Contemporary Clinical Trials Communications 3 (2016) 39–47

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

journal homepage: www.elsevier.com/locate/conctc

Research paper

Survival analysis following dynamic randomization

Xiaolong Luo [a, *], Mingyu Li [a], Gongjun Xu [b], Dongsheng Tu [c]

[a] Biometrics and Data Operations, Celgene Corporation, 300 Connell Drive, 3-7059 Berkeley Heights, NJ 07922, USA
[b] School of Statistics, University of Minnesota, USA
[c] NCIC Clinical Trials Group, Queen's University, Kingston, ON K7E 3L6, Canada

ARTICLE INFO

Article history:
Received 7 December 2015
Received in revised form
16 February 2016
Accepted 29 February 2016
Available online 10 March 2016

ABSTRACT

In this paper, we propose a method to analyze survival data from a clinical trial that utilizes a dynamic randomization for subject enrollment. The method directly accounts for dynamic subject randomization process using a marked point process (MPP). Its corresponding martingale process is used to formulate an equation for estimating the treatment effect size and for hypothesis testing. We perform simulation analyses to evaluate the outcomes of the proposed method as well as the conventional log rank method and re-randomized testing procedure.

1. Introduction

Randomized controlled trials have been a gold standard to demonstrate safety and efficacy of an experimental regimen compared with a standard regimen. They become major body of evidence for regulatory approval for marketing authorization. The randomization of subjects for clinical trials is devised to achieve two important objectives: a) to ensure that samples from treatment groups are comparable with respect to prognostic factors [8]; and b) to ensure the distributional validity of the statistics that are used for estimating and testing the treatment effect [6] and [1].

In theory, a simple randomization could ensure distributional balance within prognostic factors between treatment groups. However, with finite sample, the randomization may not necessarily reach the intended balance within each prognostic factor. In order to solve this problem, Zelen [26] proposed block randomization to ensure the balance within a few strata. When the number of strata levels increases, the block randomization may not achieve the overall balance between the treatment groups either. Taves [24] and Pocock and Simon [13] proposed a dynamic randomization scheme as a practical solution. The method has been commonly utilized in clinical trials with many well known prognostic factors, e.g., see Ref. [16].

While dynamic randomization can ensure the balance with respect to many prognostic factors between the assigned treatment groups (see Ref. [25]) and thus achieves the above objective a), the process may alter the *unconditional* distribution of the treatment assignment and make the above objective b) questionable when applying a probability distribution for the test statistics, which has been a concern in various regulatory settings. The CPMP Points to Consider on Adjustment for Baseline Covariates see Ref. [4] states a strong position against dynamic allocation. FDA guidance for industry (E9) see Ref. [11] recommends analysis to stratify factors used for dynamic randomization. There have been extensive discussion over this controversy (see for example [20]). Although some argue that conventional analyses are still appropriate when dynamic randomization was used [2], others still think a direct link between randomization and methods of statistical analysis is needed to draw reliable conclusions from clinical trial data [5].

Permutation and re-randomization based testing procedures have been used as last resort to avoid the ambiguity about the distributional property of the standard statistics following dynamic randomization, e.g., Flyer [15]; Hasegawa and Tango [28]. However, a permutation test may not be computationally practical for even moderate sample size. The re-randomization test may not be reproducible due to the use of different random number generator and the choice of replication number. More importantly, neither test is directly linked to the alternative hypothesis and there is no estimation procedure of treatment effect that is intrinsically compatible with the testing procedure. Recently, for the trials with continuous endpoints, Shao et al. [22] derived a valid t-type test based on the bootstrap method. Their results have been generalized to clinical trials with binary responses and event counts as primary endpoints [21] and to a large family of covariate-adaptive designs including dynamic randomization and tests under a linear model framework [17]. No valid procedure has been developed currently for the clinical trials with time to event endpoints and dynamic randomization.

In general, the treatment assignment mechanism via a dynamic randomization is measurable, or adaptive, with respect to the

* Corresponding author.
E-mail address: xluo@celgene.com (X. Luo).

http://dx.doi.org/10.1016/j.conctc.2016.02.004

information available prior to the randomization. Based on censored survival data, Luo et al. [12] developed the asymptotic normality for a class of statistics that are adaptive to accumulative information based on a martingale approximation. In this paper, we will apply their general theory to study the statistics that results from dynamically randomized clinical trials. In section 2, the framework of a clinical trial with dynamic randomization is described with time to an event endpoints. In section 3, Mantel-Hanzel log rank test statistics [18] and its corresponding re-randomized test [15] will be reviewed. Then, a Marked Point process (MPP) based statistics will be introduced based on [12] for both treatment effect estimation and hypothesis testing. In section 4, simulation analyses will be used to evaluate the performance of the proposed inference procedure with conventional procedures. In section 5, a real cancer trial will be used to illustrate the application of those procedures. Finally, section 6 will conclude with some discussions.

2. Clinical trials with dynamic randomization

Suppose a clinical trial starts at the calendar time $u_0 = 0$ and denote $\mathcal{T}$ a fixed but sufficiently late calendar time before which the trial will be completed. Subjects will be sequentially accrued into the trial. Denote $u_i, i \geq 1$ as the calendar time when the i-th subject arrives randomly. For simplicity, we assume that $0 < u_1 < u_2 < u_3 < \ldots$, i.e., no two subjects are enrolled at the same time and there are only two stratification factors. The results in this paper can be generalized easily to more general situations. Let (G_i, H_i) be two (discrete) baseline variables of the i-th subject. For example, $G_i \in \{1,2,\ldots,L_G\}$ may represent the site at which the subject is accrued and $H_i \in \{1,2,\ldots,L_H\}$ may represent the disease stage of the subject. Let $X_i \in \{0,1\}$ be the assigned treatment group, whose assignment will be specified later. Let Y_i be the time the subject stays in the trial so that $u_i + Y_i$ will be the calendar time at which the subject leaves the trial. Let δ_i denote the outcome variable of the subject i at his or her departure. Thus, the trial data will be the collection of $\{(u_i, G_i, H_i, X_i, Y_i, \delta_i), i = 1,2,3,\ldots\}$. We adopt a conventional survival data model and assume that there are independent random variables T_i and C_i such that

$$P(T_i > w | X_i = x) = \exp\left\{ -\int_0^w h_{x,1}(v)dv \right\}, w > 0,$$

$$P(C_i > w | X_i = x) = \exp\left\{ -\int_0^w h_0(v)dv \right\}, w > 0,$$

for some hazard functions $h_{x,1}$, $x = 0,1$, and censoring hazard function h_0. The outcome measures Y_i and δ_i are derived from $Y_i = T_i \wedge C_i$ and $\delta_i = 1(T_i \leq C_i)$.

2.1. Dynamic randomization

At any given calendar time $t > 0$, we record the information available at t as follow: the number of subjects enrolled up to time t is

$$R_t = \sum_{i \geq 1} 1(u_i \leq t),$$

and the numbers of subjects by treatment group $x = 0,1$ are

$$n_X(t;x) = \sum_{i \geq 1} 1(u_i \leq t, X_i = x).$$

The numbers of subjects by treatment group and baseline factors are

$$n_{X,G,H}(t;x,g,h) = \sum_{i \geq 1} 1(u_i \leq t, G_i = g, H_i = h, X_i = x),$$
$$g = 1,2,\ldots,L_G; h = 1,2,\ldots,L_H; x = 0,1,$$

and the corresponding marginal counts by site G and stage H are

$$n_{X,G}(t;x,g) = \sum_h n_{X,G,H}(t;x,g,h),$$
$$n_{X,H}(t;x,h) = \sum_g n_{X,G,H}(t;x,g,h),$$

and

$$n_G(t;g) = \sum_x n_{X,G}(t;x,g),$$
$$n_H(t;h) = \sum_x n_{X,H}(t;x,h).$$

As discussed by Refs. [25]; there are many ways to conduct a dynamic randomization. In a simple minimization procedure (see the application example in section 5), we can let

$$X_t = \begin{cases} 1 & \text{if } n_{X,G,H}(t-;1,g,h) < n_{X,G,H}(t-;0,g,h); \\ 0 & \text{if } n_{X,G,H}(t-;1,g,h) > n_{X,G,H}(t-;0,g,h); \\ \xi_t & \text{if } n_{X,G,H}(t-;1,g,h) = n_{X,G,H}(t-;0,g,h); \end{cases}$$

for a new subject accrued at the calendar time t with baseline $G = g$ and $H = h$, where the Bernoulli distributed $\xi_t \sim b(1,0.5)$ is an independent random variable. In the appendix, we describe another more general common approach based on balance scores and will use it for the simulation examples of Section 4.

3. Statistical inference with dynamic randomization

In this section, we briefly describe the conventional Mantel-Hanzel log rank test and the commonly used re-randomization test. Then, we will provide detail for the proposed MPP based procedure. We will use the terminology from Section 2 and assume that a clinical trial data consists of the collection of $\{(u_i, G_i, H_i, X_i, Y_i, \delta_i), i = 1,2,3,\ldots\}$.

3.1. Mantel-Hanzel log rank test

The naive log rank test would ignore the dynamic randomization process and use only the information time based data $\{(X_i, Y_i, \delta_i), i = 1,2,3,\ldots\}$.

Let $d = \sum \delta_i$ be the number of events and $Y_{(1)} < Y_{(2)} < \ldots < Y_{(d)}$ be the ordered event times. Denote $r_{x,j}$ the number of subjects at risk from the treatment group x, prior to *information* time $Y_{(j)}$, and $m_{x,j}$ the number of subjects from the treatment group x who had events at time $Y_{(j)}$. Denote $m_j = m_{1,j} + m_{0,j}$. Let

$$e_j = \frac{r_{1,j} m_j}{r_{1,j} + r_{0,j}},$$

$$v_j = \frac{r_{1,j} r_{0,j} m_j (r_{1,j} + r_{0,j} - m_j)}{(r_{1,j} + r_{0,j} - 1)(r_{1,j} + r_{0,j})^2}.$$

The Mantle-Hanzel log rank test can be calculated as

$$Z_{mh} = \frac{\sum_1^d (m_{1,j} - e_j)}{\sqrt{\sum_1^d v_j}}, \quad (1)$$

which approximates the standard normal distribution based on the rationale that the outcomes of $r_{1j}+r_{0j}$ are independent and m_{1j} follows a hyper geometric distribution with mean e_j and variance v_j (see Ref. [3]. It can be noted that, with dynamic randomization, at any given time $Y_{(j)}$, the $r_{1j}+r_{0j}$ subjects may not be statistically independent and that can raise questions on the validity of the above simple normal approximation.

3.2. Re-randomized test

A re-randomization test is performed as follow: we use the trial data

$$\{(u_i, G_i, H_i, X_i, Y_i, \delta_i), i = 1, 2, 3, \ldots\},$$

with the arrival times and subject covariates, and first calculate the log rank test statistics $Z = Z_{mh}$. Then, for each $\nu = 1,2,\ldots,N_{sim}$, where N_{sim} is a large number such as 10,000, we scramble the arrival times $u_i, i \geq 1$, use the prior treatment variables $\{(G_i,H_i), i = 1,2,3,\ldots\}$ and the dynamic randomization scheme described in Section 2 to generate the new treatment codes $\tilde{X}_{\nu,t}, t > 0$. Then, based on the re-randomized data $\{(\tilde{X}_i, Y_i, \delta_i), i = 1, 2, 3, \ldots\}$, we can calculate the "log rank" statistics Z_ν. The null hypothesis will be tested based on the observed Z and the empirical distribution of $Z_\nu, 1 \leq \nu \leq N_{sim}$. See Ref. [15] for details.

3.3. MPP based log rank test

For the proposed procedure, we use the same trial data $\{(u_i,G_i,H_i,X_i,Y_i,\delta_i), i = 1,2,3,\ldots\}$. We first define a counting measure, $p(.)$, on the combined space $[0,\mathcal{T}] \times [0,\mathcal{T}] \times R_1 \times R_2$ such that for any event times $A \subset [0,\mathcal{T}]$, entry times $B \subset [0,\mathcal{T}]$, covariates $C \subset R_1$, and outcomes $D \subset R_2$

$$\begin{aligned} &p(A \times B \times C \times D) \\ &= \sum_{i \geq 1} 1(u_i + Y_i \in A, u_i \in B, (G_i, H_i, X_i) \in C, \delta_i \in D) \\ &\equiv \int 1_{A \times B \times C \times D}(s,u,c,\delta) p(dsdudcd\delta), \end{aligned}$$

where the covariate space $R_1 = \{1,\ldots,L_G\} \times \{1,\ldots,L_H\} \times \{0,1\}$, 0 and 1 refer to the control and the treatment group. We also denote the outcome space $R_2 = \{0,1\}$, where 0 and 1 refer to the censored outcome and event respectively.

The counting measure $p(.)$ includes all information of the trial data. Most useful statistics can be written as an integral with respect to this counting measure. To avoid distraction from technical detail, we will state the main results in this section and leave most details in the appendix.

For any treatment group $x = 0,1$, we use the analog of the $r_{x,j}$, $r_{1j}+r_{0j}$ and $\frac{r_{1j}}{r_{1j}+r_{0j}}$ from the M-H log rank test. To make it more general like the weighted log rank tests [7,23]; and [12], we consider any (random) weight function $k_n(t,u,w) \to k(u,w)$ uniformly in probability, where $k_n(t,u,w)$ is $\mathcal{F}_{t-}$ measurable for each t and with uniformly bounded variation in w. Let

$$\begin{aligned} &N_x(t,w,k_n) = \int_0^{t-w} k_n(t,u,w) 1(X_u = x, Y_u \geq w) dR_u, \\ &N(t,w,k_n) = N_0(t,w,k_n) + N_1(t,w,k_n), \\ &\tilde{x}(t,w,k_n;\omega) = \frac{N_1(t,w,k_n)}{N(t,w,k_n)}, \end{aligned} \tag{2}$$

where the variable w corresponds to the event time $Y_{(j)}$ and the extra variable t indicates that only the information up to the calendar time t is used in the calculation. It is easy to see that they are all $\mathcal{F}_t$ measurable. It can be noted that, when $k_n(t,u,w)$ is independent of u, $\tilde{x}$ is independent of k_n.

Denote

$$\begin{aligned} &g_n(t,s,u,c;\omega) = k_n(t,u,s-u)[1(X_u = 1) - \tilde{x}(t,s-u,k_n;\omega)], \\ &\bar{g}(s,u,c;\omega) = k(u,s-u)[1(X_u = 1) - \bar{x}(s-u)], \end{aligned} \tag{3}$$

where $\bar{x}(w) = \lim_n \tilde{x}(t,s-u,k_n;\omega)$. Then, the statistics

$$U_n(t) = \sum_{u_i + Y_i \leq t} g_n(t, u_i + Y_i, u_i, G_i, H_i, X_i; \omega) 1(\delta_i = 1), \tag{4}$$

will be an analog of the usual weighted log rank statistics at the calendar time t. Note that, with the counting measure $p(.)$, we can write $U_n(t)$ as an integral

$$U_n(t) = \int_0^t \int_{\mathcal{X}} g_n(t,s,u,c;\omega) 1(\delta = 1) p(dsdudcd\delta).$$

Its variance estimator can be written as

$$V_n(t) = \int_0^t \int_{\mathcal{X}} g_n^2(t,s,u,c;\omega) 1(\delta = 1) p(dsdudcd\delta). \tag{5}$$

Under the assumption of proportional hazards, we assume that $h_{1,1}(w) = r h_{0,1}(w)$ for some constant $r > 0$ and all $w \geq 0$. Denote weighted cumulative events

$$\begin{aligned} &D_x(t) \\ &= \int_0^t \int_{\mathcal{X}} k_n(t,u,s-u) \frac{N_{1-x}(t,s-u,k_n)}{N(t,s-u,k_n)} 1(X_u = x, \delta = 1) p(dsdudcd\delta), \\ &\text{for } x = 0,1. \end{aligned} \tag{6}$$

We use the estimating equation (see Appendix 2 for details)

$$U_n(t) - \left[\frac{r-1}{2} D_0(t) + \frac{1}{2}\left(1 - \frac{1}{r}\right) D_1(t)\right] = 0, \tag{7}$$

to solve for the estimator of the hazard ratio r, which is denoted as $\hat{r}_n(t)$. The variance of $\hat{r}_n(t)$ can be estimated as

$$var(\hat{r}_n(t)) \approx \frac{4V_n(t)}{\left[D_0(t) + \frac{D_1(t)}{\hat{r}_n(t)^2}\right]^2}. \tag{8}$$

Finally, for testing the hypothesis of $h_{0,1}(.) = h_{1,1}(.)$, the statistics

$$Z_n(t) = \frac{U_n(t)}{\sqrt{V_n(t)}}, \tag{9}$$

converges to $N(0,1)$ as $n \to \infty$ for any large $t > 0$.

4. Simulations

In this section, we use simulation analyses to evaluate the performance of the procedures described in section 3. First, we describe the simulation of each data component as follow:

Subject Arrival: We use a Poisson process to model subject arrival. We generate independent random numbers $A_k, k = 1,2,3,..,N$ from an exponential distribution with rate $\lambda_a > 0$. Then, let

$$u_i = \sum_{k=1}^{i} A_k \tag{10}$$

be the time at which the i-th subject arrives. For each i, we generate G_i based on a distribution such that $P(G_i = g) = p_g$ for some positive constants $p_g, 1 \le g \le L_G$ and $\sum_g p_g = 1$. We also generate H_i based on another distribution such that $P(H_i = h) = q_h$ for some positive constants $q_h, 1 \le h \le L_H$ and $\sum_h q_h = 1$. In the examples below, we assume the number of subjects $N = 250$, $\lambda_a = 8$, $L_G = 8$ and $p_g, 1 \le g \le L_G$ as 1/14,1/14,1/14,2/14,2/14,2/14,2/14,3/14, $L_H = 4$ and $q_h, 1 \le h \le L_H$ as 1/6,1/6,2/6,2/6.

Treatment Assignment: After generating all subject arrival and baseline information (G and H), we use the dynamic randomization algorithm to generate the treatment group $X_i, i \ge 1$. In the examples below, we assume that $\rho = 0.5$ for 2:1 randomization, $a = b = c = 1$, and $\theta = 0.9$ as specified in the Appendix 1.

Failure Times: After generating all subject treatment assignments, we generate the failure times $T_i, i \ge 1$. Given X_i, T_i will be generated based on exponential distribution with the rate $\exp\{\alpha_g G_i + \alpha_h H_i + \beta X_i\}\lambda_0$ with baseline hazard $\lambda_0 > 0$ and treatment effect β as well as prognostic factors G and H effects (α_g and α_h). Here, $\beta < 0$ refers to positive treatment effect and non-zero α_g and α_h refer to heterogeneous populations. In the examples below, we assume that the baseline hazard $\lambda_0 = 0.15$, subgroup effects $\alpha_g = 0.01$ and $\alpha_h = 0.01$. Different treatment effect β will be specified later.

Censoring Times: The censoring time C_i will be independently generated based on uniform distribution over $[c_1, c_2]$ for some constants $c_2 > c_1 > 0$. In the examples below, we assume that $c_1 = 7$ and $c_2 = 8$.Taking $Y_i = \min\{T_i, C_i\}$ and $\delta_i = 1(T_i \le C_i)$, we complete the data simulation for $\{(U_i, G_i, H_i, X_i, Y_i, \delta_i), i = 1,2,3,\ldots\}$. For each simulation data, the hazard estimate and its variance will be calculated based on (7) and (8) and the test statistics of (9) and (1) will be calculated along with corresponding p-values p_1 and p_2 respectively. Here, we assume $k_n = 1$.

In addition, the re-randomized test under the null hypothesis that there is no treatment difference will be performed as described in section 3 for both (9) and (1), which are denoted by Refs. $p_{1,r}$ and $p_{2,r}$ respectively. In addition, the hazard ratio between the treatment (1) and control (0) will be estimated as in section 3. In the examples below, we use $N_{sim} = 10{,}000$. The simulation will be repeated for $N_{itr} = 10{,}000$ times and the corresponding calculated p-values are denoted as $p_1^k, p_2^k, p_{1,r}^k, p_{2,r}^k, k = 1, 2, \ldots, N_{itr}$. Let $\alpha = 0.025$, the empirical power will be calculated as

$$\begin{aligned} B_1 &= \frac{1}{N_{itr}} \sum_{k=1}^{N_{itr}} 1\left(p_1^k \le \alpha\right), \\ B_{1,r} &= \frac{1}{N_{itr}} \sum_{k=1}^{N_{itr}} 1\left(p_{1,r}^k \le \alpha\right), \\ B_2 &= \frac{1}{N_{itr}} \sum_{k=1}^{N_{itr}} 1\left(p_2^k \le \alpha\right), \\ B_{2,r} &= \frac{1}{N_{itr}} \sum_{k=1}^{N_{itr}} 1\left(p_{2,r}^k \le \alpha\right). \end{aligned} \tag{11}$$

4.1. Simulations under null hypothesis: $\beta = 0$

In this section, we assume there is no treatment effect in the failure time model, i.e., $\beta = 0$, and evaluate the performance of four testing procedures. Fig. 1 shows the histograms of $N_{itr} = 10{,}000$ p-values from each of four tests and they suggest reasonable resembling of uniform distributions over [0,1]. Based on one-sample Kolmogorov-Smirnov test, the p-values for testing goodness of fit with the uniform distribution are $p_1^u = 0.5945$, $p_{1,r}^u = 0.7442$, $p_2^u = 0.6617$, $p_{2,r}^u = 0.7442$ respectively. Correspondingly, the empirical powers (the actual one sided type I error in this case) based on (4) are 0.0258, 0.0279, 0.0289, and 0.0286 respectively. The mean number of events generated from 10,000 data sets is 176. Overall, all four test procedures appear acceptable in maintaining the designed type I error rate. The MPP based test keeps the type I error rate slightly better than both Mantle-Hanzel test and its re-randomized version.

4.2. Simulations under alternative hypothesis: $\beta \neq 0$

Suppose there is positive treatment effect in the failure time model, i.e., $\beta < 0$. Table 1 below shows the empirical power based on (4) for $\beta = -0.25, -0.5$ and -0.65 ($\beta = 0$ included as well).

Overall, the power of the four testing procedures appear comparable. Corresponding to slightly inflating type I error rate in the naive Mantle-Hanzel test, the MPP based test shows slightly lower power in detecting the nonzero treatment effect. In addition, the hazard ratio estimate based on (7) are close to the true parameters and their variances based on (8) are reasonable.

It can be noted that, in these simulations, we choose a large $N_{itr} = 10{,}000$ for reliability, which however leads to extensive computing time. We have tried additional limited simulation in case of smaller sample size, different censoring time interval, and differentiated subgroup effect and those analyses do not change the findings here.

5. Application

In this section, we will illustrate the above inference procedures through their application to a breast cancer trial conducted by NCIC Clinical Trials Group [19].

The trial used the minimization procedure to randomize 305 subjects with stratification factors of prior cytotoxic treatment, registration to MA. 16, presence of visceral disease, and study site. As result of the dynamic randomization, 153 were assigned to DPPE/DOX and 152 to DOX treatment groups and the treatment groups were well balanced by the stratification factors included in the randomization. Survival was a secondary endpoint of this trial. At the end of the study, 67/153 died in the DPPE/DOX group, compared with 91/152 died in the DOX group. Based on the ordinary Cox regression, the hazard ratio estimate for survival was 0.656 with standard error of 0.161. The hazard ratio estimate based on (7) with $k_n = 1$ was 0.657 with the variance estimate of 0.011. We see two point estimates were close. For the comparison of overall survival between two groups, the p-values from Mantel-Hanzle log rank test and its re-randomized test are 0.0085 and 0.0079 respectively. Based on the proposed procedure of section 3.2, the p-value is 0.0093 and its re-randomized version was 0.0079. All of them suggest favorable overall survival in the DPPE/DOX treatment arm.

6. Discussions

When implemented properly, a dynamic randomization can effectively balance many prognostic factors in controlled clinical trials [25]. The FDA guidance (E9) recommends analysis to be adjusting for factors used for the dynamic randomization but no specific methodology has been proposed. There have been substantial questions and challenges on the validity of conventional statistics without adjusting for the randomization process (see Refs. [20]), which can be part of the reasons for CPMP guidance to

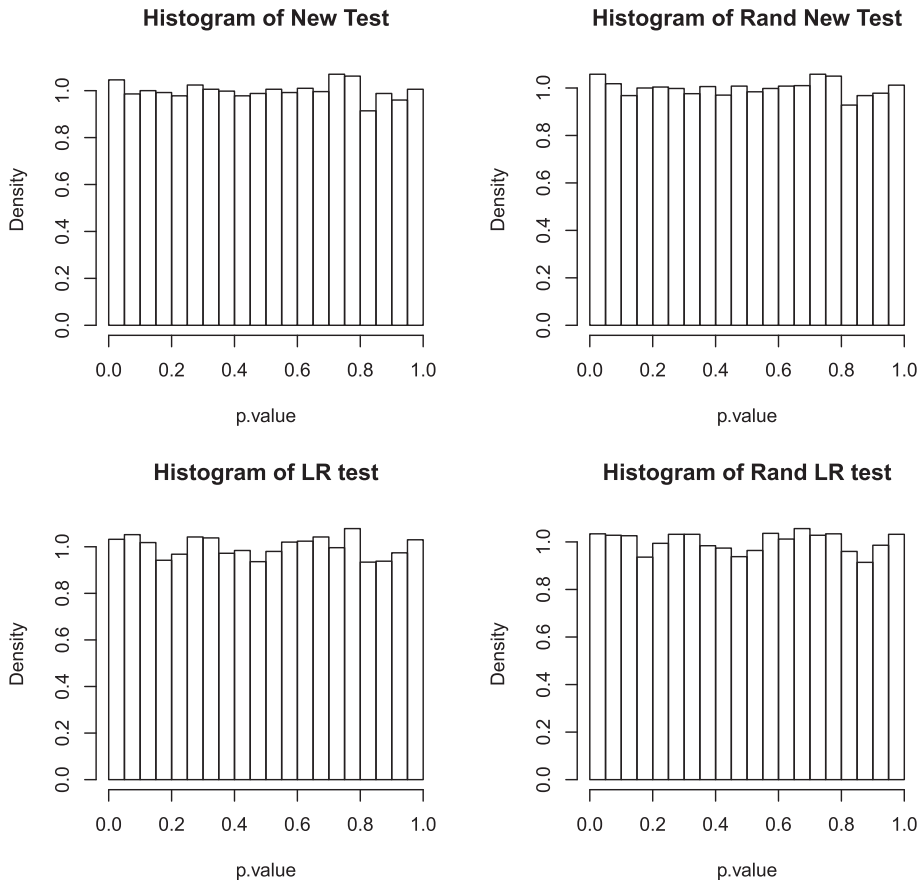

Fig. 1. P-values histograms under null hypothesis. Upper panel: MPP based procedure and its randomized version; lower panel: log rank based procedure and its randomized version.

discourage the use of such procedure. One of the FDA advisory meetings discussed the study AGLU02704 (LOTS) that used a dynamic randomization scheme to assign patients with Pompe disease to the experimental "2000L" and Placebo groups and to compare the rate change of 6-min walk distance (see Ref. [14]. The sponsor assessed the significance of the difference without considering dynamic randomization nature and obtained a p-value 0.035. However, the FDA statistician obtained a p-value 0.06 based on a re-randomization test. The intensity of the discussion around this topic highlights potentially serious impact of this controversy.

In this paper, we introduce a statistics that aligns with the stratification factors as well as the randomization process for clinical trials with time to event endpoints. It provides an estimate of the treatment effect that accounts for the randomization factors and the process, which can be used to test whether the effect is zero. In addition, the procedure is in a closed form of the observed data and thus easy to compute. We provide theoretical justification of its asymptotic distribution. We use simulation analyses to show its adequacy under moderate sample size. However, it can be noted that we do not intend to compare for the efficiency between the proposed procedure and common procedure such as the log rank test, since it is not expected that our proposed procedure would be nominally improved. We believe, as demonstrated by many simulation studies, that the log rank test is generally sufficient in controlling type I error and maintaining power, although regulatory agency has been raising concern that the log rank test does not reflect the dynamic randomization process and that its validity is unknown. Our proposed testing procedure accommodates the dynamic randomization process and can be used as a replacement for the log rank test when dynamic randomization becomes a concern. Given the additional variation being accounted for in the proposed testing procedure, it is expected that the nominal power may be slightly lower than the log rank test.

It can be noted that there is considerable similarity between the modeling framework used here and the covariate-adjusted response-adaptive design (CARA) studied by Hu and Rosenberger [9]; Zhang et al. [27]; and Hu, Zhang and He [10]. The CARA framework focuses on a target allocation with endpoints without delay. The model discussed in this paper deals with survival endpoint and dynamic randomization with high dimensional and potentially sparse stratification factors, which may not fit well in the CARA framework. In principle, CARA framework can be seen as a specific discrete case of the adaptive design framework described by Luo et al. [12]. Finally, it is interesting to note that, while there have been concerns on whether conventional procedures such as standard log rank test would be still valid with complicated dynamic randomization, similar theoretical justification may be modified via conditional argument to provide similar asymptotic result. However, such conditional argument may not work under more general adaptive design when the treatment assignment may depend on interim trial outcome such as CARA model.

Table 1
Empirical power and point estimate.

Power	$\exp\{\beta\}$			
	1.0	0.779	0.607	0.522
B_1	0.0258	0.1851	0.7003	0.904
$B_{1,r}$	0.0279	0.1898	0.7048	0.904
B_2	0.0289	0.2423	0.7623	0.93
$B_{2,r}$	0.0286	0.24	0.76	0.93
HR	1.0002	0.7793	0.604	0.52
HR Variance	0.027	0.018	0.013	0.01

Appendix 1. Dynamic randomization based on balance score

One common approach is to use balance scores defined as follow: Suppose the target randomization ratio for the treatment to the control is 1:ρ for some constant $\rho > 0$. The balance scores for the experimental arm assignment are

$$
\begin{aligned}
m_X(t;1) &= \frac{1}{1+R_{t-}}\Big\{\max\Big\{\rho(1+n_X(t-;1)), n_X(t-;0)\Big\} \\
&\quad -\min\{\rho(1+n_X(t-;1)), n_X(t-;0)\}\}, \\
m_{X,G}(t;1,g) &= \frac{1}{1+n_G(t-;g)}\Big\{\max\Big\{\rho\big(1+n_{X,G}(t-;1,g)\big), n_{X,G}(t-;0,g)\Big\} \\
&\quad -\min\{\rho\big(1+n_{X,G}(t-;1,g)\big), n_{X,G}(t-;0,g)\}\}, \\
m_{X,H}(t;1,h) &= \frac{1}{1+n_H(t-;h)}\Big\{\max\Big\{\rho\big(1+n_{X,H}(t-;1,h)\big), n_{X,H}(t-;0,h)\Big\} \\
&\quad -\min\{\rho\big(1+n_{X,H}(t-;1,h)\big), n_{X,H}(t-;0,h)\}\}
\end{aligned}
$$

and the balance scores for the control arm assignment as

$$
\begin{aligned}
m_X(t;0) &= \frac{1}{1+R_{t-}}\Big\{\max\Big\{\rho n_X(t-;1), (1+n_X(t-;0))\Big\} \\
&\quad -\min\{\rho n_X(t-;1), (1+n_X(t-;0))\}\}, \\
m_{X,G}(t;0,g) &= \frac{1}{1+n_G(t-;g)}\Big\{\max\Big\{\rho n_{X,G}(t-;1,g), \big(1+n_{X,G}(t-;0,g)\big)\Big\} \\
&\quad -\min\{\rho n_{X,G}(t-;1,g), \big(1+n_{X,G}(t-;0,g)\big)\}\}, \\
m_{X,H}(t,0,h) &= \frac{1}{1+n_H(t-;h)}\Big\{\max\Big\{\rho n_{X,H}(t-;1,h), \big(1+n_{X,H}(t-;0,h)\big)\Big\} \\
&\quad -\min\{\rho n_{X,H}(t-;1,h), \big(1+n_{X,H}(t-;0,h)\big)\}\}
\end{aligned}
$$

For any weights a,b and c, let the total balancing score be

$$\phi(t;x,g,h) = am_X(t;x) + bm_{X,G}(t;x,g) + cm_{X,H}(t;x,h), x = 0,1,$$

We can assign the treatment for a new subject accrued at the calendar time t with baseline $G=g$ and $H=h$ by

$$
\begin{aligned}
X_t = \xi_t 1(\phi(t;1,g,h) - \phi(t;0,g,h) < 0) + (1-\xi_t)1(\phi(t;1,g,h) \\
- \phi(t;0,g,h) \\
\geq 0),
\end{aligned}
$$

where the Bernoulli distributed $\xi_t \sim b(1,\theta)$ is an independent random variable. This dynamic randomization scheme will be used in the simulation examples of section 4.

Appendix 2. Technical details

Denote $\mathcal{F}_t$ as the information flow available up to time t, we have the compensator $q(dsdudcd\delta) = q_s(dudcd\delta)ds$ for the random measure $q_s(.)$ for each $s>0$ such that

$$
q_s(dudcd\delta) = \begin{cases} 1(Y_u \geq s-u, G_u = g, H_u = h, X_u = x)h_{x,1}(s-u)dR_u, \\ \text{if } s \geq u, c = (g,h,x), \delta = 1; \\ 1(Y_u \geq s-u, G_u = g, H_u = h, X_u = x)h_0(s-u)dR_u, \\ \text{if } s \geq u, c = (g,h,x), \delta = 0; \\ 0, \text{ if o.w.} \end{cases}
$$

As noted in Luo et al. [12]; the compensator $q(dsdudcd\delta)$ is mainly used for theoretical development. The actual computable statistics involves with an integral with respect to the counting measure $p(dsdudcd\delta)$, e.g.,

$$
\int_0^t \int_{\mathcal{X}} g(t,s,u,c;\omega)p(dsdudcd\delta) = \sum_{u_i+Y_i \leq t} g(t,u_i + Y_i, u_i, G_i, H_i, X_i; \omega)
$$

for any function $g(t,s,u,c;\omega)$.

Note that

$$
\begin{aligned}
&U_n(t) \\
&= \int_0^t \int_{\mathcal{X}} g_n(t,s,u,c;\omega)1(\delta=1)p(dsdudcd\delta) \\
&= \int_0^t \int_{\mathcal{X}} g_n(t,s,u,c;\omega)1(\delta=1)[p(dsdudcd\delta) - q(dsdudcd\delta)] \\
&\quad + \int_0^t \int_{\mathcal{X}} g_n(t,s,u,c;\omega)1(\delta=1)q(dsdudcd\delta)
\end{aligned}
$$

and the second term can be written as

$$
\begin{aligned}
&\int_0^t \int_{\mathcal{X}} g_n(t,s,u,c;\omega)1(\delta=1)q(dsdudcd\delta)\\
&=\sum_x \int_0^t \int_0^s g_n(t,s,u,c;\omega)1(Y_u\geq s-u,X_u=x)h_{x,1}(s-u)dR_u ds\\
&=\sum_x \int_0^t \int_0^{t-w} k_n(t,u,w)[1(X_u=1)-\tilde{x}(t,w,k_n;\omega)]1(Y_u\geq w,X_u=x)h_{x,1}(w)dR_u dw\\
&=\int_0^t h_{1,1}(w)\int_0^{t-w} k_n(t,u,w)1(Y_u\geq w,X_u=1)dR_u dw\\
&\quad-\int_0^t \tilde{x}(t,w,k_n;\omega)\sum_x h_{x,1}(w)\int_0^{t-w} k_n(t,u,w)1(Y_u\geq w,X_u=x)dR_u dw\\
&=\int_0^t \left[h_{1,1}(w)N_1(t,w,k_n)-\tilde{x}(t,w,k_n;\omega)\sum_x h_{x,1}(w)N_x(t,w,k_n)\right]dw\\
&=\int_0^t \frac{N_0(t,w,k_n)N_1(t,w,k_n)}{N(t,w,k_n)}\left(h_{1,1}(w)-h_{0,1}(w)\right)dw.
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
U_n(t)=&\int_0^t \int_{\mathcal{X}} g_n(t,s,u,c;\omega)1(\delta=1)dM_s\\
&+\int_0^t \frac{N_0(t,w,k_n)N_1(t,w,k_n)}{N(t,w,k_n)}\left(h_{1,1}(w)-h_{0,1}(w)\right)dw,
\end{aligned}
\tag{12}
$$

where $dM_s = p(dsdudcd\delta) - q(dsdudcd\delta)$.

Estimation of Treatment Effect

We can introduce a procedure to estimate the treatment effect under proportional hazards with a special weight function. We assume that $h_{1,1}(w) = rh_{0,1}(w)$ for some constant $r>0$ and all $w\geq 0$. We can simplify the term at the end of the last paragraph as

$$
\begin{aligned}
&\int_0^t \frac{N_0(t,w,k_n)N_1(t,w,k_n)}{N(t,w,k_n)}\left(h_{1,1}(w)-h_{0,1}(w)\right)dw\\
&=(r-1)\int_0^t \frac{N_0(t,w,k_n)N_1(t,w,k_n)}{N(t,w,k_n)}h_{0,1}(w)dw\\
&=(r-1)\int_0^t \frac{N_1(t,w,k_n)}{N(t,w,k_n)}\int_0^{t-w} k_n(t,u,w)1(Y_u\geq w,X_u=0)h_{0,1}(w)dR_u dw\\
&=(r-1)\int_0^t \int_0^s k_n(t,u,s-u)\frac{N_1(t,s-u,k_n)}{N(t,s-u,k_n)}1(Y_u\geq s-u,X_u=0)h_{0,1}(s-u)dR_u ds\\
&=(r-1)\int_0^t \int_{\mathcal{X}} k_n(t,u,s-u)\frac{N_1(t,s-u,k_n)}{N(t,s-u,k_n)}1(X_u=0,\delta=1)q(dsdudcd\delta)
\end{aligned}
$$

Similarly, noting that $h_{0,1}(w) = h_{1,1}(w)/r$

$$\int_0^t \frac{N_0(t,w)N_1(t,w)}{N(t,w)}\left(h_{1,1}(w) - h_{0,1}(w)\right)dw$$

$$= \left(1-\frac{1}{r}\right)\int_0^t \int_{\mathcal{X}} k_n(t,u,s-u)\frac{N_0(t,s-u,k_n)}{N(t,s-u,k_n)}1(X_u = 1, \delta = 1)q(dsdudcd\delta)$$

Denote

$$D_x(t) = \int_0^t \int_{\mathcal{X}} k_n(t,u,s-u)\frac{N_{1-x}(t,s-u,k_n)}{N(t,s-u,k_n)}1(X_u = x, \delta = 1)p(dsdudcd\delta),$$

for $x = 0, 1$.

We have

$$U_n(t) - \left[\frac{r-1}{2}D_0(t) + \frac{1}{2}\left(1-\frac{1}{r}\right)D_1(t)\right] \approx \int_0^t \int_{\mathcal{X}} g_n(t,s,u,c;\omega)1(\delta = 1)dM_s,$$

and can use the estimating equation

$$E(r;t) = 0,$$

where

$$E(r;t) = U_n(t) - \left[\frac{r-1}{2}D_0(t) + \frac{1}{2}\left(1-\frac{1}{r}\right)D_1(t)\right],$$

to solve for the estimator of the hazard ratio r, which is denoted as $\widehat{r}_n(t)$. It can be noted that the weights $\left(\frac{1}{2},\frac{1}{2}\right)$ can be replaced by any $(\lambda,1-\lambda)$, while we keep it simple by taking $\lambda = \frac{1}{2}$.Note that $\frac{dE}{dr} = -\frac{1}{2}\left(D_0(t) + \frac{D_1(t)}{r^2}\right)$. Applying the argument of M-estimation and delta method, we have

$$\int_0^t \int_{\mathcal{X}} g_n(t,s,u,c;\omega)1(\delta = 1)dM_s = E(r;t) - E(\widehat{r}_n(t);t) \approx \frac{1}{2}\left(D_0(t) + \frac{D_1(t)}{\widehat{r}_n(t)^2}\right)(\widehat{r}_n(t) - r)$$

The variance of $\widehat{r}_n(t)$ can be estimated as

$$var(\widehat{r}_n(t)) \approx \frac{4V_n(t)}{\left[D_0(t) + \frac{D_1(t)}{\widehat{r}_n(t)^2}\right]^2}$$

where

$$V_n(t) = \int_0^t \int_{\mathcal{X}} g_n^2(t,s,u,c;\omega)1(\delta = 1)p(dsdudcd\delta).$$

Comparison of Group Effects

In this section, we introduce a testing procedure to compare the survival time between two treatment groups. Note that the second term of (12) becomes 0 under the hypothesis of $h_{0,1}(.) = h_{1,1}(.)$ and thus

$$U_n(t) = \int_0^t \int_{\mathcal{X}} g_n(t,s,u,c;\omega)1(\delta = 1)dM_s$$

which, from Theorem 3.4 in Luo et al. [12]; can be approximated by

$$Q(t) = \int_0^t \int_{\mathcal{X}} \overline{g}(s,u,c;\omega)1(\delta = 1)dM_s$$

where $dM_s = p(dsdudcd\delta) - q(dsdudcd\delta)$ and $\overline{g}$ is defined in (3.3). Thus, it converges to a Gaussian process with zero mean and quadratic variation process

$$< Q > (t) = \int_0^t \int_{\mathcal{X}} \overline{g}(s,u,c;\omega)^2 1(\delta = 1)q(dsdudcd\delta).$$

In application, $< Q > (t)$ can be approximated uniformly by Ref. $V_n(t)$. Thus, the testing statistics takes the form of

$$Z_n(t) = \frac{U_n(t)}{\sqrt{V_n(t)}},$$

converges to $N(0,1)$ as $n \to \infty$ for any large $t > 0$.

References

[1] P. Armitage, Fisher, Bradford Hill, and randomization, Int. J. Epidemiol. 32 (2003) 925–928.

[2] M. Buyse, D. McEntegart, Achieving balance in clinical trials: an unbalanced view from EU regulators, Appl. Clin. Trials 13 (2004) 36–40.
[3] Christopher Jennison, Bruce W. Turnbull, Group Sequential Methods with Applications to Clinical Trials, Chapman & Hall/CRC, Boca Raton, 1999.
[4] CPMP. Points to Consider on adjustment for baseline covariates.
[5] S. Day, J. Grouin, J.A. Lewis, Achieving balance in clinical trials, Appl. Clin. Trials 14 (2005) 24–26.
[6] R.A. Fisher, The arrangement of field experiments, J Minist. Agric. G. B. 33 (1926) 503–513.
[7] T.R. Fleming, D.P. Harrington, A class of hypothesis tests for one and two sample censored survival data, Commun. Statistics 10 (1981) 763–794.
[8] A.B. Hill, The Clinical trial, New Engl. J. Med. 247 (1952) 114–119.
[9] F. Hu, W.F. Rosenberger, The Theory of Response-adaptive Randomization in Clinical Trials, John Wiley and Sons, 2006 (Wiley Series in Probability and Statistics).
[10] F. Hu, L.X. Zhang, X. He, Efficient randomized adaptive designs, Ann. Statistics. 37 (2009) 2543–2560.
[11] ICH E9, Statistical Principles for Clinical Trials, Consensus Guideline, 5 September 1998.
[12] X. Luo, G. Xu, Z. Ying, Sequential analysis of the cox model under response dependent allocation, Stat. Sin. 23 (2013) 1761–1774.
[13] S.J. Pocock, R. Simon, Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial, Biometrics 31 (1975) 103–115.
[14] Endocrinologic and Metabolic Drugs Advisory Committee Meeting, October 21, 2008, http://www.fda.gov/ohrms/dockets/ac/08/slides/2008-4389s1-00-index.htm.
[15] P.A. Flyer, A comparison of conditional and unconditional randomization tests for highly stratified designs, Biometrics 54 (No 4) (1998) 1551–1559.
[16] Sara A. Hurvitz, Luc Dirix, Judit Kocsis, Giulia V. Bianchi, Janice Lu, Jeferson Vinholes, Ellie Guardino, Chunyan Song, Barbara Tong, Vivian Ng, Yu-Waye Chu, Edith A. Perez, Phase II randomized study of trastuzumab emtansine versus trastuzumab plus docetaxel in patients with human epidermal growth factor receptor 2positive metastatic breast cancer, J. Clin. Oncol. 31 (2013) 1157–1163.
[17] W. Ma, F. Hu, L. Zhang, Testing hypotheses of covariate-adaptive randomized clinical trials, J. Am. Stat. Assoc. 110 (2015) 669–680.
[18] Nathan Mantel, Evaluation of survival data and two new rank order statistics arising in its consideration, Cancer Chemother. Rep. 50 (3) (1966) 16370.
[19] Leonard Reyno, Lesley Seymour, Dongsheng Tu, Susan Dent, Karen Gelmon, Barbara Walley, Anna Pluzanska, Vera Gorbunova, Avgust Garin, Jacek Jassem, Tadeusz Pienkowski, Janet Dancey, Laura Pearce, Mary MacNeil, Susan Marlin, David Lebwohl, Maurizio Voi, Kathleen Pritchard, Phase III study of N,N-Diethyl-2-[4-(Phenylmethyl) Phenoxy]Ethanamine (BMS-217380-01) combined with doxorubicin versus doxorubicin alone in metastatic/recurrent breast Cancer: National Cancer Institute of Canada clinical trials group study MA.19, J. Clin. Oncol. 22 (No 2) (2004) 269–276 (January 15), 2004.
[20] K.C. Rose, Dynamic allocation as a balancing act, Pharm. Stat. 3 (3) (2004) 187–191.
[21] J. Shao, X. Yu, Validity of tests under covariate-adaptive biased coin randomization and generalized linear models, Biometrics 69 (2013) 960969.
[22] J. Shao, X. Yu, B. Zhong, A theory for testing hypotheses under covariate-adaptive randomization,, Biometrika 97 (2010) 347360.
[23] R.E. Tarone, J.H. Ware, On distribution-free tests for equality for survival distributions, Biometrika 77 (1977) 147–160.
[24] D.R. Taves, Minimization: a new method of assigning patients to treatment and control group, Clin. Pharmacol. Ther. 15 (1974) 443453.
[25] D. Tu, Minimization procedure, in: Encyclopedia of Biopharmaceutical Statistics, third ed., 2010, pp. 795–798.
[26] M. Zelen, The randomization and stratification of subjects to clinical trials, J. Chronic Dis. 27 (1974) 365–375.
[27] L.X. Zhang, F. Hu, S.H. Cheung, W.S. Chan, Asymptotic properties of covariate-adjusted adaptive designs, Ann. Statistics 35 (2007) 1166–1182.
[28] T. Hasegawa, T.J. Tango, Permutation test following covariate-adaptive randomization in randomized controlled trials, Biopharm. Stat. 19 (1) (2009) 106–119, http://dx.doi.org/10.1080/10543400802527908.