Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2013)
75, *Part* 2, *pp.* 369–396

Adjusting treatment effect estimates by post-stratification in randomized experiments

Luke W. Miratrix

Harvard University, Cambridge, USA

and Jasjeet S. Sekhon and Bin Yu

University of California at Berkeley, USA

[Received September 2011. Revised July 2012]

Summary. Experimenters often use post-stratification to adjust estimates. Post-stratification is akin to blocking, except that the number of treated units in each stratum is a random variable because stratification occurs after treatment assignment. We analyse both post-stratification and blocking under the Neyman–Rubin model and compare the efficiency of these designs. We derive the variances for a post-stratified estimator and a simple difference-in-means estimator under different randomization schemes. Post-stratification is nearly as efficient as blocking: the difference in their variances is of the order of $1/n^2$, with a constant depending on treatment proportion. Post-stratification is therefore a reasonable alternative to blocking when blocking is not feasible. However, in finite samples, post-stratification can increase variance if the number of strata is large and the strata are poorly chosen. To examine why the estimators' variances are different, we extend our results by conditioning on the observed number of treated units in each stratum. Conditioning also provides more accurate variance estimates because it takes into account how close (or far) a realized random sample is from a comparable blocked experiment. We then show that the practical substance of our results remains under an infinite population sampling model. Finally, we provide an analysis of an actual experiment to illustrate our analytical results.

Keywords: Blocking; Neyman–Rubin model; Randomized trials; Regression adjustment

1. Introduction

One of the most important tools for determining the causal effect of some action is the randomized experiment, where a researcher randomly divides units into groups and applies different treatments to each group. Randomized experiments are the 'gold standard' for causal inference because, assuming proper implementation of the experiment, if a difference in outcomes is found the only possible explanations are a significant treatment effect or random chance. Analytical calculation gives a handle on the chance which allows for principled inference about the treatment effect. In the most basic analysis, a simple difference in means is used to estimate the overall sample average treatment effect (SATE), which is defined as the difference in the units' average outcome if all were treated as compared with their average outcome if they were not. This framework and estimator were analysed by Neyman in 1923 (see the English translation by

Address for correspondence: Luke W. Miratrix, Department of Statistics, Harvard University, Science Center, 1 Oxford Street, Cambridge, MA 02138-2901, USA.
E-mail: lmiratrix@stat.harvard.edu

 1369–7412/13/75369

Splawa-Neyman *et al.* (1990)) under what is now called the Neyman or Neyman–Rubin model of potential outcomes (Holland, 1986). Under this model, one need make few assumptions that are not guaranteed by the randomization itself.

Since each additional observation in an experiment sometimes comes at considerable cost, it is desirable to find more efficient estimators than the simple difference-in-means estimator to measure treatment effects. Blocking, which is when experimenters first stratify their units and then randomize treatment within predefined blocks, can greatly reduce variance compared with the simple difference estimator if the strata differ from each other. See 'A useful method' in Fisher (1926) for an early overview, Wilk (1955) for an analysis and comparison with analysis of variance or Imai *et al.* (2008) for a modern overview. Unfortunately, because blocking must be conducted before randomization, it is often not feasible owing to practical considerations or lack of foresight. Sometimes randomization may even be entirely out of the researcher's control, such as with so-called natural experiments. When blocking is not done, researchers often adjust for covariates after randomization. For example, Pocock *et al.* (2002) studied a sample of clinical trials analyses and found that 72% of these references used covariate adjustment. Keele *et al.* (2008) analysed the experimental results in three major political science journals and found that 74–95% of the references relied on adjustment. Post-stratification is one simple form of adjustment where the researcher stratifies experimental units with a pretreatment variable, estimates treatment effects within the strata and then uses a weighted average of these strata estimates for the overall average treatment effect estimate. This is the estimator that we focus on.

In this paper, we use the Neyman–Rubin model to compare post-stratification both with blocking and with using no adjustment. Neyman's framework does not require assumptions of a constant treatment effect or of identically or independently distributed disturbances, which are assumptions that are typically made when considering adjustment to experimental data without this framework (e.g. McHugh and Matts (1983)). This avenue for a robust analysis, which was revitalized by Rubin in the 1970s (Rubin, 1974), has recently had much appeal. See, for example, work on general experiments (Keele *et al.*, 2008), matched pairs (Imai, 2008) or matched pairs of clusters (Imai *et al.*, 2009). (See Sekhon (2009) for a historical review of the Neyman–Rubin model.) Also see Neyman's own treatment of blocking in the appendix of Neyman *et al.* (1935). Our estimator is equivalent to one from a fully saturated ordinary least squares regression. Freedman (2008a, b) analysed the regression-adjusted estimator under the Neyman–Rubin model without treatment-by-strata interactions and found that the asymptotic variance might be larger than if no correction were made. Lin (2012) extended Freedman's results and showed that, when a treatment-by-covariate interaction is included in the regression, adjustment cannot increase the asymptotic variance. We analyse the exact, finite sample properties of this saturated estimator. Imbens (2011) analysed estimating the treatment effect in a larger population, assuming that the given sample being experimented on is a random draw from it. However, because in most randomized trials the sample is not taken at random from the larger population of interest, we focus on estimating the treatment effect within the sample. Tsiatis *et al.* (2008) and Koch *et al.* (1998) proposed other adjustment methods that also rely on weak assumptions and that have the advantage of working naturally with continuous or multiple covariates. Because of different sets of assumptions and methods of analysis, these estimators have important differences from each other. See Section 6 for further discussion.

We derive the variances for post-stratification and simple difference-in-means estimators under many possible randomization schemes including complete randomization and Bernoulli assignment. We show that the difference between the variance of the post-stratified estimator and that of a blocked experiment is of the order of $1/n^2$ with a constant primarily dependent on the proportion of units treated. Post-stratification is comparable with blocking. Like blocking,

post-stratification can greatly reduce variance over using a simple difference-in-means estimate. However, in small samples post-stratification can substantially hurt precision, especially if the number of strata is large and the stratification variable is poorly chosen.

After randomization, researchers can observe the proportion of units actually treated in each stratum. We extend our results by deriving variance formulae for the post-stratified and simple difference estimators conditioned on these observed proportions. These conditional formulae help to explain why the variances of the estimators can differ markedly with a prognostic covariate: the difference comes from the potential for bias in the simple difference estimator when there is a large imbalance (i.e. when the observed proportions of units treated are far from what is expected). Interestingly, if the stratification variable is not predictive of outcomes the conditional mean-squared error (MSE) of the simple difference estimator usually remains the same or even goes down with greater imbalance, whereas the conditional MSE of the adjusted estimator increases. Adjusting for a poorly chosen covariate has real cost in finite samples.

The rest of the paper is organized as follows. In the next section, we set up the Neyman–Rubin model, describe the estimators and then derive the estimators' variances. In Section 3 we show that post-stratification and blocking have similar characteristics in many circumstances. In Section 4, we present our formula for the estimators' variances conditioned on the observed proportions of treated units in the strata and discuss their implications. We then align our results with those of Imbens (2011) in Section 5 by extending our findings to the superpopulation model and discussing the similarities and differences of the two viewpoints. We compare post-stratification with other forms of adjustment in Section 6, focusing on how these different approaches use different assumptions. In Section 7, we apply our method to the real data example of a large, randomized medical trial to assess post-stratification's efficacy in a real world example. We also make a hypothetical example from this data set to illustrate how an imbalanced randomization outcome can induce bias which the post-stratified estimator can then adjust for. Section 8 concludes.

The programs that were used for the simulation examples can be obtained from

`http://www.blackwellpublishing.com/rss`

2. The estimators and their variances

We consider the Neyman–Rubin model with two treatments and n units. For an example consider a randomized clinical trial with n people, half given a drug and the other half given a placebo. Let $y_i(1) \in \mathbb{R}$ be unit i's outcome if it were treated, and $y_i(0)$ its outcome if it were not. These are the *potential outcomes* of unit i. For each unit, we observe either $y_i(1)$ or $y_i(0)$ depending on whether we treat it or not. We make the assumption that treatment assignment for any particular unit has no effect on the potential outcomes of any other unit (this is typically called the stable unit treatment value assumption). In the drug example this means that the decision to give the drug to one patient would have no effect on the potential outcomes of any other patient. The treatment effect t_i for unit i is then the difference in potential outcomes, $t_i \equiv y_i(1) - y_i(0)$, which is deterministic.

Although these t_i are the quantities of interest, we cannot in general estimate them because we cannot observe both potential outcomes of any unit i and because the t_i generally differ by unit. The average across a population of units, however, is estimable. Neyman (Splawa-Neyman *et al.*, 1990) considered the overall SATE:

$$\tau \equiv \frac{1}{n} \sum_{i=1}^{n} \{y_i(1) - y_i(0)\}.$$

To conduct an experiment, randomize units to treatment and observe outcomes. Many choices of randomization are possible. The observed outcome will be one of the two potential outcomes, and which one depends on the treatment given. Random assignment gives a treatment assignment vector $T = (T_1, \ldots, T_n)$ with $T_i \in \{0, 1\}$ being an indicator variable of whether unit i was treated or not. T's distribution depends on how the randomization was conducted. After the experiment is complete, we obtain the observed outcomes Y, with $Y_i = T_i\, y_i(1) + (1 - T_i)\, y_i(0)$. The observed outcomes are random—but only because of the randomization used. The $y_i(l)$ and t_i are all fixed. Neyman first considered a *balanced complete randomization*.

Definition 1 (complete randomization of n units). Given a fixed $p \in (0, 1)$ such that $0 < pn < n$ is an integer, a *complete randomization* is a simple random sample of pn units selected for treatment with the remainder left as controls. If $p = 0.5$ (and n is even) the randomization is *balanced* in that there are the same numbers of treated units as control units.

The classic unadjusted estimator $\hat{\tau}_{\mathrm{sd}}$ is the observed *simple difference* in the means of the treatment and control groups:

$$\begin{aligned}\hat{\tau}_{\mathrm{sd}} &= \frac{1}{W(1)} \sum_{i=1}^{n} T_i Y_i - \frac{1}{W(0)} \sum_{i=1}^{n} (1 - T_i) Y_i \\ &= \sum_{i=1}^{n} \frac{T_i}{W(1)} y_i(1) - \sum_{i=1}^{n} \frac{(1 - T_i)}{W(0)} y_i(0),\end{aligned}$$

where $W(1) = \Sigma_i T_i$ is the total number of treated units, $W(0)$ is total control and $W(1) + W(0) = n$. For Neyman's balanced complete randomization, $W(1) = W(0) = n/2$. For other randomization schemes the $W(l)$ are potentially random.

We analyse the properties of various estimators on the basis of the randomization scheme used—this is the source of randomness. Fisher proposed a similar strategy for testing the 'sharp null' hypothesis of no effect (where $y_i(0) = y_i(1)$ for $i = 1, \ldots, n$); under this view, all outcomes are known and the observed difference in means is compared with its exact, known distribution under this sharp null. Neyman, in contrast, *estimated* the variance of the difference in means, allowing for the unknown counterfactual outcomes of the units to vary. These different approaches have different strengths and weaknesses that we do not discuss here. We follow this second approach.

Neyman showed that the variance of $\hat{\tau}_{\mathrm{sd}}$ is

$$\mathrm{var}(\hat{\tau}_{\mathrm{sd}}) = \frac{2}{n} \mathbb{E}[s_1^2 + s_0^2] - \frac{1}{n} S^2 \tag{1}$$

where s_l^2 are the sample variances of the observed outcomes for each group, S^2 is the variance of the n treatment effects t_i and the expectation is over all possible assignments under balanced complete randomization. We extend this work by considering an estimator that (ideally) exploits some pretreatment covariate b by using post-stratification to reduce variance.

2.1. The post-stratified estimator of the sample average treatment effect

Stratification is when an experimenter divides the experimental units into K strata according to some categorical covariate b with $b_i \in \mathcal{B} \equiv \{1, \ldots, K\}$, $i = 1, \ldots, n$. Each stratum k contains $n_k = \#\{i : b_i = k\}$ units. For example, in a cancer drug trial we might have the strata being different stages of cancer. If the strata are associated with outcomes, an experimenter can adjust a treatment effect estimate to remove the effect of random variability in the proportions of units treated. This is the idea behind post-stratification. The b_i are observed for all units and are

not affected by treatment. The strata that are defined by the levels of b have stratum-specific SATE_k:

$$\tau_k \equiv \frac{1}{n_k} \sum_{i:b_i=k} \{y_i(1) - y_i(0)\} \qquad k = 1, \ldots, K.$$

The overall SATE can then be expressed as a weighted average of these SATE_ks:

$$\tau = \sum_{k \in \mathcal{B}} \frac{n_k}{n} \tau_k. \tag{2}$$

We can view the strata as K mini-experiments. Let $W_k(1) = \Sigma_{i:b_i=k} T_i$ be the number of treated units in stratum k, and $W_k(0)$ be the number of control units. We can use a simple difference estimator for each stratum to estimate the SATE_ks:

$$\hat{\tau}_k = \sum_{i:b_i=k} \frac{T_i}{W_k(1)} y_i(1) - \sum_{i:b_i=k} \frac{(1-T_i)}{W_k(0)} y_i(0). \tag{3}$$

A post-stratification estimator is an appropriately weighted estimate of these strata level estimates:

$$\hat{\tau}_{\text{ps}} \equiv \sum_{k \in \mathcal{B}} \frac{n_k}{n} \hat{\tau}_k. \tag{4}$$

These weights echo the weighted sum of SATE_ks in equation (2). Because b and n are known and fixed, the weights are also known and fixed. We derive the variance of $\hat{\tau}_{\text{ps}}$ in this paper.

Technically, this estimator is undefined if $W_k(1) = 0$ or $W_k(0) = 0$ for any $k \in 1, \ldots, K$. We therefore calculate all means and variances conditioned on $\mathcal{D}$, the event that $\hat{\tau}_{\text{ps}}$ is defined, i.e. that each stratum has at least one unit assigned to treatment and one to control. This is fairly natural: if the number of units in each stratum is not too small the probability of $\mathcal{D}$ is close to 1 and the conditioned estimator is similar to an appropriately defined unconditioned estimator. See Section 2.2. Similarly, τ_{sd} is undefined if $W(1) = 0$ or $W(0) = 0$. We handle this similarly, letting $\mathcal{D}'$ be the set of randomizations where $\hat{\tau}_{\text{sd}}$ is defined.

Different experimental designs and randomizations give different distributions on the treatment assignment vector T and all resulting estimators. Some distributions on T would cause bias. We disallow those. Define the *treatment assignment pattern* for stratum k as the ordered vector $(T_i : i \in \{1, \ldots, n : b_i = k\})$. We assume that the randomization used has *assignment symmetry*.

Definition 2 (assignment symmetry). A randomization is *assignment symmetric* if the following two properties hold.

(a) *Equiprobable treatment assignment patterns*: all $\binom{n_k}{W_k(1)}$ ways to treat $W_k(1)$ units in stratum k are equiprobable, given $W_k(1)$.
(b) *Independent treatment assignment patterns*: for all strata j and k, with $j \neq k$, the treatment assignment pattern in stratum j is independent of the treatment assignment pattern in stratum k, given $W_j(1)$ and $W_k(1)$.

Complete randomization and Bernoulli assignment (where independent p-coin flips determine the treatment for each unit) satisfy assignment symmetry. So does blocking, where strata are randomized independently. Furthermore, given a distribution on T that satisfies assignment symmetry, conditioning on $\mathcal{D}$ maintains assignment symmetry (as do many other reasonable conditionings, such as having at least x units in both treatment and control). See the on-line supplementary material for a more formal argument. Cluster randomization or randomization

where units have unequal treatment probabilities do not, in general, have assignment symmetry. In our technical results, we assume that

(a) the randomization is assignment symmetric and
(b) we are conditioning on $\mathcal{D}$ (or $\mathcal{D}'$), the set of possible assignments where $\hat{\tau}_{\text{ps}}$ (or $\hat{\tau}_{\text{sd}}$) is defined.

The post-stratification estimator and the simple difference estimator are used when the initial random assignment ignores the stratification variable b. In a blocked experiment, the estimator that is used is $\hat{\tau}_{\text{ps}}$, but the randomization is done within the strata defined by b. All three of these options are unbiased. We are interested in their relative variances. We express the variances of these estimators with respect to the sample's (unknown) means, variances and covariance of potential outcomes divided into between-strata variation and within-stratum variation. The within-stratum variances and covariances are, for $k = 1, \ldots, K$,

$$\sigma_k^2(l) = \frac{1}{n_k - 1} \sum_{i:b_i=k} \{y_i(l) - \bar{y}_k(l)\}^2 \qquad l = 0, 1$$

and

$$\gamma_k(1,0) = \frac{1}{n_k - 1} \sum_{i:b_i=k} \{y_i(1) - \bar{y}_k(1)\} \{y_i(0) - \bar{y}_k(0)\},$$

where $\bar{y}_k(l)$ denotes the mean of $y_i(l)$ for all units in stratum k. Like many, we use $n_k - 1$ rather than n_k for convenience and cleaner formulae. The $(1, 0)$ in $\gamma_k(1, 0)$ indicates that this framework could be extended to multiple treatments.

The between-stratum variances and covariance are the weighted variances and covariance of the strata means:

$$\bar{\sigma}^2(l) = \frac{1}{n - 1} \sum_{k=1}^{K} n_k \{\bar{y}_k(l) - \bar{y}(l)\}^2 \qquad l = 0, 1$$

and

$$\bar{\gamma}(1,0) = \frac{1}{n - 1} \sum_{k=1}^{K} n_k \{\bar{y}_k(1) - \bar{y}(1)\} \{\bar{y}_k(0) - \bar{y}(0)\}.$$

The populationwide $\sigma^2(l)$ and $\gamma(1, 0)$ are analogously defined. They can also be expressed as weighted sums of the component pieces. We also refer to the '*correlation of potential outcomes*' r, where $r \equiv \gamma(1,0)/\{\sigma(0)\,\sigma(1)\}$ and the *strata level correlations* r_k, where $r_k \equiv \gamma_k(1,0)/\{\sigma_k(0)\,\sigma_k(1)\}$. An overall constant treatment effect gives $r = 1$, $\sigma(0) = \sigma(1)$, $r_k = 1$ for all k and $\sigma_k(0) = \sigma_k(1)$ for all k.

We are ready to state our main results.

Theorem 1. The strata level estimators $\hat{\tau}_k$ are unbiased, i.e.

$$\mathbb{E}[\hat{\tau}_k | \mathcal{D}] = \tau_k \qquad k = 1, \ldots, K$$

and their variances are

$$\text{var}(\hat{\tau}_k | \mathcal{D}) = \frac{1}{n_k} \{\beta_{1k}\,\sigma_k^2(1) + \beta_{0k}\,\sigma_k^2(0) + 2\,\gamma_k(1,0)\} \tag{5}$$

with $\beta_{1k} = \mathbb{E}[W_k(0)/W_k(1)|\mathcal{D}]$, the expected ratio of the number of units in control to the number of units treated in stratum k, and $\beta_{0k} = \mathbb{E}[W_k(1)/W_k(0)|\mathcal{D}]$, the reverse.

Theorem 2. The post-stratification estimator $\hat{\tau}_{\text{ps}}$ is unbiased:

$$\mathbb{E}[\hat{\tau}_{\text{ps}}|\mathcal{D}] = \mathbb{E}\left[\sum_k \frac{n_k}{n}\hat{\tau}_k|\mathcal{D}\right] = \sum_k \frac{n_k}{n}\mathbb{E}[\hat{\tau}_k|\mathcal{D}] = \sum_k \frac{n_k}{n}\tau_k = \tau.$$

Its variance is

$$\text{var}(\hat{\tau}_{\text{ps}}|\mathcal{D}) = \frac{1}{n}\sum_k \frac{n_k}{n}\{\beta_{1k}\,\sigma_k^2(1) + \beta_{0k}\,\sigma_k^2(0) + 2\,\gamma_k(1,0)\}. \tag{6}$$

See Appendix A for a proof. In essence we expand the sums, use iterated expectation and evaluate the means and variances of the treatment indicator random variables. Assignment symmetry allows for the final sum. Techniques used are similar to those found in many classic (e.g. Neyman *et al.* (1935) and Student (1923)) and recent references (e.g. Imai *et al.* (2008)).

Consider the whole sample as a single stratum and use theorem 1 to obtain immediately the following corollary.

Corollary 1. The unadjusted simple difference estimator $\hat{\tau}_{\text{sd}}$ is unbiased, i.e. $\mathbb{E}[\hat{\tau}_{\text{sd}}|\mathcal{D}'] = \tau$. Its variance is

$$\text{var}(\hat{\tau}_{\text{sd}}|\mathcal{D}') = \frac{1}{n}\{\beta_1\,\sigma^2(1) + \beta_0\,\sigma^2(0) + 2\gamma(1,0)\}, \tag{7}$$

where $\beta_1 \equiv \mathbb{E}[W(0)/W(1)|\mathcal{D}']$ and $\beta_0 \equiv \mathbb{E}[W(1)/W(0)|\mathcal{D}']$. In terms of strata level parameters, its variance is

$$\text{var}(\hat{\tau}_{\text{sd}}|\mathcal{D}') = \frac{1}{n}\{\beta_1\,\bar{\sigma}^2(1) + \beta_0\,\bar{\sigma}^2(0) + 2\,\bar{\gamma}(1,0)\} + \frac{1}{n}\sum_k \frac{n_k - 1}{n-1}\{\beta_1\,\sigma_k^2(1) + \beta_0\,\sigma_k^2(0) + 2\,\gamma_k(1,0)\}. \tag{8}$$

Conditioning $\hat{\tau}_{\text{sd}}$ on the $\mathcal{D}$ that is associated with $\hat{\tau}_{\text{ps}}$ does *not* produce an assignment symmetric randomization in the single stratum of all units, and indeed $\mathbb{E}[\hat{\tau}_{\text{sd}}|\mathcal{D}] \neq \tau$ in some cases.

For completely randomized experiments with np units treated, $\beta_1 = (1-p)/p$ and $\beta_0 = p/(1-p)$. For a balanced completely randomized experiment, equation (7) is the result that was presented in Splawa-Neyman *et al.* (1990)—see equation (1); the expectation of the sample variance is the overall variance. Then $\beta_l = 1$ and

$$\begin{aligned}\text{var}(\hat{\tau}_{\text{sd}}) &= \frac{1}{n}\{\sigma^2(1) + \sigma^2(0) + 2\gamma(1,0)\} \\ &= \frac{2}{n}\{\sigma^2(1) + \sigma^2(0)\} - \frac{1}{n}\{\sigma^2(1) + \sigma^2(0) - 2\gamma(1,0)\} \\ &= \frac{2}{n}\{\sigma^2(1) + \sigma^2(0)\} - \frac{1}{n}\text{var}\{y_i(1) - y_i(0)\}.\end{aligned}$$

2.1.1. Remarks

β_{1k} is the expectation of $W_k(0)/W_k(1)$, the ratio of control units to treated units in stratum k. For large n_k, this ratio is close to the ratio $\mathbb{E}[W_k(0)]/\mathbb{E}[W_k(1)]$ since the $W_k(l)$ will not vary much relative to their size. For small n_k, however, they will vary more, which tends to result in β_{1k} being noticeably larger than $\mathbb{E}[W_k(0)]/\mathbb{E}[W_k(1)]$. This is at the root of how the overall variance of post-stratification differs from blocking. This is discussed more formally later on and in Appendix B.

For $l = 0, 1$ the β_{lk}s are usually larger than β_l, being expectations of different variables with different distributions. For example in a balanced completely randomized experiment $\beta_1 = 1$ but $\beta_{1k} > 1$ for $k = 1, \ldots, K$ since $W_k(1)$ is random and $W(1)$ is not.

All the βs depend on both the randomization and the conditioning on $\mathcal{D}$ or $\mathcal{D}'$, and thus the variances from both equation (8) and equation (6) can change (markedly) under different randomization scenarios. As a simple illustration, consider a complete randomization of a 40-unit sample with a constant treatment effect and four strata of equal size. Let all $\sigma_k(l) = 1$ and all $r_k = 1$. Also let $\bar{\sigma}(l) = \bar{\gamma}(0,1) = 0.56$. If $p = 0.5$, then $\beta_1 = \beta_0 = 1$ and the variance of $\hat{\tau}_{\mathrm{sd}}$ is about 0.15. If $p = \frac{2}{3}$, then $\beta_1 = \frac{1}{2}$ and $\beta_0 = 2$. Equation (8) holds in both cases, but the variance in the second case will be about 10% larger owing to the larger β_0. There are fewer control units, so the estimate of the control outcome is more uncertain. The gain in certainty for the treatment units does not compensate enough. For $p = 0.5$, $\beta_{1k} = \beta_{0k} \approx 1.21$. The post-stratified variance is about 0.11. For $p = \frac{2}{3}$, $\beta_{1k} \approx 2.44$ and $\beta_{0k} \approx 0.61$. The average is about 1.52. The variance is about 14% larger than the $p = 0.5$ case. Generally speaking, the relative variances of different experimental set-ups are represented in the βs.

The correlation of potential outcomes, $\gamma_k(1,0)$, can radically impact the variance. If they are maximally negative, the variance can be 0 or nearly 0. If they are maximally positive (as in the case of a constant treatment effect), the variance can be twice what it would be if the outcomes were uncorrelated.

2.1.2 *Comparing the estimators*

Both $\hat{\tau}_{\mathrm{ps}}$ and $\hat{\tau}_{\mathrm{sd}}$ are unbiased, so their MSEs are the same as their variances. To compare $\hat{\tau}_{\mathrm{ps}}$ and $\hat{\tau}_{\mathrm{sd}}$ take the difference of their variances:

$$\begin{aligned}\mathrm{var}(\hat{\tau}_{\mathrm{sd}}|\mathcal{D}') - \mathrm{var}(\hat{\tau}_{\mathrm{ps}}|\mathcal{D}) = &\left[\frac{1}{n}\{\beta_1 \bar{\sigma}^2(1) + \beta_0 \bar{\sigma}^2(0) + 2\bar{\gamma}(1,0)\}\right] \\ &- \left[\frac{1}{n}\sum_{k=1}^{K}\left\{\left(\frac{n_k}{n}\beta_{1k} - \frac{n_k - 1}{n-1}\beta_1\right)\sigma_k^2(1) + \left(\frac{n_k}{n}\beta_{0k} - \frac{n_k - 1}{n-1}\beta_0\right)\sigma_k^2(0)\right\}\right. \\ &\left. + \frac{2}{n^2}\sum_{k=1}^{K}\frac{n - n_k}{n-1}\gamma_k(1,0)\right]. \qquad (9)\end{aligned}$$

Equation (9) breaks down into two parts as indicated by the square brackets. The first part, $\beta_1 \bar{\sigma}^2(1) + \beta_0 \bar{\sigma}^2(0) + 2\bar{\gamma}(1,0)$, is the between-strata variation. It measures how much the mean potential outcomes vary across strata and captures how well the stratification variable separates out different units, on average. The larger the separation, the more to gain by post-stratification. The second part represents the cost that is paid by post-stratification due to, primarily, the chance of random imbalance in treatment causing units to be weighted differently. This second part is non-positive and is a penalty except in some cases where the proportion of units treated is extremely close to 0 or 1 or is radically different across strata.

If the between-strata variation is larger than the cost paid then equation (9) is positive and it is good to post-stratify. If equation (9) is negative then it is bad to post-stratify. It can be positive or negative depending on the parameters of the population. In particular, if there is no between-strata difference in the mean potential outcomes, then the terms in the first square brackets are 0, and post-stratification hurts. Post-stratification is not necessarily a good idea when compared with doing no adjustment at all.

To assess the magnitude of the penalty paid compared with the gain, multiply equation (9) by n. The first term, representing the between-strata variation, is now a constant, and the scaled gain converges to it as n grows.

Theorem 3. Take an experiment with n units randomized under either complete randomization or Bernoulli assignment. Let p be the expected proportion of units treated. Without loss

of generality, assume that $0.5 \leqslant p < 1$. Let $f = \min\{n_k/n : k = 1, \ldots, K\}$ be the proportional size of the smallest stratum. Let $\sigma^2_{\max} = \max_{k,l}\{\sigma^2_k(l)\}$ be the largest variance of all the strata. Similarly define $\gamma_{\max}$. Then the scaled cost term is bounded:

$$|n\{\mathrm{var}(\hat{\tau}_{\mathrm{sd}}|\mathcal{D}') - \mathrm{var}(\hat{\tau}_{\mathrm{ps}}|\mathcal{D}')\} - \beta_1\,\bar{\sigma}^2(1) - \beta_0\,\bar{\sigma}^2(0) - 2\,\bar{\gamma}(1,0)| \leqslant C\frac{1}{n} + O\left(\frac{1}{n^2}\right)$$

with

$$C = \left\{\frac{8}{f(1-p)^2} + \frac{2p}{1-p}\right\}\sigma^2_{\max} + 2K\gamma_{\max}.$$

See Appendix B for the derivation. Theorem 3 shows us that the second part of equation (9), the harm, diminishes quickly.

Conditioning $\hat{\tau}_{\mathrm{ps}}$ on $\mathcal{D}$ and $\hat{\tau}_{\mathrm{sd}}$ on $\mathcal{D}'$ is not ideal, but $\hat{\tau}_{\mathrm{sd}}$ conditioned on $\mathcal{D}$ can be biased if the strata have unequal sizes and $p \neq 0.5$. However, owing to a similar argument to that in Section 2.2, this bias is small and equation (8) is close (i.e. within an exponentially small amount) to the MSE of $\hat{\tau}_{\mathrm{sd}}$ conditioned on $\mathcal{D}$. Thus theorem 3 holds for both estimators conditioned on $\mathcal{D}$. Indeed, theorem 3 holds unconditionally if the estimators are extended so they are reasonably defined (e.g. set to 0) when $\neg\mathcal{D}$ occurs.

If the number of strata K grows with n, as is often so when coarsening a continuous covariate, the story can change. The second terms in square brackets of equation (9) are sums over K elements. The larger the number of strata K, the more terms in the sums and the greater the potential penalty for stratification, unless the $\sigma^2_k(l)$s shrink in proportion as K grows. For an unrelated covariate, they will not tend to do so. To illustrate, we made a sequence of experiments increasing in size with a continuous covariate z unrelated to outcome. For each experiment with n units, we constructed b by cutting z into $K = n/10$ chunks. Post stratification was about 15% worse, in this case, than the simple difference estimator regardless of n. See the on-line supplementary materials for details as well as other illustrative examples. Theorem 3 captures the dependence on the number of strata through f, the proportional size of the smallest strata. If $f \propto 1/K$ then the difference will be $O(K/n)$. For example, if K grows at rate $O\{\log(n)\}$, then the scaled difference will be $O\{\log(n)/n\}$, which is nearly $O(1/n)$.

Overall, post-stratifying on variables that are not strongly related to outcome is unlikely to be worthwhile and can be harmful. Post-stratifying on variables that do relate to outcome is likely to result in large between-strata variation and thus a large reduction in variance compared with a simple difference estimator. More strata are not necessarily better, however. Simulations suggest that there is often a law of diminishing returns. For example, we made a simulated experiment with $n = 200$ units with a continuous covariate z related to outcome. We then made b by cutting z into K chunks for $K = 1, \ldots, 20$. As K increased from 1 there was a sharp drop in variance and then, as the cost due to post-stratification increased, the variance levelled off and then climbed. In this case, $K = 5$ was ideal. We did a similar simulation for a covariate z unrelated to outcome. Now, regardless of K, the $\sigma^2_k(l)$ were all about the same and the between-strata variation fairly low. As K grew, the overall variance climbed. In many cases a few moderate-sized strata give a dramatic reduction in variance, but having more strata beyond that has little effect and can even lead to an increase in $\hat{\tau}_{\mathrm{ps}}$'s variance. See the on-line supplementary material for details.

2.1.3. Estimation

Equation (6) and equation (8) are the actual variances of the estimators. In practice, the vari-

ance of an estimator, i.e. the squared standard error, would itself have to be estimated. Unfortunately, however, it is usually not possible consistently to estimate the standard errors of difference-in-means estimators owing to so-called identifiability issues as these standard errors depend on r_k, the typically unestimable correlations of the potential outcomes of the units being experimented on (see Splawa-Neyman *et al.* (1990)). One approach to estimate these standard errors consistently is to impose structure to render this correlation estimable or known; Reichardt and Gollob (1999), for example, demonstrated that quite strong assumptions must be made to obtain an unbiased estimator for the variance of $\hat{\tau}_{\mathrm{sd}}$. It is straightforward, however, to make a conservative estimate by assuming that the correlation is maximal. Sometimes there can be nice tricks—Alberto and Imbens (2008), for example, estimated these parameters for matched pairs by looking at pairs of pairs matched on covariates—but generally bounding the standard error is the best that one can do. Furthermore, the increased uncertainty and degrees-of-freedom issues from estimating the many variances composing the standard error of $\hat{\tau}_{\mathrm{ps}}$ would also have to be accounted for. Developing an appropriate method for this is an area for future work.

That being said, all terms except the $\gamma_k(1,0)$ in equation (9) are estimable with standard sample variance, covariance and mean formulae. In particular, $\bar{\gamma}(1,0)$ is estimable. By then making the conservative assumption that the $\gamma_k(1,0)$ are maximal (i.e. that $r_k = 1$ for all k so $\gamma_k(1,0) = \sigma(1)\,\sigma(0)$), we can estimate a lower bound on the gain. Furthermore, by then dividing by a similar upper bound on the standard error of the simple difference estimator, we can give a lower bound on the percentage reduction in variance due to post-stratification. We illustrate this when we analyse an experiment in Section 7.

2.2. Not conditioning on $\mathcal{D}$ changes little

Our results are conditioned on $\mathcal{D}$, the set of assignments such that $W_k(l) \neq 0$ for all $k = 1, \ldots, K$ and $l = 0, 1$. This, it turns out, results in variances that are only slightly different from not conditioning on $\mathcal{D}$.

Set $\hat{\tau}_{\mathrm{ps}} = 0$ if $\neg\mathcal{D}$ occurs, i.e. if $W_k(l) = 0$ for some k and l. Other choices of how to define the estimator when $\neg\mathcal{D}$ occurs are possible, including letting $\hat{\tau}_{\mathrm{ps}} = \hat{\tau}_{\mathrm{sd}}$—the point is that this choice does not much matter. In our case $\mathbb{E}[\hat{\tau}_{\mathrm{ps}}] = \tau \mathbf{P}\mathcal{D}$. The estimate of the treatment is shrunk by $\mathbf{P}\mathcal{D}$ towards 0. It is biased by $\tau \mathbf{P}\neg\mathcal{D}$. The variance is

$$\mathrm{var}(\hat{\tau}_{\mathrm{ps}}) = \mathrm{var}(\hat{\tau}_{\mathrm{ps}}|\mathcal{D})\mathbf{P}\mathcal{D} + \tau^2 \mathbf{P}\neg\mathcal{D}\mathbf{P}\mathcal{D}$$

and the MSE is

$$\mathrm{MSE}(\hat{\tau}_{\mathrm{ps}}) = \mathbb{E}[(\hat{\tau}_{\mathrm{ps}} - \tau)^2] = \mathrm{var}(\hat{\tau}_{\mathrm{ps}}|\mathcal{D})\mathbf{P}\mathcal{D} + \tau^2 \mathbf{P}\neg\mathcal{D}.$$

Not conditioning on $\mathcal{D}$ introduces a bias term and some extra variance terms. All these terms are small if $\mathbf{P}\neg\mathcal{D}$ is near 0, which it is: $\mathbf{P}\neg\mathcal{D}$ is $O\{n\exp(-n)\}$ (see Appendix B). Not conditioning on $\mathcal{D}$, then, gives substantively the same conclusions as conditioning on $\mathcal{D}$, but the formulae are a little more unwieldy. Conditioning on the set of randomizations where $\hat{\tau}_{\mathrm{ps}}$ is defined is more natural.

This of course applies to $\hat{\tau}_{\mathrm{sd}}$ and $\mathcal{D}'$ as well—and with a faster rate of decay since the single stratum is the entire sample. Furthermore, this also means conditioning on the 'wrong' $\mathcal{D}$ is also negligible, i.e. $\hat{\tau}_{\mathrm{sd}}$ conditioned on $\mathcal{D}$ is effectively unbiased. So the difference between conditioning on $\mathcal{D}$ and $\mathcal{D}'$ is small and, more generally, the conditioning in the theorems presented in this paper can be effectively ignored.

3. Comparing blocking with post-stratification

Let the *assignment split* W of a random assignment be the number of treated units in the strata:

$$W \equiv (W_1(1), \ldots, W_k(1)).$$

A *randomized block trial* ensures that W is constant because we randomize within strata, ensuring that a prespecified number of units are treated in each. This randomization is assignment symmetric (definition 2) and under it the probability of being defined, $\mathcal{D}$, is 1. For blocking, the standard estimate of the treatment effect has the same expression as $\hat{\tau}_{\mathrm{ps}}$, but the $W_k(l)$s are all fixed. If all blocks have the same proportion treated (i.e. $W_k(1)/n_k = W(1)/n$ for all k), $\hat{\tau}_{\mathrm{ps}}$ coincides with $\hat{\tau}_{\mathrm{sd}}$.

Because W is constant

$$\beta_{1k} = \mathbb{E}\left[\frac{W_k(0)}{W_k(1)}\right] = \frac{W_k(0)}{W_k(1)} = \frac{1-p_k}{p_k}, \tag{10}$$

where p_k is the proportion of units assigned to treatment in stratum k. Similarly, $\beta_{0k} = p_k/(1-p_k)$. Letting the subscript 'blk' denote this randomization, plug equation (10) into equation (6) to obtain the variance of a blocked experiment:

$$\mathrm{var}_{\mathrm{blk}}(\hat{\tau}_{\mathrm{ps}}) = \frac{1}{n}\sum_k \frac{n_k}{n}\left\{\frac{1-p_k}{p_k}\sigma_k^2(1) + \frac{p_k}{1-p_k}\sigma_k^2(0) + 2\gamma_k(1,0)\right\}. \tag{11}$$

Post-stratification is similar to blocking, and the post-stratified estimator's variance tends to be close to that of a blocked experiment. Taking the difference between equation (6) and equation (11) gives

$$\mathrm{var}(\hat{\tau}_{\mathrm{ps}}|\mathcal{D}) - \mathrm{var}_{\mathrm{blk}}(\hat{\tau}_{\mathrm{ps}}) = \frac{1}{n}\sum_k \frac{n_k}{n}\left\{\left(\beta_{1k} - \frac{1-p_k}{p_k}\right)\sigma_k^2(1) + \left(\beta_{0k} - \frac{p_k}{1-p_k}\right)\sigma_k^2(0)\right\}. \tag{12}$$

The $\gamma_k(1,0)$ cancelled; equation (12) is identifiable and therefore estimable.

Randomization without regard to b can have block imbalance due to ill luck: W is random. The resulting cost in variance of post-stratification over blocking is represented by the $\beta_{1k} - (1-p_k)/p_k$ terms in equation (12). This cost is small, as shown by theorem 4, as follows.

Theorem 4. Take a post-stratified estimator for a completely randomized or Bernoulli assigned experiment. Use the assumptions and definitions of theorem 3. Assume the common case for blocking of $p_k = p$ for $k = 1, \ldots, K$. Then

$$n\{\mathrm{var}(\hat{\tau}_{\mathrm{ps}}|\mathcal{D}) - \mathrm{var}_{\mathrm{blk}}(\hat{\tau}_{\mathrm{ps}})\} \leqslant \frac{8}{(1-p)^2}\frac{1}{f}\sigma_{\max}^2\frac{1}{n} + O\{\exp(-fn)\}.$$

See Appendix B for the derivation.

Theorem 4 bounds how much worse post-stratification can be compared with blocking. The scaled difference is of the order of $1/n$. The difference in variance is of order $1/n^2$. Generally speaking, post-stratification is similar to blocking in terms of efficiency. The more strata, however, the worse this comparison becomes because of the increased chance of severe imbalance with consequential increased uncertainty in the stratum level estimates. This is captured by the $1/p$-term. Many strata are generally not helpful and can be harmful if b is not prognostic.

3.1. A note on blocking

Plug equation (10) into the gain equation (equation (9)) to see immediately under what circumstances blocking has a larger variance than the simple difference estimator for a completely randomized experiment:

$$\operatorname{var}(\hat{\tau}_{\mathrm{sd}}) - \operatorname{var}_{\mathrm{blk}}(\hat{\tau}_{\mathrm{ps}}) = \frac{1}{n}\left\{ \frac{1-p}{p}\bar{\sigma}^2(1) + \frac{p}{1-p}\bar{\sigma}^2(0) + 2\bar{\gamma}(1,0) \right\} - \frac{1}{n^2}\sum_k \frac{n-n_k}{n-1}\left\{ \frac{1-p}{p}\sigma_k^2(1) + \frac{p}{1-p}\sigma_k^2(0) + 2\gamma_k(1,0) \right\}. \quad (13)$$

If $p = 0.5$, this is identical to the results in the appendix of Imai *et al.* (2008). In the worst case where there is no between-strata variation, the first term of equation (13) is 0 and so the overall difference is $O(K/n^2)$. The penalty for blocking is small, even for moderate-sized experiments, assuming that the number of strata does not grow with n. (Neyman *et al.* (1935) noted this in a footnote of his appendix where he derived the variance of a blocked experiment.) If the first term is not 0, then it will dominate for sufficiently large n, i.e. blocking will give a more precise estimate. For more general randomizations, equation (9) still holds but the βs differ. The difference in variances is still $O(1/n^2)$.

4. Conditioning on the assignment split *W*

By conditioning on the assignment split W we can break down the expressions for MSE to understand better when $\hat{\tau}_{\mathrm{ps}}$ outperforms $\hat{\tau}_{\mathrm{sd}}$. For $\hat{\tau}_{**}$ with $** \equiv$ ps, sd we have

$$\mathrm{MSE}(\hat{\tau}_{**}|\mathcal{D}) = \mathbb{E}_W[\mathrm{MSE}(\hat{\tau}_{**}|W)|\mathcal{D}] = \sum_{w\in\mathcal{W}} \mathrm{MSE}(\hat{\tau}_{**}|W=w)\,\mathbf{P}(W=w|\mathcal{D})$$

with $\mathcal{W}$ being the set of all allowed splits where $\hat{\tau}_{\mathrm{ps}}$ is defined. The overall MSE is a weighted average of the conditional MSE, with the weights being the probability of the given possible splits W. This will give us insight into when $\operatorname{var}(\hat{\tau}_{\mathrm{sd}})$ is large.

Conditioning on the split W maintains assignment symmetry and sets $\beta_{lk} = W_k(1-l)/W_k(l)$ for $k \in 1, \ldots, K$ and $\beta_l = W(1-l)/W(l)$. For $\hat{\tau}_{\mathrm{ps}}$ we immediately obtain

$$\operatorname{var}(\tau_{\mathrm{ps}}|W) = \frac{1}{n}\sum_k \frac{n_k}{n}\left\{ \frac{W_k(0)}{W_k(1)}\sigma_k^2(1) + \frac{W_k(1)}{W_k(0)}\sigma_k^2(0) + 2\gamma_k(1,0) \right\}. \quad (14)$$

Under conditioning $\hat{\tau}_{\mathrm{ps}}$ is still unbiased and so the conditional MSE is the conditional variance. $\hat{\tau}_{\mathrm{sd}}$, however, can be *biased* with a conditional MSE larger than the conditional variance if the extra bias term is non-zero. Theorem 5 shows the conditional bias and variance of $\hat{\tau}_{\mathrm{sd}}$, as follows.

Theorem 5. The bias of $\hat{\tau}_{\mathrm{sd}}$ conditioned on W is

$$\mathbb{E}[\hat{\tau}_{\mathrm{sd}}|W] - \tau = \sum_{k\in\mathcal{B}}\left[\left\{ \frac{W_k(1)}{W(1)} - \frac{n_k}{n} \right\} \bar{y}_k(1) - \left\{ \frac{W_k(0)}{W(0)} - \frac{n_k}{n} \right\} \bar{y}_k(0) \right],$$

which is not 0 in general, even with a constant treatment effect. $\hat{\tau}_{\mathrm{sd}}$'s variance conditioned on W is

$$\operatorname{var}(\hat{\tau}_{\mathrm{sd}}|W) = \sum_{k\in\mathcal{B}} \frac{W_{1k}W_{0k}}{n_k}\left\{ \frac{1}{W_1^2}\sigma_k^2(1) + \frac{1}{W_0^2}\sigma_k^2(0) + \frac{2}{W_1 W_0}\gamma_k(1,0) \right\}.$$

See Appendix A for a sketch of these two derivations. They come from an argument that is similar to the proof for the variance of $\hat{\tau}_{\mathrm{ps}}$, but with additional weighting terms.

The conditional MSE of $\hat{\tau}_{\mathrm{sd}}$ has no nice formula that we are aware of and is simply the sum of the variance and the squared bias:

$$\mathrm{MSE}(\hat{\tau}_{\mathrm{sd}}|W) = \mathrm{var}(\hat{\tau}_{\mathrm{sd}}|W) + \{\mathbb{E}[\hat{\tau}_{\mathrm{sd}}|W] - \tau\}^2. \tag{15}$$

In a typical blocked experiment, W would be fixed at W^{blk} where $W_k^{\mathrm{blk}} = n_k p$ for $k = 1, \ldots, K$. For complete randomization, $\mathbb{E}[W] = W^{\mathrm{blk}}$. We can now gain insight into the difference between the simple difference and post-stratified estimators. If W equals W^{blk}, then the conditional variance formulae for both estimators reduce to that of blocking, i.e. equation (14) and equation (15) reduce to equation (11). For $\hat{\tau}_{\mathrm{ps}}$, the overall variance for each stratum is a weighted sum of $W_k(0)/W_k(1)$ and $W_k(1)/W_k(0)$. The more unbalanced these terms, the larger the sum is. Therefore the more W deviates from W^{blk}—i.e. the more *imbalanced* the assignment is—the larger the post-stratified variance formula will tend to be. The simple difference estimator, in contrast, tends to have smaller variance as W deviates further from W^{blk} due to the greater restrictions on the potential random assignments.

$\hat{\tau}_{\mathrm{ps}}$ has no bias under conditioning, but $\hat{\tau}_{\mathrm{sd}}$ does if b is prognostic, and this bias can radically inflate the MSE. This bias increases with greater imbalance. Overall, then, as imbalance increases, the variance (and MSE) of $\hat{\tau}_{\mathrm{ps}}$ moderately increases. In contrast, for $\hat{\tau}_{\mathrm{sd}}$ the variance can moderately decrease but the bias sharply increases, giving an overall MSE that can grow quite large.

Because the overall MSE of these estimators is a weighted average of the conditional MSEs, and because under perfect balance the conditional MSEs are the same, we know that any differences in the unconditional variance (i.e. MSE) between $\hat{\tau}_{\mathrm{sd}}$ and $\hat{\tau}_{\mathrm{ps}}$ come from what happens when there is bad imbalance: $\hat{\tau}_{\mathrm{sd}}$ has a much higher MSE than $\hat{\tau}_{\mathrm{ps}}$ when there is potential for large bias and its MSE is smaller when there is not. With post-stratification, we pay for unbiasedness with a little extra variance—we are making a bias–variance trade-off that is different from that with simple differences.

The split W is directly observable and gives hints to the experimenter about the success, or failure, of the randomization. Unbalanced splits tell us that we have less certainty whereas balanced splits are comforting. For example, take a hypothetical balanced completely randomized experiment with $n = 32$ subjects: half men and half women. Compare the case where only one man ends up in treatment with the case with eight men. In the former case, a single man gives the entire estimate for average treatment outcome for men and a single woman gives the entire estimate for average control outcome for women. This seems *very* unreliable. In the latter case, each of the four mean outcomes are estimated with eight subjects, which seems more reliable. Our estimates of uncertainty should take this observed split W into account, and we can take it into account by using the conditional MSE rather than overall MSE when estimating uncertainty. The conditional MSE estimates how close one's actual experimental estimate is likely to be from the SATE. The overall MSE estimates how close such estimates will generally be to the SATE over many trials.

This idea of using all observed information is not new. When sampling to find the mean of a population, Holt and Smith (1979) argued that, for estimators adjusted by using post-stratification, variance estimates should be conditioned on the distribution of units in the strata as this gives a more relevant estimate of uncertainty. Sundberg (2003) sharpened this argument by presenting it as one of prediction. Under this view, it becomes more clear what should be conditioned on and what not. In particular, if an estimator is conditionally unbiased when conditioned on an ancillary statistic, then conditioning on the ancillary statistic increases precision. This is precisely the case when conditioning the above estimators on the observed split, assuming assignment symmetry. Similarly, in the case of sampling, Särndal *et al.* (1989) compared variance estimators for the sample totals that incorporate the mean of measured covariates compared with the population to obtain what they argued are more appropriate estimates. Pocock

et al. (2002) extended Senn (1989) and examined conditioning on the imbalance of a continuous covariate in analysis of covariance. They showed that not correcting for imbalance (measured as a standardized difference in means) gives one inconsistent control on the error rate when testing for an overall treatment effect.

5. Extension to an infinite population model

The results presented apply to estimating the treatment effect for a specific sample of units, but there is often a larger population of interest. One approach is to consider the sample to be a random draw from this larger population, which introduces an additional component of randomness capturing how the SATE varies about the population average treatment effect (PATE). See Imbens (2011). But, if the sample has not been so drawn, using this PATE model might not be appropriate. The SATE perspective should instead be used, with additional work then to generalize the results. See Hartman *et al*. (2011) or Imai *et al*. (2008). Regardless, under the PATE approach, the variances of all the estimators increase, but the substance of this paper's findings remains.

Let f_k, $k = 1, \ldots, K$, be the proportion of the population in stratum k. The PATE can then be broken down by strata:

$$\tau^* = \sum_{k=1}^{K} f_k \tau_k^*$$

with τ_k^* being the PATE in stratum k. Let the sample $\mathcal{S}$ be a stratified draw from this population holding the proportion of units in the sample to f_k (i.e. $n_k/n = f_k$ for $k = 1, \ldots, K$). (See below for different types of draws from the population.) The SATE τ depends on $\mathcal{S}$ and is therefore random. Owing to the size of the population, the sampling is close to being with replacement. An alternative view is drawing the sample with multiple independent draws from a collection of K distributions, one for each stratum. Let $\sigma_k^2(l)^*, \gamma_k^2(1,0)^*$, etc., be population parameters. Then the PATE level MSE of $\hat{\tau}_{\mathrm{ps}}$ is

$$\mathrm{var}(\hat{\tau}_{\mathrm{ps}}) = \frac{1}{n} \sum_k f_k \{(\beta_{1k} + 1)\, \sigma_k^2(1)^* + (\beta_{0k} + 1)\, \sigma_k^2(0)^*\}. \tag{16}$$

See Appendix A for the derivation. Imbens (2011) has a similar formula for the two-strata case. Compare with equation (6): all the 'correlations of potential outcomes' terms $\gamma_k(1,0)$ vanish when moving to the PATE. This is due to a perfect trade-off: the more they are correlated, the more difficult it is to estimate the SATE τ for the sample, but the easier it is to draw a sample with an SATE τ that is close to the overall PATE τ^*. Also, the variance is generally larger under the PATE view.

5.1. The simple difference estimator

For the simple difference estimator, use equation (16) with $K = 1$ to obtain

$$\mathrm{var}(\hat{\tau}_{\mathrm{sd}} | \mathcal{D}') = \frac{1}{n} \{(\beta_1 + 1)\, \sigma^2(1)^* + (\beta_0 + 1)\, \sigma^2(0)^*\}. \tag{17}$$

Now let $\bar{\sigma}^2(l)^*$ be a weighted sum of the squared differences of the strata means to the overall mean:

$$\bar{\sigma}^2(l)^* = \sum_{k=1}^{K} f_k \{\bar{y}_k^*(l) - \bar{y}^*(l)\}^2.$$

The population variances then decompose into $\bar{\sigma}^2(l)^*$ and strata level terms:

$$\sigma^2(l)^* = \bar{\sigma}^2(l)^* + \sum_{k=1}^{K} f_k \sigma_k^2(l)^*.$$

Plug this decomposition into equation (17) to obtain

$$\mathrm{var}(\hat{\tau}_{\mathrm{sd}}|\mathcal{D}') = \frac{1}{n}\left[(\beta_1+1)\left\{\bar{\sigma}^2(1)^* + \sum_{k=1}^{K} f_k \sigma_k^2(1)^*\right\} + (\beta_0+1)\left\{\bar{\sigma}^2(0)^* + \sum_{k=1}^{K} f_k \sigma_k^2(0)^*\right\}\right]. \tag{18}$$

5.2. Variance gain from post-stratification

For comparing the simple difference with the post-stratified estimator at the PATE level, take the difference of equation (18) and equation (16) to obtain

$$\begin{aligned}\mathrm{var}(\hat{\tau}_{\mathrm{sd}}|\mathcal{D}') - \mathrm{var}(\hat{\tau}_{\mathrm{ps}}|\mathcal{D}) = {} & \frac{1}{n}(\beta_1+1)\,\bar{\sigma}^2(1)^* + \frac{1}{n}(\beta_0+1)\,\bar{\sigma}^2(0)^* \\ & - \frac{1}{n}\sum_{k=1}^{K} f_k\{(\beta_{1k}-\beta_1)\,\sigma_k^2(1)^* + (\beta_{0k}-\beta_0)\,\sigma_k^2(0)^*\}.\end{aligned}$$

Similar to the SATE view, we again have a gain component (the first two terms) and a cost (the last term). For binomial assignment and complete randomization, $\beta_l \leqslant \beta_{lk}$ for all k, making the cost non-negative. There are no longer terms for the correlation of potential outcomes, and therefore this gain formula is directly estimable. The cost is generally smaller than for the SATE model owing to the missing $\gamma_k(1,0)$ terms.

5.3. Variance of blocking under population average treatment effect

For equal proportion blocking, $W_k(1) = pn_k$ and $W_k(0) = (1-p)n_k$. Using this and $\beta_{lk} + 1 = \mathbb{E}[n_k/W_k(l)]$, the PATE level MSE for a blocked experiment is then

$$\mathrm{var}_{\mathrm{blk}}(\hat{\tau}_{\mathrm{ps}}) = \frac{1}{n}\sum_k \frac{n_k}{n}\left\{\frac{1}{p}\sigma_k^2(1)^* + \frac{1}{1-p}\sigma_k^2(0)^*\right\}.$$

For comparing complete randomization (with pn units assigned to treatment) with blocked experiments, plug in the βs. The $(\beta_l - \beta_{lk})$-terms all cancel, leaving

$$\mathrm{var}_{\mathrm{blk}}(\hat{\tau}_{\mathrm{sd}}) - \mathrm{var}(\hat{\tau}_{\mathrm{ps}}) = \frac{1}{n}\frac{1}{p}\bar{\sigma}^2(1)^* + \frac{1}{n}\frac{1}{1-p}\bar{\sigma}^2(0)^* \geqslant 0.$$

Unlike from the SATE perspective, blocking can never hurt from the PATE perspective.

5.4. Not conditioning on the n_k

Allowing the n_k to vary introduces some complexity, but the gain formulae remain unchanged. If the population proportions are known, but the sample is a completely random draw from the population, the natural post-stratified estimate of the PATE would use the population weights f_k. These weights can be carried through and no problems result. Another approach is to estimate the f_k with n_k/n in the sample. In this case, we first condition on the seen vector $N \equiv n_1, \ldots, n_k$ and define a τ^N based on N. Conditioned on N, both $\hat{\tau}_{\mathrm{ps}}$ and $\hat{\tau}_{\mathrm{sd}}$ are unbiased for estimating τ^N, and we can use the above formula with n_k/n instead of f_k. Now use the tower property of expectations and variances. This results in an extra variance of a multinomial to capture how τ^N varies about τ as N varies. The variances of both the estimators will each be inflated by this extra term, which therefore cancels when looking at the difference.

6. Comparisons with other methods

Post-stratification is a simple adjustment method that ideally exploits a baseline categorical covariate to reduce the variance of an SATE estimate. Other methods allow for continuous or multiple covariates and are more general. The method that is appropriate for a given application depends on the exact assumptions that we are willing to make.

Recently, Freedman (2008a, b) studied the most common form of adjustment—linear regression—under the Neyman–Rubin model. Under this model, Freedman, for an experimental setting, showed that traditional ordinary least squares (in particular analysis of covariance) is biased (although it is asymptotically unbiased), that the asymptotic variance can be larger than with no adjustment and, worse, that the standard estimate of this variance can be quite off, even asymptotically.

Freedman's results differ from those in traditional textbooks because, in part, he used the Neyman–Rubin model with its focus on the SATE. Subsequently, Lin (2012) expanded these results and showed that ordinary least squares with all interactions cannot be asymptotically less efficient than using no adjustment and, further, that Huber–White sandwich estimators of the standard error are asymptotically appropriate. Freedman (2008a, b) and Lin (2012) focus primarily on continuous covariates rather than categorical, but their results are general. Our post-stratified estimator is identical to a fully saturated ordinary linear regression with the strata as dummy variables and all strata-by-treatment interactions—i.e. a two-way analysis-of-variance analysis with interactions. Therefore, our results apply to this regression estimator, and, in turn, all of Lin's asymptotic results apply to our $\hat{\tau}_{\mathrm{ps}}$.

Tsiatis *et al.* (2008) proposed a semiparametric method where the researcher independently models the response curve for the treatment group and the control group and then adjusts the estimated average treatment effect with a function of these two curves. This approach is particularly appealing in that concerns about data mining and pre-test problems are not an issue—i.e. researchers can search over a wide class of models looking for the best fit for each arm (assuming that they do not look at the consequent estimated treatment effects). With an analysis assuming only the randomization and the infinite superpopulation model, Tsiatis *et al.* (2008) showed that asymptotically such estimators are efficient. This semiparametric approach can accommodate covariates of multiple types: because the focus is modelling the two response curves, there is basically no limit to what information can be incorporated.

A method that does not have the superpopulation assumption is the inference method for testing for treatment effect proposed by Koch and co-workers (e.g. Koch *et al.* (1982, 1998)), who observed that, under the Fisherian sharp null of no treatment effect, one can directly compute the covariance matrix of the treatment indicator and any covariates. Therefore, using the fact that under randomization the expected difference of the covariates should be 0, one can estimate how far the observed mean difference in outcomes is from expected by using a χ^2-approximation. (One could also use a permutation approach to obtain an exact P-value.) However, rejecting Fisher's sharp null, distinct from the null hypothesis of no difference in average treatment effect, does not necessarily demonstrate an overall average effect. Nonetheless, this approach is very promising. Koch *et al.* (1982, 1998) also showed that with an additional superpopulation assumption one can use these methods to generate confidence intervals for average treatment effects.

McHugh and Matts (1983) compared post-stratification with blocking using an additive linear population model and a sampling framework, implicitly using potential outcomes for some results. They considered linear contrasts of multiple treatments as the outcome of interest, which is more general than this paper, but also imposed assumptions on the population such as

constant variance and, implicitly, a constant treatment effect. Using asymptotics, they isolated the main terms of the estimators' variance and dropped lower order terms.

Relative to post-stratification, there are three concerns with these other adjustment methods. First, many of these methods make the assumption of the infinite population sampling model that was discussed in Section 5 (which is equivalent to any model that has independent random errors, e.g. regression). The consequences of violating this assumption can be unclear. Therefore, one may prefer to estimate sample treatment effects, and then generalizing beyond the given experimental sample by using methods such as those of Hartman *et al.* (2011). Second, methods within the SATE framework that depend on a Fisherian sharp null for testing for a treatment effect have certain limitations. In some circumstances, this null may be considered restrictive and generating confidence intervals can be tricky without assuming a strong treatment effect model such as additivity. Third, asymptotic analyses may not apply when analysing small or mid-sized experiments, and experiments with such samples sizes are where the need for adjustment is the greatest.

Notwithstanding these concerns, if we are in a context where these concerns do not hold, or we have done work showing that their effect is minor, these alternative methods of adjustment depend on relatively weak assumptions and also allow for continuous covariates and multiple covariates—a distinct advantage over post-stratification. These other methods, owing to their additional modelling assumptions, may be more efficient as well. Different estimators may be more or less appropriate depending on the assumptions that we are willing to make and the covariates that we have.

Post-stratification is close in conceptual spirit to blocking. This paper shows that this conceptual relationship bears out. Blocking, however, is a stronger approach because it requires the choice of which covariates to adjust for to be determined before randomization. Blocking has the profound benefit that it forces the analyst to decide how covariates are incorporated to improve efficiency before any outcomes are observed. Therefore, blocking eliminates the possibility of searching over post-adjustment models until we are happy with the results. The importance of this feature is difficult to overstate. Blocking is, however, not always possible. In medical trials when patients are entered serially, for example, randomization must be done independently. Natural experiments, where randomization is due to processes that are outside the researchers' control, are another example that is particularly of interest in the social sciences. In these cases, post-stratification can give much the same advantages with much the same simplicity. But again, as 'Student' (W. S. Gosset) observed,

> 'there is great disadvantage in correcting any figures for position [of plots in agricultural experiments], inasmuch as it savours of cooking, and besides the corrected figures do not represent anything real. It is better to arrange in the first place so that no correction is needed' (Student, 1923).

7. Pulmonary artery catheterization data illustration

We apply our methods to evaluating pulmonary artery catheterization (PAC), an invasive and controversial cardiac monitoring device that was, until recently, widely used in the management of critically ill patients (Dalen, 2001; Finfer and Delaney, 2006). Controversy arose regarding the use of PAC when a non-random study using propensity score matching found that PAC insertion for critically ill patients was associated with increased costs and mortality (Connors *et al.*, 1996). Other observational studies came to similar conclusions leading to reduced PAC use (Chittock *et al.*, 2004). However, a randomized controlled trial found no difference in mortality between PAC and no-PAC groups (Harvey *et al.*, 2005), which substantiated the concern that the observational results were subject to selection bias (Sakr *et al.*, 2005).

The PAC trial has 1013 subjects with half treated. The outcome variable investigated here is 'qalys' or quality-adjusted life years. Higher values indicate, generally, longer life and higher quality of life. Death at the time of possible PAC insertion or shortly after receives a value of 0. Living 2 years in full health would be a 2. There is a large amount of fluctuation in these data. There is a large point mass at 0 (33% of the patients) and a long tail.

Unfortunately, the randomized controlled trial itself had observed covariate imbalance in predicted probability of death, a powerful predictor of the outcome, which calls into question the reliability of the simple difference estimate of the treatment effect. More low risk patients were assigned to receive treatment, which could induce a perceived treatment effect even if none were present. Post-stratification could help with this potential bias and decrease the variance of the estimate of treatment effect. To estimate the treatment effect by using post-stratification we first divide the continuous probability of death covariate into K K-tiles. We then estimate the treatment effect within the resulting strata and average appropriately.

This analysis is simplified for illustration. We are looking at only one of the outcomes and have dropped several potentially important covariates for clarity. Statistics on the strata for $K=4$ are listed in Table 1, including the numbers of units treated, Tx, or not, Co, for each stratum. A higher proportion of subjects in the first two groups were treated than we would expect given the randomization. Imbalance in the first group, with its high average outcome, could heavily influence the overall treatment effect estimate of $\hat{\tau}_{\mathrm{sd}}$.

We estimate the minimum gain in precision due to post-stratification by calculating point estimates of all the within- and between-strata variances and the between-strata covariance and plugging these values into equation (9). We are not taking the variability of these estimates into account. By assuming that the strata r_k are maximal, i.e. $r_k=1$ for all k, we estimate a lower bound on the reduction in variance due to post-stratification. The βs are estimated by numerical simulation of the randomization process (with 50000 trials) and are therefore exact up to uncertainty in this Monte Carlo calculation; these values do not depend on the population characteristics and so there is no sampling variability here. We show the resulting estimates for several stratifications. For $K=4$, we estimate the percentage reduction of variance, $100\% \times \{\mathrm{var}(\hat{\tau}_{\mathrm{ps}}) - \mathrm{var}(\hat{\tau}_{\mathrm{sd}})\}/\mathrm{var}(\hat{\tau}_{\mathrm{sd}})$, to be no less than 12%. If the true r_k were less than 1, the benefit would be greater. More strata appear somewhat superior, but gains level off rather quickly: Table 2.

The estimate of treatment effect changes under post-stratification. The estimates $\hat{\tau}_{\mathrm{ps}}$ hover around -0.28 for $K=4$ and higher, as compared with the -0.13 from the simple difference estimator. The post-stratified estimator appears to be correcting the bias from the random imbalance in treatment assignment.

We can also estimate the MSE for both the simple difference and the post-stratified estimator conditioned on the imbalance by plugging point estimates for the population parameters

Table 1. Strata level statistics for the PAC illustration

Strata	*Tx*	*Co*	$SD_k(1)$	$SD_k(0)$	$\hat{y}_k(1)$	$\hat{y}_k(0)$	$\hat{\tau}_k$
Low risk	136	118	5.80	5.68	5.57	5.41	0.15
Moderate risk	142	111	3.42	4.17	1.69	2.70	−1.01
High risk	106	147	3.60	3.75	1.97	2.36	−0.39
Extreme risk	122	131	3.41	3.10	1.37	1.19	0.18
Overall	506	507	4.56	4.48	2.72	2.84	−0.13

Table 2. Estimated standard errors for PAC†

K	$\hat{\tau}_{ps}$	$\hat{\tau}_{sd}$	*Results for unconditional variance*			*Results for MSE conditioned on W*				
			$\hat{\tau}_{ps}$	$\hat{\tau}_{sd}$	%	$MSE(\hat{\tau}_{ps})$	$var(\hat{\tau}_{sd})$	$bias(\hat{\tau}_{sd})$	$MSE(\hat{\tau}_{sd})$	%
2	−0.34	−0.13	0.077	0.081	5	0.077	0.076	0.207	0.118	35
4	−0.27	−0.13	0.071	0.081	12	0.072	0.070	0.137	0.089	19
10	−0.25	−0.13	0.070	0.081	14	0.071	0.069	0.119	0.083	15
15	−0.24	−0.13	0.070	0.081	14	0.070	0.067	0.115	0.081	13
30	−0.28	−0.13	0.069	0.081	15	0.068	0.064	0.148	0.086	21
50	−0.32	−0.13	0.068	0.081	15	0.066	0.061	0.190	0.097	32

†The table shows both conditioned and unconditioned estimates for different numbers of strata. '%' denotes the percentage variance reduction.

into equation (15) and equation (14). We again assume that the correlations r_k are maximal. We estimate bias by plugging in the estimated $\hat{y}_k(l)$ for the mean potential outcomes of the strata. These results are the last columns of Table 2; the percentage gain in this case is higher primarily because of the correction of the bias term from the imbalance. When conditioning on the imbalance W, the estimated MSE (i.e. variance) of the post-stratified estimator is slightly higher than the *variance* of the simple difference estimator but is substantially lower than its overall MSE. This is due to the bias correction. Because the true variances and the r_k for strata are unknown, these gains are estimates only. They do, however, illustrate the potential value of post-stratification. Measuring the uncertainty of these estimates is an area of future work.

7.1. Matched pairs estimation

We can also estimate the gains by building a fake set of potential outcomes by matching treated units to control units on observed covariates. We match as described in Sekhon and Grieve (2011). We then consider each matched pair a single unit with two potential outcomes. We use this synthetic set to calculate the variances of the estimators by using the formula from Section 2.

Matching treatment to controls and controls to treatment gives 1013 observations with all potential outcomes 'known'. The correlation of potential outcomes is 0.21 across all strata. $\tau = -0.031$. The unconditional variance for the simple difference and post-stratified estimators are 0.048 and 0.038 respectively. The percentage reduction in variance due to post-stratification is 19.6%.

We can use this data set to explore the effect of conditioning further. Assume that the treatment probability is $p = 0.5$ and repeatedly randomly assign a treatment vector and compute the resulting conditional MSE. Also compute the 'imbalance score' for the treatment vector with a χ^2-statistic:

$$\text{imbalance} \equiv \sum_k \frac{\{W_k(1) - pn_k\}^2}{pn_k}.$$

This procedure produces Fig. 1. As imbalance increases, the MSE (variance) of $\hat{\tau}_{ps}$ steadily, but slowly, increases as well. The MSE of $\hat{\tau}_{ps}$ is quite resistant to large imbalance. This is not so for $\hat{\tau}_{sd}$, however. Generally, high imbalance means high conditional MSE. This is due to the bias term which can become exceedingly large if there is imbalance between different heterogeneous

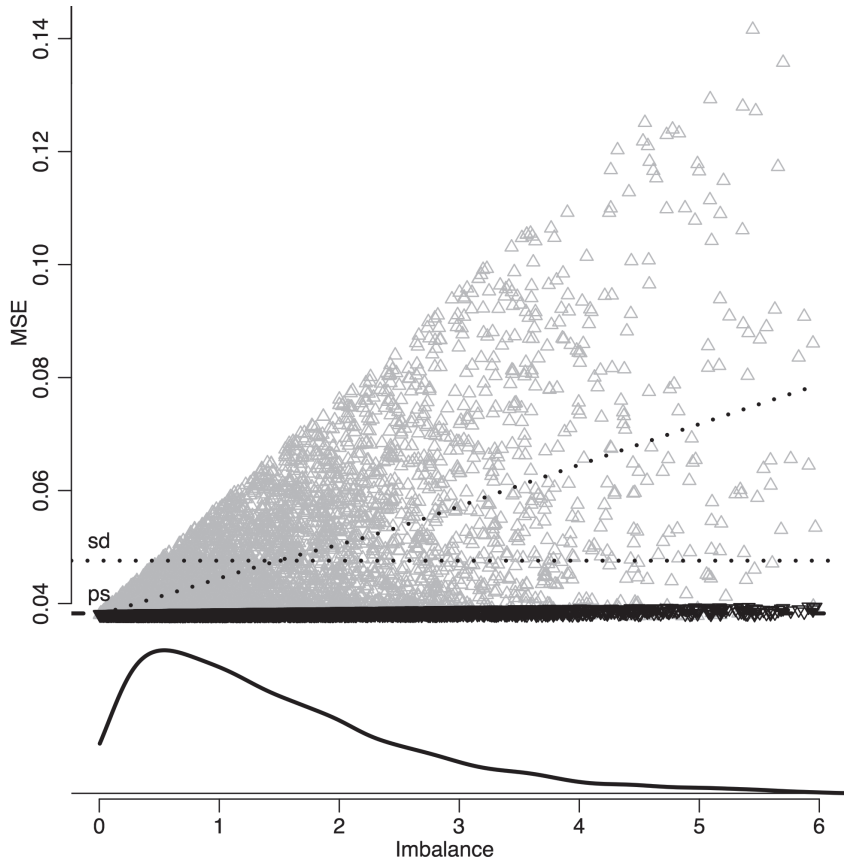

Fig. 1. PAC MSE conditioned on imbalance: the figure uses the constructed matched data set; points indicate the conditional MSE of $\hat{\tau}_{\text{ps}}$ ($\triangledown$) and $\hat{\tau}_{\text{sd}}$ ($\triangle$) given various specific splits of W; the x-axis is the imbalance score for the split; the dotted curves interpolate point clouds; the horizontal broken lines mark unconditional variances for the two estimators; the curve at the bottom is the density of the imbalance statistic

strata. Also, for a given imbalance, the simple difference estimator can vary widely depending on whether stratum level bias terms are cancelling out or not. This variability is not apparent for the post-stratified estimator, where only the number of units treated drives the variance; the post-stratified points cluster closely to their trend line.

The curve at the bottom of Fig. 1 shows the density of the realized imbalance score: there is a good chance of a fairly even split with low imbalance. In these cases, the variance of $\hat{\tau}_{\text{sd}}$ is smaller than the unconditional formula would suggest. If the randomization turns out to be 'typical' the unconditional variance formula would be conservative. If the imbalance is large, however, the unconditional variance may be overly optimistic. This chance of large imbalance with large bias is why the unconditioned MSE of $\hat{\tau}_{\text{sd}}$ is larger than that of $\hat{\tau}_{\text{ps}}$.

The observed imbalance for the actual assignment was about 2.37. The conditional MSE is 0.083 for $\hat{\tau}_{\text{sd}}$ and 0.039 for $\hat{\tau}_{\text{ps}}$, which is a 53% reduction in variance. The conditional MSE for the simple difference estimator is 75% larger than its unconditional MSE owing to the bias that is induced by the imbalance. We would be overly optimistic if we were to use $\text{var}(\hat{\tau}_{\text{sd}})$ as a measure of certainty, given the observed, quite imbalanced, split W. For the post-stratified estimator, however, the conditional variance is only about 1% higher than the unconditional; the degree

of imbalance is not meaningfully impacting the precision. Generally, with post-stratification, the choice of using an unconditional or conditional formula is less of a concern.

7.2. Discussion

The PAC randomized controlled trial has a strong predictor of outcome. Using it to post-stratify substantially increases the precision of the treatment effect estimate. Furthermore, post-stratification mitigates the bias that is induced by an unlucky randomization. When concerned about imbalance, it is important to calculate conditional standard errors—not doing so could give overly optimistic estimates of precision. This is especially true when using the simple difference estimator. The matched pairs investigation shows this starkly; $\hat{\tau}_{\text{sd}}$'s conditional MSE is 75% larger than the unconditional.

8. Conclusions

Post-stratification is a viable approach to experimental design in circumstances where blocking is not feasible. If the stratification variable is determined beforehand, post-stratification is nearly as efficient as a randomized block trial would have been: the difference in variances between post-stratification and blocking is a small $O(1/n^2)$. However, the more strata, the larger is the potential penalty for post-stratification. There is no guarantee of gains.

Conditioning on the observed distribution of treatment across strata allows for a more appropriate assessment of precision. Most often the observed balance will be good, even in moderate-sized experiments, and the conditional variance of both the post-stratified and the simple difference estimator will be smaller than estimated by the unconditional formula. However, when balance is poor, the conditional variance of the estimators, especially for the simple difference estimator, may be far larger than what the unconditional formula would suggest. Furthermore, in the unbalanced case, if a truly prognostic covariate is available post-stratification can significantly improve the precision of one's estimate. For a covariate that is unrelated to outcome, however, a simple difference estimator can be superior.

When viewing a post-stratified or a blocked estimate as an estimate of the PATE, under the assumption that the sample is a draw from a larger population, our findings generally hold although the potential for decreased precision is reduced. However, in most cases the sample in a randomized trial is not such a random draw. We therefore advocate for viewing the estimators as estimating the SATE, not the PATE.

Problems arise when stratification is determined after treatment assignment. The results of this paper assume that the stratification is based on a fixed and defined covariate b. However, in practice covariate selection is often done after the fact in part because, as is pointed out by Pocock *et al.* (2002), it is often quite difficult to know which of a set of covariates are significantly prognostic *a priori*. But variable selection invites 'fishing expeditions', which undermine the credibility of any findings. Doing variable selection in a principled manner is still notoriously difficult and is often poorly implemented; Pocock *et al.* (2002), for example, found that many clinical trial analyses select variables inappropriately. Tsiatis *et al.* (2008) summarized the controversy in the literature and, in an attempt to move away from strong modelling, and to allow for free model selection, proposed a semiparametic approach as a solution.

Beach and Meier (1989) suggested that, at minimum, all potential covariates for an experiment be listed in the original protocol. Call these z. In our framework, variable selection is then to *build* a stratification b from z *and* T after having randomized units into treatment and control. Stratification b (now B) is random as it depends on T. Questions immediately arise: how do we define the variance of the estimator? Can substantial bias be introduced by the strata

building process? The key to these questions probably depends on appropriately conditioning on both the final, observed, strata and the process of constructing B. This is an important area of future work.

Acknowledgements

We thank Peter Aronow, Winston Lin, Terry Speed and Jonathan Wand for helpful comments. Luke Miratrix is grateful for the support of a Graduate Research Fellowship from the National Science Foundation. This work is supported in part by National Science Foundation grant SES-0835531 (CDI) and Army Research Office grant W911NF-11-1-0114. The authors are responsible for any errors.

Appendix A

A.1. Proof of theorem 1

The proof of theorem 1 is based on iterated expectations and a large amount of unpleasant algebra. The following outline shows the highlights. We leave the conditioning on $\mathcal{D}$ implicitly in the expectations for cleaner presentation. See the on-line supplementary material for a version with more detail. We first set up a few simple expectations. Under assignment symmetry,

$$\mathbb{E}\left[\frac{T_i}{W_k(1)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{T_i}{W_k(1)}|W_k(1)\right]\right] = \mathbb{E}\left[\frac{1}{n_k}\right] = \frac{1}{n_k}.$$

Rearrange $\beta_{1k} \equiv \mathbb{E}[W_k(0)/W_k(1)] = n_k\,\mathbb{E}[1/W_k(1)] - 1$ to obtain $\mathbb{E}[1/W_k(1)] = (\beta_{1k}+1)/n_k$ and

$$\mathbb{E}\left[\frac{T_i^2}{W_k^2(1)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{T_i}{W_k^2(1)}|W_k(1)\right]\right] = \frac{1}{n_k}\mathbb{E}\left[\frac{1}{W_k(1)}\right] = \frac{\beta_{1k}+1}{n_k^2}. \tag{19}$$

These derivations are easier if we use $\alpha_{1k} \equiv \mathbb{E}[1/W_k(1)]$, but the βs are more interpretable and lead to a nicer final formula. There are analogous formulae for the control unit terms and cross-terms. We use these relationships to compute means and variances for the strata level estimators.

A.2. Unbiasedness

The strata level estimators are unbiased:

$$\begin{aligned}\mathbb{E}[\hat{\tau}_k] &= \mathbb{E}\left[\sum_{i:b_i=k}\frac{T_i}{W_k(1)}y_i(1) - \sum_{i:b_i=k}\frac{1-T_i}{W_k(0)}y_i(0)\right]\\ &= \sum_{i:b_i=k}\mathbb{E}\left[\frac{T_i}{W_k(1)}\right]y_i(1) - \sum_{i:b_i=k}\mathbb{E}\left[\frac{1-T_i}{W_k(0)}\right]y_i(0)\\ &= \sum_{i:b_i=k}\frac{1}{n_k}y_i(1) - \sum_{i:b_i=k}\frac{1}{n_k}y_i(0) = \tau_k.\end{aligned}$$

A.3. Variance

$\mathrm{var}(\hat{\tau}_k) = \mathbb{E}[\hat{\tau}_k^2] - \tau_k^2$. Expand τ_k^2 into three parts $a' - b' + c'$:

$$\tau_k^2 = \underbrace{\left\{\sum_{i:b_i=k}\frac{1}{n_k}y_i(1)\right\}^2}_{a'} - \underbrace{2\left\{\sum_{i:b_i=k}\frac{1}{n_k}y_i(1)\right\}\left\{\sum_{i:b_i=k}\frac{1}{n_k}y_i(0)\right\}}_{b'} + \underbrace{\left\{\sum_{i:b_i=k}\frac{1}{n_k}y_i(0)\right\}^2}_{c'}.$$

Similarly, expand the square of $\mathbb{E}[\hat{\tau}_k^2]$ to obtain $a - b + c$. Simplify these parts with algebra and relationships such as shown in equation (19). We then obtain, for example,

$$a = \mathbb{E}\left[\sum_{i:b_i=k} \frac{T_i}{W_k(1)} y_i(1)\right]^2$$
$$= \frac{\beta_{1k}+1}{n_k^2} \sum_{i:b_i=k} y_i^2(1) + \frac{-\beta_{1k}+n_k-1}{n_k^2(n_k-1)} \sum_{i\neq j} y_i(1)\, y_i(1).$$

Parts b and c are similar.

The variance is then $\text{var}(\hat{\tau}_k) = a - a' - b + b' + c - c'$, a sum of several ugly differences. Algebra, and recognizing formulae for the sample variances and covariances, gives

$$a - a' = \frac{\beta_{1k}}{n_k}\sigma_k^2(1),$$
$$b' - b = \frac{2}{n_k}\gamma_k(1,0)$$

and

$$c - c' = \frac{\beta_{0k}}{n_k}\sigma_k^2(0).$$

Sum these differences to obtain equation (5).

A.4. Proof of theorem 2

The mean is immediate. For the variance, observe that

$$\text{var}(\hat{\tau}_{\text{ps}}) = \mathbb{E}\left[\left\{\sum_{k=1}^{K} \frac{n_k}{n}(\hat{\tau}_k - \tau_k)\right\}^2\right]$$
$$= \sum_{k=1}^{K}\left(\frac{n_k}{n}\right)^2 \mathbb{E}[(\hat{\tau}_k - \tau_k)^2] + \sum_{k\neq r} \frac{n_k n_r}{n^2}\mathbb{E}[(\hat{\tau}_k - \tau_k)(\hat{\tau}_r - \tau_r)].$$

The first sum is what we want. The second is 0 since, using the tower property and assignment symmetry,

$$\mathbb{E}[\mathbb{E}[(\hat{\tau}_k - \tau_k)(\hat{\tau}_r - \tau_r)|W]] = \mathbb{E}[\mathbb{E}[(\hat{\tau}_k - \tau_k)|W]\mathbb{E}[(\hat{\tau}_r - \tau_r)|W]] = \mathbb{E}[0 \times 0] = 0.$$

A.5. Proof of theorem 5

Calculate the MSE of $\hat{\tau}_{\text{sd}}$ conditioned on the split W with a slight modification to the above derivation. Define a new estimator that is a weighted difference in means:

$$\hat{\alpha}_k \equiv A_k \sum_{i:b_i=k} \frac{T_i}{W_k(1)} y_i(1) - B_k \sum_{i:b_i=k} \frac{1-T_i}{W_k(0)} y_i(0)$$

with A_k and B_k constant. $\hat{\alpha}_k$ is an unbiased estimator of the difference in means weighted by A_k and B_k:

$$\mathbb{E}[\hat{\alpha}_k] = \mathbb{E}\left[A_k \sum_{i:b_i=b} \frac{T_i}{W_k(1)} y_i(1) - B_k \sum_{i:b_i=b} \frac{T_k}{W_i(0)} y_i(0)\right] = A_k\, \bar{y}_k(1) - B_k\, \bar{y}_k(0).$$

Now follow the derivation of the variance of $\hat{\tau}_k$ propagating A_k and B_k through. These are constant and they come out, giving

$$\text{var}(\hat{\alpha}_k) = \frac{1}{n_k}\left\{A_k^2 \beta_{1k}\, \sigma_k^2(1) + B_k^2 \beta_{0k}\, \sigma_k^2(0) + 2A_k B_k\, \gamma_k(1,0)\right\}.$$

Expand $\hat{\tau}_{\text{sd}}$ into strata terms:

$$\hat{\tau}_{\text{sd}} = \sum_{k=1}^{K}\left\{\frac{W_{1k}}{W_1}\sum_{i:b_i=k}\frac{T_i}{W_{1k}} y_i(1) - \frac{W_{0k}}{W_0}\sum_{i:b_i=k}\frac{1-T_i}{W_{0k}} y_i(0)\right\} = \sum_{k=1}^{K}\hat{\alpha}_k$$

with $A_k = W_{1k}/W_1$ and $B_k = W_{0k}/W_0$. Conditioning on W makes the A_k and the B_k constants, $\beta_{1k} = W_{0k}/W_{1k}$ and $\beta_{0k} = W_{1k}/W_{0k}$. Assignment symmetry ensures that, conditional on W, the stratum assignment patterns are independent, so the $\hat{\alpha}_k$ are as well, and the variances then add:

$$\operatorname{var}(\hat{\tau}_{\mathrm{sd}}|W) = \sum_{k=1}^{K} \operatorname{var}(\hat{\alpha}_k|W).$$

The bias is $\mathbb{E}[\hat{\tau}_{\mathrm{sd}}|W] - \tau$ with

$$\mathbb{E}[\hat{\tau}_{\mathrm{sd}}|W] = \sum_{k=1}^{K} \mathbb{E}[\hat{\alpha}_k|W] = \sum_{k=1}^{K} A_k\, \bar{y}_k(1) - B_k\, \bar{y}_k(0).$$

Expand τ as in equation (2) and rearrange terms.

A.6. Extending to population average treatment effect

First, decompose the variance:

$$\operatorname{var}(\hat{\tau}_{\mathrm{ps}}|\mathcal{D}) = \mathbb{E}_{\mathcal{S}}[\operatorname{var}(\hat{\tau}_{\mathrm{ps}}|\mathcal{S},\mathcal{D})|\mathcal{D}] + \operatorname{var}_{\mathcal{S}}(\mathbb{E}[\hat{\tau}_{\mathrm{ps}}|\mathcal{S},\mathcal{D}]|\mathcal{D}).$$

The first term is simply the expectation of equation (6): the SATE variance formula. Since $\mathcal{S}$ is random, so are the $\sigma_k^2(l)$, etc. The expectation of these quantities over $\mathcal{S}$ gives the population parameters as they are unbiased estimators. The βs are all constant, and $\mathcal{D}$ is independent of $\mathcal{S}$. Therefore $\mathbb{E}_{\mathcal{S}}[X|\mathcal{D}] = \mathbb{E}_{\mathcal{S}}[X]$ and

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}[\operatorname{var}(\hat{\tau}_{\mathrm{ps}}|\mathcal{S},\mathcal{D})|\mathcal{D}] &= \mathbb{E}_{\mathcal{S}}\left[\frac{1}{n}\sum_k \frac{n_k}{n}\{\beta_{1k}\,\sigma_k^2(1) + \beta_{0k}\,\sigma_k^2(0) + 2\,\gamma_k(1,0)\}\right] \\ &= \frac{1}{n}\sum_k \frac{n_k}{n}\{\beta_{1k}\,\sigma_k^2(1)^* + \beta_{0k}\,\sigma_k^2(0)^* + 2\,\gamma_k(1,0)^*\}. \end{aligned} \tag{20}$$

The second term is

$$\begin{aligned}\operatorname{var}(\mathbb{E}[\hat{\tau}_{\mathrm{ps}}|\mathcal{S},\mathcal{D}]) &= \operatorname{var}(\tau) \\ &= \operatorname{var}\left(\sum_{k=1}^{K} \frac{n_k}{n}\tau_k\right) \\ &= \frac{n_k^2}{n^2}\sum_{k=1}^{K} \operatorname{var}(\bar{y}_{k1} - \bar{y}_{k0}) \\ &= \frac{n_k^2}{n^2}\sum_{k=1}^{K} \frac{1}{n_k}\{\sigma_k^2(1)^* + \sigma_k^2(0)^* - 2\,\gamma_k(1,0)^*\}. \end{aligned} \tag{21}$$

Sum equation (20) and equation (21) to obtain the PATE level MSE.

Appendix B

β_{lk} can be approximated by $\mathbb{E}[W_k(1-l)]/\mathbb{E}[W_k(l)]$. For example, in the complete-randomization case $\beta_{1k} \approx (1-p)/p$. Generally, the βs are larger than their approximations. They can be less, but only by a small amount. For complete randomization and Bernoulli assignment, the difference between the βs and their approximations is bounded by the following theorem.

Theorem 6. Take an experiment with n units randomized under either complete randomization or Bernoulli assignment. Let p be the expected proportion of units treated. Let $\mathcal{D}$ be the event that $\hat{\tau}_{\mathrm{ps}}$ is defined. Let $p_{\max} = \max(p, 1-p)$ and $n_{\min}$ be the smallest strata size. Then $\beta_{1k} - (1-p)/p$ is bounded above:

$$\begin{aligned}\beta_{1k} - \frac{1-p}{p} &\leqslant \frac{4}{p^2}\frac{1}{n_k} - \frac{1}{p}\frac{1}{n_k+1} + \max\left\{\left(\frac{n_k}{2} - \frac{4}{p^2 n_k}\right)\exp\left(-\frac{p^2}{2}n_k\right), 0\right\} + 2n_k\,K(p_{\max})^{n_{\min}} \\ &= \frac{4}{p^2}\frac{1}{n_k} + O\{n_k \exp(-n_{\min})\}. \end{aligned}$$

Furthermore, it is tightly bounded below:

$$\beta_{1k} - \frac{1-p}{p} \geqslant -\frac{2}{p}(1-p)^{n_k} - 2n_k K(p_{\max})^{n_{\min}} = -O\{n_k \exp(-n_{\min})\}.$$

Similar results apply to the β_{0k} and β_l.

Proof. Start without conditioning on $\mathcal{D}$. $W_{1k} = \Sigma T_i$ with $T_i \in \{0, 1\}$. For Bernoulli assignment, the T_i are independent identically distributed Bernoulli variables with probability p of being 1. For completely randomized experiments, the W_{1k} are distributed according to a hypergeometric distribution, i.e. as the number of white balls drawn in n_k draws without replacement from an urn of n balls with np white balls. Regardless, $\mathbb{E}[W_{1k}] = n_k p$.

Define $Y_{n_k} \equiv n_k / W_{1k} \times \mathbf{1}_{\{W_{1k}>0\}}$. Owing to the indicator function, $Y_{n_k} \leqslant n_k$. Given $\mathcal{D}$, the event that *all* strata level estimators are well defined, $Y_{n_k} = n_k / W_{1k}$, so

$$\beta_{1k} - \frac{1-p}{p} = \mathbb{E}\left[\frac{W_{0k}}{W_{1k}}|\mathcal{D}\right] - \frac{1-p}{p} = \mathbb{E}\left[\frac{n_k}{W_{1k}}|\mathcal{D}\right] - \frac{1}{p} = \mathbb{E}[Y_{n_k}|\mathcal{D}] - \frac{1}{p}.$$

We first show that the probability of $\neg\mathcal{D}$ is very small, which will allow for approximating the expectation of the conditioned Y_{n_k} with the unconditioned. If $n_{\min}$ is the size of the smallest strata, then

$$\begin{aligned}
\mathbf{P}\neg\mathcal{D} &\leqslant \sum_{k=1}^{K} \mathbf{P}(W_{1k} = 0 \text{ or } W_{0k} = 0) \\
&\leqslant 2K \max_{l=0,1;k=1,\ldots,K} \mathbf{P}(W_{lk} = 0) \\
&\leqslant 2K(p_{\max})^{n_{\min}}.
\end{aligned}$$

Expand the expected value of Y as

$$\mathbb{E}[Y_{n_k}] = \mathbb{E}[Y_{n_k}|\mathcal{D}]\mathbf{P}\mathcal{D} + \mathbb{E}[Y_{n_k}|\neg\mathcal{D}]\mathbf{P}\neg\mathcal{D}.$$

Use this and the bound $Y_{n_k} \leqslant n_k$ to obtain

$$\begin{aligned}
\left|\mathbb{E}[Y_{n_k}|\mathcal{D}] - \mathbb{E}[Y_{n_k}]\right| &= \left|\mathbb{E}[Y_{n_k}|\mathcal{D}] - \mathbb{E}[Y_{n_k}|\mathcal{D}]\mathbf{P}\mathcal{D} - \mathbb{E}[Y_{n_k}]|\neg\mathcal{D}]\mathbf{P}\neg\mathcal{D}\right| \\
&= \left|\mathbb{E}[Y_{n_k}]|\mathcal{D}](1 - \mathbf{P}\mathcal{D}) - \mathbb{E}[Y_{n_k}|\neg\mathcal{D}]\mathbf{P}\neg\mathcal{D}\right| \\
&= \left|\mathbb{E}[Y_{n_k}|\mathcal{D}] - \mathbb{E}[Y_{n_k}|\neg\mathcal{D}]\right| \mathbf{P}\neg\mathcal{D} \\
&\leqslant n_k \mathbf{P}\neg\mathcal{D} = 2n_k K(p_{\max})^{n_{\min}}. \qquad (22)
\end{aligned}$$

This shows that $\mathbb{E}[Y_{n_k}|\mathcal{D}]$ is quite close to $\mathbb{E}[Y_{n_k}]$, i.e.

$$\mathbb{E}[Y_{n_k}] - \frac{1}{p} - 2n\,K(p_{\max})^{n_{\min}} \leqslant \beta_1 - \frac{1-p}{p} \leqslant \mathbb{E}[Y_{n_k}] - \frac{1}{p} + 2n\,K(p_{\max})^{n_{\min}}.$$

Now we need the following lemma to bound $\mathbb{E}[Y_{n_k}] - 1/p$.

Lemma 1. Let W be a binomial (n, p) random variable or a hypergeometric (n, w, N) random variable, i.e. a sample of size n from coin flips with probability of heads p or an urn with $N = nc$ balls, $c > 1$, of which $w = ncp$ are white. Then for $Y = (n/W)\mathbf{1}_{\{W>0\}}$:

$$-\frac{2}{p}(1-p)^n \leqslant \mathbb{E}[Y] - \frac{1}{p} \leqslant \frac{4}{p^2}\frac{1}{n} - \frac{1}{p}\frac{1}{n+1} + \max\left\{\left(\frac{n}{2} - \frac{4}{p^2 n}\right)\exp\left(-\frac{p^2}{2}n\right), 0\right\}.$$

See the on-line supplementary material for a proof, which uses results from Hoeffding (1963). Use lemma 1 on $\mathbb{E}[Y_{n_k}]$. This gives our stated bounds.

B.1. Remark on lemma 1

Numerical calculation shows that the constants of the $1/n$-term are overly large, but the rate of $1/n$ appears to be correct. Fig. 2 shows a log–log-plot of the actual percentage increase of $\mathbb{E}[Y]$ over $1/p$ for several values of p and n along with the calculated bounds. When the exponential term becomes negligible, the bound appears to be about 4, 7 and 31 times bigger for $p = 0.1, 0.5, 0.9$ respectively, i.e. the constants on the $1/n$-term are overstated by this much. For low p, the exponential terms can remain for quite some time

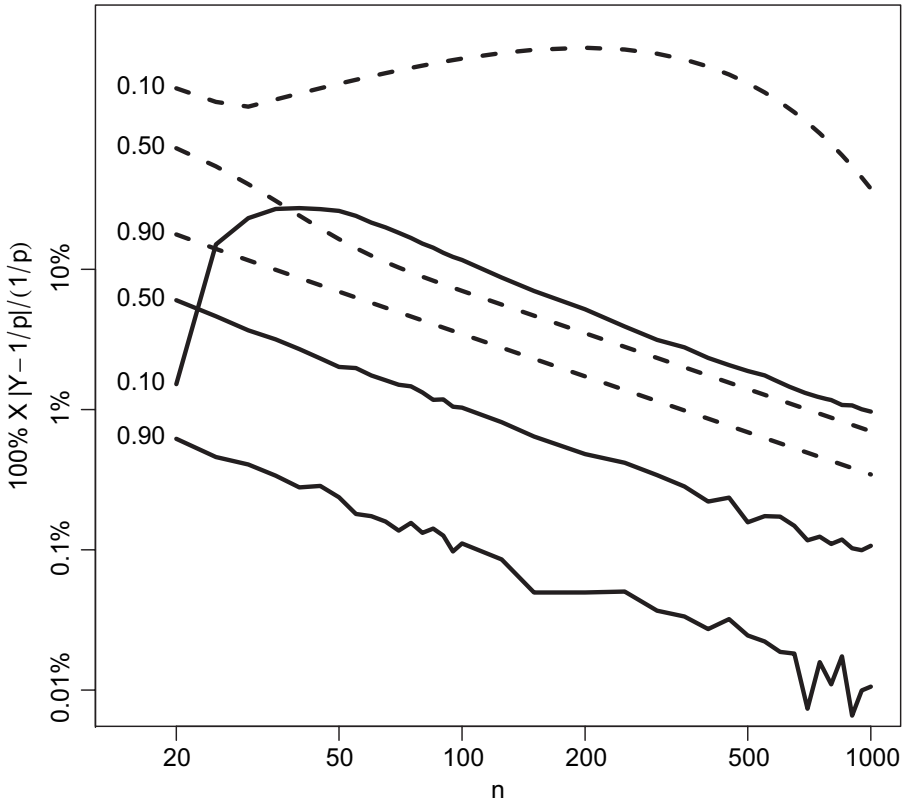

Fig. 2. Log–log-plot comparing actual percentage difference with a given bound: the percentage difference is calculated as $100\% \times (\mathbb{E}[Y] - 1/p)/(1/p)$, with Y as defined in lemma 1; three probabilities of assignment are shown ($p = 0.1, 0.5, 0.9$); the actual differences are computed with Monte Carlo sampling; Y is generated with a Bernoulli distribution

in the bound and there is significant bias in actuality due to the high chance of 0 units being assigned to treatment. The log–log-slope is -1, suggesting the $1/n$-relationship.

B.2. Proof of theorem 3

Assume the conditions stated for theorem 3 and consider equation (9). Replace all σs and γs with $\sigma^2_{\max}$ and $\gamma^2_{\max}$. Replace all β_{l0} with $\tilde{\beta}_0$, the largest such β for some stratum k, and the same for $\tilde{\beta}_1$. Collapse the sums to obtain

$$\text{scaled cost} \leqslant \left(\tilde{\beta}_0 - \frac{n-K}{n-1}\beta_0\right)\sigma^2_{\max} + \left(\tilde{\beta}_1 - \frac{n-K}{n-1}\beta_1\right)\sigma^2_{\max} + 2\frac{K-1}{n-1}\gamma_{\max}.$$

Then,

$$\begin{aligned}\left|\tilde{\beta}_0 - \frac{n-K}{n-1}\beta_0\right| &\leqslant |\tilde{\beta}_0 - \beta_0| + \left|\frac{n-K}{n-1}\beta_0 - \beta_0\right| \\ &\leqslant \frac{4}{(1-p)^2}\frac{1}{fn} + \frac{K-1}{n-1}\frac{p}{1-p} + O\left(\frac{1}{n^2}\right).\end{aligned}$$

Because the lower bound is so tight, we do not need to double the bound from theorem 6 for bounding the difference $|\tilde{\beta} - \beta_0|$. Because the β_1-expression will be smaller at the end, we can simply double the β_0-expression. This gives the bound.

B.3. Proof of theorem 4

The proof of theorem 4 is handled the same way as for theorem 3 but is more direct.

References

Alberto, A. and Imbens, G. (2008) Estimation of the conditional variance in paired experiments. *Ann. Econ. Statist.*, **91–92**, 175–187.

Beach, M. L. and Meier, P. (1989) Choosing covariates in the analysis of clinical trials. *Contr. Clin. Trials*, **10**, 161S–173S.

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

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

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

Finfer, S. and Delaney, A. (2006) Pulmonary artery catheters. *Br. Med. J.*, **333**, 930–931.

Fisher, R. A. (1926) The arrangement of field experiments. *J. Min. Agric. Gt Br.*, **33**, 503–513.

Freedman, D. A. (2008a) On regression adjustments in experiments with several treatments. *Ann. Appl. Statist.*, **2**, 176–196.

Freedman, D. A. (2008b) On regression adjustments to experimental data. *Adv. Appl. Math.*, **40**, 180–193.

Hartman, E., Grieve, R. D., Ramsahai, R. and Sekhon, J. S. (2011) From SATE to PATT: combining experimental with observational studies.

Harvey, S., Harrison, D., Singer, M., Ashcroft, J., Jones, C., 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.

Hoeffding, W. (1963) Probability inequalities for sums of bounded random variables. *J. Am. Statist. Ass.*, **58**, 13–30.

Holland, P. W. (1986) Statistics and causal inference. *J. Am. Statist. Ass.*, **81**, 945–960.

Holt, D. and Smith, T. M. F. (1979) Post stratification. *J. R. Statist. Soc.* A, **142**, 33–46.

Imai, K. (2008) Variance identification and efficiency analysis in randomized experiments under the matched-pair design. *Statist. Med.*, **27**, 4857–4873.

Imai, K., King, G. and Nall, C. (2009) The essential role of pair matching in cluster-randomized experiments, with application to the Mexican universal health insurance evaluation. *Statist. Sci.*, **24**, 29–53.

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.

Imbens, G. W. (2011) Experimental design for unit and cluster randomized trials. *Conf. International Initiative for Impact Evaluation, Cuernavaca*.

Keele, L. J., McConnaughy, C. and White, I. (2008) Adjusting experimental data. *Working Paper*.

Koch, G., Amara, I., Davis, G. and Gillings, D. (1982) A review of some statistical methods for covariance analysis of categorical data. *Biometrics*, **38**, 563–595.

Koch, G., Tangen, C., Jung, J. and Amara, I. (1998) Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them. *Statist. Med.*, **17**, 1863–1892.

Lin, W. (2012) Agnostic notes on regression adjustments to experimental data: reexamining Freedman's critique. *Ann. Appl. Statist.*, to be published.

McHugh, R. and Matts, J. (1983) Post-stratification in the randomized clinical trial. *Biometrics*, **39**, 217–225.

Neyman, J., Iwaszkiewicz, K. and Kołodziejczyk, St. (1935) Statistical problems in agricultural experimentation (with discussion). *J. R. Statist. Soc.*, suppl., **2**, 107–180.

Pocock, S. J., Assmann, S. E., Enos, L. E. and Kasten, L. E. (2002) Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statist. Med.*, **21**, 2917–2930.

Reichardt, C. S. and Gollob, H. F. (1999) Justifying the use and increasing the power of a t test for a randomized experiment with a convenience sample. *Psychol. Meth.*, **4**, 117–128.

Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, **66**, 688–701.

Sakr, Y., Vincent, J., Reinhart, K., Payen, D., Wiedermann, C., Zandstra, D. and Sprung, C. (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.

Särndal, C.-E., Swensson, B. and Wretman, J. H. (1989) The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, **76**, 527–537.

Sekhon, J. S. (2009) Opiates for the matches: matching methods for causal inference. *A. Rev. Polit. Sci.*, **12**, 487–508.

Sekhon, J. S. and Grieve, R. D. (2011) A matching method for improving covariate balance in cost-effectiveness analysis. *Hlth Econ.*, **21**, 695–714.
Senn, S. J. (1989) Covariate imbalance and random allocation in clinical trials. *Statist. Med.*, **8**, 467–475.
Splawa-Neyman, J., Dabrowska, D. M. and Speed, T. P. (1990) On the application of probability theory to agricultural experiments: Essay on principles, Section 9. *Statist. Sci.*, **5**, 465–472.
Student (1923) On testing varieties of cereals. *Biometrika*, **3–4**, 271–293.
Sundberg, R. (2003) Conditional statistical inference and quantification of relevance. *J. R. Statist. Soc.* B, **65**, 299–315.
Tsiatis, A. A., Davidian, M., Zhang, M. and Lu, X. (2008) Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Statist. Med.*, **27**, 4658–4677.
Wilk, M. B. (1955) The randomization analysis of a generalized randomized block design. *Biometrika*, **42**, 70–79.

Supporting information

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

'Supplementary material for Adjusting treatment effect estimates by post-stratification in randomized experiments'.