Why Match in Individually and Cluster Randomized Trials?

Laura B. Balzer∗  Maya L. Petersen†
Mark J. van der Laan‡

∗University of California, Berkeley, lbbalzer@hsph.harvard.edu
†University of California, Berkeley, mayaliv@berkeley.edu
‡University of California, Berkeley, laan@berkeley.edu

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

http://biostats.bepress.com/ucbbiostat/paper294

Copyright ©2012 by the authors.
The decision to match individuals or clusters in randomized trials is motivated by both practical and statistical concerns. Matching protects against chance imbalances in baseline covariate distributions and is thought to improve study credibility. Matching is also implemented to increase study power. This article compares the asymptotic efficiency of the pair-matched design, where units are matched on baseline covariates and the treatment randomized within pairs, to the independent design, where units are randomly paired and the treatment randomized within pairs. We focus on estimating the average treatment effect and use the efficient influence curve to understand the information provided by each design for estimation of this causal parameter. Our theoretical results indicate that the pair-matched design is asymptotically less efficient than its non-matched counterpart. Our simulations confirm these results asymptotically and in finite samples. Our approach is estimator-independent, avoids all parametric modeling assumptions, and applies equally to individually randomized and cluster randomized trials.
1 Introduction

For almost sixty years matching has inspired a heated debate in the literature. Matching is an intuitive design strategy to ensure groups are comparable at baseline. In observational studies, matching remains a widely accepted method to reduce bias, since the process removes the association between the covariate and the exposure at the start of the study. In experimental studies, matching helps to avoid chance imbalances in baseline covariate distributions between treatment groups. In other words, it helps to prevent “freak samples” that can occur randomly in small trials (Billewicz, 1964). Matching can decrease the variation in the outcome and may, but is not guaranteed to, increase study efficiency (Cochran, 1953; Billewicz, 1964; Youkeles, 1963; Mathen, 1963; Billewicz, 1965; Miettinen, 1968; McKinlay, 1977; Wacholder and Weinberg, 1982; Martin et al., 1993; Klar and Donner, 1997). Effective matching requires a comprehensive understanding of the relationship between the matching variables and the outcome as well as a large pool of candidate units. Matching can also be costly in terms of time, money and information lost (e.g. discarding unmatchable units). Furthermore, pair-matching halves the number independent units and thereby reduces the degrees of freedom of common estimators. Finally, covariate adjustment during the analysis may be an attractive alternative, as it has to been shown to increase study power without bias even in small trials (Pocock et al., 2002; Tsiatis et al., 2008; Zhang et al., 2008; Moore and van der Laan, 2009; Rosenblum, 2011; Rubin and van der Laan, 2011).

Despite these concerns, matching remains a popular design strategy in cluster randomized trials. These studies are characterized by a public health intervention delivered to groups of people. For example, media campaigns are delivered to entire cities, educational programs to schools, safety training to workplaces, and water quality improvements to communities. Due to their expense, these studies tend to be small, usually involving fewer than 30 clusters. Often, researchers use measured covariates (e.g. baseline prevalence, geographic region and cluster size) to create matched pairs of clusters. They then randomize the treatment within those pairs. Matching tends to be less common in individually randomized trials, but does exist. For example, the treatment may be randomized between twins or within the same individual over time (i.e. crossover studies). Ophthalmic trials are also known for randomizing the exposure to one eye the patient and the control to the other.

In this paper, we contrast the pair-matched design, where units are matched on baseline covariates and the treatment randomized within pairs, and the independent design, where units are randomly paired and the treatment randomized within pairs. These designs imply a specific data generating experiment, from which we can define the statistical model and the parameter of interest. Throughout we focus on estimating the average treatment effect. Our approach, however, is quite general and can be easily applied to other causal and non-causal quantities. We provide estimator-independent results on design efficiency. We also avoid all parametric modeling assumptions and complex criteria regarding the direction and strength of the correlation between the matching variable and the outcome. Analogue to Rose and van der Laan (2009), this article uses the efficient influence curve to investigate pair-matching individuals and clusters in randomized trials. The variance of the efficient influence curve is a direct measure of the information provided by a design for estimation of the target parameter. Specifically, the variance establishes a lower-bound on the asymptotic efficiency of all reasonable estimators (Bickel et al., 1993). Our theoretical results indicate that when matching on measured covariates, the pair-matched design is asymptotically less efficient than its non-matched counterpart. Our simulations echo these results and are not limited to dichotomous variables and category matching. Instead, they explore (i) binary and continuous outcomes, (ii) matching exactly or on a summary measure of continuous variables, (iii) effect modification, and (iv) right censoring. Overall, the simulations suggest that the pair-matched design is less efficient than the independent design both asymptotically and in finite samples.
2 Brief Literature Review

In this section, we present a brief review of pair-matching in randomized trials. Our review is not intended to be exhaustive. Instead, we give a concise introduction to cluster randomized trials. Then we consider the major arguments for and against matching.

2.1 Cluster Randomized Trials

Excellent introductions to cluster randomized trials are given by Murray (1998), Donner and Klar (2000) and Hayes and Moulton (2009). A review of the recent advances in the design and analysis of these studies is given by Campbell et al. (2007). For the sake of discussion, we define a cluster as a pre-existing group of individuals, characterized by some geographic, political or administrative criteria. Examples of clusters include communities, neighborhoods, schools, workplaces and households. Cluster or group randomized trials are dedicated to understanding the effect of an intervention implemented at the cluster-level. For example, the impact of a community-based educational program for improved personal and household hygiene on diarrheal morbidity was estimated with a community randomized trial involving 18 villages in rural Zaire (Haggerty et al., 1994). Likewise, the effect of treating STDs at community health centers on HIV incidence was evaluated with a pair-matched community randomized trial, occurring in the Mwanza region of Tanzania (Grosskurth et al., 1995). These examples help to illustrate the main reasons for conducting a cluster randomized trial, as opposed to an individually randomized trial: (i) implementation occurs naturally or conveniently at the cluster-level, (ii) avoidance of contamination, and (iii) capturing intervention effects among cluster members who are not direct recipients of the treatment (Hayes and Moulton, 2009).

An important challenge to cluster randomized trials is their limited sample size. Due to their expense and feasibility, most studies have few clusters (e.g. less than 30). While there may be thousands of individuals in each cluster, the precision of the effect estimate is largely determined by the number of independent units. Individuals within clusters are correlated and thereby not independent. Much attention has been given to understanding the variance within clusters and the variance between clusters. (See, for example, Freedman et al. (1997); Klar and Donner (1997); Murray (1998); Donner and Klar (2000); Bloom (2006); Campbell et al. (2007); Hayes and Moulton (2009).) However, quantities, such as the intracluster correlation coefficient or the coefficient of variation, are not necessary to understand the efficiency gained or lost by pair-matching.

2.2 Matching

A gentle introduction to matching is given by Costanza (1995) and a more advanced one by Rothman et al. (2008). Matching is an intuitive design strategy to ensure the comparability of study groups and to reduce variability in the outcome. In observational studies, matching refers to the systematic sampling of units from the reference group (e.g. control or unexposed) in order to balance the distributions of known confounders with that of the index group (e.g. case or exposed). In experimental studies, matching refers to the pairing or stratification of units, according to baseline covariates, and then the randomization of treatment within those pairs or blocks. There are several types of matching and many algorithms for implementation. Throughout we focus on simple pair-matching, where two units are matched on baseline covariates and the treatment randomized within the pair. With such a procedure, the treated and control observations within a matched pair are no longer independent.

Sometimes the pairing is natural and adjusts for many measured as well as unmeasured covariates. For example, to understand the efficacy and safety of a new treatment of glaucoma, researchers...
randomized laser treatment to one eye and the standard medication to the other eye of each patient (Glaucoma Laser Trial Research Group, 1991). Likewise, crossover trials, where the same patient is randomized over time to both treatments, can be considered a special type of pair-matching (assuming no carryover or temporal effects). More often, however, matches are artificially created according to measured covariates. For example, to understand the effect of antioxidant supplementation on platelet function, researchers pair-matched men according to smoking status, baseline outcome status and entry time (Salonen et al., 1991). Likewise, to investigate the impact of targeting younger children on reducing undernutrition, researchers pair-matched communities on access to healthcare, external sponsorship, as well as other geographic and ecological criteria (Ruel et al., 2008). It has long been acknowledged that such pairing requires a comprehensive understanding of the relationship between the selected covariates and the outcome and that our knowledge is always imperfect and incomplete (Cochran, 1953; Billewicz, 1964).

2.2.1 Matching to Improve Study Credibility

In observational studies, matching remains a popular method to control for confounding and improve study validity (Miettinen, 1970; Kupper et al., 1981; Karon and Kupper, 1982). In experimental studies, matching on important determinants of the outcome ensures these covariates are evenly distributed between treatment groups and thereby is said to improve the “face validity” of the study (Klar and Donner, 1997; Murray, 1998; Donner and Klar, 2000; Bloom, 2006; Hayes and Moulton, 2009). We contend that this perceived credibility is not a statistical or scientific notion. Treatment randomization guarantees there is no confounding. There may be an imbalance in important determinants of the outcome, and data sparsity may limit the ability of the most sophisticated analysis methods to control for these covariates. Nonetheless, the unadjusted estimator of the treatment effect is unbiased, even in small samples. This is commonly misunderstood but important concept. (See, for example, Rothman (1977).)

2.2.2 Matching to Improve Study Efficiency

In the absence of confounding concerns, the statistical argument for matching must rely on potential gains in study power (Miettinen, 1968, 1970). Matching can increase precision by decreasing the variability of the outcome within a matched pair. Early studies, however, suggested that the matched design was not guaranteed to be more efficient than its non-matched counterpart (Cochran, 1953; Youkeles, 1963; Billewicz, 1965; Chase, 1968; Miettinen, 1968). For example, Cochran (1953) showed the variance reduction obtained from pair-matching depended on the correlation between the matching variable and the outcome. Today tables of “break-even matching correlations” remain prominent in the literature (Martin et al., 1993; Diehr et al., 1995; Klar and Donner, 1997; Hayes and Moulton, 2009). These tables indicate when the unadjusted t-test and the paired t-test have the same power. Unfortunately, use of such tables requires in depth knowledge of the relationship between the matching variables and the outcome. Hayes and Moulton (2009), among others, are skeptical of the use of prior data on correlations, as estimates depend on temporal trends, sample sizes and the matching algorithm implemented. Likewise, Martin et al. (1993) warn against extrapolating correlations based on individual-level information to community-level relationships.

Even if one had strong a priori knowledge about the likely correlation between matching variables and the outcome in the target population, matching algorithms often discard units otherwise included in a random sample. McKinlay (1977), in particular, appreciated the information lost through the exclusion of these units. Furthermore, finding effective matches in cluster randomized trials may be especially difficult as the pool of possible candidates is generally small (Martin et al.,
Matching also induces a dependence between units. Thereby, the pair-matched study has half as many independent units as its non-matched, non-paired counterpart. The consequences of censoring may also be more drastic in pair-matched trials, where the dropout of a single unit can preclude the use of its companion in the analysis. Overall, pair-matching has the potential to make the randomized trial less powerful, as the variance is a direct function of the sample size (i.e. the number of independent units).

Most importantly, matching is a form of adjustment. Units are selected to be similar on baseline covariates, and estimators taking the difference in the outcome within matched pairs (e.g. the paired \( t \)-test) are controlling for these covariates. Consequently, unadjusted estimators in completely randomized trials have appeared less efficient than paired estimators in matched trials (Wacholder and Weinberg, 1982; Martin et al., 1993; Freedman et al., 1997; Klar and Donner, 1997; Imai, 2008; Imai et al., 2009; Hayes and Moulton, 2009). This apparent efficiency gain is “misleading” (McKinlay, 1977), as a fair comparison of design efficiency warrants adjusting for baseline covariates in the non-matched trial. In this paper, we evaluate impact of pair-matching on study efficiency, while respecting that the covariates used to create matched pairs can also be used to obtain an unbiased, adjusted estimate. Our results are estimator-independent and free from parametric modeling assumptions.

3 The Data, Model and Target Parameter

In this section, we describe the observed data, the statistical model and the target parameter for a completely randomized trial. We then consider the resulting efficient influence curve and its importance. We begin with a motivating example.

3.1 The Data Generating Experiment and Statistical Model

An estimated 195 million children are suffering from undernutrition, which not only attenuates physical and cognitive development but also plays a role in over a third of the deaths in children under five (United Nations Children’s Fund (UNICEF), 2009). Suppose we are interested in estimating the causal effect of ready-to-use therapeutic food (RUTF) on recovery from undernutrition. RUTF is peanut butter-type paste, fortified with milk proteins and essential nutrients, and does not require water for use (World Health Organization (WHO) et al., 2007). We propose a randomized trial to contrast the effect of RUTF with the standard supplement on the marginal proportion of children recovering from undernutrition. Specifically, we plan to randomize individual children to receive RUTF or the standard supplement, and then measure their nutritional status at two months.

Let us consider the data generating experiment for the proposed randomized trial. First, an undernourished child is randomly sampled. Then we measure his or her baseline covariates \( W \). These characteristics could include sex, age, weight, height, breastfeeding history and access to potable water. Next the treatment \( A \) is randomized. Let \( A \) be a binary variable, equalling one if the child received RUTF and zero if the child received the standard supplement. Finally, the outcome \( Y \) is a binary indicator, equalling one if the child recovered from undernutrition and zero otherwise. This data generating experiment is described by the following structural causal model...
(SCM) (Pearl, 2009):

\[ U = (U_W, U_A, U_Y) \sim P_U \]
\[ W = f_W(U_W) \]
\[ A = f_A(U_A) \]
\[ Y = f_Y(W, A, U_Y) \]

where the exogenous variables \( U = (U_W, U_A, U_Y) \) are not measured and the endogenous variables \( X = (W, A, Y) \) are observed. This SCM translates our knowledge of the data generating experiment into non-testable causal assumptions. Specifically, each endogenous variable is a deterministic function \( f_{X_j} \) of its parents. The observed baseline covariates \( W \) are some function of unmeasured background variables \( U_W \); the treatment \( A \) is randomly allocated, and the observed outcome \( Y \) is some unspecified function of baseline covariates \( W \), the treatment \( A \), and unmeasured error \( U_Y \). Since the treatment is randomized, \( A \) is known to be independent of \( W \) and the unmeasured error contributing to the treatment assignment \( U_A \) (e.g. a coin flip) is independent of the other errors. There is no unmeasured confounding.

Suppose we repeat the above experiment (sampling a child, measuring baseline covariates, randomizing the treatment and measuring the outcome) \( n \) times. Then the observed data would consist of \( n \) independent observations of random variable \( O = (W, A, Y) \), each identically distributed as \( P_0 \). The distribution \( P_0 \) is an element of the statistical model \( \mathcal{M} \), which specifies the collection of possible data generating distributions. (Throughout we use subscript 0 to denote the truth.) The statistical model is implied by the SCM and is semiparametric. In other words, we are assuming the observed data are generated by the above causal model, which only assumes the time-ordering \( W \rightarrow A \rightarrow Y \) and the randomization of the treatment. These assumptions are known to hold by design, and the set of possible distributions for the observed data is otherwise unrestricted.

### 3.1.1 The Analogous Cluster Randomized Trial

Alternatively, we could have conducted a cluster randomized trial to investigate the effect of a school-based supplementary feeding program on the marginal proportion of children recovering from undernutrition. For example, schools could be randomized to administer RUTF or the standard supplement to their students at lunchtime. Then the baseline covariates \( W \) could include geographic region, proximity to potable water, number of students, as well as summary measures of the distributions of sex, age, anthropometric indices, and socioeconomic status. The treatment \( A \) would still be a binary variable, indicating if the school was randomized to the intervention or to the control. Finally, the outcome \( Y \) would represent the proportion of children in each school recovering from undernutrition. The data generating experiment would be (i) randomly sampling a school, (ii) measuring baseline covariates, (iii) randomly allocating the treatment, and (iv) measuring the outcome at two months. Thereby, the above SCM applies equally to the cluster randomized trial. In other words, the individually randomized trial and the cluster randomized trial yield analogous data structures, whose generation can be described by the same causal model. The observed data is \( O = (W, A, Y) \) and distributed as \( P_0 \) in a semiparametric statistical model \( \mathcal{M} \). The only restriction on the possible probability distributions is that the treatment is randomized. Consequently, the following discussion applies to both individual and cluster randomized trials.

### 3.2 Counterfactuals and Identifiability of the Target Parameter

In the ideal world, the outcomes under both the treatment and the control would be observed. These counterfactual outcomes, denoted \( Y(a) \), are post-intervention random variables, corresponding with
setting $A = a$ on the SCM (Pearl, 2009). For the individually randomized trial, this framework allows us to ask, what would have been the marginal proportion of recoveries if all children had received RUTF compared to the marginal proportion if all children had received the control? Likewise, for the cluster randomized trial, we can compare the expected proportion of recoveries if all schools had administered RUTF compared to the expected proportion if none of the schools had administered it. In both scenarios, the target causal parameter is the difference in the means of the counterfactual distributions:

$$
\Psi^F(P_{XF}) = E_{XF}[Y(1)] - E_{XF}[Y(0)]
$$

where $XF = (W, Y(1), Y(0))$ represents the full data resulting from the ideal experiment and is distributed as $P_{XF}$. This target causal parameter is known as the average treatment effect or as the causal risk difference. Throughout, we use these names interchangeably.

In reality, we only observe one counterfactual outcome, corresponding with the observed treatment. Nonetheless, the average treatment effect can be identified from the observed data. In particular, given the randomization of the treatment, the target statistical parameter is

$$
\Psi(P_0) = E_0(Y|A = 1) - E_0(Y|A = 0) = E_0[E_0(Y|A = 1, W) - E_0(Y|A = 0, W)] = E_0[\bar{Q}_0(1, W) - \bar{Q}_0(0, W)]
$$

where $\bar{Q}_0(A, W) = E_0(Y|A, W)$ represents the conditional mean outcome, given the intervention and baseline covariates. We note that conditioning on baseline characteristics is not necessary for identifiability. Randomization guarantees there is no confounding. Stratification, however, allows us to adjust for chance imbalances in covariate distributions, reduces variation in the treatment-specific mean, and increases study power (Pocock et al., 2002; Tsiatis et al., 2008; Zhang et al., 2008; Moore and van der Laan, 2009; Rosenblum, 2011; Rubin and van der Laan, 2011). We also note that the positivity assumption is immediately satisfied, as the probability of receiving RUTF or the standard supplement is simply 0.5. (Also known as the experimental treatment assignment, this assumption requires there to be a positive probability of receiving each treatment in all stratas of baseline covariates. For a review this assumption and its consequences for estimation, see Petersen et al. (2012).) Overall, the target causal parameter has been identified as a well-defined and interesting parameter of the observed data.

### 3.3 The Efficient Influence Curve of the Target Parameter

The efficient influence curve is a fundamental quantity in statistical learning. It is the canonical gradient of the pathwise derivative of $\Psi$ at $P_0$ (Bickel et al., 1993). It is a mean zero function of the unit data and an element of the tangent space $T(P) \subset L^2_0(P)$. We acknowledge that readers may be unfamiliar with the gradients, influence curves and tangent spaces. We note, however, that it is not necessary to understand the mathematical theory behind these functions. Instead, the efficient influence curves for many statistical parameters (e.g. risk differences, risk ratios, odds ratios, survival functions) have already been derived (Bickel et al., 1993; van der Laan and Robins, 2003; van der Laan and Rose, 2011).

In particular, the efficient influence curve for the target parameter $\Psi(P_0) = E_0[\bar{Q}_0(1, W) - \bar{Q}_0(0, W)]$ is
\( \tilde{Q}_0(0, W) \) is the following function of the observed data

\[
D^*(P_0)(O) = D^*_Y(P_0) + D^*_W(P_0)
\]

with

\[
D^*_Y(P_0) = \left( \frac{I(A = 1)}{P_0(A = 1|W)} - \frac{I(A = 0)}{P_0(A = 0|W)} \right) (Y - \bar{Q}_0(A, W))
\]

\[
D^*_W(P_0) = \bar{Q}_0(1, W) - \bar{Q}_0(0, W) - \Psi(P_0)
\]

where \( I(\cdot) \) is the indicator function and \( P_0(A|W) \) is the conditional probability of treatment, given baseline covariates. We remind the reader that \( \bar{Q}_0(A, W) = E_0(Y|A, W) \) represents the conditional expectation of the outcome, given the treatment and baseline covariates. We also note that in a randomized trial the probability of receiving the intervention is independent of the covariates: \( P_0(A = 1|W) = P_0(A = 1) = 0.5 \). The first term \( D^*_Y(P_0) \) is the difference between the observed outcome and the expected outcome, weighted by the probability of the observed treatment. It is a mean zero function of the unit data \( O = (W, A, Y) \). The second term \( D^*_W(P_0) \) is the difference in the treatment-specific means minus the target parameter, which averages this difference over the baseline covariate distribution. It is a mean zero function of only \( W \). These two components of the efficient influence curve are orthogonal.

The variance of the efficient influence curve \( D^*(P_0) \) establishes a lower bound on the asymptotic variance of all regular, asymptotically linear estimators of the target parameter (Bickel et al., 1993). In the limit, these estimators are normally distributed with variance specified by their influence curve. In other words, an estimator’s influence curve is the foundation for statistical inference (e.g. hypothesis testing and confidence interval construction). More importantly, an estimator achieves the smallest possible variance if and only if its influence curve equals the efficient influence curve. (See Appendix A.4 in van der Laan and Rose (2011) for a self-contained proof.) Consequently, the efficient influence curve has been used in the construction of optimal double-robust estimating functions (Robins and Rotnitzky, 1992; van der Laan and Robins, 2003) and the targeted minimum loss estimator (van der Laan and Rubin, 2006; van der Laan and Rose, 2011). However, the goal of this paper is not to consider various estimators of the target parameter. Instead, we wish to evaluate the efficiency of pair-matching in randomized trials. As the variance of the efficient influence curve provides lower bound on the variance of all reasonable estimators, we can compare its variance under matched sampling and under independent sampling to understand the information provided in each design for estimation of the target parameter.

4 Design Alternatives

In the previous section, we presented a motivating example and two randomized trials that could be used to estimate the average treatment effect. First, we considered an individually randomized trial: (i) randomly sampling a child from the target population, (ii) measuring baseline covariates, (iii) randomly allocating the treatment, and (iv) measuring the outcome at two months. We also considered the analogous cluster randomized trial. In both studies, the resulting point treatment data structure was \( O = (W, A, Y) \) and distributed as \( P_0 \) in a semiparametric model \( \mathcal{M} \). We considered \( O \) as a random coarsening of the time-ordered full data structure \( X^F = (W, Y(1), Y(0)) \) and identified the causal risk difference from the observed data: \( \Psi(P_0) = E_0(Q_0(1, W) - Q_0(0, W)) \).

Finally, we considered the efficient influence curve of \( \Psi \) at \( P_0 \) and discussed its importance. In this section, we consider two alternative data generating experiments: the Independent Design and the Pair-Matched Design.
4.1 The Independent Design

Before conducting either trial, suppose we carry out a series of Monte Carlo simulations, where we generate the observed data $O$ according to the SCM (Eq. 1). Suppose these simulations suggest the vast majority of the trials will have an imbalance in the number of units receiving the treatment and the number receiving the control. We want to ensure that exactly half of the children (schools) receive RUTF and exactly half the control. Therefore, we decide to randomly sample two units and randomly allocate the treatment within that pair. The experimental unit then becomes a pair of observations:

$$O' = (O_1, O_0) = ((W_1, Y_1), (W_0, Y_0)) \sim P'_0$$

where the subscript $a$ denotes the treatment assignment. Specifically, $O_1$ represents the observed data for the child (school) receiving RUTF, while $O_0$ denotes the observed data for the companion child (school) receiving the standard supplement. Now the aforementioned SCM no longer describes the experiment that would generate the observed data. Instead, we can think of the data generating experiment as sampling a treated observation from the conditional distribution of $(W, Y)$, given $A = 1$, and then independently sampling a control observation from the conditional distribution of $(W, Y)$, given $A = 0$. Repeating this experiment $n$ times, the observed data set is comprised of $n$ independent, identically distributed (i.i.d.) observations of random variable $O' = (O_1, O_0) \sim P'_0$. The statistical model $M'$, describing the collection of possible probability distributions of $O'$, is semiparametric and implied by the underlying statistical model $M$. In particular, the marginal distributions of $(W, Y)$ are completely determined by the population distribution $P_0 \in M$. We refer to this data generating experiment as the “Independent Design” and note it is equivalent to pair-matching on the empty set.

4.2 The Pair-Matched Design

Now suppose that despite balanced treatment allocation, we are convinced that pair-matching children or schools on important determinants of the outcome is necessary for study credibility. For example, in the individually randomized trial, we might be worried if, by chance, a greater number of children suffering from severe acute malnutrition were randomized to the control. Likewise, policy makers and funders might be skeptical of the results of the cluster randomized study if, by chance, the majority of the control schools were in an urban slum. Consequently, we decide to pair-match units on a subset of measured baseline covariates and randomize the treatment within pairs. Let $M \subset W$ represent the selected matching variables, which are believed to be predictive of the outcome. These matching factors can be discrete or continuous; they can also be summary measures of the baseline covariates. Then the following data structure would be observed

$$O' = (O_1, O_0) = ((W_1, M_1, Y_1), (W_0, M_0 = M_1, Y_0)) \sim P'_0$$

As before, the observed data for the treated unit $O_1$ is sampled from the conditional distribution of $(W, M, Y)$, given $A = 1$. Now, however, the observed data for the control $O_0$ is sampled from the conditional distribution of $(W, M, Y)$, given $A = 0$ and $M = M_1$, where $M_1$ is the observed value of the matching variable in its treated companion:

$$(W_1, M_1, Y_1) \sim (W, M, Y | A = 1)$$

$$(W_0, M_0, Y_0) \sim (W, M, Y | A = 0, M = M_1)$$

For the individually randomized trial, the experimental unit now consists of a treated child and a control child, who share the value of matching variable $M_1 = M_0$. Analogously, for the cluster
randomized trial, the experimental unit is a pair of schools, sharing matching covariates. Then the observed data consist of \( n \) i.i.d. observations \( O'_1, \ldots, O'_n \) and are distributed as \( P'_0 \) in the semiparametric model \( M' \). Again, the marginal distributions of \((W, Y)\) are specified by the underlying distribution \( P_0 \in M \). We refer to this data generating experiment as the “Pair-Matched Design”.

5 The Asymptotic Efficiencies of the Independent Design and the Pair-Matched Designs

In the previous section, we considered two alternative data generating experiments. The Independent Design guaranteed treatment was evenly allocated. The Pair-Matched Design ensured balanced treatment allocation and that the covariates \( M \subset W \) were evenly distributed across study arms. In both cases, the experimental unit was a pair of observations, and the observed data were \( O' = (O_1, O_0) = ((W_1, M_1, Y_1), (W_0, M_0, Y_0)) \sim P'_0 \). In the Independent Design, the matching set is empty: \( M = \{\emptyset\} \). In this section, we use weights to map the efficient influence curve for the target parameter at the non-paired data distribution \( P_0 \) into the corresponding efficient influence curve at the paired data distribution \( P'_0 \). Then we compare the relative efficiency of the designs through the variances of their influence curves. This approach focuses on the target parameter of interest, avoids any parametric modeling assumptions and is estimator-independent.

5.1 Mapping Functions of Non-paired Data into Functions of Paired Data

A general weighting scheme can correct for the pairing of observations (van der Laan, 2008; Rose and van der Laan, 2009). The mapping was originally developed to adjust estimators and inference for biased sampling in case-control designs, but is widely applicable to many paired data structures. Specifically, any function \( F \) of the underlying random variable \( O = (W, A, Y) \sim P_0 \) can be mapped into a function \( F' \) of the observed random variable \( O' = (O_1, O_0) \sim P'_0 \) by

\[
F'(O') \equiv q_0 F(O_1) + \bar{q}_0(M) F(O_0) \\
= q_0 F(W_1, M_1, Y_1) + \bar{q}_0(M) F(W_0, M_0, Y_0)
\]

where \( q_0 = P_0(A = 1) \) is the marginal probability of being allocated the treatment,

\[
\bar{q}_0(M) = q_0 \frac{P_0(A = 0|M)}{P_0(A = 1|M)}
\]

is a ratio of treatment probabilities, and \( F(O_a) \) is the function evaluated \( A = a \).

Now let us map the efficient influence curve for the target parameter \( \Psi \) under sampling from \( P_0 \) into its weighted counterpart for paired sampling from \( P'_0 \). In both the Independent Design and the Pair-Matched Design, the treatment is randomized within pairs. Therefore, the weight for the treated observation \( O_1 \) is \( q_0 = 0.5 \), and the weight for the control observation \( O_0 \) is also \( \bar{q}_0(M) = 0.5 \). (Had the treatment not been randomized, the weights for the control observations would not be equal under the two designs.) As a result, the efficient influence curves for target parameter are equivalent under independent and matched sampling:

\[
D^*_I(O') = D^*_M(O') = 0.5D^*(O_1) + 0.5D^*(O_0)
\]  

(2)

Here, the superscript \( I \) denotes the Independent Design and the superscript \( M \) denotes the Pair-Matched Design. \( D^*(O_a) \) is the efficient influence curve at the underlying distribution \( P_0 \) evaluated...
5.2 The Variance of the Efficient Influence Curve

The information contained in the Independent Design for estimating the average treatment effect is determined by the variance of the weighted efficient influence curve:

\[ \text{Var}[D^*(O')] = 0.25 \text{Var}[D^*(O_1)] + 0.25 \text{Var}[D^*(O_0)] \]

The covariance term \( \text{Cov}[D^*(O_1), D^*(O_0)] \) is zero as the treated child (school) and the control child (school) within a pair are independently sampled. On the other hand, the variance of the most efficient estimator in the Pair-Matched Design is determined by

\[ \text{Var}[D^{*M}(O')] = 0.25 \text{Var}[D^*(O_1)] + 0.25 \text{Var}[D^*(O_0)] + 0.5 \text{Cov}[D^*(O_1), D^*(O_0)] \]

Here, the covariance term is non-zero, as the observations share characteristics \( M_1 = M_0 \). Since the marginal distribution of \( O_a \) is the same under both the designs, we can re-express the variance of the efficient influence curve under matched sampling as

\[ \text{Var}[D^{*M}(O')] = \text{Var}[D^{*I}(O')] + 0.5 \text{Cov}[D^*(O_1), D^*(O_0)] \]  

Thereby, the Pair-Matched Design will be asymptotically more efficient if and only if the covariance between \( D^*(O_1) \) and \( D^*(O_0) \) is negative.

Recall that the efficient influence curve is the sum of two orthogonal pieces: \( D^* = D^*_Y + D^*_W \). Therefore, the covariance of \( D^*(O_1) \) with \( D^*(O_0) \) can also be expressed as a sum of covariances:

\[ \text{Cov}[D^*(O_1), D^*(O_0)] = \text{Cov}[D^*_Y(O_1), D^*_Y(O_0)] + \text{Cov}[D^*_W(O_1), D^*_W(O_0)] \]

Consequently, the efficiency gained or lost by pair-matching depends on the magnitude and direction of the following expectations

\[ \text{Cov}[D^*_Y(O_1), D^*_Y(O_0)] = E_0 \left[ \left( \frac{Y_1 - \bar{Q}_0(1, W_1)}{P_0(A = 1|W_1)} \right) \left( \frac{Q_0(0, W_1) - \bar{Q}_0(0, W_1) - \Psi(P_0)}{P_0(A = 0|W_0)} \right) \right] \]

\[ \text{Cov}[D^*_W(O_1), D^*_W(O_0)] = E_0 \left[ (\bar{Q}_0(1, W_1) - \bar{Q}_0(0, W_1) - \Psi(P_0))(\bar{Q}_0(1, W_0) - \bar{Q}_0(0, W_0) - \Psi(P_0)) \right] \]

The covariance of the \( D^*_Y \) terms will typically be negative, while the covariance of the \( D^*_W \) terms will typically be positive.

Let us consider the ideal situation where the measured baseline covariates \( W \) are strongly and independently predictive of the outcome. Let us further assume that we are able to match exactly on all of them: \( M = W \). Then the deviations between the observed outcome \( Y_a \) and the expected outcome \( \bar{Q}(a, W_a) \) are going to be small. For example, imagine the outcome is a simple function of the treatment, covariates and error: \( Y = \beta_0 + \beta_1 A + \beta_2 W + \beta_3 AW + \epsilon, \) with \( \epsilon \sim N(0, \sigma_\epsilon^2) \). Then the
covariance of the $D_Y^*$ terms is exactly $-4E(e^2) = -4\sigma^2$. As more of the variation in the outcome is explained by the covariates, this residual shrinks to zero. Now let us consider the covariance of the $D_W^*$ components of the efficient influence curve. When matching on all baseline covariates, the $D_W^*$ terms will be exactly equal at $O_1$ and $O_0$:

$$E_0[D_W^*(O_1)D_W^*(O_0)] = E_0[\bar{Q}_0(1,w) - \bar{Q}_0(0,w) - \Psi(P_0)^2] \geq 0$$

In many cases, the difference in the treatment-specific means $\bar{Q}_0(1,w) - \bar{Q}_0(0,w)$ will diverge from the target parameter $\Psi$, which averages over the covariate distribution. This divergence will grow as the matching variables have a stronger effect on the outcome. Thus, when matching on strong predictors of the outcome, the covariance of the $D_Y^*$ components shrinks to zero, while the covariance of the $D_W^*$ components will often be large and positive. Consequently, the sum of the covariances is easily positive and the Pair-Matched Design asymptotically less efficient than the Independent Design.

5.3 An Alternative Data Generating Experiment

In the previous subsection, the variance of the efficient influence curve indicated that the Pair-Matched Design was more efficient if and only if the covariance term was negative. This conclusion might seem counterintuitive. Indeed, there appears to be a consensus in the literature that matching on a strong predictor of the outcome should induce a positive correlation between outcomes within pairs and thereby reduce the variance of the estimator (e.g. Martin et al. (1993); Diehr et al. (1995); Klar and Donner (1997); Hayes and Moulton (2009)). In this subsection, we explore the foundation of this claim.

Let us consider a data generating experiment, where we randomly assign the treatment $A$ and measure the outcome $Y$ at the end of followup. As no baseline covariates are measured, the observed data would be $O_r = (A, Y)$ and distributed as $P_r$, known to be an element of the semiparametric statistical model $M_r$. Here, we use subscript $r$ to emphasize this data structure, probability distribution and statistical model are reduced compared to those with measured covariates $W$. Given the randomization of the intervention, the average treatment effect can still be identified as an element of $P_r$ by

$$\Psi_r(P_r) = E(Y|A = 1) - E(Y|A = 0) = \bar{Q}(1) - \bar{Q}(0)$$

This target statistical parameter is the unadjusted difference in the treatment-specific means. Its efficient influence curve at $P_r$ is

$$D_r^*(P_r)(O_r) = \left( \frac{I(A = 1)}{P(A = 1)} - \frac{I(A = 0)}{P(A = 0)} \right) (Y - \bar{Q}(A))$$

where $P(A)$ is the marginal probability of treatment and equal to 0.5 in a randomized trial.

Now consider the analogous paired experiments. In the Independent Design, researchers would randomly pair two units, randomize the treatment and measure the outcomes. Alternatively, in the Pair-Matched Design, researchers would match units on unmeasured characteristics, randomize the treatment, and record the outcomes. For example, consider the opthalmic trials, where eyes are matched exactly on all unmeasured characteristics of the patient, the treatment randomized, and the outcomes measured. In both designs, the observed data are

$$O'_r = (O_1, O_0) = (Y_1, Y_0) \sim P'_r$$
As before, $Y_1$ represents the observed outcome under treatment, and $Y_0$ represents the observed outcome under the control. Random variable $O'_r$ is distributed as $P'_r$, known to be an element of a semiparametric statistical model $M'_r$. Once more, $M'_r$ is implied by the underlying statistical model $M_r$.

Intuitively, we expect that matching on important determinants of the outcome (even if not measured) will reduce the variability within pairs and thereby the variance of the estimator. In other words, we expect an ophthalmic trial randomizing the treatment within eyes of a single patient to be more efficient than another trial randomizing the treatment to eyes of different patients. To be more rigorous, we need to map the efficient influence curve at the non-paired, reduced distribution more efficient than another trial randomizing the treatment to eyes of different patients. To be more words, we expect an opthalmic trial randomizing the treatment within eyes of a single patient to be matched on unmeasured covariates. As an example, we discussed opthalmic trials, in which eyes are naturally matched on all baseline covariates of the patient. Another example is twin studies, where the treatment is randomized within monozygotic pairs (Hunter et al., 2000). It is very

where the treatment is randomized within monozygotic pairs (Hunter et al., 2000). It is very

Once more, the Pair-Matched Design is asymptotically less efficient in many cases. Matching often induces a positive covariance between observations within a pair and decreases the precision the design (Eq. 3).

For comparison, we also considered the less common data generating experiment, where units are paired on unmeasured covariates. As an example, we discussed ophthalmic trials, in which eyes are naturally matched on all baseline covariates of the patient. Another example is twin studies, where the treatment is randomized within monozygotic pairs (Hunter et al., 2000). It is very

as the marginal distribution of $Y_0$ is determined by the underlying distribution $P_r$. Consequently, the Pair-Matched Design is asymptotically more efficient when units are matched on unmeasured predictors of the outcome. We emphasize that the function $D_r^*(P'_r)(O'_r)$ is the efficient influence curve if and only if the observed data are $O'_r = (Y_1, Y_0) \sim P'_r$. In other words, the above discussion applies to the case where there are no measured baseline covariates.

$\bar{Y}_r$ is distributed as $Y_r$.

Then the variance of the efficient influence curve in the Independent Design at $P'_r$ is

$$\text{Var}[D_r^I(O'_r)] = \text{Var}[Y_1 - \bar{Q}(1)] + \text{Var}[Y_0 - \bar{Q}(0)]$$

Again, the covariance is zero, as the treatment-specific outcomes are independently sampled. On the other hand, the variance of the efficient influence curve in the Pair-Matched Design at $P'_r$ is

$$\text{Var}[D_r^M(O'_r)] = \text{Var}[Y_1 - \bar{Q}(1)] + \text{Var}[Y_0 - \bar{Q}(0)] - 2\text{Cov}[(Y_1 - \bar{Q}(1), (Y_0 - \bar{Q}(0))]$$

Here, the covariance between observations within a pair is non-zero as the units are known to be similar on important (but unmeasured) determinants of the outcome. Furthermore, the covariance will be positive, whenever the matching variables are predictive of the outcome. Expressing the efficiency of the Pair-Matched Design in terms of the efficiency of the Independent Design yields

$$\text{Var}[D_r^M(O'_r)] = \text{Var}[D_r^I(O'_r)] - 2\text{Cov}[(Y_1 - \bar{Q}(1), (Y_0 - \bar{Q}(0))]$$

(5) as the marginal distribution of $Y_0$ is determined by the underlying distribution $P_r$. Consequently, the Pair-Matched Design is asymptotically more efficient when units are matched on unmeasured predictors of the outcome. We emphasize that the function $D_r^*(P'_r)(O'_r)$ is the efficient influence curve if and only if the observed data are $O'_r = (Y_1, Y_0) \sim P'_r$. In other words, the above discussion applies to the case where there are no measured baseline covariates.

### 5.4 Recap of Theoretical Results

In most pair-matched randomized trials, the data generating experiment involves measuring baseline covariates $W$, matching two units on a subset of these characteristics $M \subset W$, randomizing the treatment $A$ within the pair, and measuring the outcomes $Y$ at the end of followup. The resulting data structure is $O' = ((W_1, M_1, Y_1), (W_0, M_0, Y_0))$ and distributed as $P'_r \in M'_r$. By examining the variance of the efficient influence curve for the target parameter, we concluded that the Pair-Matched Design is asymptotically less efficient in many cases. Matching often induces a positive covariance between observations within a pair and decreases the precision the design (Eq. 3).

For comparison, we also considered the less common data generating experiment, where units are paired on unmeasured covariates. As an example, we discussed ophthalmic trials, in which eyes are naturally matched on all baseline covariates of the patient. Another example is twin studies, where the treatment is randomized within monozygotic pairs (Hunter et al., 2000). It is very
difficult to imagine an analogous cluster randomized trial. Nonetheless, we do not disregard the possibility nor the validity of such an experiment. In any case, when baseline covariates are not measured, the observed data are simply $O'_r = (Y_1, Y_0)$ and distributed as $P'_r \in M'_r$. The variance of the efficient influence curve indicated that the Pair-Matched Design is asymptotically more efficient for estimation of the unadjusted difference in treatment-specific means. Matching on strong but unmeasured predictors of the outcome induces a positive covariance and thereby increases the precision of the Pair-Matched Design (Eq. 5).

6 Simulation Studies

In this section, we explore the theoretical results. Specifically, we simulate data for an individually randomized and a cluster randomized trial and then evaluate the information generated from different designs for estimation of the average treatment effect. We acknowledge that these simulations do not reflect the complexities of real data. Instead, their purpose is to concretely illustrate the theoretical concepts presented above.

6.1 Simulation Methodology

Suppose there is a single baseline covariate of interest $W$. For simplicity, let $W$ be drawn from a normal distribution with mean 0 and variance 0.2$^2$. Let the intervention $A$ be binary and randomized as $P_0(A|W) = P_0(A) = 0.5$. Suppose the individual-level outcome is also binary $Y \in \{0, 1\}$. Then the corresponding cluster-level outcome is the proportion $Y \in [0, 1]$. In either case, let us define the function $h(A, W)$ of the baseline covariate $W$ and the treatment $A$:

$$h(A, W) = \text{expit} \left[ \beta_0 + \beta_1 A + \beta_2 W + \beta_3 AW \right]$$

where the expit($x$) is the inverse logistic function and equal to $\frac{1}{1+\exp(-x)}$. For an individually randomized trial, the binary outcome is drawn from a Bernoulli distribution with probability given by $h(A, W)$:

$$Y \sim \text{Bernoulli}(p = h(A, W))$$

For the cluster randomized trial, suppose there are $J = 10$ individuals within each cluster. Then $J$ binary outcomes are drawn from a binomial distribution with probability $p = h(A, W)$, and the cluster-level outcome $Y$ is taken as the average. We also consider a cluster randomized trial with $J = 2500$ individuals subsampled. This allows us to examine the effect of subsample size on the relative efficiency of the designs. Alternatively, the proportion of individuals with the outcome can be directly simulated as

$$Y = h(A, W) + \epsilon$$

where $\epsilon \sim N(0, 0.01^2)$.

By varying the regression coefficients $\beta_0$ and $\beta_1$ in function $h(A, W)$, we can understand how the asymptotic efficiency is impacted by the rarity of the outcome. To simulate a rare outcome, we set these coefficients to yield cumulative incidences of 1% in the control and 0.6% in the treated. For a more common outcome, these coefficients are set to yield cumulative incidences of 25% in the control and 18.75% in the treated. These values were chosen to correspond with realistic incidences and effect sizes in public health. Also, to understand the effects of matching on a variable strongly, moderately or weakly predictive of the outcome, we set the coefficient $\beta_2$ to 4, 3, and 2, respectively. Finally, we consider scenarios without an interaction, $\beta_3 = 0$, and with a weak interaction, $\beta_3 = -0.75$. Among treated observations, this interaction term reduces, but does not reverse, the effect of the baseline covariate on the outcome.
Our interest is not to understand the asymptotic variance of the completely randomized design: $O = (W,A,Y) \sim P_0$. Instead, we want to simulate paired data. Specifically, the data generating experiment for the Independent design is sampling $(W, M, Y)$ conditional on $A = 1$ and then sampling $(W, M, Y)$ conditional on $A = 0$. The data generating experiment for the Pair-Matched Design is sampling $(W, M, Y)$ given $A = 1$, and then sampling $(W, M, Y)$ given $A = 0$ and $M = M_1$. We consider two versions of matching. In the first, we match the control unit to the treated exactly on the baseline covariate: $W$. In the second, we match only on quartiles of the baseline covariate: $M_1 = \Phi(W_1) = M_0$. In either case, the observed data are

$$O' = (O_1, O_0) = ((W_1, M_1, Y_1), (W_0, M_0, Y_0)) \sim P'_0 \in \mathcal{M}'$$

For each design and each scenario, we calculated the true variance of the efficient influence curve. Specifically, we generated $N = 50,000$ pairs of observations and computed the sample variance of the influence curve values. We then compared the relative efficiency of the designs with

$$RE_{true} = \frac{\text{Var}[D^{rI}(O')]}{\text{Var}[D^{rM}(O')]}$$

where $D^{rI}(O')$ is the efficient influence curve at $P'_0$ in the Independent Design and $D^{rM}(O')$ is the efficient influence curve at $P'_0$ in the Pair-Matched Design.

We also considered the relative efficiency of the designs, when we failed to include the baseline covariate $W$ in the specification of the data, statistical model and target parameter. Indeed, for each simulated data set, we evaluated the efficient influence curve as if the observed data were only $O'_r = (Y_1, Y_0) \sim P'_r \in \mathcal{M}_r'$ and as if the target statistical parameter were $\Psi_r(P_r) = \bar{Q}(1) - \bar{Q}(0)$. Under this misspecification, the relative efficiency of the designs is given by

$$RE_{misp} = \frac{\text{Var}[D^{rI}(O'_r)]}{\text{Var}[D^{rM}(O'_r)]}$$

where $D^{rI}(O'_r)$ is the efficient influence curve at $P'_r$ in the Independent Design and $D^{rM}(O'_r)$ is the efficient influence curve at $P'_r$ in the Pair-Matched Design. We note that this would be the correct efficiency comparison for the less common data generating experiment, where units are paired on unmeasured and unobserved baseline covariates. This, however, is not the data generating experiment for the simulation. All statistical programing and computing was done R version 2.14.1 (R Development Core Team, 2011).

6.2 Simulation Results

In Table 1 we explore the relative efficiency of the designs when researchers are able to exactly match the continuous baseline covariate: $W_1 = W_0$. This represents the ideal scenario for the Pair-Matched Design. The first column indicates the type of study: individually randomized or cluster randomized. Clust$^{10}$ indicates the outcome $Y$ is the average of the simulated outcomes of $J = 10$ individuals; Clust$^{2500}$ indicates the outcome $Y$ is the average of the simulated outcomes of $J = 2500$ individuals, and Clust* indicates the cluster-level outcome is directly simulated according to Eq. 6. The second column indicates whether the outcome was rare or common. The remaining columns give the asymptotic efficiency of the Independent Design relative to the Pair-Matched Design under various scenarios. Specifically, $RE_{misp}$ represents the relative efficiency of the designs when we have ignored the baseline covariates and thereby misspecified the observed data as $O'_r = (Y_1, Y_0) \sim P'_r$. Conversely, $RE_{true}$ represents the true relative precision of the designs and is the variance of the efficient influence curve at $P'_0$ under independent sampling divided by the variance of the efficient
<table>
<thead>
<tr>
<th>Type</th>
<th>Outcome</th>
<th>No interaction</th>
<th>Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Strong</td>
<td>Moderate</td>
</tr>
<tr>
<td>Indv</td>
<td>Rare</td>
<td>$RE_{misp}$</td>
<td>0.987</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$RE_{true}$</td>
<td>0.977</td>
</tr>
<tr>
<td></td>
<td>Common</td>
<td>$RE_{misp}$</td>
<td>1.116</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$RE_{true}$</td>
<td>1.004</td>
</tr>
<tr>
<td>Clust$^{10}$</td>
<td>Rare</td>
<td>$RE_{misp}$</td>
<td>1.083</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$RE_{true}$</td>
<td>0.991</td>
</tr>
<tr>
<td></td>
<td>Common</td>
<td>$RE_{misp}$</td>
<td>2.114</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$RE_{true}$</td>
<td>0.996</td>
</tr>
<tr>
<td>Clust$^{2500}$</td>
<td>Rare</td>
<td>$RE_{misp}$</td>
<td>6.602</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$RE_{true}$</td>
<td>0.610</td>
</tr>
<tr>
<td></td>
<td>Common</td>
<td>$RE_{misp}$</td>
<td>62.822</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$RE_{true}$</td>
<td>0.609</td>
</tr>
<tr>
<td>Clust$^*$</td>
<td>Rare</td>
<td>$RE_{misp}$</td>
<td>1.809</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$RE_{true}$</td>
<td>0.943</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$RE_{true}$</td>
<td>0.650</td>
</tr>
</tbody>
</table>

Table 1: The asymptotic efficiency of the Independent Design relative to the Pair-Matched Design, when matching exactly on a continuous baseline covariate. The first column refers to the type of trial and the method of simulation. The second column indicates whether the outcome was rare or common. The third column denotes whether the efficiency comparison is made under misspecification or the correct specification of the data, model and target parameter.

$RE_{misp}$ is the variance of the efficient influence curve at misspecified $P_{0}'$ under independent sampling divided by the variance of the efficient influence curve under matched sampling at misspecified $P_{0}'$. $RE_{true}$ is the ratio of these variances at the true data generating distribution $P_{0}$. The remaining columns give the relative efficiencies for the six scenarios examined. The strength of the relationship of $W$ on $Y$ is varied with $\beta_2$, and the interaction is varied with $\beta_3$ equal to 0 or -0.75.

For the individually randomized trial, matching does not appear to induce meaningful gains or losses in the asymptotic efficiency. The Pair-Matched Design yields approximately the same amount of information for estimating the average treatment effect as the Independent Design. When the observed baseline covariates are excluded from the data, model and target parameter, the matched design appears at most 11.6% more efficient than its non-matched counterpart. However, when the true data generating experiment for the simulation is respected, this difference is inconsequential and reduced to a maximum gain and loss of 0.4%. This is not surprising as the variability in the outcomes outweighs the positive and negative contributions of the covariance terms. Indeed, the covariance term is at most 1.67% of the total variance of the efficient influence curve at $P_{0}'$ in the Pair-Matched Design. (See Table 2 for the ratio of the covariance of $D^*(O_1)$ and $D^*(O_0)$ to the total variance of the efficient influence curve in the Pair-Matched Design. Recall the covariance is zero in the Independent Design.)

Similar precision results are seen for the cluster randomized trial with a subsample of $J = 10$ individuals (Table 1). When we misspecify the data generating experiment by ignoring measured baseline covariates, the Pair-Matched Design appears at most 2.114 times more efficient than the Independent Design. Conversely, when we correctly specify the observed data, the design employing matching is up to 3% less efficient than its non-matched counterpart. Again, it is not surprising
<table>
<thead>
<tr>
<th>Type</th>
<th>Outcome</th>
<th>No interaction</th>
<th>Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Strong</td>
<td>Moderate</td>
</tr>
<tr>
<td>Indv</td>
<td>Rare</td>
<td>Cov:Var\text{mis}p</td>
<td>0.0004</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cov:Var\text{true}</td>
<td>0.0167</td>
</tr>
<tr>
<td></td>
<td>Common</td>
<td>Cov:Var\text{mis}p</td>
<td>0.0605</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cov:Var\text{true}</td>
<td>-0.0109</td>
</tr>
<tr>
<td>Clust\text{10}</td>
<td>Rare</td>
<td>Cov:Var\text{mis}p</td>
<td>0.0489</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cov:Var\text{true}</td>
<td>-0.0135</td>
</tr>
<tr>
<td></td>
<td>Common</td>
<td>Cov:Var\text{mis}p</td>
<td>0.5521</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cov:Var\text{true}</td>
<td>0.0061</td>
</tr>
<tr>
<td>Clust\text{2500}</td>
<td>Rare</td>
<td>Cov:Var\text{mis}p</td>
<td>2.9950</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cov:Var\text{true}</td>
<td>0.7304</td>
</tr>
<tr>
<td></td>
<td>Common</td>
<td>Cov:Var\text{mis}p</td>
<td>31.5072</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cov:Var\text{true}</td>
<td>0.7796</td>
</tr>
<tr>
<td>Clust\text{*}</td>
<td>Rare</td>
<td>Cov:Var\text{mis}p</td>
<td>0.4086</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cov:Var\text{true}</td>
<td>0.0948</td>
</tr>
<tr>
<td></td>
<td>Common</td>
<td>Cov:Var\text{mis}p</td>
<td>27.6952</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cov:Var\text{true}</td>
<td>0.6947</td>
</tr>
</tbody>
</table>

Table 2: The ratio of the covariance term to the total variance of the efficient influence curve for the Pair-Matched Design, when matching exactly on a continuous covariate. The first column refers to the type of trial and the method of simulation. The second column indicates whether the outcome was rare or common. The third column denotes whether the comparison is made under misspecification or the correct specification of the data, model and target parameter. Cov:Var\text{mis}p is the ratio of the covariance term to the total variance of the efficient influence curve at the misspecified $P_r'$. Cov:Var\text{true} is the ratio of the covariance term to the total variance of the efficient influence curve at the true data generating distribution $P_0'$. The remaining columns give the ratios for the six scenarios examined. The strength of the relationship of $W$ on $Y$ is varied with $\beta_2$, and the interaction is varied with $\beta_3$ equal to 0 or -0.75.

that matching has a little effect when only $J = 10$ individuals are subsampled, as the variability in the cluster-level outcome is large. For example, the covariance term is at most 5.74\% of the total variance at the true $P_0'$ in the Pair-Matched Design (Table 2).

The efficiency results are quite dramatic for the cluster randomized trial with $J = 2500$ individuals subsampled (Table 1). The Pair-Matched Design appears unequivocally more efficient under misspecification ($RE_{\text{mis}p} > 1$), and is unequivocally less efficient under the correct specification ($RE_{\text{true}} < 1$). In particular, the matched design appears up to 62.822 times more efficient than its non-matched counterpart, assuming the efficient influence curve for $O_r' = (Y_1, Y_0)$. Conversely, it is 64.2\% less efficient in the same scenario, assuming the efficient influence curve for the true observed data $O'$. Indeed, for one simulation (matching on a strong predictor of a common outcome in the presence of interaction), the Independent Design is 89.4\% more efficient than the Pair-Matched Design. In all scenarios, the covariance term is large and positive (Table 2). It contributes up to 93.73\% of the total variance at the true $P_0'$ in the Pair-Matched Design.

Similar precision results are seen for the cluster randomized trial, where the proportion of individuals with the outcome is directly simulated (Table 1). The Pair-Matched Design appears up to 56.667 times more efficient, when the baseline covariates are excluded from specification of the data, model and target parameter. Under the correct specification, the matched design is up to 54.7\% less efficient than its non-matched counterpart. Again, the covariance term is positive in all scenarios (Table 2). It contributes up to 90.55\% of the total variance at the true $P_0'$ in the Pair-Matched Design.
Table 3: The asymptotic efficiency of the Independent Design relative to the Pair-Matched Design, when matching on a coarsening of the baseline covariates. The first column refers to the type of trial and the method of simulation. The second column indicates whether the outcome was rare or common. The third column denotes whether the efficiency comparison is made under misspecification or the correct specification of the data, model and target parameter. \( RE_{mis} \) is the variance of the efficient influence curve at misspecified \( P_r \) under independent sampling divided by the variance of the efficient influence curve under matched sampling at misspecified \( P_r \). \( RE_{true} \) is the ratio of these variances at the true data generating distribution \( P_0^* \). The remaining columns give the relative efficiencies for the six scenarios examined. The strength of the relationship of \( W \) on \( Y \) is varied with \( \beta_2 \), and the interaction is varied with \( \beta_3 \) equal to 0 or -0.75. In the Pair-Matched Design, units were matched on quartiles \( M_1 = \Phi(W_1) = M_0 \).

<table>
<thead>
<tr>
<th>Type</th>
<th>Outcome</th>
<th>No interaction</th>
<th>Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Strong</td>
<td>Moderate</td>
</tr>
<tr>
<td>Indv</td>
<td>Rare</td>
<td>( RE_{mis} )</td>
<td>1.004</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( RE_{true} )</td>
<td>0.998</td>
</tr>
<tr>
<td></td>
<td>Common</td>
<td>( RE_{mis} )</td>
<td>1.087</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( RE_{true} )</td>
<td>0.994</td>
</tr>
<tr>
<td>Clust(^{10})</td>
<td>Rare</td>
<td>( RE_{mis} )</td>
<td>1.045</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( RE_{true} )</td>
<td>0.992</td>
</tr>
<tr>
<td></td>
<td>Common</td>
<td>( RE_{mis} )</td>
<td>1.802</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( RE_{true} )</td>
<td>0.994</td>
</tr>
<tr>
<td>Clust(^{2500})</td>
<td>Rare</td>
<td>( RE_{mis} )</td>
<td>2.110</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( RE_{true} )</td>
<td>0.725</td>
</tr>
<tr>
<td></td>
<td>Common</td>
<td>( RE_{mis} )</td>
<td>6.143</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( RE_{true} )</td>
<td>0.628</td>
</tr>
<tr>
<td>Clust(^{*})</td>
<td>Rare</td>
<td>( RE_{mis} )</td>
<td>1.366</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( RE_{true} )</td>
<td>0.967</td>
</tr>
<tr>
<td></td>
<td>Common</td>
<td>( RE_{mis} )</td>
<td>6.013</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( RE_{true} )</td>
<td>0.669</td>
</tr>
</tbody>
</table>

The last two sets of simulations illustrate the following trends in design efficiency. When the measured baseline covariates are excluded, the Pair-Matched Design appears most efficient when the covariates are strongly predictive of the outcome and in the absence of the interaction. When we correctly specify the data generating experiment by including the measured covariates, the Pair-Matched Design tends to be least efficient when matching on covariates strongly predictive of the outcome and in the presence of the interaction. We also explored the relative efficiency when units were matched only on quartiles of baseline covariates, \( M_1 = \Phi(W_1) = M_0 \). The simulation results, given in Table 3, are less dramatic, but the overall trends hold. When variability in the outcome outweighs the positive or negative contributions of the covariance term, the designs yield approximately the same amount of information for estimating the average treatment effect. Otherwise, the Pair-Matched Design appears asymptotically more efficient when the data, model and target parameter are misspecified and is asymptotically less efficient under the correct specification.

### 6.3 Finite Sample Simulations

Our previous results focused on the asymptotic efficiency lost by pair-matching. Understanding the finite sample performance, however, is essential for the application of this work to cluster randomized trials, which usually have very limited sample sizes. Consequently, we simulated 5000 cluster randomized trials, each limited to \( n = 10 \) or \( n = 20 \) independent units. As before, the
baseline covariate $W$ was drawn from $N(0, 0.2^2)$ and the treatment $A$ randomized. We focused on the cluster randomized trial where $J = 2500$ individual outcomes were drawn from a binomial distribution with probability $p = h(A, W)$ and the cluster-level outcome $Y$ was the average. We further supposed that with such a limited sample size, researchers would only be able to measure and to match on quartiles of the baseline covariate. Let $W$ be a four-level categorical variable, representing the observed quartiles of $W$. Then the data generating experiment for the Independent Design was sampling $(W, Y)$ given $A = 1$ and independently sampling $(W, Y)$ given $A = 0$. The data generating experiment for the Pair-Matched Design was sampling $(W, Y)$ given $A = 1$ and then sampling $(W, Y)$ given $A = 0$ and $W = W_1$. In both cases, the following data structure was simulated

$$O' = (O_1, O_0) = ((W_1, Y_1), (W_0, Y_0)) \sim P_0'$$

where $W_1 = W_0$ under matching. Again, we varied the rarity of the outcome, the presence of an interaction, and the baseline covariate’s effect on the outcome.

For each trial, we estimated the target parameter using weights to map standard estimators into their paired counterparts (van der Laan, 2008). Specifically, with $n$ i.i.d. observations of the non-paired data $O = (W, A, Y) \sim P_0$, a simple substitution of $\Psi(P_0) = E_0[\bar{Q}_0(1, W) - \bar{Q}_0(0, W)]$ is

$$\Psi_n(P_n) = \frac{1}{n} \sum_{i=1}^n \bar{Q}_n(1, W_i) - \bar{Q}_n(0, W_i)$$

where $P_n$ denotes the empirical probability distribution, $i$ indexes independent observations, and $\bar{Q}_n(A, W)$ is an estimate of the treatment-specific mean outcome, given baseline covariate category $W$. Then with $n$ i.i.d. observations of the paired data $O'$, the weighted substitution estimator is

$$\Psi_n(P'_n) = \frac{1}{n} \sum_{i=1}^n \bar{q}_0 \{ \bar{Q}_n(1, W_{1i}) - \bar{Q}_n(0, W_{1i}) \} + \bar{q}_0 \{ \bar{Q}_n(1, W_{0i}) - \bar{Q}_n(0, W_{0i}) \}$$

where weights $\bar{q}_0$ and $\bar{q}_0$ are 0.5, given the randomization of the treatment.

To estimate the conditional mean function $\bar{Q}_n(A, W)$, we performed three simple regressions. In Model 1, we regressed the observed outcome $Y$ on the treatment $A$ and observed covariate $W$. In Model 2, we created dummy variables for the categories of $W$, and regressed the observed outcome $Y$ on the treatment $A$ and these indicators. In Model 3, we created dummy variables for each combination of treatment and covariates, and regressed the observed outcome $Y$ on these indicators. In all cases, we used logistic regression to guarantee the predicted outcomes and effect estimate remained bounded between zero and one. We consider these regression functions as working models, in that we do not assume the true data generating distribution $P_0$ is an element of the parametric statistical model. In fact, we know these models are misspecified as the outcome $Y$ was simulated according to $h(A, W)$, which is a function of the treatment $A$ and the unobserved, continuous covariate $W$. Nonetheless, when a generalized linear model for the conditional mean function includes an intercept and a treatment indicator, the resulting estimate of the causal effect will be unbiased and locally efficient in randomized trials (Scharfstein et al., 1999). In other words, the substitution estimators implementing parametric models 1 and 2 are targeted maximum likelihood estimators and guaranteed to solve the efficient score equation (Moore and van der Laan, 2009; Rosenblum, 2011; Rubin and van der Laan, 2011). For comparison, we also estimated the target parameter as the average of the unadjusted difference of the outcomes within pairs:

$$\Psi_n(P'_n) = \frac{1}{n} \sum_{i=1}^n Y_{1i} - Y_{0i}$$

http://biostats.bepress.com/ucbbiostat/paper294
This would be the non-parametric maximum likelihood estimator, if the observed data were \( O_r = (Y_1, Y_0) \) and we were interested in the unadjusted risk difference: \( \Psi_r(P_r) = \bar{Q}(1) - \bar{Q}(0) \). Finally, to determine the finite sample performance, we calculated the variance of each estimator over 5,000 simulations. The ratio of the estimator’s variance under independent sampling compared to matched sampling is a measure of the information contained in each design for estimating \( \Psi \) in very finite samples.

6.3.1 Finite Sample Results

A graph of the relative efficiency of each estimator is given in Fig. 1 for a rare outcome (1% in the control and 0.6% in the treated) and in Fig. 2 for a more common outcome (25% in the control and 18.75% in the treated). The y-axis represents the estimator’s variance under independent sampling divided by the estimator’s variance under matched sampling. The x-axis represents different simulation scenarios varying the effect of the baseline covariate on the outcome and the presence of the interaction.

As expected, the Pair-Matched Design yields more efficient estimators of the unadjusted risk difference \( \Psi_r \). The matched design appears at most 6.2 times more efficient than its non-matched counterpart (for a common outcome and a sample size of 10). This comparison is unfair, as matching is a form of adjustment. When the substitution estimator (Eq. 7) is implemented, the Independent Design often yields more efficient estimators. For the rare outcome (Fig. 1), the relative efficiencies of the adjusted estimators implementing Models 1 and 2 are less than or equal to one in all scenarios. In other words, pair-matching harms estimator precision in very small samples of \( n = 10 \) or \( n = 20 \) clusters. The Independent Design is up to 25.2% more efficient. When the outcome is common (Fig. 2) and the regression model contains an intercept, the treatment and the covariate (Model 1), the Independent Design is always more efficient. Indeed, the Pair-Matched Design is up to 21.6% less efficient with a sample size of 10 and 23.9% less efficient with a sample size of 20. When the outcome is common and a less parsimonious model is fit with indicators for each strata of \( W \) (Model 2), the Pair-Matched Design tends to be less efficient, especially in the presence of the weak interaction.

Our attempt at estimating the conditional mean function with a saturated parametric model failed. The algorithm automatically drops strata without support. Consequently, the regression model becomes unsaturated and makes nonsensical extrapolations. Therefore, the Pair-Matched Design, which guarantees support in observed covariate categories, yielded a more efficient estimator when the sample size was limited to 10. However, the estimator’s performance was substantially improved by increasing the number of experimental units to 20. Then for rare outcomes the Pair-Matched Design is less efficient, and for more common outcomes is less efficient in the presence of a weak interaction.

7 Right Censoring and Pair-Matching in Randomized Trials

In this section, we consider the effect of missing outcomes on the validity and efficiency of pair-matching in randomized trials. Let us again consider the examples seeking to estimate the causal effect of RUTF versus the standard supplement on the marginal proportion of children recovering from undernutrition. For the individually randomized trial, suppose a child is more likely to be admitted to a health center and thereby lost to followup, if he or she is suffering from severe acute malnutrition at baseline. If this censoring is differential between treatment groups, then the study is subject to bias. The children whose outcomes are observed differ systematically from the children whose outcomes are missing. For the cluster randomized trial, suppose an external organization
Figure 1: The relative efficiency of estimators in finite samples when the outcome is rare. We explore four estimators of the average treatment effect in trials limited to $n = 10$ and $n = 20$ independent units. The y-axis gives the sample variance of the estimator in the Independent Design relative to the Pair-Matched Design over 5000 trials. The x-axis denotes different simulation scenarios, varying the presence of the interaction and the strength of the effect of $W$ on $Y$. In each simulation, the cluster-level outcome was taken as the average of the $J = 2500$ individual-level outcomes. In the Pair-Matched Design, units were matched exactly on continuous covariate $W$. 

http://biostats.bepress.com/ucbbiostat/paper294
Figure 2: The relative efficiency of estimators in finite samples when the outcome is common. We explore four estimators of the average treatment effect in trials limited to $n = 10$ and $n = 20$ independent units. The y-axis gives the sample variance of the estimator in the Independent Design relative to the Pair-Matched Design over 5000 trials. The x-axis denotes different simulation scenarios, varying the presence of the interaction and the strength of the effect of $W$ on $Y$. In each simulation, the cluster-level outcome was taken as the average of the $J = 2500$ individual-level outcomes. In the Pair-Matched Design, units were matched exactly on continuous covariate $W$. 
began providing RUTF to control schools in rural areas. Since these schools no longer follow their assigned treatment, they are artificially censored. The probability of being observed would depend on the treatment assignment and geographic region.

7.1 Data, Model and Target Parameter

Let us adopt the data structure, statistical model and target parameter to incorporate censoring. In the individually randomized trial, a child is randomly sampled; his or her baseline covariates $W$ measured, and the treatment $A$ randomly assigned. After two months, the child is either recovered from undernutrition or not: $Y \in \{0, 1\}$. The outcome is only observed, however, if the child was followed until the end of the study. Let $\Delta$ be an indicator, equalling one if the child’s outcome was recorded. Then the observed outcome can be represented with $\tilde{Y} = \Delta Y$. By definition, $\tilde{Y}$ equals the true outcome $Y$ if the child was not censored ($\Delta = 1$) and equals zero otherwise. The data generating experiment for the cluster randomized trial is analogous. Both can be described by the following SCM (Pearl, 2009)

$$U = (U_W, U_A, U_\Delta, U_Y) \sim P_U$$
$$W = f_W(U_W)$$
$$A = f_A(U_A)$$
$$\Delta = f_\Delta(W, A, U_\Delta)$$
$$\tilde{Y} = \Delta f_Y(W, A, U_Y)$$

As before, the baseline covariates $W$ are some function of unmeasured variables; the treatment $A$ is randomly allocated, and the outcome $Y$ is generated by some unspecified function of the measured covariates, the treatment and background factors. However, we only observe the outcome if $\Delta = 1$. The missingness indicator, itself, is a function of the baseline covariates, the treatment and unmeasured error. Repeating this experiment $n$ times yields $n$ i.i.d. copies of random variable

$$O = (W, A, \Delta, \tilde{Y} = \Delta Y) \sim P_0 \in \mathcal{M}$$

The statistical model $\mathcal{M}$ for the set of possible data distributions is implied by the SCM and is semiparametric. In particular, the SCM specifies our knowledge of the treatment mechanism: $A$ is independent of $W$. We have not, however, placed any restrictions on the censoring mechanism.

In the ideal world, the outcomes under the treatment and under the control would be observed for all units. Let $Y(a, 1)$ represent the counterfactual outcome, intervening on the SCM to set treatment $A = a$ and to prevent censoring $\Delta = 1$. Then the target causal parameter is a function of this joint intervention. It is difference in the expected outcome had all units received RUTF compared to the expected outcome had all units received the standard supplement and there was no censoring:

$$\Psi_F(P_{X^F}) = E_{X^F}[Y(1, 1)] - E_{X^F}[Y(0, 1)]$$

where $X^F = (W, Y(1, 1), Y(0, 1))$ denotes the full data and is distributed as $P_{X^F}$. This causal parameter corresponds to the average treatment effect in the absence of censoring. With the coarsening at random (CAR) assumption (Heitjan and Rubin, 1991; Jacobsen and Keiding, 1995; Gill et al., 1997), the full data parameter can be identified from the observed data. Under CAR, the joint intervention $(A, \Delta)$ is conditionally independent from the full data $X^F$, given the measured baseline covariates $W$. (We note this assumption only serves to enrich our interpretation of the target statistical parameter. If one is hesitant to believe CAR, an interesting and well-defined statistical parameter can still be specified.)
The target statistical parameter is the marginal difference in the intervention-specific mean outcomes:

\[
\Psi(P_0) = E_0[E_0(\tilde{Y} | A = 1, \Delta = 1, W) - E_0(\tilde{Y} | A = 0, \Delta = 1, W)] \\
= E_0[Q_0(1, 1, W) - Q_0(0, 1, W)]
\]

where \( Q_0(a, 1, W) \) represents the conditional expectation of the outcome, given the joint intervention \( (A = a, \Delta = 1) \) and baseline covariates. Here, conditioning on baseline covariates is necessary for identifiability and can improve study efficiency. We also require that the conditional probability of being assigned each treatment is positive and that the conditional probability of being observed is positive:

\[
0 < P_0(A = 1|W) < 1 \\
0 < P_0(\Delta = 1|A, W) < 1
\]

The first assumption is immediately satisfied by the randomization of the treatment: \( P_0(A|W) = P_0(A) = 0.5 \). The second assumption is required for the target parameter to be well-defined. For example, if all severely malnourished children dropped out of the control arm, there would be no hope in estimating the average treatment effect.

The efficient influence curve for the target parameter \( \Psi(P_0) \) is the following function of the observed data

\[
D^*(P_0)(O) = D_Y^*(P_0) + D_W^*(P_0)
\]

with

\[
D_Y^*(P_0) = \left( \frac{I(A = 1, \Delta = 1)}{P_0(A, \Delta|W)} - \frac{I(A = 0, \Delta = 1)}{P_0(A, \Delta|W)} \right) (\bar{Y} - Q_0(A, 1, W))
\]

\[
D_W^*(P_0) = Q_0(1, 1, W) - Q_0(0, 1, W) - \Psi(P_0)
\]

where \( P_0(A, \Delta|W) \) is the conditional distribution of the joint intervention, given baseline covariates. We remind the reader that the variance of \( D^*(P_0) \) establishes a lower bound on the variance of all reasonable estimators of the target parameter (Bickel et al., 1993). In other words, an estimator is efficient if and only if its influence curve equals the efficient influence curve. By mapping the efficient influence curve for the completely randomized design into the efficient influence curves for the Independent Design and the Pair-Matched Design, we can study the information provided when the outcome is subject to censoring.

### 7.2 Design Alternatives & Asymptotic Efficiencies

The Independent Design guarantees the treatment allocation is balanced. The Pair-Matched Design guarantees the treatment allocation is balanced and that matching variables \( M \subset W \) are balanced between study groups. For both designs, the data generating experiment could be described as sampling one treated observation from the conditional distribution of \( (W, M, \Delta, \bar{Y}) \) given \( A = 1 \) and then sampling one control observation from this conditional distribution given \( A = 0 \) and \( M = M_1 \), where \( M_1 \) is the value of the matching variable in the treated unit and equal to the empty set \( \{\emptyset\} \) in the Independent Design. Repeating this experiment \( n \) times yields \( n \) i.i.d. copies of random variable

\[
O' = (O_1, O_0) = (W_1, M_1, \Delta_1, \bar{Y}_1), (W_0, M_0, \Delta_0, \bar{Y}_0) \sim P_0'
\]

with \( M_0 = \{\emptyset\} \) in the Independent Design and \( M_1 = M_0 \) in the Pair-Matched Design. The statistical model \( \mathcal{M}' \) is semiparametric and is implied by the underlying statistical model \( \mathcal{M} \). The marginal distributions of \( (W, M, \Delta, \bar{Y}) \) are determined by the population distribution \( P_0 \).
As before, the efficient influence curve for target parameter $\Psi$ under sampling from $P_0$ can be mapped into its weighted counterpart for paired sampling from $P'_0$. Specifically, randomizing the treatment $A$ yields weights of 0.5, and the efficient influence curves under independent and matched sampling are equivalent:

$$D^*I(O') = D^*M(O') = 0.5D^*(O_1) + 0.5D^*(O_0)$$

with

$$D^*(O_1) = \frac{\Delta_1(\bar{Y}_1 - \bar{Q}_0(1, 1, W_1))}{P_0(A = 1, \Delta_1|W_1)} + \bar{Q}_0(1, 1, W_1) - \bar{Q}_0(0, 1, W_1) - \Psi(P_0)$$

$$D^*(O_0) = \frac{-\Delta_0(\bar{Y}_0 - \bar{Q}_0(0, 1, W_0))}{P_0(A = 0, \Delta_0|W_0)} + \bar{Q}_0(1, 1, W_0) - \bar{Q}_0(0, 1, W_0) - \Psi(P_0)$$

Recall that the superscript $I$ denotes the Independent Design, the superscript $M$ denotes the Pair-Matched Design, and $D^*(O_a)$ is the efficient influence curve at $P_0$ evaluated at $A = a$. Once more, the precision gained or lost by pair-matching depends on the magnitude and direction of the covariance term:

$$Var[D^*I(O')] = 0.25Var[D^*(O_1)] + 0.25Var[D^*(O_0)]$$

$$Var[D^*M(O')] = 0.25Var[D^*(O_1)] + 0.25Var[D^*(O_0)] + 0.5Cov[D^*(O_1), D^*(O_0)]$$

The covariance between $D^*(O_1)$ and $D^*(O_0)$ can be written as the sum of the following covariances

$$Cov[D^*_Y(O_1), D^*_Y(O_0)] = E_0 \left[ \left( \frac{\Delta_1(\bar{Y}_1 - \bar{Q}_0(1, 1, W_1))}{P_0(A = 1, \Delta_1|W_1)} \right) \left( \frac{-\Delta_0(\bar{Y}_0 - \bar{Q}_0(0, 1, W_0))}{P_0(A = 0, \Delta_0|W_0)} \right) \right]$$

$$Cov[D^*_W(O_1), D^*_W(O_0)] = E_0 \left[ (\bar{Q}_0(1, 1, W_1) - \bar{Q}_0(0, 1, W_1) - \Psi(P_0))(\bar{Q}_0(1, 1, W_0) - \bar{Q}_0(0, 1, W_0) - \Psi(P_0)) \right]$$

If both outcomes are observed, the covariance of the $D^*_Y$ terms is often negative. The numerator is the deviation between the observed outcome and the expected outcome given the intervention and covariates. It shrinks to zero as the measured covariates explain more about the outcome. The denominator represents the conditional distribution of the joint intervention. It can be factorized as $P_0(a|W_a)P_0(\Delta_0|a, W_a) = 0.5P_0(\Delta_0|a, W_a)$. Therefore, the censoring mechanism plays an important role in determining the magnitude of the covariance of the $D^*_Y$ terms. Nonetheless, $Cov[D^*_Y(O_1), D^*_Y(O_0)]$ is always zero, if one or more of the outcomes is missing within a pair. On the other hand, the $D^*_W$ term is the difference in the expected outcomes minus the target parameter and always contributes whether the outcome is observed or not. In many cases, the covariance of the $D^*_W$ terms will be large and positive, as the target parameter $\Psi$ averages over the distribution of baseline covariates. Consequently, the overall covariance of $D^*(O_1)$ and $D^*(O_0)$ is easily positive and the Pair-Matched Design asymptotically less efficient than the Independent Design.

### 7.3 The Importance of Measuring Baseline Covariates

Suppose we did not measure any baseline covariates and proceeded with randomized trial. The observed data under independent or matched sampling would be

$$O'_r = (O_1, O_0) = (\Delta_1, \bar{Y}_1), (\Delta_0, \bar{Y}_0)$$

The semiparametric statistical model $M'_r$ is implied by the underlying statistical model $M_r$ for the distribution of reduced data structure $O'_r = (A, \Delta, \bar{Y})$. In order to identify the causal parameter,
we must assume that missingness is only a function of the assigned treatment \( A \). In other words, we must assume there is no factor (e.g. baseline malnutrition status) that affects both dropout and the outcome. This assumption is often unrealistic, and if it fails to hold, the resulting estimand is biased.

Nonetheless, let us assume that missingness is completely random or only a function of the assigned treatment. Then the average treatment effect in the absence of censoring is identified from the reduced data \( O_r = (A, \Delta, \bar{Y}) \sim P_r \) as

\[
\Psi_r(P_r) = E(\bar{Y}|A = 1, \Delta = 1) - E(\bar{Y}|A = 0, \Delta = 1) = \bar{Q}(1,1) - \bar{Q}(0,1)
\]

This target statistical parameter is the unadjusted difference in the intervention-specific mean outcomes. The efficient influence curve for \( \Psi_r \) at \( P_r \) is

\[
D_r^*(P_r)(O_r) = \left( \frac{I(A = 1, \Delta = 1)}{P(A, \Delta)} - \frac{I(A = 0, \Delta = 1)}{P(A, \Delta)} \right)(\bar{Y} - \bar{Q}(A,1))
\]

Then the efficient influence curve for \( \Psi_r \) under paired sampling from \( P_r^* \) is

\[
D_r^{*M}(O_r^*) = D_r^{*M}(O_r^*) = \frac{\Delta_1(\bar{Y}_1 - \bar{Q}(1,1))}{P(\Delta_1|A_1)} - \frac{\Delta_0(\bar{Y}_0 - \bar{Q}(0,1))}{P(\Delta_0|A_0)}
\]

The asymptotic efficiency of the Independent Design compared to the Pair-Matched Design for estimating the average treatment effect when the outcome is subject to missingness, once again, depends on the direction and magnitude of a covariance term:

\[
Var[D_r^{*M}(O_r^*)] = Var[D_r^{*M}(O_r^*)] - 2Cov\left[ \frac{\Delta_1(\bar{Y}_1 - \bar{Q}(1,1))}{P(\Delta_1|A_1)}, \frac{\Delta_0(\bar{Y}_0 - \bar{Q}(0,1))}{P(\Delta_0|A_0)} \right]
\]

Therefore, the Pair-Matched Design will be asymptotically more efficient when units are matched on unmeasured predictors of the outcome. Both designs will be biased, however, if censoring depends on unmeasured characteristics that also affect the outcome.

### 7.4 Simulation Study Incorporating Missingness

We adopted the previously discussed simulations to understand the asymptotic efficiency of pair-matching in randomized trials when the outcome is subject to censoring. We again focused on the cluster randomized trial where the outcome \( \bar{Y} \) is the average of \( J = 2500 \) individual outcomes, drawn from a binomial distribution with probability \( p = h(A, W) \). The outcome was only observed, however, if \( \Delta = 1 \). We considered three scenarios for the censoring mechanism:

- Random (non-informative): \( P(\Delta = 1) = 0.7 \)
- Dependent only on the treatment: \( P(\Delta = 1|A) = 0.65 + 0.1A \)
- Informative: \( P(\Delta = 1|A, W) = \expit[\logit(0.7) + 3AW - 5(1 - A)W] \)

The data generating experiment for the Independent Design was sampling \( (W, M, \Delta, \bar{Y}) \) given \( A = 1 \) and independently sampling \( (W, M, \Delta, \bar{Y}) \) given \( A = 0 \). The data generating experiment for the Pair-Matched Design was sampling \( (W, M, \Delta, \bar{Y}) \) given \( A = 1 \) and then sampling \( (W, M, \Delta, \bar{Y}) \) given \( A = 0 \) and \( M = M_1 \). We considered matching exactly on the continuous baseline covariate: \( W_1 = W_0 \). Table 4 gives the asymptotic efficiency of the Independent Design relative to the Pair-Matched Design when the outcome is subject to missingness.
Table 4: The asymptotic efficiency of the Independent Design relative to the Pair-Matched Design when the outcome is subject to missingness. The first column refers to the type of censoring. The second column indicates whether the outcome was rare or common. The third column denotes whether the efficiency comparison is made under misspecification or the correct specification of the data, model and target parameter. $RE_{misp}$ is the variance of the efficient influence curve at misspecified $P'_r$ under independent sampling divided by the variance of the efficient influence curve under matched sampling at misspecified $P'_r$. $RE_{true}$ is the ratio of these variances at the true data generating distribution $P'_0$. The remaining columns give the relative efficiencies for the six scenarios examined. Specifically, the strength of the relationship of $W$ on $Y$ is varied with $\beta_2$, and the interaction is varied with $\beta_3$ equal to 0 or -0.75. In the Pair-Matched Design, units were matched exactly on continuous covariate $W$.

Once again, it is crucial to include the measured baseline covariates in the specification of the data, model and target parameter. When the measured covariates are excluded, the matched design appears most efficient when the covariates are strongly predictive of the outcome and in the absence of the interaction. Indeed, the Pair-Matched Design appears over three times more precise than the Independent Design, assuming the efficient influence curve at $P'_r$. When we correctly specify the data generating experiment by including measured covariates, the matched design is least efficient when matching on a covariate strongly predictive of the outcome and in the presence of the interaction. The Independent design is nearly twice as precise at the Pair-Matched Design at $P'_0$. These results hold when we vary the rarity of the outcome and the type of censoring. When censoring is informed by both the treatment and covariates, the unadjusted statistical parameter $\Psi_r(P'_r)$ is not equivalent to the target causal parameter. Consequently, the relative efficiency, assuming the efficient influence curve at $P'_r$, is not given.

8 Discussion and Conclusion

For nearly sixty years, researchers have been debating the merits of matching in experimental studies. Despite warnings, matching is often implemented as a way to improve study credibility and to increase study power. In this paper, we compared the relative efficiency of the Independent Design to the Pair-Matched Design. In the first data generating experiment, units are randomly paired and the treatment randomized within the pair. In the second, units are matched on baseline covariates and then the treatment randomized within the pair. The Independent Design guarantees that treatment allocation is balanced. The Pair-Matched Design ensures not only that the treatment allocation is balanced but also that important baseline covariates are balanced between the study
groups.

We used the efficient influence curve to evaluate the information contained in each design for estimation of the average treatment effect. The variance of the efficient influence curve establishes a lower bound on the asymptotic variance of all regular, asymptotically linear estimators (Bickel et al., 1993). Therefore, the ratio of the variance of this function in the Independent Design to the Pair-Matched Design is a direct measure of design efficiency. Our theoretical and simulation results indicate that pair-matching on measured baseline covariates is asymptotically less efficient than simply adjusting for these covariates in the analysis. Our theoretical and simulation results also indicate that in the uncommon situation when units can be matched on unmeasured predictors of the outcome, the Pair-Matched Design is asymptotically more efficient. These results hold when we vary the rarity of the outcome, the presence of an interaction, the strength of the effect of the baseline covariate on the outcome, and subject the outcome to missingness. The finite sample simulations, limiting the number of independent units to 10 or 20, support these findings.

Overall, it is of utmost importance to carefully describe the data generating experiment. For many public health interventions, it is difficult to imagine a pair-matched trial, where units can be matched on unmeasured covariates that are predictive of the outcome. Indeed, many pair-matched cluster randomized trials employ an algorithm to pair clusters on the basis of measured covariates. It is critical to include these measured covariates in the definition of the data, statistical model and target parameter. Misspecification of these quantities can lead to incorrect conclusions regarding the efficiency of pair-matching in randomized trials.

Acknowledgment

This work was supported by NIH grant R01 AI074345-05.

References


