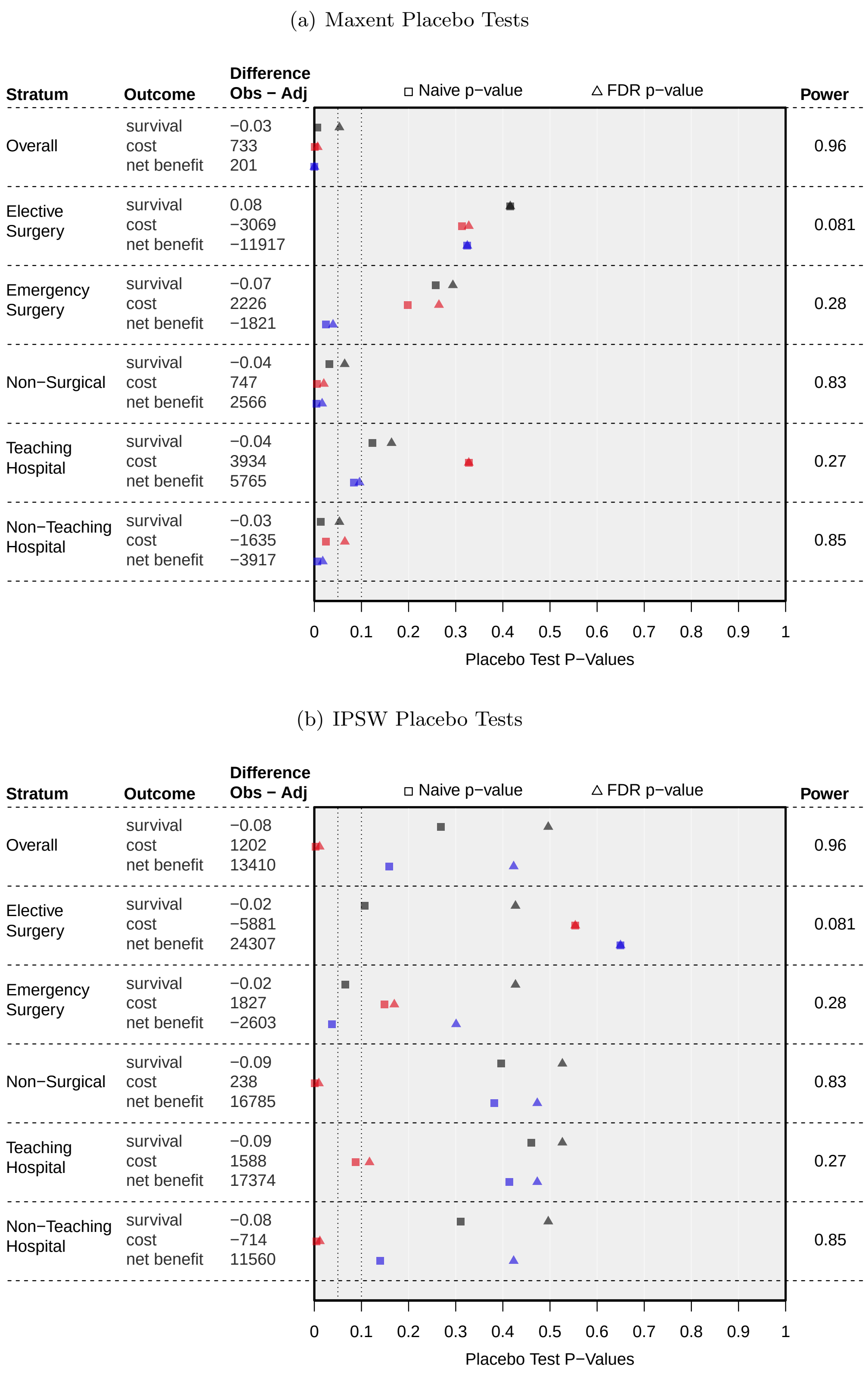
for the observable characteristics of the PAC patients in the NRS. To recognize the uncertainty in the estimation of the weights, standard errors for both SATT and PATT were estimated using subsampling (Politis and Romano, 1994). Abadie and Imbens (2006) show that the bootstrap is not valid for estimating the standard error of a matching estimator, but suggest that subsampling (Politis and Romano, 1994) is a valid estimating strategy. 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,m, in the RCT. We used 1000 bootstrap replicates. [7]

IPSW and MaxEnt provided weights that were reasonably stable (see Appendix F), with mean weights of 1, and maxima of 4 (IPSW) and 8 (MaxEnt); no individual stratum was given an extreme weight. Figure 3 reports the covariate balance, based on t-test p-values, for the PAC patients in the NRS versus the RCT, after reweighting with either approach. While IPSW achieved good balance for some covariates, other prognostic factors such as attendance at teaching hospital, and ICU size were still imbalanced. By contrast, MaxEnt balanced all covariate means between the PAC patients in the RCT and the NRS.

6.2 Results of the Placebo Tests

For each approach, we reported placebo tests that tested the underlying assumptions for estimating 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 Figures 4(a) and 4(b), for

[7] One of the driving 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, by nature of the RCT design, where the true propensity of each individual to be assigned a particular treatment is random, there are many potential matches for each unit. Therefore, even within each iteration of the bootstrap, the probability of a close match for each unit is high. It may not be surprising then that the Bickel and Sakov (2008) algorithm selects $m = n$.


(a) Maxent Placebo Tests

(b) IPSW Placebo Tests

Figure 4: Placebo Tests

The column labelled “Difference” presents the difference between the observed outcomes for the population treated and the observed outcomes of the adjusted experimental treated, on the scale of the outcome. The column labelled “Power” presents the power of the equivalence t-test for each stratum.

all three endpoints (survival rates (black), cost (red) and INB (blue)). We present the equivalence based placebo test p-values (represented by squares) for the overall estimate and each subgroup, and allow for multiple comparisons, by presenting p-values with a false discovery rate correction using the Benjamini-Hochberg method (represented by triangles) (Benjamini and Hochberg, 1995). After IPSW, there are still large differences in hospital survival rates between the PAC patients in the adjusted RCT treated group and the NRS. For the survival rate and net benefit endpoints, IPSW fails almost all the placebo tests. The placebo tests following MaxEnt report small survival differences between the PAC patients in the RCT (reweighted) versus the NRS settings, overall and for most subgroups. For the overall sample 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. 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 weighting, and as there is also sufficient power, and so the placebo tests are passed.

Figure 5: Population Treatment Effects on Hospital Survival Rates

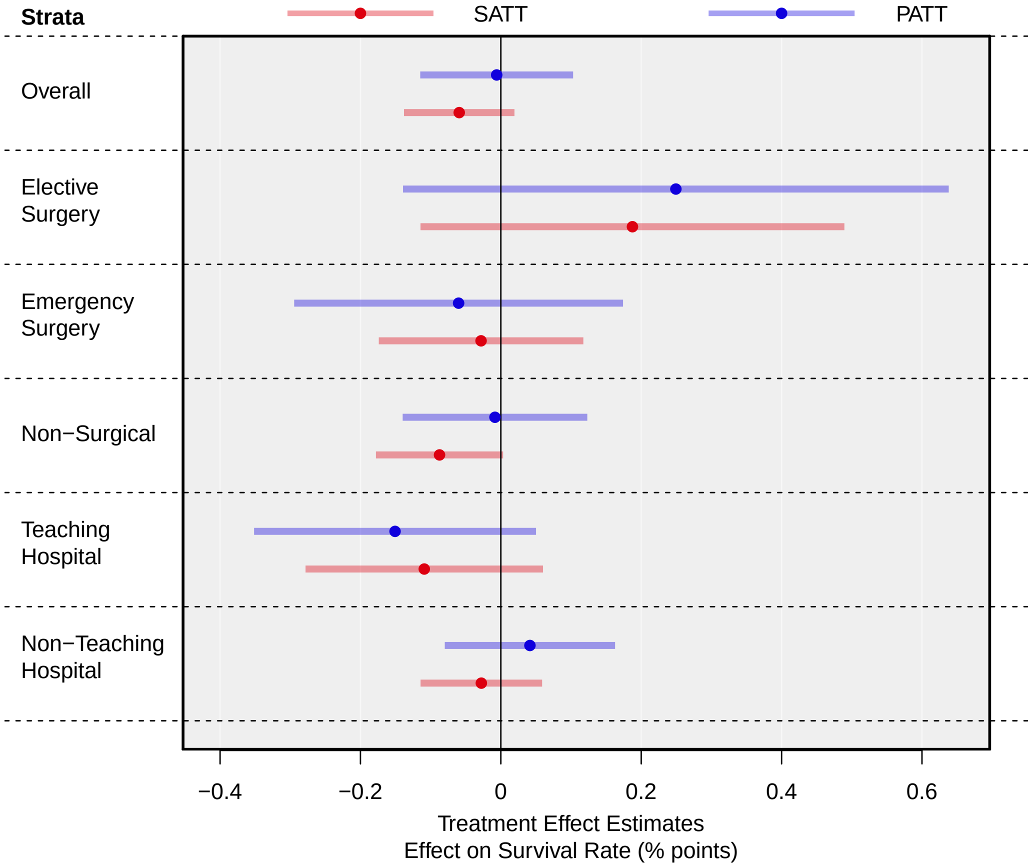
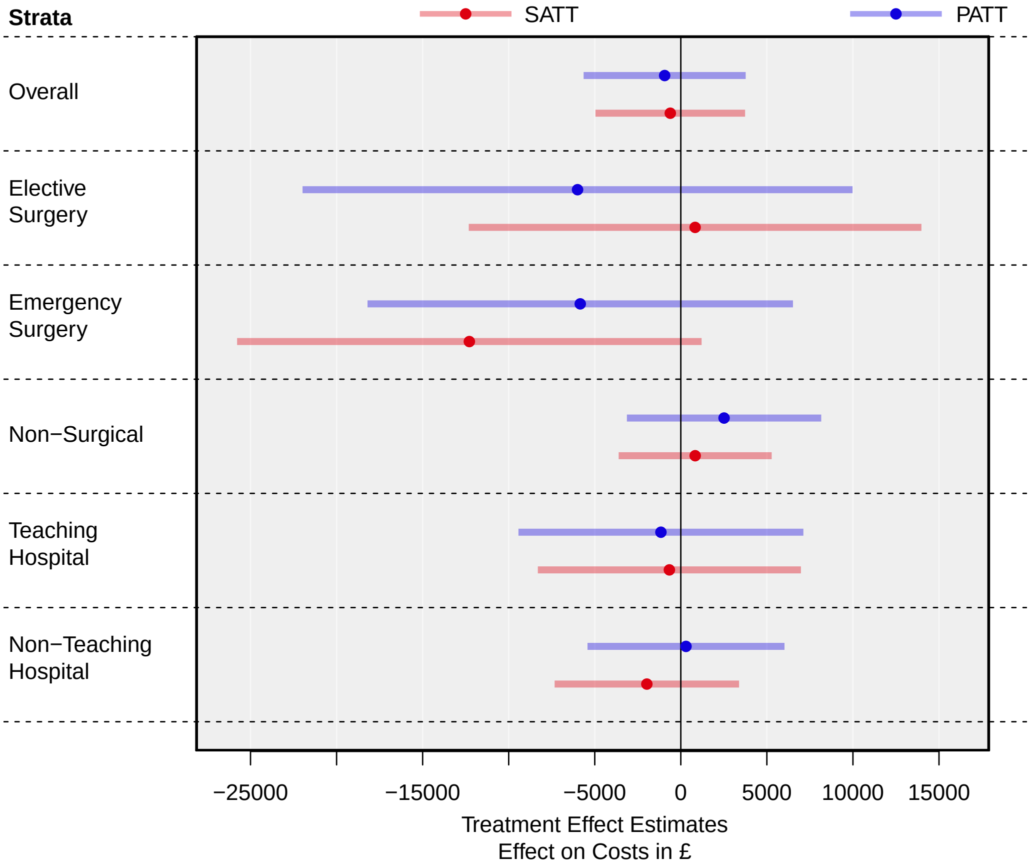


Figure 6: Population Treatment Effects on Costs (£)

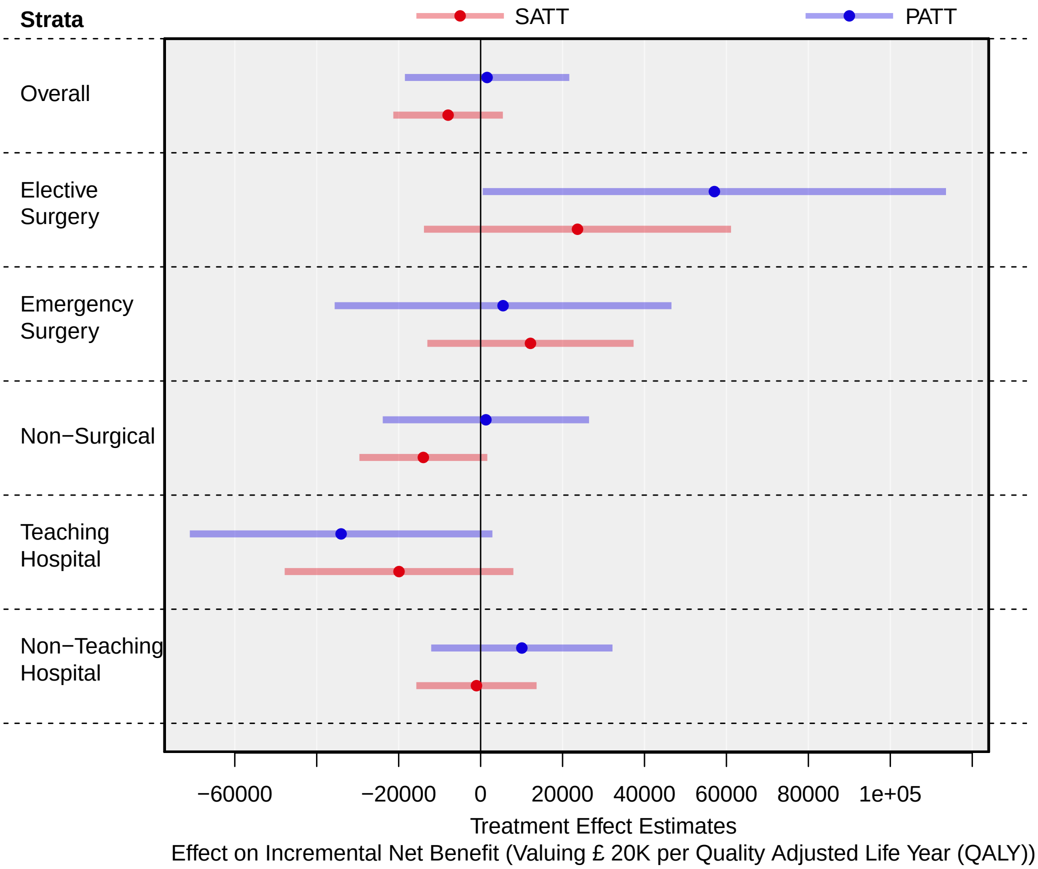


6.3 Population Estimates in the PAC example

We report SATT estimated from the matched RCT data, and PATT after using the MaxEnt weights to adjust the SATT estimates, but as IPSW did not pass the overall placebo tests, these results are not presented. The 95% confidence intervals (CIs) reported use the standard errors obtained by subsampling (Figures 5–7). 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 reported a negative INB for the SATT, but for the PATT, the INB was positive. This finding suggests that 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.

Theorem 1 focuses on how to estimate PATT from RCT data, however the theorem can be applied to estimate other causal quantities of interest. This section discusses an estimator that makes slightly weaker assumption by using the population treated observations directly, but no placebo test is then available.

Figure 7: Population Treatment Effects on net monetary benefits



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

$$\tau_{PATT} = \mathbb{E}(Y_i|S_i = 0, T_i = 1) - \mathbb{E}_{01}\{\mathbb{E}(Y_i|W_i^{C_T}, S_i = 1, T_i = 0)\}, \tag{8}$$

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:

$$\tau_{PATT_{DID}} = \mathbb{E}_{01}\{\mathbb{E}(Y_i|W_i^T, S_i = 1, T_i = 1) - \mathbb{E}(Y_i|W_i^T, S_i = 1, T_i = 0)\} \tag{9}$$

$$-[\mathbb{E}_{01}\{\mathbb{E}(Y_i|W_i^T, S_i = 1, T_i = 1)\} - \mathbb{E}(Y_i|S_i = 0, T_i = 1)], \tag{10}$$

assuming $W_i^T = W_i^{C_T}$. 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.

Appendix B discusses alternative population estimands of interest to policy makers.

8 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 that follow directly from the conceptual framework and can assess whether the requisite assumptions are satisfied. These placebo tests contrast the reweighted RCTs endpoints, with those of the target population represented, for example by a NRS. The general framework is illustrated with estimation strategies that reweight the matched RCT data, according to IPSW or maximum entropy reweighting, but could also exploit alternative estimation strategies including double-robust estimators (Green and Kern, 2012). 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 (Stuart et al., 2011; Green and Kern, 2012; Imai, King and Stuart, 2008; Kline and Tamer, 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. Here, the target population is best defined by those who would meet the criteria treatment 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. 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.

In our illustration the sample average treatment effects were estimated from one RCT, and individual-level data from a single NRS were taken to represent the target population. More generally, the framework can assess the underlying assumptions with reweighting individual participant data from several RCTs to a target population. An advantage of having individual-level data for the target population is that the individual stratum from the RCT can be reweighted, not just according to the population means, but also to other moments of the distribution, such as the variance. While the proposed framework still applies to situations where only aggregated information is available for the target population, this implies further constraints to the reweighting approach. Firstly, the RCT data can only be reweighted according to aggregate statistics for the covariates reported. Secondly, while MaxEnt reweighting can be undertaken with summary data (Mattos and Viega, 2004), IPSW requires individual-level data from the NRS. More generally, population treatment effects can be required for several target populations, for example according to geographical location or time period. The proposed framework can then assess whether or not the requisite assumptions are met for each setting of policy interest.

This framework complements the move to RCTs with pragmatic designs, which requires that the participants and treatments included, represent those in the target population (Tunis, Stryer and Clancy, 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. Where the placebo test results suggested that the underlying assumptions were violated, for some subgroups, for example teaching hospitals, the divergent point estimates suggested this was because of unobserved differences between the settings. For other subgroups, for example patients having elective surgery, the point estimates were similar, but the small sample size meant the placebo test had insufficient power. More generally RCTs apply restrictive exclusion criteria, or treat according to more rigid treatment protocols than would be applied in routine practice (Rothwell, 2005). Here it would be anticipated that the assumptions pertaining to both the consistency of treatment, and unobserved differences in the populations would be violated; the placebo test would indicate the likely bias in the treatment effects estimated for the target population.

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, and the systematic differences between the RCT and the target population (Greenland, 2005), but also the uncertainty in estimating the weights. A recommended approach is to undertake subsampling (Politis and Romano, 1994), and use the algorithm described in Bickel and Sakov (2008) to select the size of the subsample. It is anticipated that, as in the PAC example, when the treatment effects are estimated for the population rather than the sample, there will be increased uncertainty. An advantage of the proposed approach is that this additional uncertainty in estimating population treatment effects is made explicit. Future studies should anticipate the additional uncertainty at the design stage when developing the sampling strategy, for example in estimating the sample size required for estimating treatment effects for the population of interest.

The proposed framework warrants careful consideration in other settings for estimating population treatment effects, and the paper raises the following areas for further research. First, the framework could be considered further in assessing the underlying assumptions, when estimating other parameters of interest (e.g. PATC, PATE), for alternative target populations, and with NRS data available at different levels of aggregation. Second, further research is required to consider how the proposed framework could be useful 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. Third, the current paper illustrates an approach for reweighting evidence from head-to-head RCTs, but the framework can extend to those settings where indirect or mixed treatment comparisons are required, and there is a common comparator, for example usual care. Here the placebo tests can assess whether the underlying assumptions for estimating population treatment effects are met, by contrasting the reweighted endpoints from each RCT with those of the target population, for the common comparator (e.g. usual care). Fourth, there will be settings where the requisite assumptions for estimating population treatment effects are not met. Further research is required to examine how beest to reduce the inevitable bias in the estimated population treatment effects.

References

Abadie, Alberto and Guido Imbens. 2006. "On the Failure of the Bootstrap for Matching Estimators." Working Paper.

Benjamini, Yoav and Yosef 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, Peter J. and Anat 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, Leo. 2001. "Random Forests." *Machine Learning* 45(1):5–32.

Chipman, Hugh A., Edward I. George and Robert E. McCulloch. 2006. "BART: Bayesian Additive Regression Trees.".

Chittock, Dean R, Vinay K Dhingra, Juan J Ronco, James A Russell, Dave M Forrest, Martin Tweeddale and John C Fenwick. 2004. "Severity of illness and risk of death associated with pulmonary artery catheter use." *Critical Care Medicine* 32:911–915.

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

Cole, Stephen R and Elizabeth 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, Alfred F, Theodore S Speroff, Neal V Dawson, Charles Thomas, Frank E Harrell, Douglas Wagner, Norman Desbiens, Lee Goldman, Albert W Wu, Robert M Califf, William J Fulkerson, Humberto Vidaillet, Steven Broste, Paul Bellamy, Joanne Lynn and William 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, James E. 2001. "The Pulmonary Artery Catheter—Friend, Foe, or Accomplice?" *Journal of the American Medical Association* 286:348–350.

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

Diamond, Alexis and Jasjeet 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, Simon and Anthony Delaney. 2006. "Pulmonary artery catheters as currently used, do not benefit patients." *British Medical Journal* 333:930–1.

Gheorghe, Adrian, Tracy E. Roberts, Jonathan C. Ives, Benjamin R. Fletcher and Melanie Calvert. 2013. "Centre Selection for Clinical Trials and the Generalisability of Results: A Mixed Methods Study." *PLOS ONE* 8(2).

Green, Donald P. and Holger L. Kern. 2012. "Modeling Heterogeneous Treatment Effects in Large-Scale Experiments Using Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76(3):491–511.

Greenhouse, Joel B, Eloise E Kaizar, Kelly Kelleher, Howard Seltman and William 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, Sander. 2005. "Multiple-bias modelling for analysis of observational data." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 168(2):267–306.

Harrison, David A, Anthony R Brady and Kathy 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, Erin K. and F. Daniel Hidalgo. 2011. "What's the Alternative?: An equivalence approach to Placebo and Balance Tests." Working Paper.

Harvey, SE, CA Welch, DA Harrison and MA Singer. 2008. "Post hoc insights from PAC-Man—the UK Pulmonary artery catheter trial." *Critical Care* 35(6):1714–1721.

Harvey, Sheila, David A Harrison, Mervyn Singer, Joanne Ashcroft, Carys M Jones, Diana Elbourne, William Brampton, Dewi Williams, Duncan Young and Kathryn 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.

Heckman, James J and Edward Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica* 73(3):669–738.

Heckman, James J and Sergio Urzua. 2009. Comparing IV with Structural Models: What Simple IV Can and Cannot Identify. IZA Discussion Papers 3980 Institute for the Study of Labor (IZA).

Hoch, Jeff S, Andrew H Briggs and Andy 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.

Imai, Kosuke, Dustin Tingley and Teppei Yamamoto. 2013. "Experimental Design for Identifying Causal Mechanisms." *Journal of the Royal Statistical Society, Series A* 176(1):5–51.

Imai, Kosuke, Gary King and Elizabeth A. Stuart. 2008. "Misunderstandings among Experimentalists and Observationalists about Causal Inference." *Journal of the Royal Statistical Society, Series A* 171(2):481–502.

Imbens, Guido. 2009. Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009). NBER Working Paper 14896.

Kang, Joseph D. Y. and Joseph 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, Leslie. 1992. "Weighting for unequal P_i." *Journal of Official Statistics* 8:183–200.

Kline, Brendan and Elie Tamer. 2011. "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.

Liaw, Andy and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2(3):18–22.

Mattos, Rogerio and Alvaro Viega. 2004. "Entropy Optimization: Computer implementation of the maxent and minxent principles.".

Miratrix, Luke W., Jasjeet S. Sekhon and Bin 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, Nandita and Alka Indurkhya. 2005. "A propensity score approach to estimating the cost-effectiveness of medical therapies from observational data." *Health Economics* 14:805–815.

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

Nixon, Richard M and Simon G Thompson. 2005. "Incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations." *Health Economics* 14:1217—1229.

Politis, Dimitrius N. and Joseph P. Romano. 1994. "Large Sample Confidence Regions Based on Subsamples Under Minimal Assumptions." *The Annals of Statistics* 22:2031–2050.

Porter, Kristin E., Susan Gruber, Mark J. van der Laan and Jasjeet S. Sekhon. 2011. "The Relative Performance of Targeted Maximum Likelihood Estimators." *The International Journal of Biostatistics* 7(1).
URL: *http://www.bepress.com/ijb/vol7/iss1/31/*

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

Sakr, Yasser, Jean-Louis Vincent, Konrad Reinhart, Didier Payen, Christian J Wiedermann, Durk F Zandstra and Charles 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, Jasjeet S. 2009. "Opiates for the Matches: Matching Methods for Causal Inference." *Annual Review of Political Science* 12:487–508.

Sekhon, Jasjeet 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, Jasjeet S. and Richard Grieve. 2012. "A Non-Parametric Matching Method for Bias Adjustment with Applications to Economic Evaluations." *Health Economics* 21(6):695–714.

Shadish, William R, Thomas D Cook and Donald Thomas Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Shannon, Claude E. 1948. "A Mathematical Theory of Communication." *The Bell Systems Technical Journal* 27:379–423, 623–656.

SR, Cole and CE Frangakis. 2009. “The consistency statement in causal inference a definition or an assumption?” *Epidemiology* 20(1):3–5.

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, Elizabeth A., Stephen R. Cole, Catherine P. Bradshaw and Philip J. Leaf. 2011. “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, Sean R, Daniel B Stryer and Carolyn M Clancy. 2003. “Practical clinical trials.” *JAMA: the journal of the American Medical Association* 290(12):1624–1632.

Turner, Rebecca M, David J Spiegelhalter, Gordon Smith and Simon G Thompson. 2009. “Bias modelling in evidence synthesis.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 172(1):21–47.

Wellek, Stefan. 2010. *Testing Statistical Hypotheses of Equivalence and Noninferiority.*

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

Willan, Andrew R, Andrew H Briggs and Jeffrey S Hoch. 2004. “Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data.” *Health Economics* 13:461–475.

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

Willan, Andrew R., Eric Bingshu Chen, Richard 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.

A Proof of Theorem 1

Proof. From (2) and (4)

$$\begin{aligned}\mathbb{E}(Y_{i01}|S_i=0,T_i=1) &= \mathbb{E}(Y_{i11}|S_i=0,T_i=1)\\ &= \mathbb{E}_{01}\{\mathbb{E}(Y_{i11}|W_i^T,S_i=1,T_i=1)\}\\ &= \mathbb{E}_{01}\{\mathbb{E}(Y_i|W_i^T,S_i=1,T_i=1)\}.\end{aligned} \tag{11}$$

From (3) and (5)

$$\begin{aligned}\mathbb{E}(Y_{i00}|S_i=0,T_i=1) &= \mathbb{E}(Y_{i10}|S_i=0,T_i=1)\\ &= \mathbb{E}_{01}\{\mathbb{E}(Y_{i10}|W_i^{C_T},S_i=1,T_i=0)\}\\ &= \mathbb{E}_{01}\{\mathbb{E}(Y_i|W_i^{C_T},S_i=1,T_i=0)\}.\end{aligned} \tag{12}$$

The result follows by substituting Eqs. (11) and (12) in the quantity of interest τ_{PATT} in Equation (1). Without strong ignorability of sample assignment, from (6),

$$\begin{aligned}&\mathbb{E}(Y_{i01}|S_i=0,T_i=1)-\mathbb{E}(Y_{i00}|S_i=0,T_i=1)\\ &= \mathbb{E}_{01}\{\mathbb{E}(Y_{i11}|W_i^T,S_i=0,T_i=1)\}-\mathbb{E}_{01}\{\mathbb{E}(Y_{i10}|W_i^{C_T},S_i=0,T_i=1)\}\\ &= \mathbb{E}_{01}\{\mathbb{E}(Y_{i11}|W_i^T,S_i=1,T_i=1)\}-\mathbb{E}_{01}\{\mathbb{E}(Y_{i10}|W_i^{C_T},S_i=1,T_i=1)\},\end{aligned}$$

and the result follows from randomization. □

B 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

$$(Y_{i01},Y_{i11}) \perp\!\!\!\perp S_i|(W_i^T,T_i=0) \qquad \text{and} \qquad (Y_{i00},Y_{i10}) \perp\!\!\!\perp S_i|(W_i^{C_T},T_i=0), \tag{13}$$

respectively. Additionally, if Equation (3) in assumption 1, $(Y_{i00},Y_{i10}) \perp\!\!\!\perp S_i|(W_i^{C_T},T_i=0)$ and assumption 4 hold then $\mathbb{E}(Y_i|S_i=0,T_i=0)=\mathbb{E}_{00}\{\mathbb{E}(Y_i|W_i^{C_T},S_i=1,T_i=0)\}$. Therefore the mean outcomes would be the same for the control group in the target population and the adjusted RCT, adjusted such that $W_i^{C_T}$ 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 $\tau_{PATE}=\mathbb{E}(Y_{i01}-Y_{i00}|S_i=0)$. In such a case, assumptions 1–4, as well as $(Y_{i01},Y_{i11}) \perp\!\!\!\perp S_i|(W_i^{C_T},T_i=0)$ and $(Y_{i00},Y_{i10}) \perp\!\!\!\perp S_i|W_i^{C_T},T_i=0$, are sufficient for identification. Assuming $W_i^T=W_i^{C_T}$, these assumptions and randomisation imply that $Y_{ist} \perp\!\!\!\perp (S_i,T_i)|W_i^T$, which means that the potential outcomes for units with the same W_i^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 randomisation, it can be shown that

$$\mathbb{E}(Y_{ist}|S_i=0)=\mathbb{E}_{W_i^{C_T}|S_i=0}\{\mathbb{E}(Y_i|W_i^{C_T},S_i=1,T_i=t)\} \tag{14}$$

$$\tau_{PATE}=\mathbb{E}_{W_i^{C_T}|S_i=0}\{\mathbb{E}(Y_i|W_i^{C_T},S_i=1,T_i=1)-\mathbb{E}(Y_i|W_i^{C_T},S_i=1,T_i=0)\}, \tag{15}$$

for $t=0,1$. Equation 14 implies that the mean outcome in the target population is the same as in the adjusted RCT $T_i=t$ group, adjusted such that $W_i^{C_T}$ 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 $T_i = t$ 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.

C Equivalence Tests

Equivalence tests begin with the null hypothesis:

$$H_0 : \tfrac{\mu_{\text{adj samp}} - \mu_{\text{pop}}}{\sigma} \geq \epsilon_U \qquad \text{or} \qquad \tfrac{\mu_{\text{adj samp}} - \mu_{\text{pop}}}{\sigma} \leq \epsilon_L$$

$$\text{versus}$$

$$H_1 : \epsilon_L < \tfrac{\mu_{\text{adj samp}} - \mu_{\text{pop}}}{\sigma} < \epsilon_U$$

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:

$$|T| < C_{\alpha;m,n}(\epsilon)$$

$$\text{with}$$

$$C_{\alpha;m,n}(\epsilon) = F^{-1}(\alpha; df_1 = 1, df_2 = N - 2, \lambda^2_{nc} = mn\epsilon^2/N)^{\frac{1}{2}}$$

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 $\lambda^2_{nc} = mn\epsilon^2/N$. 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).

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 .[8] In this context this ensures that individuals in the intervention 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 population of interest 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.

[8] 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).

E Covariates used in Estimation

Table 2: Covariates used in GenMatch Estimation

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 unit, teaching hospital indicator, mechanical ventilator at admission, base excess
Additional Covariates used for Matching	
	physiology score, admission diagnosis, gender, 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 3: Covariates used in MaxEnt Estimation

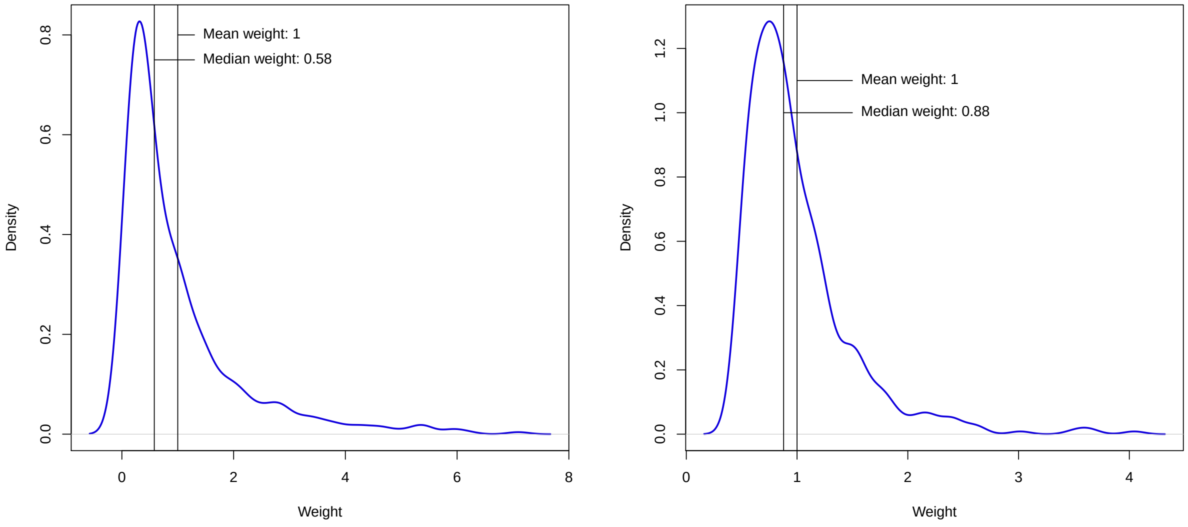
MaxEnt Margins	
	age, elective surgery indicator, emergency surgery indicator, teaching hospital indicator, gender, baseline probability of death, mechanical ventilator at admission, chemical measure of decline, history variables on cardiac, respiratory, liver, and immune measures, categorical variables 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 × age categories, unit size × teaching hospital indicator, teaching hospital indicator × base excess categories, mechanical ventilation × base excess categories, teaching hospital indicator × mechanical ventilator at admission, unit size × mechanical ventilator at admission, gender × teaching hospital, teaching hospital × age categories, gender × age categories, emergency surgery indicator × gender, elective surgery indicator × gender, teaching hospital indicator × history variables on cardiac, respiratory, liver, and immune measures, age categories × base excess × gender, gender × history variables on cardiac and respiratory measures, mechanical ventilation at admission × history variables on cardiac, respiratory, and renal measures

Table 4: Covariates used in IPSW Estimation

Propensity Score Covariates
gender, age, categorical age variables, elective surgery indicator, emergency surgery indicator, 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

F Weights

Figure 8: Weight Distributions



(a) Maxent Weights

(b) IPSW Weights

G Bootstrapped Placebo Tests

The MaxEnt equivalence based placebo tests were bootstrapped along with the *p*-values and standard errors for the SATT and PATT estimates. The following table shows the bootstrapped *p*-values for the survival, cost, and net-benefit placebo tests. As discussed in section C, the equivalence tests here are designed with the null hypothesis that adjusted RCT treated are not equivalent to the treated in the target population. Therefore, the bootstrapped *p*-value is calculated as the frequency with which the adjusted RCT estimate falls outside a 0.2 standard deviation range around the mean population treated value. The following table presents the bootstrapped *p*-values and associated Benjamini-Hochberg FDR correction.

Table 5: Bootstrapped MaxEnt Placebo Test *p*-values

			Survival		**Cost**		**Net Benefit**	
Stratum	**RCT treated** *n*	**NRS treated** *n*	***p*-value**	**BH FDR *p*-value**	***p*-value**	**BH FDR *p*-value**	***p*-value**	**BH FDR *p*-value**
Overall	506	1051	0.034	0.18	0	0	0.0041	0.03
Elective Surgery	32	98	0.46	0.46	0.35	0.35	0.46	0.46
Emergency Surgery	142	243	0.38	0.43	0.17	0.22	0.19	0.21
Non-surgical	332	710	0.088	0.18	0.012	0.050	0.018	0.035
Teaching Hospital	110	447	0.22	0.29	0.26	0.30	0.12	0.16
Non-Teaching Hospital	396	604	0.054	0.18	0.066	0.11	0.0083	0.033

H Endpoints and p-values

Table 6: Treatment effect estimates using MaxEnt method. Standard errors are calculated using subsampling, and FDR p-values use the Benjamini-Hochberg method.

	SATT	PATT	PATT	PATT
	Estimate	Estimate	Naive p-value	FDR p-value
Survival Estimates				
Overall	-0.059	-0.0059	0.72	0.97
Elective Surgery	0.13	0.36	0.047	0.38
Emergency Surgery	-0.098	-0.13	0.41	0.97
Non-Surgical	-0.069	0.016	0.77	0.97
Teaching Hospital	-0.036	-0.0051	0.85	0.97
Non-Teaching Hospital	-0.013	0.093	0.41	0.97
Cost Estimates				
Overall	-611.9	-939.0	0.83	0.99
Elective Surgery	-87.5	-5437.5	0.95	0.99
Emergency Surgery	-6280.3	-5654.9	0.42	0.99
Non-Surgical	99.5	1294.7	0.48	0.99
Teaching Hospital	4385.9	1607.1	0.74	0.99
Non-Teaching Hospital	229.9	2023.3	0.88	0.99
Cost-Effectiveness Estimates				
Overall	-7936.7	1580	0.63	0.92
Elective Surgery	16641	68092	0.016	0.13
Emergency Surgery	-4800.1	-1893.3	0.84	0.92
Non-Surgical	-7116.5	7587.7	0.58	0.92
Teaching Hospital	-13928	-4322	0.71	0.92
Non-Teaching Hospital	941.6	18791	0.27	0.92