ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2015)
178, *Part* 3, *pp.* 757–778

From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects

Erin Hartman,

University of California at Berkeley, USA

Richard Grieve

London School of Hygiene and Tropical Medicine, UK

Roland Ramsahai

University of Cambridge, UK

and Jasjeet S. Sekhon

University of California at Berkeley, USA

[Received December 2013. Final revision August 2014]

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 the assumptions that are 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 to adjust the RCT data. This approach is considered in a cost-effectiveness analysis of a clinical intervention: pulmonary artery catheterization.

Keywords: Causal inference; Cost-effectiveness studies; External validity; Observational studies; Placebo tests; Randomized controlled trials

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 maximize internal validity. However, policy makers require evidence on the relative effectiveness and cost-effectiveness of interventions for target populations that usually differ from 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,

Address for correspondence: Richard Grieve, Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, 15–17 Tavistock Place, London, WC1H 9SH, UK.
E-mail: Richard.Grieve@lshtm.ac.uk

2009; Heckman and Vytlacil, 2005; Hotz *et al.*, 2005; Imbens, 2009). In RCTs, treatment protocols and interventions differ from those administered routinely, and trial participants—e.g. 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 the conditions under which RCTs can identify population treatment effects in combination with observational data, and we develop methods to test whether 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 rigour 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 must 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 various research designs for estimating population treatment effects. These research designs can be used with a variety of 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), which is an invasive and controversial cardiac monitoring device that is used in critical care. Although 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 (Diamond and Sekhon, 2013; Sekhon and Grieve, 2012), to create matched strata within the RCT. We use maximum entropy 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 that are required for identifying population average treatment effects. Section 4 describes the placebo tests for checking the underly-

ing assumptions, and 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

PAC is a cardiac monitoring device that is 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 that 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 1014 subjects who were 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 cost-effectiveness analysis 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 case mix programme database. The database contains information on case mix, patient outcomes and resource use for about 1.5 million admissions to 250 critical care units in the UK (Harrison *et al*., 2004). A total of 57 units from the case mix programme collected additional prospective data on PAC use for consecutive admissions between May 2003 and December 2004. Over this time period, 10 units recorded no PAC use and were excluded from this analysis, as were units participating in the RCT (the PAC-Man study). The RCT data that were used exclude one participant for whom no end point data were available. The NRS applied the same inclusion and exclusion criteria for individual patients as the corresponding PAC-Man study, which resulted in a sample of 1052 PAC cases and 31447 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 use only information from the 1052 patients who received PAC in routine clinical practice, and from 1013 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 that are common to both the RCT and the NRS settings, and includes those covariates which are expected to modify the effect of PAC. For a centre to participate, the PAC-Man study required that local clinicians were in equipoise about the potential benefits of the intervention (Harvey

Table 1. Baseline characteristics and end points for the PAC-Man study, and for patients in the NRS who received PAC†

	Results for RCT		*Results for NRS, PAC, n = 1052*
	No PAC, n = 507	*PAC, n = 506*	
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 (standard deviation) baseline probability of death	0.55 (0.23)	0.53 (0.24)	0.52 (0.26)
Mean (standard deviation) 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)
Intensive care unit size (beds)			
5 or fewer	57 (11.2)	59 (11.7)	79 (7.5)
6–10	276 (54.4)	272 (53.8)	433 (41.2)
11–15	171 (33.7)	171 (33.8)	303 (28.8)
End points			
Deaths in hospital	333 (65.9)	346 (68.4)	623 (59.3)
Mean hospital cost (£)	19078	18612	19577
Standard deviation hospital cost (£)	28949	23751	24378

†The numbers in parentheses are percentages unless stated otherwise.

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 in a measure of cost-effectiveness such as the incremental net monetary benefit (Willan *et al.*, 2003; Willan and Lin, 2001). Net monetary benefits can be calculated by weighting each life year by using a quality adjustment anchored on a scale from 0 (death) to 1 (perfect health), to report quality-adjusted life years for each treatment. Then net monetary benefits for each treatment group can be calculated by multiplying the quality-adjusted life year by an appropriate threshold willingness to pay for a quality-adjusted life year gain (e.g. the threshold recommended by the National Istitute for Health and Care Excellence in England and Wales is £20000–30000 to gain a quality-adjusted life year), and subtracting the cost. Finally, the incremental net monetary benefit of the new treatment can be estimated by contrasting the mean net monetary benefits for each alternative.

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 that are required for estimating PATT in this context.

3. Identifying the population average treatment effect on the treated from a randomized controlled trial

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 identifying 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. When there is non-random attrition, causal effects even for the experimental sample cannot be estimated without additional assumptions. We define S_i as a sample indicator, taking value s, and T_i as a treatment indicator taking 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 $W_i^{\mathrm{C_T}}$ 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, in the RCT sample is defined as

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

where the expectation is over the (random) units $S=1$ (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 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, may differ from SATE. When treament assignment is ignorable, they are

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

where $t=0$ for τ_{SATC} and $t=1$ for τ_{SATT}. SATT estimates the average treatment effect conditionally on the distribution of potential outcomes under treatment, and SATC estimates the average treatment effect conditionally 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. Note that the treatment effects that are discussed here refer to infinite populations and samples, whereas Imai *et al.* (2008) referred to treatment effects in infinite populations as superpopulation effects.

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 conditionally on the distribution of potential outcomes under control, and the population average treatment effect on treated, PATT, as the treatment effect conditionally on the distribution of potential outcomes under treatment:

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

Our main quantity of interest is equation (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.

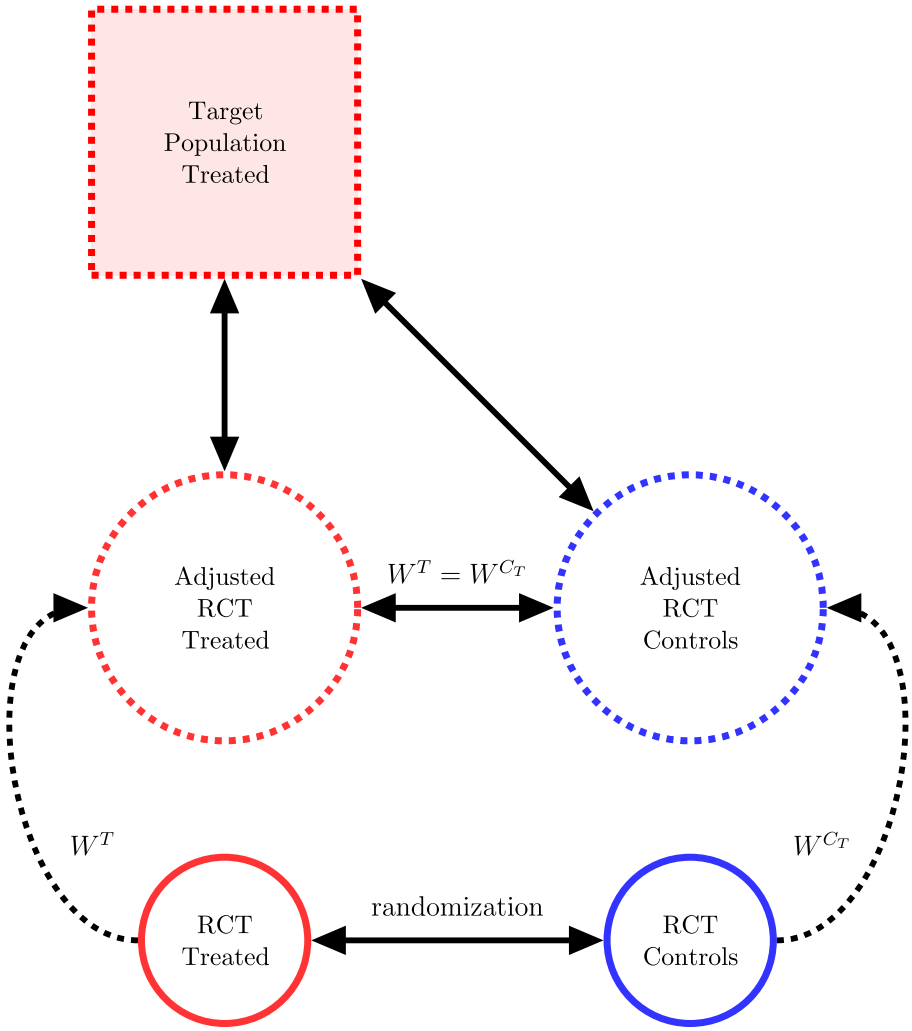

Fig. 1. Schematic diagram showing adjustment of sample effect to identify population effect: ↔, exchangeability of potential outcomes; ⇢, adjustment of the covariate distribution

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. Fig. 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, e.g. 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 to 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).

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

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

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

for $s = 0, 1$. The expectation $\mathbb{E}_{01}\{\cdot\}$ is a weighted mean of the W^{T} specific means, $\mathbb{E}(Y_{s1} | W^{\mathrm{T}}, S = 1, T = 1)$, with weights according to the distribution of W^{T} in the treated target population, $\Pr(W^{\mathrm{T}} | S = 0, T = 1)$. Essentially, on the right-hand 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. Fig. 1 illustrates this process with the single arrow from the RCT treated in the full red circle, to the adjusted group in the broken red circle. The adjustment can be performed with the weighting methods that are discussed in Section 5.

The right-hand side of equation (4) is the expectation in the adjusted RCT treated group, depicted as the broken red circle in Fig. 1. The left-hand side of equation (4) is the expectation in the treatment group in the target population, depicted as the broken red square in Fig. 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 Fig. 1, the double arrow between the broken 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).

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

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

Assumption 3 implies that

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

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

The right-hand side of equation (5) is the expectation in the adjusted RCT control group, which is depicted as the broken blue circle in Fig. 1. The left-hand side of equation (5) is the

expectation in the treated group in the target population, which is depicted as the broken red square in Fig. 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)

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

where L_j is the treatment and sample assignment vector for unit j. This is a stable unit treatment value assumption, 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 a proof given in Appendix A.

Theorem 1. Assuming that consistency and the stable unit treatment value assumption hold, if

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

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

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

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

From theorem 1, it is possible to identify τ_{PATT} from the adjusted RCT data alone. In Fig. 1, the adjusted experimental controls and treated units are exchangeable only if $W_i^{\mathrm{T}} = W_i^{C_{\mathrm{T}}}$. As Fig. 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 $W_i^{\mathrm{T}} = W_i^{\mathrm{C_T}}$, 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 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 that is subsequently chosen. From Section 3, if equation (2) in assumption 1, assumption 2 and assumption 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

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

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

A placebo test can be used to highlight the failure of several underlying assumptions, but it 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 end points, such as cost, that have a high variance. Cost-effectiveness analyses typically have both these features.

To address this concern, Hartman and Hidalgo (2011) introduced 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 end points for the treatment group in the RCT are not equivalent to those for the treatment group in the target population'. This null hypothesis of non-equivalence is only rejected if there is sufficient power. This alleviates the issues of confounding the notion of statistical equivalence with a tests relationship to sample size discussed in Imai *et al.* (2008). Hence, a low p-value would offer support for the identification strategy. The advantage of the test proposed 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 predefined. 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 uses weighting methods, such as inverse propensity score weighting (IPSW) (Stuart *et al.*, 2011) and maximum entropy weighting (Kullback, 1997;

Jaynes, 1957), which rely on ancillary information, e.g. from an NRS, to reweight the RCT data. The other prominent approach is to estimate the response surface by using the RCT data and to extrapolate this response surface to the target population, and includes methods such as the Bayesian additive regression tree (BART) method (Chipman *et al.*, 2010), classification and regression trees (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 $W^{\mathrm{C_T}}$ can either use weights from the first class of estimators, or predicted values from a response surface model. Either way, to identify the population estimand of interest, the estimation strategy must pass the placebo tests proposed.

Although 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 SATT overall and by prespecified 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 randomized controlled trial

We create matched pairs within the RCT data, by matching controls to treated units within the RCT by using genetic matching to maximize the balance between the randomized groups (Diamond and Sekhon, 2013; Sekhon, 2011). We recommend including, in the matching algorithm, those covariates that are expected to influence not only the end points, but also the selection of patients into the RCT. Covariates that are related to the selection into the RCT are part of the conditioning sets W^{T} and $W^{\mathrm{C_T}}$, and therefore care should be taken to ensure that these covariates are balanced.

5.2. Weighting methods

We focus on a reweighting approach, maximum entropy, that can be applied when either summary or individual data are available for the target population. This approach goes back to at least Jaynes (1957), Kullback (1997) and Ireland and Kullback (1968) and has much in common with method-of-moments estimators (e.g. Hansen (1982) and Hellerstein and Imbens (1999)). In brief, this approach does not assume that the propensity score is correctly specified; nor does it make additional assumptions about the distribution of weights. Under maximum entropy, the cell weights, marginal distributions or other population moments for the conditioning covariates, W^{T}, are used as constraints. It ensures that the weights chosen for the matched pairs sum to 1, but simultaneously satisfy the maximum entropy constraints given by the population characteristics. See appendix D in the on-line supporting information for more details. IPSW is considered in the on-line supporting information (appendix H).

6. Empirical example: pulmonary artery catheterization

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

6.1. Matching and weighting in the pulmonary artery catheterization example

We used genetic matching to create matched pairs within the RCT data, by matching a control

unit to each treated unit. The matching algorithm included those covariates that are expected to influence the selection of patients into the RCT and the end points. (See Table E.1 in the on-line supporting information appendix E.) The loss function was specified to require that balance, according to t-tests and Kolmogorov–Smirnov tests, was not made worse on those covariates that were expected to be of high prognostic importance after matching. Genetic matching matched 1–1 with replacement using a population size of 5000. Matching was repeated within each subgroup to report SATT at the subgroup level. Note that the aggregated subgroup estimates may not be equivalent to the overall estimate because different matches are used for the overall and subgroup estimates. Variance estimates were calculated conditionally 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 was well balanced after matching according to both t-tests and Kolmogorov–Smirnov tests, as shown in Fig. F.1 in the on-line supporting information appendix F.

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

We use maximum entropy 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 that are listed in Table E.2 in the on-line supporting information appendix E. For each covariate that was 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 by using subsampling (Politis and Romano, 1994). Abadie and Imbens (2008) showed that the bootstrap is not valid for estimating the standard error of a matching estimator but noted that subsampling (Politis and Romano, 1994) is valid. We used the algorithm that was 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. (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 $m = n$.)

Maximum entropy provided weights that were reasonably stable (see Fig. G.1 in the on-line supporting information appendix G), with a mean weight of 1 and a maximum 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 end points for the PAC patients in the NRS with the adjusted means for the PAC patients in the RCT. The results are reported in Fig. 2 for all three end points: survival rates (black), cost (red) and incremental net monetary benefit (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 by using the Benjamini–Hochberg method (Benjamini and Hochberg, 1995). Following maximum

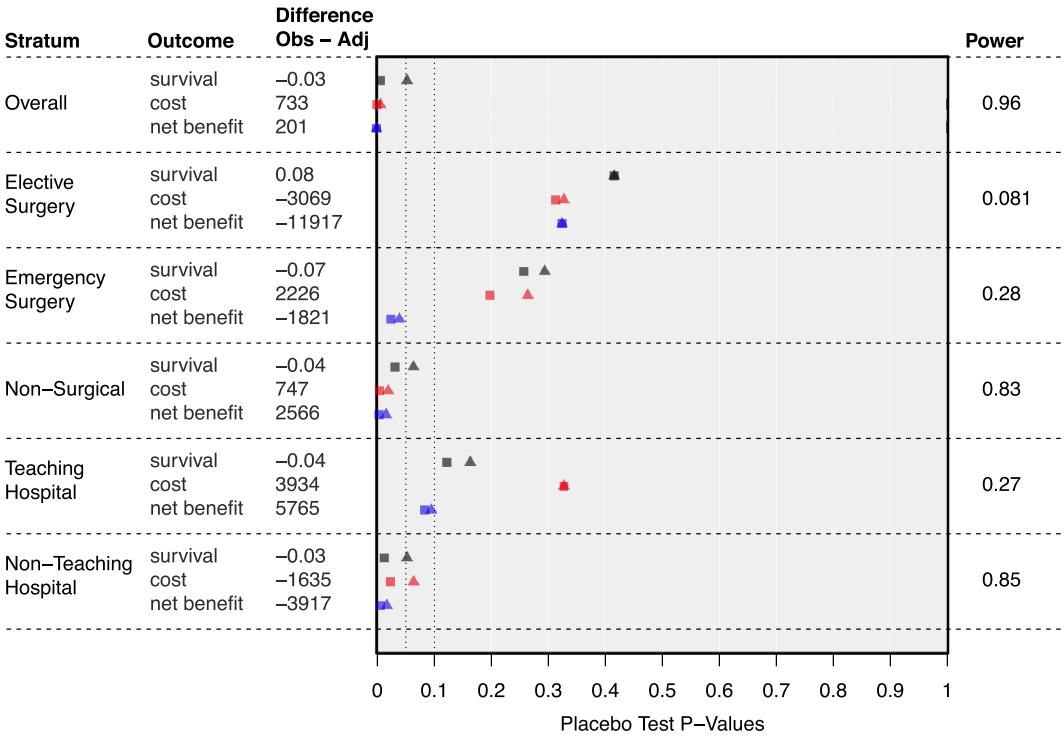

Fig. 2. Maximum entropy placebo tests: results of the equivalence tests comparing the mean outcome of NRS treated to the reweighted mean of the RCT treated; the column labelled 'Difference' presents the difference between the observed outcomes for the PAC group in the NRS and the PAC group in the RCT after reweighting; the *p*-values presented are before (■, ■, ■) and after (▲, ▲, ▲) false discovery rate adjustment; the column labelled 'Power' presents the power of the equivalence *t*-test for each stratum

entropy, 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. For some subgroups (teaching hospitals and 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 the on-line supporting information appendix H.

6.3. Population estimates in the pulmonary artery catheterization example

We report SATT estimated from the matched RCT data, and PATT after using the maximum entropy weights to adjust the SATT-estimates. The 95% confidence intervals are obtained by using subsampling (Figs 3–5). For the overall group, the PATT- and SATT-estimates are similar for each end point. 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 incremental net monetary benefit for SATT but, for PATT, the estimated incremental net monetary benefit was positive. This finding suggests that, for non-teaching hospitals in the target population, PAC was relatively cost effective. However, the confidence intervals for each

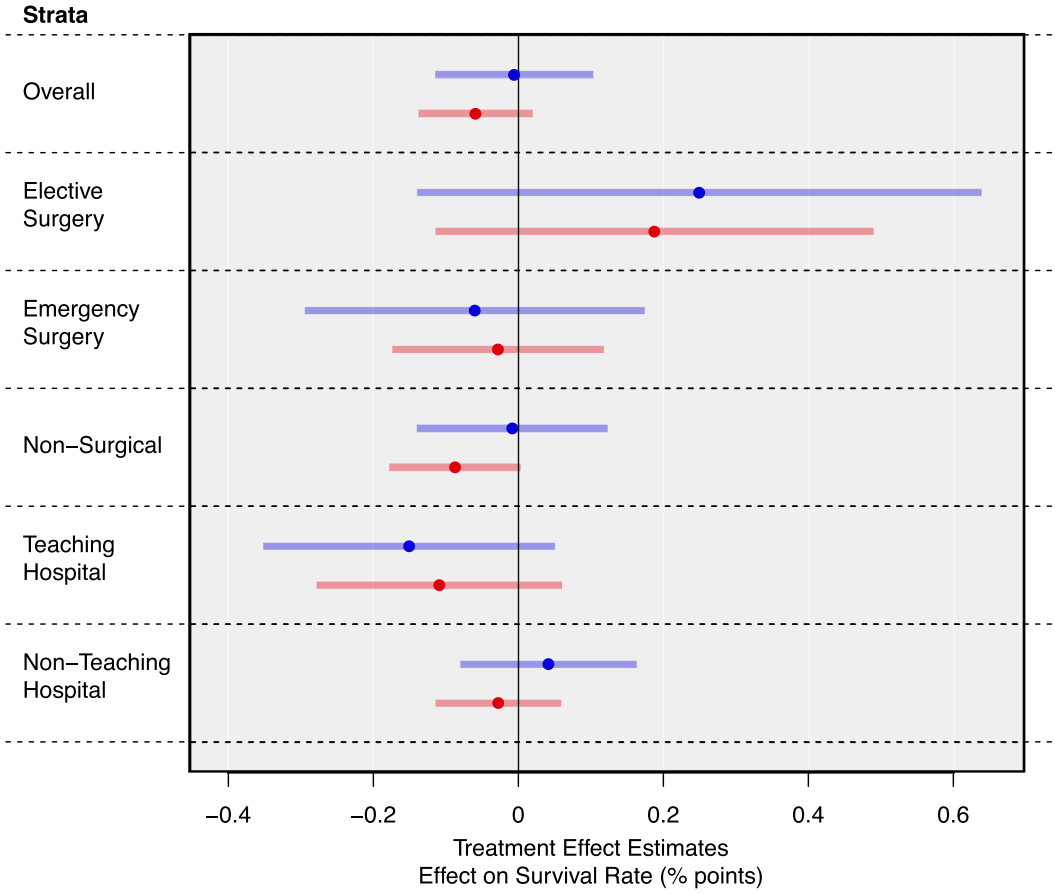

Fig. 3. Population treatment effects on hospital survival rates: ●, SATT; ●, PATT

estimate overlapped zero, and, in general, the confidence intervals for the PATT-estimates were wider than those for the corresponding SATT-estimates.

6.4. Response surface models

An alternative estimation strategy is to use response surface models to estimate covariate–end point relationships in the RCT, and to use these estimates to predict population treatment effects in the target population. For example, in the case of ordinary least squares 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 gains in efficiency relative to weighting approaches, especially if not all the covariates that are included in the adjustment are predictive of potential outcomes.

The placebo tests proposed 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.

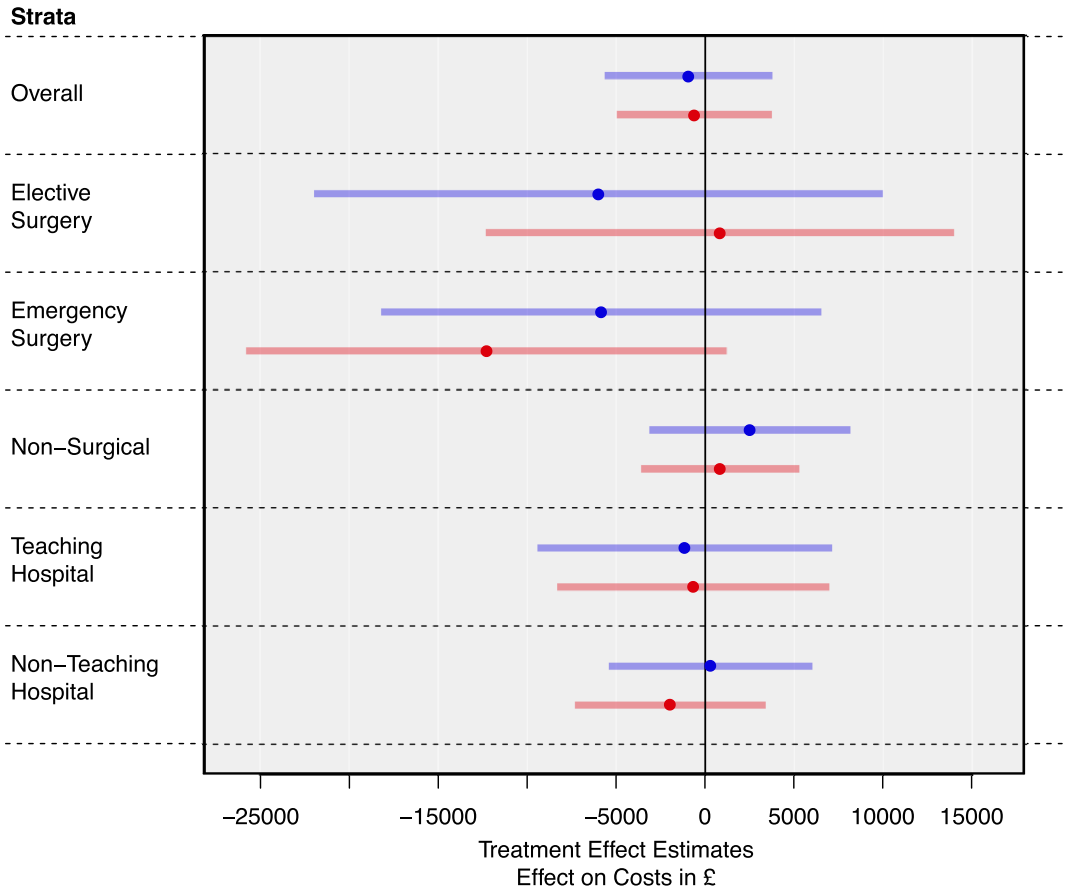

Fig. 4. Population treatment effects on costs: ●, SATT; ●, PATT

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

We apply the BART method in the PAC example, by estimating a response surface model on the patients who were randomized to receive PAC. We estimate a model for the relationship between the baseline characteristics and each end point (mortality, cost and net monetary benefit). We then predict the outcomes that the target population would have had, if it 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 Fig. I.5 in the on-line supporting information appendix I, this response surface modelling approach provided estimates that did not pass the requisite placebo tests for

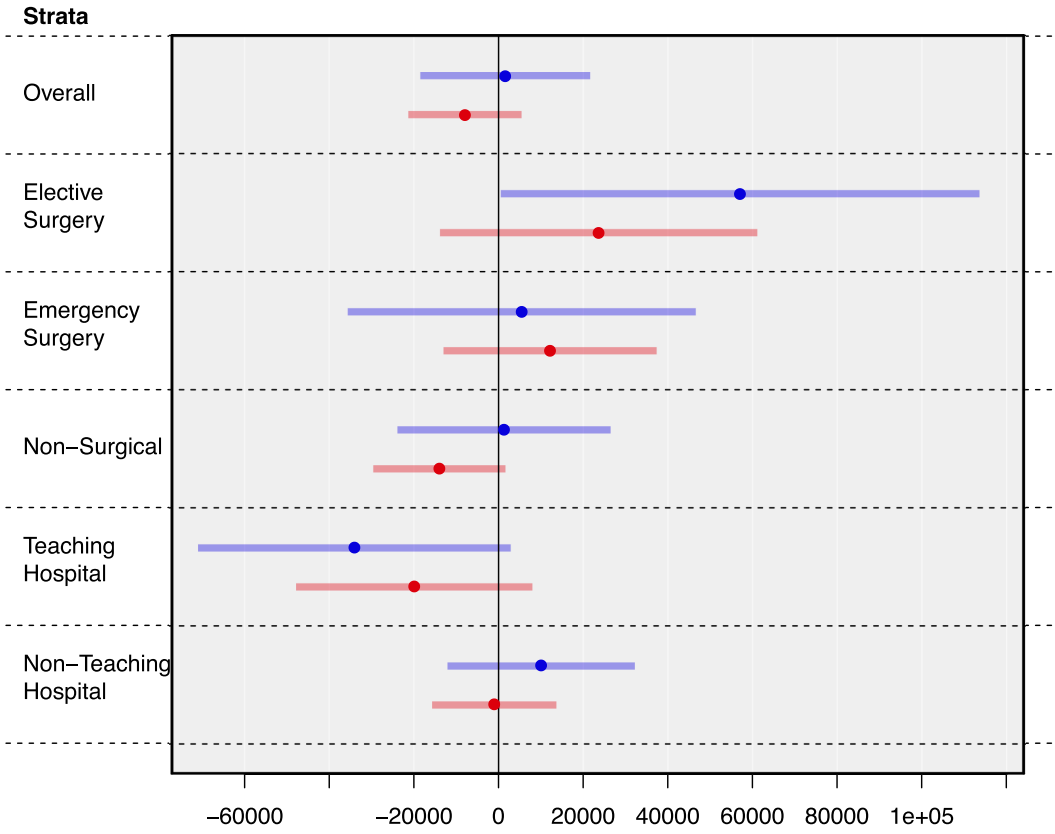

Fig. 5. Population treatment effects on incremental net monetary benefits calculated by valuing each quality-adjusted life year gain at a threshold of £20000 per quality-adjusted life year: ●, SATT; ●, PATT

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 that is 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

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

from equation (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. We can see this by rewriting equation (8) as

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

$$-[\mathbb{E}_{01}\{\mathbb{E}(Y|W^{\mathrm{T}}, S=1, T=1)\} - \mathbb{E}(Y|S=0, T=1)], \tag{10}$$

assuming that $W^{\mathrm{T}} = W^{\mathrm{C_T}}$. 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 a placebo test is no longer available to check whether the identifying assumptions hold. Hence, although 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) showed that all the estimands that we consider (e.g. PATE, PATT and PATC) are weighted averages of marginal treatment effects. The marginal treatment effect 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) showed that if, conditionally 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 marginal treatment effect conditions on unobserved utility. In our case, essential heterogeneity cannot occur in the RCT because there is full compliance, but the issue can arise 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 that we offer would be sensitive to this problem if it were present. In short, our approach is for the case where the marginal treatment effect is just a function of observed covariates, and we offer specification tests to help to assess whether this indeed is so.

Hotz *et al*. (2005) examined how the efficacy of worker training programmes differs from one location to another. They offered a formalization that is similar to ours, but there are key differences because of the set-up that they examined: they formalized the comparison of two randomized trials undertaken in different locations. The set-up 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 conducted placebo tests that contrast outcomes for controls across settings. They then conducted a placebo test contrasting end points for treatment *versus* treatment to assess whether treatment was homogeneous (conditional on passing the control–control placebo). This interpretation of the placebos and the set-up pertains to the setting with RCTs undertaken in different locations. Our theorem, placebo tests and estimators differ because we have an RCT and observational data on the target population.

Allcott (2011, 2014), and Allcott and Mullainathan (2012) found that treatment effects vary greatly across the experimental locations that they considered, 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) showed that in their setting non-experimental estimates have poor external validity, and that is worse than when non-experimental effects are predicted by using experimental results from another location. This is consistent with our set-up where we reweight the experimental estimand, and we do not resort to the observational estimator which is available to us.

Although 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 end points with those of the target population provided, for example, by an 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 that are 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 that are 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 PATT. The framework can be applied to other situations, e.g., 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 PATC. In these settings, the framework proposed 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 to ensure that the treatments that are 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 and 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 maximum entropy reweighting supported the underlying assumptions that are required for estimating PATT, overall and for some subgroups. For those subgroups, e.g. teaching hospitals where the placebo test results showed that 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 approach proposed encourages future studies to recognize fully 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 expected 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 end points 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 that are required to identify the complier average causal effect for the target population.

Acknowledgements

We gratefully acknowledge the constructive comments from the Associate Editor and two reviewers. We also thank Sheila Harvey (London School of Hygiene and Tropical Medicine), David Harrison (Intensive Care National Audit Research Centre) and Kathy Rowan (Intensive Care National Audit Research Centre) for access to data from the PAC-Man cost-effectiveness analysis and the Intensive Care National Audit Research Centre case mix programme database, and Chris McCabe and Katherine Stevens (Sheffield Centre for Health and Related Research) 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. This report is independent research supported by the National Institute for Health Research (Senior Research Fellowship, Dr Richard Grieve SRF-2013-06-016). The views expressed in this publication are those of the author(s) and not necessarily those of the National Health Service or the National Institute for Health Research. The authors are responsible for any errors.

Appendix A: Proof of theorem 1

From equations (2) and (4)

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

From equations (3) and (5)

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

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

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

and the result follows from randomization.

Appendix B: Equivalence tests

Equivalence tests begin with the null hypothesis:

$$H_0: \frac{\mu_{\text{adj samp}}-\mu_{\text{pop}}}{\sigma} \geqslant \varepsilon_{\mathrm{U}} \quad \text{or} \quad \frac{\mu_{\text{adj samp}}-\mu_{\text{pop}}}{\sigma} \leqslant \varepsilon_{\mathrm{L}}$$

versus

$$H_1: \varepsilon_{\mathrm{L}} < \frac{\mu_{\text{adj samp}}-\mu_{\text{pop}}}{\sigma} < \varepsilon_{\mathrm{U}}$$

where $\mu_{\text{adj samp}}$ is the true mean of the reweighted sample treated and μ_{pop} is the true mean of the population treated, and σ is the pooled standard deviation of the two groups. We define $\varepsilon_{\mathrm{L}}=0.2$ and $\varepsilon_{\mathrm{U}}=0.2$, 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 hypothesis of non-equivalence if

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

with

$$C_{\alpha;m,n}(\varepsilon)=F^{-1}(\alpha;\mathrm{df}_1=1,\mathrm{df}_2=N-2,\lambda_{nc}^2=mn\varepsilon^2/N)^{1/2}$$

where $C_{\alpha;m,n}(\varepsilon)$ is the square root of the inverse F-distribution with level α, degrees of freedom 1, $N-2$, and non-centrality parameter $\lambda_{nc}^2=mn\varepsilon^2/N$. 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 that is defined in Wellek (2010).

Appendix C: Identifiability of alternative causal quantities

The main population treatment effect that is 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

$$\begin{aligned}&(Y_{01},Y_{11}) \perp\!\!\!\perp S|(W^{\mathrm{T}},T=0),\\ &(Y_{00},Y_{10}) \perp\!\!\!\perp S|(W^{\mathrm{C_T}},T=0)\end{aligned} \tag{13}$$

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 assumptions required. However, this is not necessary to apply theorem 1 because expression (13) is not assumed in the current analysis.

In circumstances where the estimand of interest is PATE then the estimand of interest is the effect in the entire target population, where $\tau_{\text{PATE}} = \mathbb{E}(Y_{01} - Y_{00}|S=0)$. 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

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

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

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 that are required to justify equations (14) and (15). Stuart *et al.* (2011) verified 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) still to hold. This occurs if treatment consistency is violated but the potential outcomes in the target population and RCT differ by some constant.

The assumptions that are 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 that are used in the analysis here are weaker and do not imply equation (14), such a test is not done.

References

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

Allcott, H. (2011) Social norms and energy conservation. *J. Publ. Econ.*, **95**, 1082–1095.

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

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

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B, **57**, 289–300.

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

Breiman, L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Chipman, H. A., George, E. I. and McCulloch, R. E. (2010) Bart: Bayesian additive regression trees. *Ann. Appl. Statist.*, **4**, 266–298.

Chittock, D. R., Dhingra, V. K., Ronco, J. J., Russell, J. A., Forrest, D. M., Tweeddale, M. and Fenwick, J. C. (2004) Severity of illness and risk of death associated with pulmonary artery catheter use. *Crit. Care Med.*, **32**, 911–915.

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

Cole, S. R. and Stuart, E. A. (2010) Generalizing evidence from randomized clinical trials to target populations the actg 320 trial. *Am. J. Epidem.*, **172**, 107–115.

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

Dalen, J. E. (2001) The pulmonary artery catheter—friend, foe, or accomplice? *J. Am. Med. Ass.*, **286**, 348–350.

Deaton, A. (2009) Instruments of development: randomization in the tropics, and the search for the elusive keys to economic development. *Working Paper 14690*. National Bureau of Economic Research, Cambridge.

Diamond, A. and Sekhon, J. S. (2013) Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies. *Rev. Econ. Statist.*, **95**, 932–945.
Finfer, S. and Delaney, A. (2006) Pulmonary artery catheters as currently used, do not benefit patients. *Br. Med. J.*, **333**, 930–931.
Gheorghe, A., Roberts, T. E., Ives, J. C., Fletcher, B. R. and Calvert, M. (2013) Centre selection for clinical trials and the generalisability of results: a mixed methods study. *PLOS ONE*, **8**, no. 2, article e56560.
Greenhouse, J. B., Kaizar, E. E., Kelleher, K., Seltman, H. and Gardner, W. (2008) Generalizing from clinical trial data: a case study, the risk of suicidality among pediatric antidepressant users. *Statist. Med.*, **27**, 1801–1813.
Greenland, S. (2005) Multiple-bias modelling for analysis of observational data (with discussion). *J. R. Statist. Soc.* A, **168**, 267–306.
Hansen, L. P. (1982) Large sample properties of generalized method of moments estimators. *Econometrica*, **50**, 1029–1054.
Harrison, D. A., Brady, A. R. and Rowan, K. (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. *Crit. Care*, **8**, 99–111.
Hartman, E. K. and Hidalgo, F. D. (2011) What's the alternative?: an equivalence approach to placebo and balance tests. *Working Paper*. Department of Political Science, University of California at Berkeley, Berkeley.
Harvey, S., Harrison, D. A., Singer, M., Ashcroft, J., Jones, C. M., Elbourne, D., Brampton, W., Williams, D., Young, D. and Rowan, K. (2005) An assessment of the clinical effectiveness of pulmonary artery catheters in patient management in intensive care (pac-man): a randomized controlled trial. *Lancet*, **366**, 472–477.
Harvey, S., Welch, C., Harrison, D. and Singer, M. (2008) Post hoc insights from pac-man—the UK pulmonary artery catheter trial. *Crit. Care*, **35**, 1714–1721.
Heckman, J. J., Ichimura, H., Smith, J. and Todd, P. (1998) Characterizing selection bias using experimental data. *Econometrica*, **66**, 1017–1098.
Heckman, J. J. and Urzua, S. (2009) Comparing iv with structural models: What simple iv can and cannot identify. *Discussion Paper 3980*. Institute for the Study of Labor, Bonn.
Heckman, J. J., Urzua, S. and Vytlacil, E. (2006) Understanding instrumental variables in models with essential heterogeneity. *Rev. Econ. Statist.*, **88**, 389–432.
Heckman, J. J. and Vytlacil, E. (2005) Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, **73**, 669–738.
Hellerstein, J. K. and Imbens, G. W. (1999) Imposing moment restrictions from auxiliary data by weighting. *Rev. Econ. Statist.*, **81**, 1–14.
Hoch, J. S., Briggs, A. H. and Willan, A. R. (2002) Something old, something new, something borrowed, something blue: a framework for the marriage of econometrics and cost-effectiveness analysis. *Hlth Econ.*, **11**, 415–430.
Hotz, V. J., Imbens, G. and Mortimer, J. H. (2005) Predicting the efficacy of future training programs using past experiences at other locations. *J. Econmetr.*, **125**, 241–270.
Imai, K., King, G. and Stuart, E. A. (2008) Misunderstandings between experimentalists and observationalists about causal inference. *J. R. Statist. Soc.* A, **171**, 481–502.
Imai, K., Tingley, D. and Yamamoto, T. (2013) Experimental designs for identifying causal mechanisms (with discussion). *J. R. Statist. Soc.* A, **176**, 5–51.
Imbens, G. (2009) Better late than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *Working Paper 14896*. National Bureau of Economic Research, Cambridge.
Ireland, C. T. and Kullback, S. (1968) Contingency tables with given marginals. *Biometrika*, **55**, 179–188.
Jaynes, E. T. (1957) Information theory and statistical mechanics. *Phys. Rev.*, **106**, 620–630.
Kang, J. D. Y. and Schafer, J. L. (2007) Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statist. Sci.*, **22**, 523–539.
Kish, L. (1992) Weighting for unequal P_i. *J. Off. Statist.*, **8**, 183–200.
Kline, B. and Tamer, E. (2011) Using observational vs. randomized controlled trial data to learn about treatment effects. Department of Economics, Northwestern University, Evanston. (Available from `http://dx.doi.org/10.2139/ssrn.1810114`.)
Kullback, S. (1997) *Information Theory and Statistics*. New York: Wiley.
Liaw, A. and Wiener, M. (2002) Classification and regression by randomforest. *R News*, **2**, 18–22.
Miratrix, L. W., Sekhon, J. S. and Yu, B. (2013) Adjusting treatment effect estimates by post-stratification in randomized experiments. *J. R. Statist. Soc.* B, **75**, 369–396.
Mitra, N. and Indurkhya, A. (2005) A propensity score approach to estimating the cost-effectiveness of medical therapies from observational data. *Hlth. Econ.*, **14**, 805–815.
Mojtabai, R. and Zivin, J. G. (2003) Effectiveness and cost-effectiveness of four treatment modalities for substance disorders: a propensity score analysis. *Hlth Serv. Res.*, **38**, 233–259.
National Research Council (2013) *Frontiers in Massive Data Analysis*. Washington DC: National Academies Press.
Nixon, R. M. and Thompson, S. G. (2005) Incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. *Hlth Econ.*, **14**, 1217–1229.
Politis, D. N. and Romano, J. P. (1994) Large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.*, **22**, 2031–2050.

Porter, K. E., Gruber, S., van der Laan, M. J. and Sekhon, J. S. (2011) The relative performance of targeted maximum likelihood estimators. *Int. J. Biostatist.*, **7**, no. 1.
Rothwell, P. M. (2005) External validity of randomised controlled trials: to whom do the results of this trial apply? *Lancet*, **365**, 82–93.
Sakr, Y., Vincent, J.-L., Reinhart, K., Payen, D., Wiedermann, C. J., Zandstra, D. F. and Sprung, C. L. (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**, 2722–2731.
Sekhon, J. S. (2009) Opiates for the matches: matching methods for causal inference. *A. Rev. Politi. Sci.*, **12**, 487–508.
Sekhon, J. S. (2011) Matching: multivariate and propensity score matching with automated balance search. *J. Statist. Softwr.*, **42**, 1–52.
Sekhon, J. S. and Grieve, R. (2012) A nonparametric matching method for covariate adjustment with application to economic evaluation (genetic matching). *Hlth Econ.*, **21**, 695–714.
Shadish, W. R., Cook, T. D. and Campbell, D. T. (2002) *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.
Stevens, K., McCabe, C., Jones, C., Ashcroft, J., Harvey, S. and Rowan, K. On behalf of the PAC-Man Study Collaboration (2005) The incremental cost effectiveness of withdrawing pulmonary artery catheters from routine use in critical care. *Appl. Hlth Econ. Hlth Poly*, **4**, 257–264.
Stuart, E. A., Cole, S. R., Bradshaw, C. P. and Leaf, P. J. (2011) The use of propensity scores to assess the generalizability of results from randomized trials. *J. R. Statist. Soc.* A, **174**, 369–386.
Tunis, S. R., Stryer, D. B. and Clancy, C. M. (2003) Practical clinical trials. *J. Am. Med. Ass.*, **290**, 1624–1632.
Turner, R. M., Spiegelhalter, D. J., Smith, G. and Thompson, S. G. (2009) Bias modelling in evidence synthesis. *J. R. Statist. Soc.* A, **172**, 21–47.
Wellek, S. (2010) *Testing Statistical Hypotheses of Equivalence and Noninferiority*. Boca Roton: CRC Press.
Willan, A. R. and Briggs, A. H. (2006) *Statistical Analysis of Cost-effectiveness Data*. Hoboken: Wiley.
Willan, A. R., Briggs, A. H. and Hoch, J. S. (2004) Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data. *Hlth Econ.*, **13**, 461–475.
Willan, A. R., Chen, E. B., Cook, R. J. and Lin, D. Y. (2003) Incremental net benefit in randomized clinical trials with quality-adjusted survival. *Statist. Med.*, **22**, 353–362.
Willan, A. R. and Lin, D. Y. (2001) Incremental net benefit in randomized clinical trials. *Statist. Med.*, **20**, 1563–1574.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Additional appendices for "From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects"'.