Research Article

Statistics in Medicine

Received 21 August 2011, Accepted 8 February 2012 Published online 25 April 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5363

Balancing continuous and categorical baseline covariates in sequential clinical trials using the area between empirical cumulative distribution functions

Yunzhi Lin[a] and Zheng Su[b*†]

Covariate adaptive allocation is often adopted in sequential clinical trials to maintain the balance of baseline covariates that could potentially confound the outcome of a trial. Several allocation methods exist in the literature that can handle both continuous and categorical covariates. We propose a minimization approach to maintaining the balance of multiple continuous and categorical covariates in sequential clinical trials, which uses the area between the empirical cumulative distribution functions of the observed covariate values as the imbalance metric. Numerical results based on extensive simulation studies and a real dataset show that the proposed approach produces more accurate estimates of the treatment effect and leads to more powerful trials than the existing approaches for trials with binary, continuous, and time-to-event outcomes. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: clinical trial; randomization; baseline covariates; imbalance; cumulative distribution function; minimization

1. Background

Sequential clinical trials are frequently conducted to determine whether an experimental treatment is superior or non-inferior to a control treatment. Lachin *et al.* [1] emphasized that the conventional randomization method via flipping an unbiased coin may result in imbalance in groups sizes and in the distribution of some key baseline covariates when the sample size is relatively small. Covariate adaptive allocation is often adopted in sequential clinical trials to maintain the balance of baseline covariates that could potentially confound the outcome of a trial, and the key idea is that a new participant is sequentially assigned to a treatment group depending on the specific covariates and the assignments for previously enrolled patients. Various allocation approaches exist in the literature that are designed to balance the group sizes and categorical covariates (see, e.g., [2, 3]).

Ciolino *et al.* [4] studied the loss in power of ignoring the imbalances in continuous covariates in randomized clinical trials. In their approach, imbalance in the observed covariate values is measured by the ratio of the sum of ranks. There are several approaches available in the literature for balancing both continuous and categorical covariates in a sequential clinical trial. Frane [5] proposed a p-value based approach where imbalances in the covariates are measured by the p-values that correspond to testing whether the median values of the covariates in the two treatment groups are identical. Su [6] used the largest difference in the three pairs of quartiles of the two distributions of the observed covariate values to quantify the imbalance level of a continuous covariate. Endo *et al.* [7] proposed to use the Kullback–Leibler divergence (KLD) of the two probability density functions (PDFs) of the observed covariate values in the two treatment groups as the imbalance metric.

[a] *University of Wisconsin–Madison, Madison, WI, U.S.A.*
[b] *Genentech Inc., South San Francisco, CA 94080, U.S.A.*
[*] *Correspondence to: Zheng Su, Genentech Inc., South San Francisco, CA 94080, U.S.A.*
[†] *E-mail: su.zheng@gene.com*

We structure the rest of the paper as follows. In Section 2, we discuss details and major limitations of the existing metrics for continuous covariates and introduce a new imbalance metric, which applies to both continuous and categorical covariates. We describe the tools used to assess the performance of the proposed approach and the existing ones in Section 3. We provide simulation results of the proposed approach and the existing ones in Section 4, which show that the proposed approach produced more accurate estimates of the treatment effect and led to more powerful trials than the existing approaches for trials with binary, continuous, and time-to-event outcomes in all the simulation scenarios considered. We applied the proposed approach to a real dataset and summarized the results in Section 5. Again, the proposed approach outperformed the existing ones. We provide some discussions and concluding remarks in Section 6.

2. Methods

In the approach of Ciolino *et al.* [4], imbalance in the observed covariate values is measured by the ratio of the sum of ranks. For each covariate, the pooled covariate values are ranked and the ratio of the sum of the ranks in the experimental treatment group to the sum of the ranks in the control group is calculated as the measurement of imbalance. Frane's [5] method first temporarily assigns a new patient to one treatment group and then calculates the p-value for each of the covariates to determine how balanced the median values are if this patient were to be assigned into this group. Next, the smallest of all the p-values will be determined which is used to represent the imbalance level of these covariates. Similarly, the method will calculate the smallest p-value by temporarily assigning this new patient to the other treatment group. Eventually, this patient will be allocated with a higher probability to the group where the resulting smallest p-value is relatively larger, which represents less imbalance in the covariates. Su [6] used the largest difference in the three pairs of quartiles of the two distributions of the observed covariate values to quantify the imbalance level of a continuous covariate.

There are major limitations with the imbalance metrics used for continuous covariates in these approaches. In the approaches of Frane [5] and Ciolino $et al.$ [4], having a large p-value or a ratio of 1 for the sum of ranks of the covariate values does not necessarily mean that the distributions of a covariate are balanced between the two treatment groups. For example, suppose 10 patients have been assigned to one treatment group with covariate values $(6, 7, \ldots, 15)$ and another 10 patients have been assigned to the other group with covariate values $(1, 2, \ldots, 5, 16, 17, \ldots, 20)$. A Wilcoxon rank sum test produces a p-value that is equal to 1, which represents perfect balance in Frane's [5] approach. Similarly, the ratio of the sum of ranks is equal to 1, which represents perfect balance in the approach of Ciolino *et al.* [4]. However, the two distributions have markedly different variances. A p-value based on a rank test or the ratio of the sum of ranks reflects the difference in the medians of the two distributions but does not reflect the difference in the variance of the distributions. In Su's [6] approach, the distributions of the covariate values are considered perfectly balanced if these two distributions have equal quartiles. However, the two distributions can still be markedly different even with the same quartiles. A good imbalance metric should have the same bounded range of values for all the covariates such that they contribute equally to the overall balance of a trial when the sum of the imbalance values is used to quantify the balance of all the covariates. In addition, the metric should reflect the difference in the whole distribution functions instead of just several values of the functions.

In the approach of Endo *et al.* [7], the sum of the KLDs for all the covariates is used to quantify the imbalance level of all the covariates. One major limitation with the KLD is that it needs to be calculated on the basis of the PDFs of the distributions of the covariate values, and it is well known that estimating a PDF is much more difficult than estimating a cumulative distribution function (CDF) because of the important monotonicity property of a CDF. In addition, because a KLD is not bounded and it is calculated on the basis of the ratio of the PDFs, a PDF value that is close to 0 in the denominator can lead to an extreme outlier value for the KLD, which can dominate the sum of the KLDs. Suppose that a total of $m + n$ patients have been allocated with m patients receiving the experimental treatment and n patients receiving the control treatment. Denote the values of a continuous covariate observed in the experimental and control groups by $X = (x_1, \ldots, x_m)$ and $Y = (y_1, \ldots, y_n)$, respectively. Denote the PDFs of the observed covariate values in the two treatment groups by $f_1(x)$ and $f_2(x)$, respectively. The KLD from $f_1(x)$ to $f_2(x)$ is defined as

$$D_{\mathrm{KL}}(f_1 || f_2) = \int f_1(x) \log \frac{f_1(x)}{f_2(x)} \, \mathrm{d}x.$$

Because the KLD is not symmetric (i.e., $D_{\mathrm{KL}}(f_1||f_2)$ is not necessarily equal to $D_{\mathrm{KL}}(f_2||f_1)$), Endo *et al.* [7] proposed to use

$$D(f_1, f_2) = D_{\mathrm{KL}}(f_1||f_2) + D_{\mathrm{KL}}(f_1||f_2) = \int (f_1(x) - f_2(x))(\log f_1(x) - \log f_2(x))\,\mathrm{d}x$$

as the imbalance metric for the distributions of the observed covariate values. Realizing the difficulty in obtaining reliable estimates of PDFs with relatively small sample sizes, Endo *et al.* [7] assumed that the values of all the continuous covariates follow normal distributions, for which a closed form of the KLD is readily available. More specifically, suppose $f_i, i = 1, 2$ is the PDF of a normal distribution with mean μ_i and variance σ_i^2, we have

$$D(f_1, f_2) = \frac{1}{2}((\mu_1 - \mu_2)^2 + (\sigma_1^2 + \sigma_2^2))\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right) - 2.$$

Endo *et al.* [7] proposed to substitute the unknown means and variances by their sample counterparts when calculating $D(f_1, f_2)$. In practice, the distribution of a continuous covariate may not be normal (e.g., uniform or multimodal), and as a result, the normality assumption as required in the approach of Endo *et al.* [7] can often be violated.

Motivated by the ultimate goal of having identical distribution functions at the end of a trial, we herein propose to use the normalized area between the two empirical CDFs (ECDFs) to quantify the imbalance level in the distributions of a covariate. Denote the ECDFs for the treatment and control groups by $\hat{F}(t)$ and $\hat{G}(t)$, respectively, where

$$\hat{F}(t) = \frac{1}{m}\sum_{i=1}^{m} I\{x_i \leqslant t\}, \quad \hat{G}(t) = \frac{1}{n}\sum_{i=1}^{n} I\{y_i \leqslant t\},$$

and let $A(\hat{F}(t), \hat{G}(t))$ denote the area between $\hat{F}(t)$ and $\hat{G}(t)$. Because $A(\hat{F}(t), \hat{G}(t))$ is not unit less (i.e., it depends on the unit used for the observed covariate values), we propose to use the normalized area $\tilde{A}(\hat{F}(t), \hat{G}(t))$ as the imbalance metric for $\hat{F}(t)$ and $\hat{G}(t)$, where

$$\tilde{A}(\hat{F}(t), \hat{G}(t)) = \frac{A(\hat{F}(t), \hat{G}(t))}{\max(x_1, \ldots, x_m, y_1, \ldots, y_n) - \min(x_1, \ldots, x_m, y_1, \ldots, y_n)}.$$

The normalized area $\tilde{A}(\hat{F}(t), \hat{G}(t))$ is bounded by 0 and 1 with 0 representing perfect balance (i.e., the two ECDFs are identical) and 1 representing the worst possible imbalance.

The same metric can be applied to categorical covariates. For a categorical covariate with $K \geqslant 2$ different categories, suppose that for the m patients receiving the experimental treatment, the percent of patients in the K categories are $p_1, \ldots, p_K$ and, for the n patients receiving the control treatment, the percent of patients in the K categories are $q_1, \ldots, q_K$. The ideal balance is achieved when $p_1 = q_1, \ldots, p_K = q_K$, and as a result, all the K comparisons should contribute to the quantification of the imbalance level for this categorical covariate. We can treat a categorical covariate with K different categories as K binary (dummy) variables $(B_1, \ldots, B_K)$, with each binary variable representing one of the categories. If a subject is in category i, we have $B_i = 1$ and $B_j = 0$ for $j \neq i$. As a result, the observed probabilities for $B_i = 1$ in the two treatment groups are p_i and q_i and the area between the ECDFs for B_i is $|p_i - q_i|$. The sum of the areas for the K categories is $\sum_{i=1}^{K} |p_i - q_i|$. Given that

$$\sum_{i=1}^{K} |p_i - q_i| \leqslant \sum_{i=1}^{K} (p_i + q_i) = 2,$$

with equality holding if and only if at least one of p_i and q_i is equal to 0 for any i, we propose to use the normalized area

$$\tilde{A} = \sum_{i=1}^{K} |p_i - q_i|/2$$

to quantify the imbalance in the distributions of a categorical covariate. Again, $\tilde{A}$ is bounded by 0 and 1 with 0 representing perfect balance and 1 representing the worst possible imbalance.

We thus have introduced a metric for both continuous and categorical covariates, which is bounded by 0 and 1 for any covariate. The following procedure, named TAM (an acronym for total area minimization), can then be used in a sequential clinical trial with both continuous and categorical covariates.

1. Determine the covariates that need to be balanced for the clinical trial.
2. Determine the desired balance level for the group sizes. For a 1:1 allocation scheme, it will be important to have similar number of patients in both groups at the end of the trial. Let $C = |S_t - S_c|$ denote the difference in the number of patients in the two groups, where S_t and S_c denote the number of patients in the treatment group and the control group, respectively. A binary imbalance score is usually desired for group sizes (i.e., group sizes are either balanced or not balanced). In this case, the group sizes may be considered imbalanced if $C > c$, where c is a constant specified by the study team.
3. Start the trial with simple randomization until there is at least one subject in each of the two treatment groups, following which a minimization algorithm can be applied. Use the sum of the areas between the ECDFs for all the covariates as the imbalance metric, which is minimized as the study progresses. A weighted sum of the areas can be considered if it is more important to achieve balance for some covariates than others as determined by the study team.
4. To determine whether to allocate a new patient to either the treatment group or the control group, the study team can first assess whether it affects the balance in group sizes by assigning this patient to one of the two groups. In cases where one option will result in imbalance in the group sizes and the other will not, the patient should be assigned to the group that will keep the group sizes balanced. Otherwise, the sum of the areas by temporarily assigning the patient to each of the two groups can be compared, and eventually, the patient will be assigned to the group that corresponds to a smaller total area. If non-deterministic allocation is desired, a new patient can be assigned with a higher probability (e.g., 90%) to the group that corresponds to the smaller total area via a biased coin. When deterministic allocation is replaced by biased coin randomization, the proposed approach is similar to adding Efron's [8] biased coin randomization to the minimization method of Pocock and Simon [3]. The method of Pocock and Simon [3] only applies to categorical variables, whereas the proposed approach can handle both continuous and categorical variables without having to categorize the continuous variables. In cases where there is a tie in the sum of the areas by temporarily assigning the patient to each of the two groups, the patient will be assigned randomly with an unbiased coin.

3. Tools to assess performance

We can examine the performance of the proposed method and several other randomization methods from three aspects: the amount of imbalance, the loss of information due to imbalance, and the impact on the inference of the treatment effect. These properties are assessed by means of three tools: (1) the overall imbalance score based on the area between the ECDFs; (2) the loss function introduced by Smith [9]; and (3) the precision of the estimation and the power attained.

3.1. Area between the ECDFs

The amount of imbalance for a categorical factor can be easily defined, for instance, as the difference in the number or percentage of subjects between treatments within each category. Herein, we will use the newly introduced metric of area between the ECDFs as the imbalance measure. As discussed in Section 2, this metric has the benefit of allowing direct comparisons of algorithms with imbalances of both categorical and continuous covariates measured by the same standard. In the simulation study, we will use the sum of the imbalances across all prognostic factors as the overall imbalance score.

3.2. Loss function

A useful tool to evaluate the performance of randomization schemes is the loss function introduced by Smith [9] and later adopted by many others (see, e.g., [10–12]). This important measure quantifies the loss of efficiency due to imbalance in the prognostic covariates. Consider a two-treatment study with n patients. If the outcome of interest is continuous and normally distributed with variance σ^2, we can write the regression model as

$$E(Y) = \alpha\Delta_n + X_n\beta,$$

where α and β represent the effects of the treatment and the prognostic factors, Δ_n is the $n \times 1$ vector of allocation with elements $+1$ and -1, and X_n is the $n \times q$ design matrix composed of an intercept and $q-1$ prognostic factors. Note that categorical covariates here will be coded as dummy variables. The variance of the least squares estimator of α can be written in the form

$$\mathrm{Var}(\hat{\alpha}) = \frac{\sigma^2}{n - \Delta_n^t X_n (X_n^t X_n)^{-1} X_n^t \Delta_n} = \frac{\sigma^2}{n - L_n},$$

where $L_n = \Delta_n^t X_n (X_n^t X_n)^{-1} X_n^t \Delta_n$ is the loss function. The variance is minimal when the design is exactly balanced, that is, L_n is zero if all the covariates have the same ECDFs in both treatment groups. L_n is a loss measuring the effect of imbalance on the variance of the treatment effect estimate. Atkinson [10, 11] and Heritier *et al.* [12] have extensively studied the properties of the loss function for various designs and sample sizes.

3.3. *Precision of estimation and power*

To evaluate the impact of imbalance on the precision and power of testing for the treatment effect in trials with various outcomes, we compare the performance of the proposed method and the existing ones in three widely used models: (1) a linear regression model with continuous outcomes; (2) a logistic regression model with binary outcomes; and (3) a Cox proportional hazards model with time-to-event outcomes. More specifically, the linear model has the form $Y = \alpha\Delta + \sum_{i=1}^{c} \beta_i X_i + e$, the logistic model has the form $\log(\text{odds}) = \alpha\Delta + \sum_{i=1}^{c} \beta_i X_i$, and the Cox proportional hazards model has the form $h(t|\Delta, X_1, \ldots, X_c) = h_0(t)e^{\alpha\Delta + \sum_{i=1}^{c} \beta_i X_i}$, where Y is the outcome, α is the effect of the treatment, Δ is the indicator for treatment assignment, β_i is the fixed effect of baseline covariate i, X_i is the value of covariate i, c is the total number of covariates, $e \sim N(0, 1)$, and $h_0(t)$ is the baseline hazard function.

We consider various treatment effect sizes, including an effect size of zero to evaluate the type I error rate. For each trial, we record the estimate of the treatment effect and its p-value, adjusting for the baseline covariates. We calculate the mean squared error of the estimated treatment effect and the statistical power (or type I error rate) as the proportion of simulations with the treatment effect p-value < 0.05.

4. Simulation study

We evaluate the randomization methods in the context of trials with two treatment groups, active and control. We consider three prognostic factors, one continuous and two categorical, and generated with the distributions shown in Table I. This design represents a trial with a typical number of factors and different possible values for each factor. We list in Table I the β coefficients of the three covariates for simulating trial outcomes in the power calculation. For the results to be general enough to give some insight into the allocation methods under a variety of circumstances, we will examine a series of trials for up to 200 patients.

4.1. *Allocation methods*

We will compare the performance of the proposed method with three other allocation methods: (1) the simple randomization method; (2) the p-value based approach by Frane [5]; and (3) the minimization scheme using quartiles by Su [6]. We will not compare the proposed method with schemes that are

Table I. Probability distribution and effect size of the prognostic factors used for simulation.

Prognostic factor	Attribute	Distribution	Effect size
X1	Continuous	Uniform $U(0, 2)$	1.0
X2	Categorical	Two categories with probabilities 0.5 and 0.5	0.5
X3	Categorical	Three categories with probabilities 0.5, 0.3 and 0.2	0.3, 0.6

designed for only categorical but not continuous covariates (e.g., the minimization approach by Pocock and Simon [3]). We also do not include in the simulation study the KLD method by Endo *et al.* [7] as it would require a normality assumption for the continuous covariates, which is often violated.

For all three covariate adaptive methods, we start the algorithm when there is at least one subject in each of the two treatment groups. Both the proposed method and Frane's p-value method first check the overall trial imbalance before balancing the covariates, and here, we set the imbalance threshold at 3, meaning the algorithm will only proceed to check covariate imbalance if the difference in the overall group sizes is less than or equal to 3. In Su's minimization scheme, the group size is considered imbalanced if the number of patients between the treatment and control groups differ by more than 3. A continuous covariate is considered imbalanced if any pair of its quartiles differs by more than 10%, and a categorical covariate is considered imbalanced if the number of patients for any of its categories differs by more than 2. The weighted overall imbalance score is defined as a weighted sum of the imbalance indicator for the overall trial and each of the factors. The weight vector here is chosen to be $w = c(4, 3, 3, 3)$, which reflects that it is of higher priority to maintain the balance for the overall trial and that the three covariates are considered of equal importance. The new patient is assigned to the treatment group that will result in a lower overall imbalance score.

4.2. Results

4.2.1. Overall imbalance score. Figure 1 shows the average overall imbalance score based on the total area of 10,000 simulated trials for up to 200 patients. The proposed method achieves the lowest overall imbalance score among all the methods studied. For instance, for a clinical trial with 100 patients, the average total area is about 0.059, which is 68% lower than that of Frane's p-value approach (average

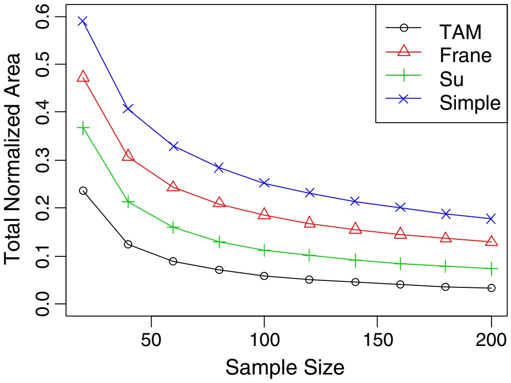

Figure 1. The average value of total normalized area versus sample size for the four methods compared: the proposed method (TAM); the p-value approach by Frane (Frane); the minimization scheme using quartiles by Su (Su); and the simple randomization (Simple). Two-arm trial with three prognostic factors.

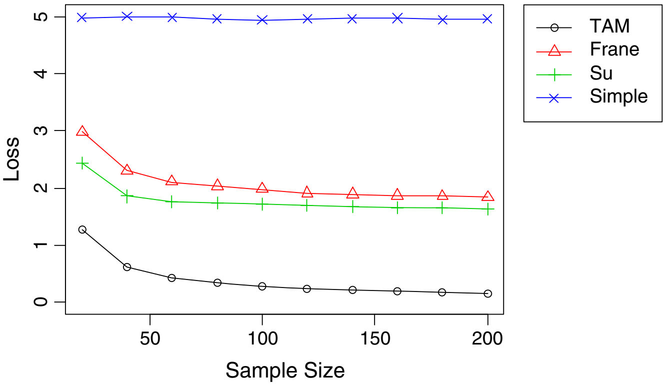

Figure 2. The average value of loss versus sample size for the four methods compared; two-arm trial with three prognostic factors.

Table II. Type I error rate under various sample sizes for the four methods compared: the proposed method (TAM); the p-value approach by Frane (Frane); the minimization scheme using quartiles by Su (Su); and the simple randomization (Simple).

Panel (a)	Linear regression model, three prognostic factors with pre-specified fixed effects			
	$N = 50$	$N = 100$	$N = 150$	$N = 200$
TAM (%)	4.8	4.8	4.9	4.9
Frane (%)	5.1	5.0	5.3	5.2
Su (%)	5.3	4.9	5.0	4.9
Simple (%)	4.8	4.5	4.6	5.1
Panel (b)	**Logistic regression model, three prognostic factors with pre-specified fixed effects**			
	$N = 50$	$N = 100$	$N = 150$	$N = 200$
TAM (%)	4.9	5.0	4.7	4.9
Frane (%)	5.0	5.7	5.5	5.2
Su (%)	5.0	5.1	5.0	5.0
Simple (%)	4.7	5.2	5.5	5.0
Panel (c)	**Cox regression model, three prognostic factors with pre-specified fixed effects**			
	$N = 50$	$N = 100$	$N = 150$	$N = 200$
TAM (%)	5.1	4.8	5.1	4.6
Frane (%)	5.1	4.9	5.0	4.4
Su (%)	5.1	5.0	5.2	4.6
Simple (%)	5.6	5.0	5.2	5.2

Table III. Power for all methods under various sample sizes and treatment effect sizes.

Panel (a) Linear model

	Power ($N = 50$)			Power ($N = 100$)			Power ($N = 150$)			Power ($N = 200$)		
Effect size	0.2	0.4	0.6	0.2	0.4	0.6	0.2	0.4	0.6	0.2	0.4	0.6
TAM (%)	12.0	27.5	54.2	18.1	51.2	85.4	24.0	68.9	96.3	30.2	80.8	99.6
Frane (%)	10.5	26.3	52.9	16.1	49.8	83.0	22.8	67.0	94.5	29.2	79.9	98.5
Su (%)	10.8	26.2	52.6	16.0	49.0	83.4	22.4	67.4	94.7	29.3	79.0	98.5
Simple (%)	10.1	25.2	50.5	15.4	47.2	82.5	22.1	66.4	94.3	27.5	77.7	97.8

Panel (b) Logistic regression model

	Power ($N = 50$)			Power ($N = 100$)			Power ($N = 150$)			Power ($N = 200$)		
Effect size	0.6	0.8	1	0.6	0.8	1	0.6	0.8	1	0.6	0.8	1
TAM (%)	8.8	11.8	15.2	16.9	25.4	36.1	23.2	37.6	49.9	29.1	47.2	63.4
Frane (%)	8.7	10.8	13.9	16.1	25.4	34.8	22.9	35.8	49.0	28.0	47.1	61.4
Su (%)	8.8	11.0	13.2	15.0	24.7	35.3	22.3	35.2	48.1	28.6	44.9	61.6
Simple (%)	7.3	10.2	13.0	15.0	23.4	34.7	21.5	35.0	47.1	27.5	44.5	60.4

Panel (c) Cox regression model

	Power ($N = 50$)			Power ($N = 100$)			Power ($N = 150$)			Power ($N = 200$)		
Effect size	0.6	0.8	1	0.6	0.8	1	0.6	0.8	1	0.6	0.8	1
TAM (%)	11.9	24.3	42.9	15.3	39.3	68.0	18.5	51.9	83.9	22.0	64.3	93.5
Frane (%)	10.2	22.1	39.4	13.7	38.1	66.2	18.1	51.5	82.4	21.5	63.5	91.8
Su (%)	11.5	21.9	37.8	14.0	36.3	67.4	18.2	50.7	82.6	21.7	63.8	91.7
Simple (%)	10.1	21.8	37.2	12.5	35.2	64.9	18.0	51.2	81.7	21.3	62.0	91.4

total area = 0.186) and 48% lower than that of Su's minimization approach (average total area = 0.112). For all four methods, the level of imbalance decreases as the number of subjects increases. As expected, simple randomization has the worst imbalance at all sample sizes.

4.2.2. Loss function. Figure 2 shows the average loss of 10,000 simulated trials for up to 200 patients. For the proposed method, the loss is about 1.3 at 20 patients and rapidly declines to almost zero as the sample sizes increases. Both the p-value based approach by Frane and the minimization scheme using quartiles by Su have much greater loss than the proposed approach. The simple randomization approach has the largest loss value for all sample sizes considered.

4.2.3. Type I error rate, power, and precision of estimation. Table II shows that, for all three models, all the randomization methods considered resulted in type I error rates around the conventional 5% level across different sample sizes, assuming no treatment effect. This demonstrated that covariate adaptation during the trial does not incur inflated type I error rates. Table III shows that the proposed method achieved the highest power for all the sample sizes and effect sizes considered.

Figure 3 shows the mean squared errors from 10,000 simulated trials based on the linear model, the logistic regression model, and the Cox model, respectively. The proposed method achieved the lowest

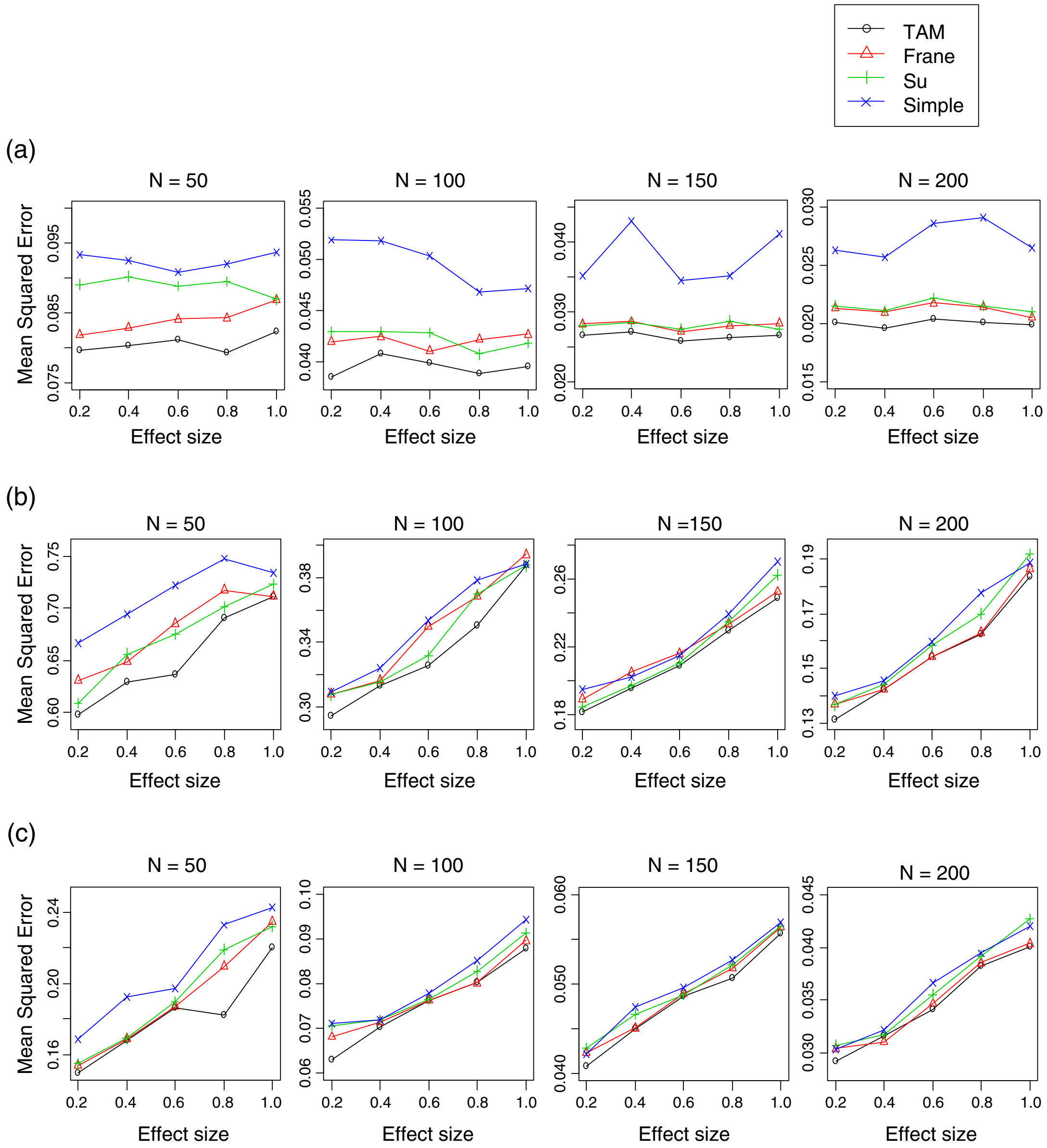

Figure 3. Average observed squared errors across treatment effect sizes (range from 0.2 to 1.0), for sample sizes of 50, 100, 150, and 200. (a) Linear regression model; (b) logistic regression model; (c) Cox regression model.

mean squared errors for all the sample sizes and treatment effect sizes considered. The advantage of the proposed approach is more prominent when the sample size is smaller. The p-value approach by Frane and the quartile approach by Su produced similar mean squared errors. Again, the simple randomization method exhibited the largest mean squared errors among all the allocation methods studied.

In summary, the proposed approach adequately controls the type I error rate, increases the power of a trial, and produces more accurate estimates of the treatment effect compared with the existing approaches in the literature.

Atkinson [13] defined the selection bias of a treatment allocation algorithm as the difference in the average number of correct and incorrect guesses of the next treatment to be allocated. The proposed approach is completely predictable (unless there is a tie in the sum of the areas by temporarily assigning the patient to each of the two groups) if a clinician is aware of both the treatment assignments for the previously allocated patients and all the values of their stratification variables. The simple randomization approach, on the other hand, is completely unpredictable and thus has no selection bias. It is important to note that the gain in power and accuracy in the treatment effect estimate with the proposed approach is achieved by having no randomization. Even though Figure 2 shows that the proposed approach and the simple randomization approach have the smallest and largest loss, respectively, the trend observed for loss is reversed for selection bias as the proposed approach and the simple randomization approach have the largest and smallest selection bias among the four approaches considered.

5. A real example

In this section, we compare the performance of the proposed method with that of the existing ones on a real dataset from patients with burn wounds. Infection of a burn wound is a common complication resulting in extended hospital stays and potentially death of burned patients, and control of infection remains a prominent component of burn management. Ichida *et al.* [14] compared a routine bathing care method (initial surface decontamination with 10% povidone–iodine followed with regular bathing with Dial soap (The Dial Corporation, Scottsdale, Arizona, United States.)) with a body-cleansing method using 4% chlorhexidine gluconate in terms of reducing the risk of infection for burned patients. In this study, several baseline covariates that are potentially prognostic of infection risk were considered in the analyses. The categorical covariates were gender, race (white and non-white), and type of burn (four categories), and the continuous covariate was severity of the burn as measured by percentage of total surface area of body burned. In this study, the average percentage of burn was 24.7% with a range of 2–95%, and the distribution of this continuous covariate was highly non-normal as demonstrated by normality tests. We can find the data for this study in [15].

When designing a sequential clinical trial comparing a novel treatment with the standard of care for the control of infection risk in patients with burn wounds, these prognostic factors will need to be balanced. We considered various sample sizes for a clinical trial (20–200 patients) and sampled from these patients as in the study of Ichida *et al.* [14] to compare the performance of various randomization methods. For each sample size, a total of 10,000 simulated trials were considered. We intended to determine what the covariate imbalance level would have been had these patients been enrolled in a sequential clinical trial using one of these randomization methods. Figures 4 and 5 show that the proposed approach resulted in

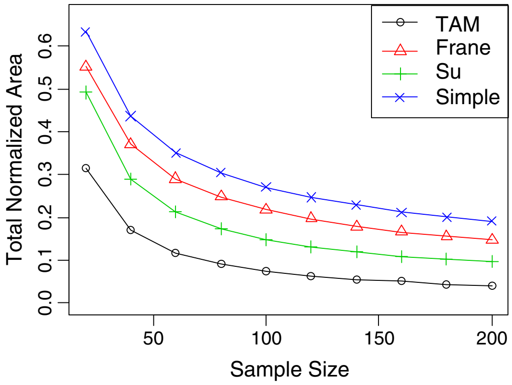

Figure 4. The average value of the total normalized area versus sample size; burn management example.

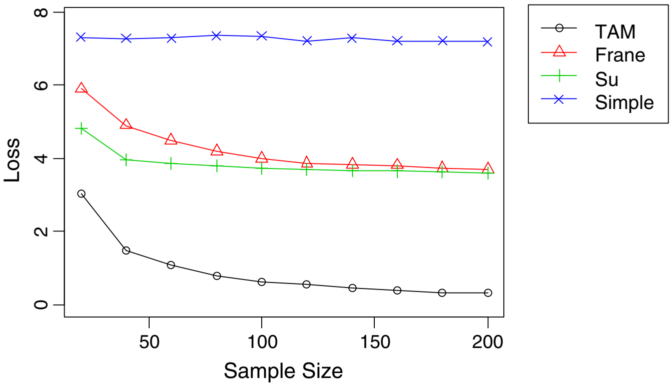

Figure 5. The average value of loss versus sample size; burn management example.

the smallest average total area and loss value for all the sample sizes considered. More specifically, the loss value converged to zero as the sample size increased for the proposed method, which shows that it can achieve the optimal power for testing the treatment effect when the sample size is relatively large.

6. Discussions and conclusions

Covariate adaptive allocation is often adopted in sequential clinical trials to maintain the balance of baseline covariates that could potentially confound the outcome of a trial. Several allocation methods exist in the literature that can handle both continuous and categorical covariates. However, there are potential issues with the imbalance metrics used for continuous covariates in these existing approaches. In this paper, we introduce a new imbalance metric, which applies to both continuous and categorical covariates. This imbalance metric is defined as the normalized area between the two ECDFs of the observed covariate values in both treatment groups. As a result of the normalization, each covariate has the same range of imbalance metric values from 0 to 1 with 0 representing perfect balance and 1 representing extreme imbalance. In addition, the proposed approach makes no assumptions on the distributions of the covariate values. Numerical results based on simulated trials and a real dataset show that the proposed approach produced more accurate estimates of the treatment effect and led to more powerful trials than the existing approaches for trials with binary, continuous, and time-to-event outcomes. When assessing a new allocation scheme for a sequential clinical trial, researchers should assess whether the scheme adequately controls the type I error rate, how its power compares with those of the existing approaches, and whether it produces the most accurate effect size estimate as measured by the mean squared errors for the analysis endpoints/approaches to be used at the end of the trial. The proposed approach may be preferred over the existing ones when there are one or more continuous covariates that need to be balanced in a sequential clinical trial.

The proposed approach is similar to the two-sample Kolmogorov–Smirnov test in the sense that they both compare the difference between the two ECDFs of the two samples. Different from the Kolmogorov–Smirnov test, which looks at the maximum deviance between the two ECDF curves, the proposed method looks at the area between the two curves. Our experience with an algorithm that uses the maximum deviance between the two ECDF curves as the imbalance metric is that it performed slightly worse than the proposed method. We think the main reason is that the maximum deviance between the two ECDFs is not sufficient to capture how different the two distribution functions are. For example, the same maximum deviance may correspond to two curves with a systematic shift or it may be that the two curves are nearly identical except at one specific location. These two scenarios may lead to quite different areas between the two curves as used in the proposed approach.

Senn *et al.* [16] nicely summarized the current status of an ongoing debate among statisticians on whether minimization-based approaches should be adopted in sequential clinical trials. One major concern with a minimization-based approach is that there is a loss of randomness in the allocation algorithm as the allocation of a new patient is determined by the treatment assignments of previously enrolled patients and their covariate values. The proposed approach can be modified to reduce the loss of randomness. For example, patients enrolled at the beginning of a trial can be randomized with an unbiased coin, and minimization can be introduced after a certain number or percent of patients have been randomized (the number or percent of patients can be random within a certain given range). After minimization has been triggered, it is also possible that not all future patients need to be allocated on the basis of the minimization algorithm. Instead, future patients can be randomly selected to either be randomized with an unbiased coin or be allocated according to the minimization algorithm. We refer interested readers to [16] for a detailed discussion on the limitations of minimization-based methods.

Acknowledgement

The authors thank the anonymous reviewers for their insightful comments and suggestions, which have led to a significantly improved paper.

References

1. Lachin JM, Matts JP, Wei LJ. Randomization in clinical trials: conclusions and recommendations. *Controlled Clinical Trials* 1988; **9**:365–374.
2. Taves DR. Minimization: a new method of assigning patients to treatment and control groups. *Clinical Pharmacology and Therapeutics* 1974; **15**:443–453.
3. Pocock SJ, Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* 1975; **31**:103–115.
4. Ciolino J, Zhao W, Martin R, Palesch Y. Quantifying the cost in power of ignoring continuous covariate imbalances in clinical trial randomization. *Contemporary Clinical Trials* 2011; **32**:250–259.
5. Frane JW. A method of biased coin randomization, its implementation, and its validation. *Drug Information Journal* 1998; **32**:423–432.
6. Su Z. Balancing multiple baseline characteristics in randomized clinical trials. *Contemporary Clinical Trials* 2011; **32**:547–550.
7. Endo A, Nagatani F, Hanada C, Yoshimura A. Minimization method for balancing continuous prognostic variables between treatment and control groups using Kullback-Leibler divergence. *Contemporary Clinical Trials* 2006; **27**:420–431.
8. Efron B. Forcing a sequential experiment to be balanced. *Biometrika* 1971; **58**:403–417.
9. Smith RL. Properties of biased coin designs in sequential clinical trials. *The Annals of Statistics* 1984; **12**:1018–1034.
10. Atkinson AC. Optimum biased-coin designs for sequential treatment allocation with covariate information. *Statistics in Medicine* 1999; **18**:1741–1752.
11. Atkinson AC. The distribution of loss in two-treatment biased-coin designs. *Biostatistics* 2003; **4**:179–193.
12. Heritier S, Gebski V, Pillai A. Dynamic balancing randomization in controlled clinical trials. *Statistics in Medicine* 2005; **24**:3729–3741.
13. Atkinson AC. The comparison of designs for sequential clinical trials with covariate information. *Journal of the Royal Statistical Society Series A* 2002; **165**:349–373.
14. Ichida JM, Wassell JT, Keller MD, Ayers LW. Evaluation of protocol change in burn-care management using the Cox proportional hazards model with time-dependent covariates. *Statistics in Medicine* 1993; **12**:301–310.
15. Klein JP, Moeschberger ML. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer: New York, 1997.
16. Senn S, Anisimov VV, Fedorov VV. Comparisons of minimization and AtkinsonŠs algorithm. *Statistics in Medicine* 2010; **29**:721–730.