The role of pairwise matching in experimental design for an incidence outcome

Adam Kapelner¹, Abba M. Krieger² and David Azriel³

Queens College, CUNY, The Wharton School of the University of Pennsylvania and Technion Israel Institute of Technology

Summary

We consider the problem of evaluating designs for a two-arm randomised experiment with an incidence (binary) outcome under a non-parametric general response model. Our two main results are that the a priori pair matching design is (1) the optimal design as measured by mean squared error among all block designs which includes complete randomisation. And (2), this pair-matching design is minimax, that is, it provides the lowest mean squared error under an adversarial response model. Theoretical results are supported by simulations and clinical trial data where we demonstrate the superior performance of pairwise matching designs under realistic conditions.

Key words: binary response; experimental design; incidence endpoint; incidence outcome; logistic regression; restricted randomisation.

1. Introduction

Our goal is to examine the role of experimental design when estimating a treatment effect via the difference-in-means estimator for the average treatment effect in a two-arm treatment-control randomised controlled trial (RCT). Each of the 2n subjects is assigned, that is, administered a value wi to the treatment group (wi = +1) or a control group (wi = −1) and all values together is called an assignment w = [w1, . . . , w2n]⊤. Each subject’s covariates xi, a vector of real values with length d, are known beforehand and considered fixed. At the completion of the study, an incidence response y = [y1, . . . , y2n]⊤ ∈ {0, 1}2n is collected (e.g. cardiac event vs. no cardiac event). We will assume the covariates and the treatment are related to the probability of a positive response pi = Pr(Yi = 1 | xi, wi) but we do not assume a functional form for this probability function. This setting of ‘assign subjects first and assess response later’ is called ‘non-sequential’ and was the classic assignment setting studied by Fisher (1925) when he was assigning fertiliser treatments to agricultural
The setting is still of great importance today. For example, they occur in clinical trials as ‘many phase I studies use ‘banks’ of healthy volunteers ... and ... in most cluster randomised trials, the clusters are identified before treatment is started’ (Senn 2013, p. 1440).

The practitioner has a choice before the experiment begins that can affect the efficiency of the estimation: the experimental design—the set of allocations drawn when assigning the subjects to either treatment or control. This is a well-studied problem when the response is continuous, but less well-studied when not. The naive design allows for every assignment \( w \in \{+1, -1\}^{2n} \) and is called the ‘Bernoulli trial’ (Imbens & Rubin 2015, section 4.3). This is an unpopular design in the non-sequential setting as it allows for differing numbers of subjects in the two experimental arms. To avoid this possibility, the experimenter frequently restricts the two arms to have the same number of subjects and this we call ‘balanced complete randomisation design’ (BCRD). Designs that are even more restrictive are popular in practice, for example, blocking (BL Fisher 1925) and rerandomisation (Student 1938; Morgan & Rubin 2012). Less popular in practice are pairwise matching (PM Greevy et al. 2004) and optimal single-assignment designs (Bertsimas, Johnson & Kallus 2015).

The fundamental question ‘which of these myriad restricted designs are truly optimal?’ was recently answered in the case of continuous response. Kallus (2018) demonstrates that the minimax variance design for the difference-in-means estimator is contingent upon the space of the response function as specified by its functional norm. Under this unified theory, the BL design emerges to be minimax under the supremum norm, the pairwise matching design emerges to be minimax under the Lipschitz norm, the rerandomisation design emerges to be minimax under a linear norm and others. However, without knowledge of the space to which the response function belongs, there is ‘no free lunch’ and the classic Fisherian design of complete randomisation is vindicated as the minimax strategy. To our knowledge, no systematic overarching theory has been developed in our setting of an incidence response.

Design in non-linear models, especially logistic regression has a long literature but focused on settings that are tangential to our setting. Schein & Ungar (2007) develop a rubric for evaluating active learning designs. These designs evaluate how the estimator performs under many possible future subjects and selects the subject and its assignment which are optimal to collect a response. Park, Mancenido & Montgomery (2019); Johnson & Montgomery (2009); Mancenido et al. (2019) and many others before them take a classic optimal design approach in response surface methodology (Box & Wilson 1951) employing a holistic metric of the estimation of all design parameters via \( D \)-optimality or other classic metrics of optimality (Myers, Myers & Carter Jr 1994). The design problem in these works are different than in our setting, as they study the situation where the \( x \)s could be determined by the experimenter. Their approach (1) requires an explicit probability model for the response meaning an explicit link function \( \phi \) (e.g. the expit, probit, cloglog, cauchit functions) between the probability and the embedded function of the parameters (e.g. the linear model \( \beta_0 + \beta^T x_i + \beta_T w_i \)). And (2) the criterions themselves frequently are recursively dependent on knowledge of the \( d + 2 \) unknown parameters, the \( \beta_j \)s. To address this situation, many authors subjectively posit values for the \( d + 2 \) unknown parameters then proceed to create locally optimal designs based on these posited \( \beta_j \)s. This procedure can be arbitrarily inaccurate if the initial \( \beta_j \)s prove to be at odds with the true parameter values.
(Abdelbasit & Plackett 1983, section 1.1). Other authors sequentially run the experiment to minimise the chance of an egregious disparity between initial and actual parameter values (Section 1.2, ibid). And others use a Bayesian approach by specifying a prior on the $\beta_j$s and marginalising their effect out using expectation (Section 1.3, ibid).

Herein, we are interested in maximising the efficient estimation of only one parameter (the average treatment effect), we do not demand an explicit $\phi$ and we do not wish to rely on initial estimates of the $d + 2$ parameters. In our set-up, the allocation vector $w$ is to be determined, as opposed to the $x$s as in the classical setting. Furthermore, we demand the non-sequential setting where we must provide the entire $w$ vector of individual assignments without the luxury of observing any part of $y$. To our knowledge, our setting is relatively unexplored.

Although unexplored, our setting is widespread as there are countless clinical trials run annually where the primary outcome is an incidence metric. By way of example, Bjermer et al. (2003) measured incidence of asthma exacerbation, Julius et al. (2004) measured incidence of cardiac mortality, Zietman et al. (2010) measured incidence of failure to improve prostate outcome. Each of these examples employed the restricted design of block randomisation. A cursory search of other contemporary clinical trials with primary incidence outcomes confirms that BL is the default design.

Herein, we demonstrate that the PM design of Greevy et al. (2004) always outperforms BCRD and BL in mean squared error and that PM is the minimax design over all possible designs when the probability model is adversarial, seeking to create the highest error in estimation. Section 2 sets up our problem, introduces the designs we consider and explains how these designs produce assignments $w$. Section 3 records our theoretical results, Section 4 provides simulation evidence of our theoretical results (including an example using clinical trial data) and Section 5 concludes.

### 2. Our model and designs

We assume a population of $2N$ subjects where each of the $i$ subjects has an independent (across subjects) Bernoulli-distributed response when administered treatment and when administered control

$$Y_{T,i} \overset{ind}{\sim} \text{Bern}(p_i(x_i, w_i = +1)),$$

$$Y_{C,i} \overset{ind}{\sim} \text{Bern}(p_i(x_i, w_i = -1)).$$

The responses of subject $i$ are assumed independent of the responses of subject $j$ for all $i \neq j$. As the $p_i$’s are assumed affixed to the subjects, their $x_i$s are necessarily affixed as well. We define the population parameter to be the mean difference of the Bernoulli parameters across treatment and control for all $2N$ subjects, that is,

$$\tau = \frac{1}{2N}(\rho_T - \rho_C)^\top 1_{2N}. $$

which is known as the population ‘mean risk difference’ where $\rho_T = [p_1(x_1, +1), \ldots, p_{2N}(x_{2N}, +1)]^\top$ and $\rho_C = [p_1(x_1, -1), \ldots, p_{2N}(x_{2N}, -1)]^\top$.

We now define our sampling mechanism formally as $S \sim \text{Multin}(2N, 2n, (2N)^{-1}1_{2N})$, a $2N$-length multinomial random variable whose realisations $s$ have a 1 for indices that are present in the sample and 0 otherwise.
We define the sample mean risk difference parameter as the mean in a given sample \( s \) as
\[
\tau_s = \frac{1}{2n}(\rho_T - \rho_C)^\top s.
\] (3)
Since \( E(S) = n/N \mathbf{1}_{2N} \), then \( E_S(\tau_S) = \tau \), that is, its expectation over all possible samples is unbiased.

The experimental design \( W \) is formally a multivariate shifted-and-scaled Bernoulli that produces vectors of assignments \( w \in \{-1, +1\}^{2n} \) assumed to be uniform over its support, \( \Pr(W = w_j) = \Pr(W = w_k) \) for all \( j, k \). We make two restrictions on the designs considered. First, (A1) ensures that all assignments produce an equal number of treatment and control assignments, that is, \( w^\top \mathbf{1} = 0 \) for all \( w \) in the support of \( W \). Second, (A2) we assume each individual subject has equal probability of being assigned to either arm, that is, \( E(W) = \mathbf{0}_{2n} \). Note that (A2) is a weaker assumption than \( \Pr(W = w) = \Pr(W = -w) \), which is common in the design literature and implies (A2). A design that addresses observed covariate imbalance among the two treatment groups will necessarily depend on the \( x_i \)s in the sample and thus \( W \) is conditional on \( s \), but we omit this conditioning in our notation for simplicity.

To sum up our assumptions, there are three sources of the randomness in the responses listed in the order of realisation: (R1) at the beginning of the study, \( s \) is drawn from \( S \) to create the sample of \( 2n \) subjects from the population of all \( 2N \) subjects; (R2) the treatment assignments \( w \) is drawn from \( W \) to allocate subjects to the two treatment groups and (R3) at the conclusion of the study, \( y \) is drawn from \( Y \) to provide the endpoint measures of the subjects via the Bernoulli process in (1). (R2) is termed the randomisation model while (R1, R3) is termed the invoked population model (see Rosenberger & Lachin 2016, Chapter 6 for details).

The non-parametric estimator for \( \tau \) for a sample \( s \) we consider is
\[
\hat{\tau} = \frac{1}{n} W^\top Y,
\] (4)
where \( Y \) denotes the responses in sample \( s \). A realisation of this estimator is the familiar classic difference-in-means estimate \( \bar{y}_T - \bar{y}_C \) where \( \bar{y}_T \) and \( \bar{y}_C \) denote the average of the responses in the treatment and control group respectively.

We show in Section A.1 of the supporting information found online that \( \hat{\tau} \) is unbiased for \( \tau \) when taking the expectation over (R1), (R2) and (R3). Thus, its mean squared error (MSE) is its variance, which is derived in Section A.2 of the supporting information found online to be
\[
\text{MSE}(\hat{\tau}) = \frac{1}{4n^2} \left( E_S(v^\top \Sigma v) + 2 E_S(p_T^\top (1 - p_T) + p_C^\top (1 - p_C)) \right) + V_{\rho_T, \rho_C},
\] (5)
where \( p_T \) and \( p_C \) are the subvectors consisting of entries in \( \rho_T \) and \( \rho_C \) when \( s_i = 1 \), \( v = p_T + p_C \), \( \Sigma = \text{var}(W) \), the variance–covariance matrix of all the assignments \( w \) produced by the design and \( V_{\rho_T, \rho_C} \) is a variance over all samples from the population and depends only on the population treatment probabilities. The term \( p_T^\top (1 - p_T) + p_C^\top (1 - p_C) \) has a nice interpretation—it is the sum of the sample subjects’ variances of the Bernoullis for both arms’ incidence responses. The only term that depends on the design is the underbraced quadratic form; this term is of special interest and will be the objective function to
be compared among designs in Section 3. To minimise the MSE, one can minimise this term for each draw of $s$ and that is the focus of the rest of this paper. Thereby, we hereon condition on $S = s$.

As the $2n \times 2n$ variance–covariance matrix of the design plays a fundamental role, we will explain its values for the designs we consider here. For any design, the diagonal values are all 1 as $E(W_i) = 0$ by property (A2) and $W_i^2 = 1$ for all $i$ since $W_i \in \{-1, +1\}$. Thus, experimental designs differ in their degree of dependence between assignments $W_i$ and $W_j$ codified by the off-diagonal elements.

Consider the BL design with $B$ blocks where each are equally sized of size $n_b = 2n/B$ for all $b \in \{1, \ldots, B\}$ subjects each. The variance–covariance matrix for BL, $\Sigma_{BL,B}$, is a block-diagonal matrix with $B$ blocks each of size $n_b \times n_b$ as the subjects between blocks are independent. The off-diagonal entries within blocks are $-1/(n_b - 1)$ because if one subject in the block is assigned to the treatment arm, this makes it a bit more probable that the other subjects in the block are assigned to the control arm as the number of treatment and control subjects must be equal within the block. The BCRD design can then be thought of as one large block and thus all off-diagonal entries in its variance–covariance matrix, $\Sigma_{BL,1}$, are $-1/(2n - 1)$. And the PM design can be thought of as the case where there are $n_b = 2$ subjects per block for all $b$ with $B = n$ total blocks and thus its variance–covariance matrix $\Sigma_{BL,n}$ is block-diagonal with $2 \times 2$ blocks with off-diagonal entries of $-1/(2 - 1) = -1$.

How are these blocks created in BL? To obtain the theoretical results of Section 3, we must assume these blocks are created with an optimal match structure $M^*$ created from $v$ (see Krieger, Azriel & Kapelner 2023, section 2.2.3). The structure is formally a set of $B$ tuples of sizes $n_1, \ldots, n_B$, each set indicating the subjects indices of the subjects belonging to each block. This optimal match structure is created by first sorting the values of $v_i$ and recording the order of sorted subject indices. Then the $B$ tuples would fill up in order. For example if the blocksize is homogeneous with all $n_b = 4$, the design $W$ would view sorted subject numbers 1, 2, 3, 4 as a 4-tuple and randomise with 1/6 probability between the permutations $\langle+1,+1,-1,-1\rangle, \langle+1,-1,+1,-1\rangle, \ldots, \langle-1,-1,+1,+1\rangle$. Analogously, sorted subject numbers 5, 6, 7, 8 would be a 4-tuple and randomised in the same fashion, etc. In the PM design, the design $W$ would view sorted subject numbers 1, 2 as a pair and randomise with 1/2 probability between $\langle+1,-1\rangle$ and $\langle-1,+1\rangle$. Then sorted subject numbers 3, 4 would be a pair and randomised in the same fashion, etc. Thus in PM, $M^* = \{[i_1^*,j_1^*], \ldots, [i_n^*,j_n^*]\}$ whose elements specify the indices of the $n$ optimal pairs.

How can the match structure $M^*$ be created in practice if $v$ is unknown? In the case of $d = 1$ (one covariate $x$ is measured for each subject), if we assume the functional form of $p_i(x_i, w_i)$ is monotonic in $x$ then either $x_1 \leq x_2 \leq \ldots \leq x_{2n}$ implies $v_1 \leq v_2 \leq \ldots \leq v_{2n}$ or $v_{2n} \leq v_{2n-1} \leq \ldots \leq v_1$. In this case ordering the subjects by their covariate value will sort subjects by their $v_i$ values. This monotonicity is standard for instance when the probability is assumed to be a function with a linear term in $x$ embedded in a monotonic link function $\phi$ with range $(0, 1)$, i.e.,

$$p_i(x_i, w_i) = \phi(\beta_0 + \beta_1 x_i + \beta_T w_i).$$

What if this monotonic assumption cannot be assumed? For example, in the case where the embedded function in (6) has a quadratic term in the one covariate, then ordering by the $x_i$’s will not order the subjects by their $v_i$ values. However, sorting by the $v_i$ values
is necessary only to create the optimal pairwise match structure $M^*$. An approximate pairwise match structure $M$, albeit suboptimal, would likely still perform well. We explore this setting in the simulations of Section 4.

If $d > 1$, knowledge of the $v_i$s is equivalent to knowledge of the full functional form of $p_i(x_i, w_i)$. For instance, in the generalised linear model $\phi(\beta_0 + \beta^\top x_i + \beta_T w_i)$, one would need to know the values of the parameters $\beta_0, \beta$ and $\beta_T$. These parameters are unknown and we find ourselves ironically in the setting of those who design experiments using $D$-optimality as we wrote about above in Section 1. Once again, in practice, an approximate match structure $M$ would likely still perform well. We explain how we obtain this approximate match structure and explore the MSE performance of these designs in the case of more than one covariate in Section 4.

3. Theoretical results

For the first two results, assume the subjects are sorted by their unknown values of $v_i$. We first prove in Section A.3 of the supporting information found online the following result.

Theorem 1. (For any sample size, PM is optimal among all block designs, including BCRD). Under the model of (1), the MSE for PM, the block design with $n$ blocks, is lower than the MSE for any block design with less than $n$ blocks where the blocksize is even. Thus, for $B < n$ and all blocksizes even but not necessarily equally sized, $\text{MSE}(\hat{\tau}_{PM}) < \text{MSE}(\hat{\tau}_{BL(B)})$ for any $v$, where $\hat{\tau}_{PM}, \hat{\tau}_{BL(B)}$ are the estimator of (4) under matching and the block design with $B$ blocks respectively.

Although PM is optimal among block designs, block designs do not span the space of all possible experimental designs. We believe it is impossible to solve for the ‘optimal design’ among the entire space of designs, that is, to compute the measure corresponding to $W_* = \arg\min_{W \in W} \{\text{MSE}(\hat{\tau}_W)\}$ where $W$ denotes the space of all designs that satisfy assumptions (A1) and (A2) from Section 2. First of all, the MSE is a function of the design only through $\Sigma$. So at best, we could theoretically find the variance–covariance matrix of the optimal design, that is, $\Sigma_* = \arg\min_{\Sigma \in S} \{v^\top \Sigma v\}$ where $S$ denotes the space of all variance–covariance matrices of a multivariate Bernoulli whose realisations satisfy assumptions (A1) and (A2) from Section 2. Once $\Sigma_*$ is located, it corresponds to very many different equally optimal designs $\{W_*\}$ as the multivariate Bernoulli random variable model has $2^{2n} - 1$ parameters with non-unique second moments (Teugels 1990, section 2.3). Furthermore, the optimal design would be conditional on the unknown value of $v$.

Instead, we prove what is tractable: a theorem about the minimax design under an adversarial response model. In Section A.4 in the supporting information found online we demonstrate the following result.

Theorem 2. (PM is minimax.). Under the model of (1),

$$PM \in \arg\min_{W \in W} \max_{v \in V} \{\text{MSE}(\hat{\tau}_W)\},$$

where $\hat{\tau}_W$ denotes the estimator of (4) under an arbitrary design $W$, $W$ denotes the space of all designs that satisfy assumptions (A1) and (A2) from Section 2 and $V$ is the
space of all sorted vectors \( v = p_T + p_C \), the sum of the two probability parameter vectors, \( \mathcal{V} = \{ v : 0 \leq v_1 \leq v_2 \leq \ldots \leq v_{2n} \leq 2 \} \).

If the subjects are instead randomly sorted with respect to their \( v_i \) values, we have

**Remark 1.** (PM is robust to suboptimal matching.). If the matches are randomly assigned, then the MSE of PM is the same as the MSE of BCRD.

There is a limit to this robustness. If the matches are adversarial, then PM can perform worse than BCRD. **Remark 2** below states a sufficient and necessary conditions for PM to have higher MSE than BCRD, and is proved in Section A.6 of the supporting information found online.

**Remark 2.** (Sufficient and necessary condition for PM to perform worse than BCRD.). We have that

\[
\text{MSE}(\hat{\tau}_{BCRD}) - \text{MSE}(\hat{\tau}_{PM}) = \frac{1}{4n^2} \left( \frac{1}{2n-1} \sum_{i<j} (v_i - v_j)^2 - \sum_{k=1}^{n} (v_{i_k} - v_{j_k})^2 \right),
\]

where \( k = 1, \ldots, n \) and the pairs \( (i_k, j_k) \) are in the match set \( M \). Therefore, BCRD outperforms PM iff

\[
\frac{1}{n(2n-1)} \sum_{i<j} (v_i - v_j)^2 \leq \frac{1}{n} \sum_{k=1}^{n} (v_{i_k} - v_{j_k})^2.
\]  \((7)\)

To interpret the condition in (7), notice that the left-hand side is the average squared distance over all pairs \( v_i, v_j \) with \( i < j \) and the right-hand side is the average over the pairs in the match set \( M \). It follows that BCRD yields lower MSE if the pairs in the match set are more distant than an average pair, and otherwise PM is better.

Comparing BCRD to PM, as in **Remark 2** is equivalent to quantifying the quality of the match through the R-squared one gets in the analysis of variance where the covariate is the \( n \) level categorical variable denoting the match set. The total sum of squares is \( \sum_{i=1}^{2n} (v_i - \bar{v})^2 \) and the error sum of squares is

\[
\sum_{k=1}^{n} \left( v_{i_k} - \frac{v_{i_k} + v_{j_k}}{2} \right)^2 + \sum_{k=1}^{n} \left( v_{j_k} - \frac{v_{i_k} + v_{j_k}}{2} \right)^2 = \frac{1}{2} \sum_{k=1}^{n} (v_{i_k} - v_{j_k})^2.
\]  \((8)\)

Therefore, the value of R-squared is

\[
1 - \frac{1}{2} \frac{\sum_{k=1}^{n} (v_{i_k} - v_{j_k})^2}{\sum_{i=1}^{2n} (v_i - \bar{v})^2}.
\]  \((9)\)

By **Remark 1**, BCRD is equivalent to creating a match set randomly. The expected value of this R-squared for BCRD is, by (17) in the supporting information found online,

\[
1 - \frac{1}{2} \frac{\sum_{i=1}^{2n} (v_i - \bar{v})^2}{\sum_{i=1}^{2n} (v_i - \bar{v})^2} = \frac{n - 1}{2n - 1},
\]  \((10)\)
that is, it does not depend on $v$ and increases to $1/2$ as $n$ increases. This implies that the expected adjusted R-squared is zero. Comparing Equations (9) and (10) is equivalent to the condition in (7).

It is possible for PM to perform better or worse than BCRD depending on the quality of the matches. The ideal match pairs the smallest $v$ with the second smallest $v$, the third smallest $v$ with the fourth smallest $v$ and so on. The worst possible match tends to put the largest $v$ with the smallest $v$, the second largest $v$ with the second smallest $v$ and so on. It is easy to see what occurs if the values of $v$ are equally spaced. In this case, the R-squared for PM can vary anywhere from 0 to 1 as compared to an R-squared near 1/2 for BCRD.

In practice it is unlikely to know exactly the order of the $v$s, but one often has some sense for which individuals are likely to be at risk. Even limited information can lead to a gain over BCRD. As a simple example, which is easy to analyse, assume that the experimenter knows which individuals are above the median of the risks (i.e. the $v$s). Assume further that there is no knowledge about the ordering within the two halves. If we assume that the $v$s are iid uniform, then a simple calculation demonstrates that this reduces the expectation of the critical term $v^\top \Sigma v$ by a factor of four.

It is important to note that when employing PM in the real world, one attempts to construct pairs to be as similar as possible by minimising the pairwise distance function between the subjects’ $x_i$s (this common approach is in fact how we illustrate the superior performance of PM over BL and BCRD in the simulation Sections of 4.1 and 4.2 for $d > 1$). And thus the condition of (7) will likely never occur in the real world. It follows that in practice, PM could be employed without fear. This result echoes Pashley & Miratrix (2022) who conclude that ‘it is hard to go too far wrong’.

Furthermore, if the covariates are not related to the response, the following remark shows that all designs are equally performant. Its proof is found in Section A.7.

**Remark 3.** (All designs are equal if covariates are uninformative.) If $v$ is constant, the conditional MSE given $s$ of any experimental design is $1/(4n^2)\left(2(p_T^\top (1 - p_T) + p_C^\top (1 - p_C))\right)$.

### 4. Simulation results

#### 4.1. Simulated data

We begin by simulating the case of one covariate $x$ measured per subject ($d = 1$). We explore two different response models of (6) by using $\phi = \expit$; this corresponds to the classic logistic regression case with no model misspecification:

$$p_i(x_i, w_i) = \frac{\exp(\beta_0 + \beta_1 x_i + \beta_T w_i)}{1 + \exp(\beta_0 + \beta_1 x_i + \beta_T w_i)},$$

(11)

We also simulate under an analogous cauchit model. As the cauchit response simulation for is similar to the simulation for, we defer its details and its results to Section B in the supporting information found online.

In the case of $d = 1$, we are guaranteed to have the optimal match structure $M^*$ by sorting the $x_i$ values since the probability function is monotonic in $x$ (see discussion in Section 2). This also provides optimal block structure for any number of blocks.

We generate the values of $x$ using the standard logistic distribution quantiles evenly spaced between 0.005 and 0.995 with spacing varying by each sample size which we set to be $2n \in \{64, 128, 256\}$. Stacking the $x_i$’s row-wise gives us the one-column matrix $X$. 

© 2023 Statistical Society of Australia.
We set the response parameters to be \( \beta_0 = 4, \beta_1 = 2 \) and \( \beta_T = 1 \). Given \( X \) and these \( \beta \)'s, we can compute \( p_T \) and \( p_C \) which allows us to compute the risk difference, our main parameter of interest \( \tau_s \) of (3) precisely. These parameter values were selected to induce separation between \( p_T \) and \( p_C \) and to ensure their values are not too close to zero or one in order prevent numerical issues during the simulation. They were additionally selected in an attempt to emphasise the role of the quadratic form component of the MSE (see (5)) in order that the estimated design differences will be more salient.

We then simulate allocation vectors for four designs: BCRD, BL (with \( B = 8 \) blocks for all sample sizes), rerandomisation (which we denote R) and PM. We generate \( N_{\text{sim}} = 500,000 \) assignments \( w \) from each design for each sample size. (For R, we first draw \( 100N_{\text{sim}} \) assignments from BCRD, sort them by the standardised difference of the averages between the treatment and control groups, then take the top 1% which is \( N_{\text{sim}} \) total assignments). For each of the \( N_{\text{sim}} \) assignments in each design, we compute \( p_T \) and \( p_C \) via (11) which allows us to draw the random responses \( y_i \)'s from independent Bernoulli realisations of (1). Using the responses and the assignment we can then compute the risk difference estimate, log odds ratio and the estimate of \( \beta_T \) from a logistic regression (only the latter requires the covariate information). The first estimate, \( \hat{\tau} \) defined in (4), was the subject of our theoretical investigation herein. We include the other estimates in this simulation only to provide intuition about their theoretical performance which we leave to future work. (Note that the true log odds ratio parameter does not need computation as it is \( \beta_T \) and independent of \( X \) by construction in (11), our data generating process). We then average all estimates over the \( N_{\text{sim}} \) replicates to generate average estimates. We use the average estimates to compute the estimated mean squared error using the true parameter values explained previously.

The mean squared error estimates for \( d = 1 \) appear in the left-most column of Figure 1. We can see that the PM design barely outperforms the BL design. But both designs outperform BCRD by a large margin. R performs between BL and BCRD (R cannot perform as well as PM as its goal is to enforce parity between \( x_T \) and \( x_C \) which is different than minimising the critical MSE term of (5), \( v^\top \Sigma v \). The results for BCRD, BL and PM are expected by Theorem 1. This performance pattern also extends to the log odds ratio estimator. Additionally, for the logistic regression estimate result (row 3), we observe that BCRD is not at an extreme disadvantage. This is likely due to the fact that logistic regression employs a posteriori adjustments for covariate imbalance which reduces estimation error. For example, in the setting of continuous responses when comparing the OLS estimator for continuous response to the classic difference-in-mean estimator, the estimation error of the former is an entire order in \( n \) smaller than the latter (see Kapelner et al. 2021, equations 7 and 14). Combining both PM with logistic regression estimation we leave to future work.

The case of \( d = 1 \) is unrealistic as real clinical trials have more than one characteristic measured per subject. To understand design performance in this more realistic case, we now simulate \( d = 2 \) and \( d = 5 \). We employ the same \( \beta \)'s but duplicate the previous \( \beta_1 \) coefficient for all \( d \) covariates, that is, \( \beta_0 = 4, \beta = 21_d \) and \( \beta_T = 1 \). We will discuss how \( X \) is generated in the coming paragraphs. We first must discuss the problem of how to create the blocks for BL design and the pairs for the PM design when \( d \geq 2 \). In \( d = 1 \), the order of the \( v_i \) elements correspond to the order of the \( x_i \) elements but in \( d \geq 2 \), this is not the case, so we need to use the information about the \( x_i \)'s to approximate the order of the \( v_i \)'s. Thus the theorems of Section 3 do not apply when \( d \geq 2 \) in a strict sense. The simulations
Figure 1. Performance results on the simulated data for the logistic response model of (11) for all sample sizes, number of covariates, estimates and the four designs BCRD, BL, R and PM. The top row shows the average estimates for \( \hat{\tau} \), the second row shows the average estimates for the log odds ratio (LOR) and the last row shows the average treatment coefficient estimates in the logistic regression.

herein provide intuition about the theoretical performance in the case of the BL and PM designs when the blocks are imperfectly constructed.

For the PM design, we must generate \( n \) pair matches. To do so, we employ the optimal non-bipartite matching algorithm (see Lu et al. 2011) using the R package \texttt{nbpMatching} (Beck, Lu & Greevy 2016). This algorithm requires a specified distance function between two subjects’ covariate vectors, \( x_i \) and \( x_j \), to generate a distance matrix and then remarkably solves the minimum sum of all pair distance problem in polynomial time. We employ the Mahalanobis distance which was recommended by Rubin (1979), the first work that demonstrated the robustness of matching in regression with a non-linear response model. As this algorithm returns the same matches for a distance function scaled by a multiplicative constant, we employ the proportional between-subjects Mahalanobis distance, \((x_i - x_j)^\top \hat{\Sigma}^{-1} (x_i - x_j)\) where \( \hat{\Sigma}^{-1} \) is the \( d \times d \) sample variance–covariance matrix of all \( 2n \) subjects’ \( d \) covariate vectors (Stuart 2010, section 2.2).

For the BL design, we wish to retain \( B = 8 \) blocks for all sample sizes and number of covariates. For \( d = 2 \) we use four blocks for the first covariate and two blocks for the second. For \( d = 5 \), we use two blocks for each of the first three covariate and do not block on the remaining covariates. This is standard in practice; experimenters block on a subset of covariates that are a priori conjectured to have the most pronounced effect on the response. There is an additional problem: blocks created from continuous covariate data will likely be heterogeneously sized (i.e. \( n_b \) will vary block–block) and the sizes may be uneven. Since
our theoretical results are proven for even-sized blocks, one way to enforce this setting is to enforce homogeneity. To do so, we generate the homogeneous block designs first and then populate the $X$ matrix after. To do so, the first covariate is always generated in order of the quantiles of the standard logistic distribution as explained above for the $d = 1$ case. For the $d = 2$ case, the second covariate is generated as well via the quantiles of the standard logistic distribution, shuffled, spliced in half, ordered, spliced in half again to create the four blocks, then reshuffled. An analogous procedure is followed for the third covariate when $d = 5$.

The R design must select the top 1% of $w$ vectors from a set of $100N_{sim}$ BCRD vectors. We define the ‘top’ as those allocations that have the lowest proportional total Mahalanobis distance defined as $(\bar{x}_T - \bar{x}_C)^T \hat{\Sigma}_X^{-1} (\bar{x}_T - \bar{x}_C)$, where $\bar{x}_T$ and $\bar{x}_C$ denote the average covariate vector over all $n$ treatment subjects and $n$ control subjects respectively. This metric differs from the between-subjects Mahalanobis distance we discussed in the paragraph about the PM design. Here, the metric is defined on all $2n$ subjects; there, the metric is defined on pairs of subjects.

Performance results can be seen in columns 2–3 of Figure 1. We mostly observe the same results as the $d = 1$ setting except the poor performance of BCRD and BL is more pronounced. We omit the results for the logistic regression estimates for $d > 1$ which were unstable.

### 4.2. Clinical data

In this section, we use clinical trial data from Foster et al. (2010), a 12-week, multi-centre, double-blind, placebo-controlled sequential RCT investigating whether amitriptyline, an antidepressant drug, can effectively treat painful bladder syndrome. The endpoint we chose was the final pain reading after 12 weeks. Since this was a continuous endpoint, we coerced it to incidence by thresholding at the value of the sample median. The original RCT employed a Bernoulli trial design where each subject received the treatment independent with probability 50%. The 18 covariates assessed for each patient in the study (and their data types) were: age (continuous), gender (binary), the presence of sexually transmitted disease (binary), lives with a partner (binary), baseline urinary tract infection (binary), patient’s race is white (binary), patient’s race is Hispanic (binary), level of education (ordinal with five levels), level of employment (ordinal with five levels), baseline symptom index (continuous), baseline anxiety and depression (continuous), baseline mental quality of life (continuous), baseline problems index (continuous), baseline symptom inventory (continuous), baseline urination frequency (continuous), baseline bladder-associated pain (continuous) and baseline bladder urgency (continuous). This study’s result was negative; the investigators found no statistically significant effect of the treatment over the placebo.

We sought to compare the performance of the different designs we considered as in Section 4.1. One way to do this comparison is via a parametric bootstrap resampling procedure. We sample $w$s from the designs and then $y$s from a probability model such as a logistic regression fitted to the entire dataset and the different $w$s. We calculate our estimators and compare. This simulation would lack verisimilitude if this probability model was wrong but given that we cannot perform new prospective studies under different designs, it is a good compromise for this work.

© 2023 Statistical Society of Australia.
Figure 2. Sample average MSE of $\hat{\beta}$ under the PM design and the BCRD design as a function of sample size for all settings of $d$. Error bars are unavailable as simulation estimates are highly dependent on each other as they are functions of much of the same data.

After dropping any patients with missing covariates or responses, the total number of subjects considered was $2n = 224$ where 116 subjects were administered the treatment, 108 were administered the placebo. Due to the endpoint’s discreteness, the thresholding on the sample median did not split the data evenly; 132 subjects’ responses were coded as zero and 92 responses were coded as 1.

As in the previous section, we simulate $d \in \{1, 2, 5\}$. We first fit a stepwise logistic regression to all 18 variables to find the importance order of the patient characteristics. In order of importance, the most important variable were baseline symptom inventory, baseline mental quality of life, baseline symptom inventory, baseline bladder-associated pain and lives with a partner. A regression on the most important variable yields high statistical significance of the first covariate, a regression the top $d = 2$ yields high statistical significance for the first covariate and borderline statistical significance for the second covariate and a regression on the top $d = 5$ yields high statistical significance for the first covariate and borderline statistical significance for the fifth covariate. Over 1,000 random training-test splits, the out-of-sample area under the curve values for each model are 0.68, 0.69 and 0.70, respectively indicating poor discrimination of the response (Hosmer Jr, Lemeshow & Sturdivant 2013, p. 177). Furthermore, it indicates that most of the predictive power is in the first covariate; the addition of the second covariate over the first and the second to fifth over the first and the third to fifth over the first and second does not add much predictive power.

We use these three regression models as probability models in this parametric bootstrap study. We also add a treatment effect to the models of $\beta_T = 1$ only to induce greater separation between the two designs. Because we assume a probability model, we precompute the true value of $\beta_T$ which we seek to estimate. To understand the sample-size dependence of performance, we simulate under $2n_0 \in \{40, 60, \ldots, 200, 220\}$. For a specific sample size, we draw a $w$ under PM, a $w$ under BCRD, then we draw a $y$ under PM’s allocation and finally a $y$ under BCRD’s allocation. We then compute $\hat{\beta}$. We repeat this process approximately 100,000 times for each $n_0, d$ for a total of approximately 3,000,000 faux clinical studies. MSE results were then aggregated at the values of $n_0$ and $d$ to average the squared error results and displayed in Figure 2.

The $d = 1$ and $d = 2$ cases are expected given our theory and the previous section’s simulation results. Even though the covariates are mostly uninformative, they are...
informative enough to the extent that Remark 3 does not apply. The matching $M$ seems to approximate the optimal matching well: the efficiency ratio of the BCRD estimator to the PM estimator at a sample size of 200 is approximately 2. This implies that PM can save 50% of the sample size in a trial relative to BCRD. However, at $d = 5$, we observe no performance edge of PM over BCRD. This is likely due to the matches being poor as there are now three of the five covariates completely uninformative and only one that is informative. Hence, we are likely matching randomly and observing no gain in PM over BCRD (as expected by Remark 1).

5. Discussion

The PM design occupies a central role when estimating a treatment’s risk difference in incidence response models as demonstrated by both theoretical analysis and multiple simulations. This result is not surprising as PM’s role in providing robust estimation under a non-linear response model has a long literature in observational studies (Stuart 2010). Also, PM was shown to be the minimax design in the case of continuous response when the response model belongs to a space of Lipschitz functions (Kallus 2018, section 2.3.2), which is clearly true of popular probability link models. PM being minimax differs with the analogous result for continuous response models as our theoretical results apply to probability response models of a general form (not only Lipschitz functions). The Bernoulli trial is the minimax design in the case of complete ignorance about the response function in the continuous setting as shown in (Kallus 2018, section 2.1) and Efron (1971, section 5); BCRD is the minimax design when assuming structure about the subjects’ covariates’ affect on the response $Wu$ (1981, theorems 1 and 3).

We make explicit our design recommendation of PM. For the case of $d = 1$, one first sorts the subjects in order of the value of the one measured covariate (or composite risk metric which may be common in many clinical settings). For $d > 1$, our theoretical conclusion requires sorting the subject by order of the unknown $v_i$s which is impossible in practice. However, the simulations of Section 4 show that non-bipartite pair matching using the Mahalanobis distance metric of the covariate vector pairs can approximate matching on the unknown $v_i$s. The intuitive reason why this performs well is that for any continuous function, $x_i \approx x_j$ implies that $f(x_i) \approx f(x_j)$. This approximation should be especially good in the usually assumed response model of $f(x) = \phi(\beta_0 + \beta^\top x)$ with $\phi$ being a link function with a slow-moving gradient throughout most of the input space. Section 4.2 demonstrates that even in a typical clinical setting where covariates are not too relevant to a noisy endpoint, the PM design still provides improvements in estimator efficiency. This efficiency persists only for matching on a few previously known influential variables. As the number of variables increases, and the additional variables are less important, PM loses its performance edge over BCRD as the matches become more and more random relative to the underlying true $v$ as expected by Remark 1. We do not observe evidence of adversarial matches (worse-than-random matches) in our simulations which would yield worse performance than BCRD (Remark 2). In conclusion, Remarks 1 and 2 imply that if the matches of the $v_i$s (unknown to the investigator) are better than average, PM will outperform BCRD. These better-than-average matches are realistic when the researcher knows a priori that are a few dominant covariates that affects the response. Since this is the case in our real-world example that
is why PM exhibits significant practical performance gains over BCRD. Thus, this work is not merely of theoretical interest.

There are many extensions of this work. Most glaring, we only did a theoretical analysis of the difference-in-means estimator for the risk difference parameter. We could theoretically explore the performance of the other many common estimators for risk ratio, odds ratio, log odds ratio (with and without covariate-adjustment post-assignment using a logistic regression for instance). Intuitively, we believe there are results analogous to Theorem 1 in the settings of the non-parametric log odds ratio estimator and the covariate-adjusted log odds ratio estimator (the logistic regression coefficient) as we observed empirically (see Figures 1 and 2) that the PM design outperforms BL which outperforms BCRD in estimator error. Furthermore, we can investigate paired estimators for all estimators mentioned above and our difference-in-means estimator. Even in our setting of the difference-in-means estimator for risk difference, there is more work to be done: we would like to establish theory for the case of $d > 1$ under the practice recommendation above of optimal non-bipartite matching via the Mahalanobis distance metric. This is important as we know that getting the matches ‘wrong’ results in performance worse than BCRD (in unshown simulations) and we would like to have a measure of risk to know when to revert to BCRD. We can also explore the performance of the pairing on-the-fly method of Kapelner et al. (2021) in the sequential allocation setting which is important as the vast majority of clinical trials are sequential. Lastly, this work was limited to estimation and assessing the uncertainty of the estimate to conduct inference is not straightforward. Extending the recent work of Pashley & Miratrix (2021) to both our estimator and binary outcome would be a good place to begin.

Supporting information

Additional supporting information may be found in the online version of this article at http://wileyonlinelibrary.com/journal/anzs.

Data S1. Supporting Information

References

ABDELBASIT, K.M. & PLACKETT, R. (1983). Experimental design for binary data. Journal of the American Statistical Association, 78, 90–98.

BECK, C., LU, B. & GREEVY, R. (2016). nbpmatching: Functions for optimal non-bipartite matching. r package version 1.5.1.

BERTSIMAS, D., JOHNSON, M. & KALLUS, N. (2015). The power of optimization over randomization in designing experiments involving small samples. Operations Research, 63, 868–876.

BJERMER, L., BISGAARD, H., BOUSQUET, J., et al. (2003). Montelukast and fluticasone compared with salmeterol and fluticasone in protecting against asthma exacerbation in adults: one year, double blind, randomised, comparative trial. Bmj, 327, 891.

BOX, G.E. & WILSON, K.B. (1951). On the experimental attainment of optimum conditions. Journal of the Royal Statistical Society, Series B, 13, 1–45.

EFRON, B. (1971). Forcing a sequential experiment to be balanced. Biometrika, 58, 403–417.

FISHER, R.A. (1925). Statistical methods for research workers. Edinburgh: Edinburgh Oliver & Boyd.

FOSTER, H., HANNO, P., NICKEL, J., et al. (2010). Effect of amitriptyline on symptoms in treatment naïve patients with interstitial cystitis/painful bladder syndrome. The Journal of urology, 183, 1853–1858.

GREEVY, R., LU, B., SILBER, J.H. & ROSENBAUM, P. (2004). Optimal multivariate matching before randomization. Biostatistics, 5, 263–275.

© 2023 Statistical Society of Australia.
HOSMER, D.W., Jr., LEMESHOW, S. & STURDIVANT, R.X. (2013). Applied logistic regression, 3rd edn. Hoboken: John Wiley & Sons.

IMBENS, G.W. & RUBIN, D.B. (2015). Causal inference in statistics, social, and biomedical sciences. New York: Cambridge University Press.

JOHNSON, R.T. & MONTGOMERY, D.C. (2009). Choice of second-order response surface designs for logistic and poisson regression models. International Journal of Experimental Design and Process Optimisation, 1, 2–23.

JULIUS, S., KJELDSEN, S.E., WEBER, M., et al. (2004). Outcomes in hypertensive patients at high cardiovascular risk treated with regimens based on valsartan or amlodipine: the value randomised trial. The Lancet, 363, 2022–2031.

KALLUS, N. (2018). Optimal a priori balance in the design of controlled experiments. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80, 85–112.

KAPELNER, A., KRIEGER, A.M., SKLAR, M., SHALIT, U. & AZRIEL, D. (2021). Harmonizing optimized designs with classic randomization in experiments. The American Statistician, 75, 195–206.

KRIEGER, A.M., AZRIEL, D.A. & KAPELNER, A. (2023). Better experimental design by hybridizing binary matching with imbalance optimization. Canadian Journal of Statistics, 51, 275-292. https://doi.org/10.1002/cjs.11685

LU, B., GREEVY, R., XU, X. & BECK, C. (2011). Optimal nonbipartite matching and its statistical applications. The American Statistician, 65, 21–30.

MANCENDIO, M.V., PAN, R., MONTGOMERY, D.C. & ANDERSON-COOK, C.M. (2019). Comparing d-optimal designs with common mixture experimental designs for logistic regression. Chemometrics and Intelligent Laboratory Systems, 187, 11–18.

MORGAN, K.L. & RUBIN, D.B. (2012). Rerandomization to improve covariate balance in experiments. The Annals of Statistics, 40, 1263–1282.

MYERS, W.R., MYERS, R.H. & CARTER, W.H., Jr. (1994). Some alphabetic optimal designs for the logistic regression model. Journal of Statistical Planning and Inference, 42, 57–77.

PARK, A.R., MANCENDIO, M.V. & MONTGOMERY, D.C. (2019). Separation in d-optimal experimental designs for the logistic regression model. Quality and Reliability Engineering International, 35, 776–787.

PASHLEY, N.E. & MIRATRIX, L.W. (2021). Insights on variance estimation for blocked and matched pairs designs. Journal of Educational and Behavioral Statistics, 46, 271–296.

PASHLEY, N.E. & MIRATRIX, L.W. (2022). Block what you can, except when you shouldn’t. Journal of Educational and Behavioral Statistics, 47, 69–100.

ROSENBERGER, W.F. & LACHIN, J.M. (2016). Randomization in clinical trials: theory and practice, 2nd edn. Hoboken: John Wiley & Sons.

RUBIN, D.B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. Journal of the American Statistical Association, 74, 318–328.

SCHEIN, A.I. & UNGAR, L.H. (2007). Active learning for logistic regression: an evaluation. Machine Learning, 68, 235–265.

SENN, S. (2013). Seven myths of randomisation in clinical trials. Statistics in Medicine, 32, 1439–1450.

STUART, E.A. (2010). Matching methods for causal inference: A review and a look forward. Statistical science, 25, 1.

STUDENT (1938). Comparison between balanced and random arrangements of field plots. Biometrika, 29, 363–378.

TEUGELS, J.L. (1990). Some representations of the multivariate Bernoulli and binomial distributions. Journal of multivariate analysis, 32, 256–268.

WU, C.F. (1981). On the robustness and efficiency of some randomized designs. The Annals of Statistics, 9, 1168–1177.

ZIETMAN, A.L., BAE, K., SLATER, J.D., et al. (2010). Randomized trial comparing conventional-dose with high-dose conformal radiation therapy in early-stage adenocarcinoma of the prostate: long-term results from proton radiation oncology group/American college of radiology 95-09. Journal of clinical oncology, 28, 1106.