Contemporary Clinical Trials Communications 1 (2015) 28–31

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

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

A novel design for randomized immuno-oncology clinical trials with potentially delayed treatment effects

Pei He [a, *], Zheng Su [b]

[a] Genentech Inc. South San Francisco, CA, 94080, USA
[b] Deerfield Institute, 780 Third Avenue, 37th Floor, New York, NY, 10017, USA

ARTICLE INFO

Article history:
Received 5 June 2015
Received in revised form
6 August 2015
Accepted 24 August 2015
Available online 10 November 2015

Keywords:
Clinical trial design
Immuno-oncology
Change point
Non-proportional hazards

ABSTRACT

The semi-parametric proportional hazards model is widely adopted in randomized clinical trials with time-to-event outcomes, and the log-rank test is frequently used to detect a potential treatment effect. Immuno-oncology therapies pose unique challenges to the design of a trial as the treatment effect may be delayed, which violates the proportional hazards assumption, and the log-rank test has been shown to markedly lose power under the non-proportional hazards setting. A novel design and analysis approach for immuno-oncology trials is proposed through a piecewise treatment effect function, which is capable of detecting a potentially delayed treatment effect. The number of events required for the trial will be determined to ensure sufficient power for both the overall log-rank test without a delayed effect and the test beyond the delayed period when such a delay exists. The existence of a treatment delay is determined by a likelihood ratio test with resampling. Numerical results show that the proposed design adequately controls the Type I error rate, has a minimal loss in power under the proportional hazards setting and is markedly more powerful than the log-rank test with a delayed treatment effect.

© 2015 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

The Cox semi-parametric proportional hazards model [1] is widely adopted in clinical trials with time-to-event outcomes to compare an experimental treatment with the best supportive care (BSC). A key assumption in the Cox model is that the hazard ratio is a constant over time, which may be violated as there can be a lag period before the experimental treatment starts to exhibit a beneficial effect. This is particularly the case for immuno-oncology therapies, with one example being ipilimumab, a fully human monoclonal antibody that blocks CTLA-4 to promote immunity. Ipilimumab demonstrated statistically significant improvements in overall survival (OS) in two Phase 3 randomized controlled trials in patients with metastatic melanoma. In both trials a delayed treatment effect of about 4 months was observed based on the Kaplan–Meier (K–M) curves for overall survival [2,3].

The log-rank test is frequently used to detect a potential treatment effect in randomized time-to-event trials, which has been shown to markedly lose power under the non-proportional hazards setting [4]. Weighted log-rank tests [e.g. Refs. [5–7]] have been proposed to account for a delayed separation of the K–M curves. One challenge with a weighted log-rank test at the design stage of a trial is the choice of the weight function, which determines the amount of weights assigned to observations at various times. In addition, when there is not a delayed treatment effect the log-rank test is the asymptotically most powerful nonparametric test under the proportional hazards setting [8].

The rest of the paper is structured as follows. In Section 2, we propose a two-stage design and analysis approach for immuno-oncology clinical trials. The number of events required for a trial will be determined to ensure sufficient power for both the overall log-rank test without a delayed treatment effect and the test of a treatment effect beyond the delayed period when such a delay exists. The existence of a treatment delay is determined by a likelihood ratio test with resampling [9]. Numerical results are given in Section 3, which show that the proposed design adequately controls the Type I error rate, has a minimal loss in power under the proportional hazards setting and is markedly more powerful than the log-rank test with a delayed treatment effect. Some discussions and concluding remarks are given in Section 4.

2. Methods

In this section, we first briefly review the sequential testing

* Corresponding author.
E-mail address: he.pei@gene.com (P. He).

http://dx.doi.org/10.1016/j.conctc.2015.08.003
2451-8654/© 2015 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

approach proposed by He et al. [9] for detecting potentially multiple change points in the proportional hazards model. We simplify their methodology by focusing on detecting only one change point, which will serve as the first step in the analysis of an immuno-oncology trial in our proposed approach.

Suppose there are a total of N patients recruited into an immuno-oncology trial with time-to-event endpoints. The proportional hazards model has the form $h_1(t) = e^{\beta} h_0(t)$, where $h_1(t)$ and $h_0(t)$ are the respective hazard functions for the treatment and BSC arms. Let $X_1, \ldots, X_N$ denote the independent and identically distributed survival times, $C_1, \ldots, C_N$ be the censoring times which are assumed to be independent of the survival times, and $Z_i = 1$ or $0, i = 1, \ldots, N$, for the treatment arm or BSC arm, respectively. Only the pairs $(T_i, \delta_i), i = 1, \ldots, N$ are observed, where $T_i = \min(X_i, C_i)$ and $\delta_i = I\{X_i \le C_i\}$.

The partial likelihood of the log hazard ratio β is

$$L(\beta) = \prod_{i=1}^{N} \left(\frac{e^{\beta Z_i}}{\sum_{j:T_j \ge T_i} e^{\beta Z_j}} \right)^{\delta_i}, \tag{1}$$

and the log partial likelihood of β is

$$l(\beta) = \sum_{i=1}^{N} \delta_i \left\{ \beta Z_i - \log \left(\sum_{j:T_j \ge T_i} e^{\beta Z_j} \right) \right\}. \tag{2}$$

The hypotheses are $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$. The Wald statistic has the form $(-\ddot{l}(\hat{\beta}))^{1/2}(\hat{\beta} - \beta_0)$, where $\hat{\beta}$ maximizes the log partial likelihood $l(\beta)$ and $\ddot{l}$ denotes the second derivative of l. The null hypothesis H_0 can be evaluated based on the asymptotic normality of. $(-\ddot{l}(\hat{\beta}))^{1/2}(\hat{\beta} - \beta_0)$.

The change point model proposed by He et al. [9] assumes the hazard ratio function $\beta(t)$ to be a step function as follows:

$$\beta(t) = \begin{cases} \beta_1 & 0 \le t \le \tau_1 \\ \beta_2 & \tau_1 < t \le \tau_2 \\ \vdots & \\ \beta_{k+1} & t > \tau_k, \end{cases}$$

where $0 = \tau_0 < \tau_1 < \ldots < \tau_{k+1} = \infty$ denote the change points. Here we use a simplified version of their model that assumes only one change point τ to be clinically meaningful. Therefore the hazard ratio function $\beta(t)$ has a simpler form:

$$\beta(t) = \begin{cases} \beta_1 & 0 \le t \le \tau \\ \beta_2 & t > \tau. \end{cases}$$

By replacing β with $\beta(t)$ in (2), we obtain a combined form of the log partial likelihood for β_1, β_2 and τ:

$$l(\beta_1, \beta_2, \tau) = \sum_{i=1}^{N} \delta_i \left\{ 1_{\{T_i \le \tau\}} \left(\beta_1 Z_i - \log \left(\sum_{s:T_s \ge T_i} e^{\beta_1 Z_s} \right) \right) + 1_{\{T_i > \tau\}} \left(\beta_2 Z_i - \log \left(\sum_{s:\ T_s \ge T_i} e^{\beta_2 Z_s} \right) \right) \right\}. \tag{3}$$

The corresponding likelihood ratio statistic LR_τ is as follows:

$$\mathrm{LR}_\tau = -2\log \frac{\sup_\beta l_{H_0}(\beta)}{\sup_{\beta_1, \beta_2, \tau_1} l_{H_1}(\beta_1, \beta_2, \tau_1)} \tag{4}$$

The following resampling procedure is adopted to estimate the distribution of LR_τ under the null hypothesis:

1. Calculate the Breslow [10] estimate of the baseline cumulative hazard function $\widehat{\Lambda}(t)$ and obtain the K–M estimate of the censoring distribution $\widehat{S}_c(t)$. The survival functions for the BSC arm and the experimental arm can be estimated by $\widehat{S}_{BSC}(t) = \exp\{-\widehat{\Lambda}(t)\}$ and $\widehat{S}_{exp}(t) = \widehat{S}_{BSC}^{\exp(\widehat{\beta})}$, respectively.
2. Generate a total of B (e.g. $B = 2000$) simulated trials with the survival functions $\widehat{S}_{BSC}(t)$ and $\widehat{S}_{exp}(t)$, which LR_τcorrespond to a true model of no change points, and the censoring distribution $\widehat{S}_c(t)$. Obtain the likelihood ratio statistics $\mathrm{LR}_\tau^b, b = 1, \ldots, B$ for each resampled trial.
3. Reject the null hypothesis if LR_τ, the likelihood ratio statistic calculated from the original trial, is larger than the $(1-\alpha) \times 100$ th percentile of $\{\mathrm{LR}_\tau^b, b = 1, \ldots, B\}$ where α controls the false discovery rate under the null hypothesis.

Intuitively, when analyzing an immuno-oncology trial a natural first step is to determine whether a change point in the hazard ratio function exists. If a change point is not detected, one should proceed with the standard log-rank test to determine whether a treatment effect exists. If a change point is detected, which implies a delayed treatment effect, one should assess whether there is a statistically significant treatment effect beyond the change point and also ensure that there is not a statistically significant and clinically meaningful effect favoring the control arm before the change point. In this two-stage analysis approach it is important to ensure that the overall Type I error rate is not inflated.

Given the desired properties above we propose the following two-stage analysis approach:

1. Apply the likelihood ratio and resampling approach of He et al. [9] to determine whether a change point in the hazard ratio function exists with false discovery rate α_1.
2. If a change point is not detected the standard log-rank test is used to assess whether a treatment effect exists with Type I error rate α_2. If a change point is detected, a log-rank test for the observations beyond the change point (or equivalently the score test based on the proportional hazards assumption) is conducted with type I error rate $\alpha_1 + \alpha_2 = \alpha$ to determine whether a treatment effect exists, where α is the desired overall type I error rate. When a treatment effect is detected the same test should be applied to the observations before the change point to ensure no early harm caused by the experimental treatment.

Theoretically, the proposed approach controls the overall Type I error rate, which is split between the two analysis steps. Intuitively, under the null hypothesis of no treatment effect α_1 controls the probability that the analysis will not use the log-rank test, and even if we conservatively assume that the null hypothesis is always rejected when a change point is incorrectly detected the overall Type I error rate is still controlled by $\alpha_1 + \alpha_2 = \alpha$. In practice, we may set $\alpha_1 = 1\%$ and $\alpha_2 = 4\%$ when the overall Type I error rate is set to be 5%. By having $\alpha_2 = 4\%$the proposed two-stage approach will have minimal power loss under a proportional hazards alternative. For example, a design with 90% power at the 5% Type I error level will have 88% power at the 4% Type I error level, and a loss of 2% in this setting as a tradeoff can translate to a substantial power gain under the non-proportional hazards setting, which is demonstrated in Section 3.

The proposed approach can be further simplified if there is a strong prior belief that the treatment will not cause early harm to patients. In this case, β_1 in the approach of He et al. [9] can be set to be 0. Intuitively, with this added assumption there is one less variable to estimate in the change point detection algorithm, which increases the power for detecting the change point when one exists

Table 1
Power analyses of the proposed design and Logrank design under various scenarios.

	τ	Exp (β_1)	Exp (β_2)	Proposed design	Logrank design
Null case	NA	1.0	1.0	0.048	0.049
Proportional hazard case	NA	0.75	0.75	0.886	0.903
Non-proportional case 1	4	1.0	0.5	0.994	0.975
Non-proportional case 2	5	1.0	0.5	0.978	0.890
Non-proportional case 3	6	1.0	0.5	0.937	0.738
Non-proportional case 4	7	1.0	0.5	0.878	0.576
Non-proportional case 5	8	1.0	0.5	0.712	0.321
Non-proportional case 6	4	0.9	0.5	0.998	0.996
Non-proportional case 7	5	0.9	0.5	0.992	0.977
Non-proportional case 8	6	0.9	0.5	0.986	0.938
Non-proportional case 9	7	0.9	0.5	0.955	0.876
Non-proportional case 10	8	0.9	0.5	0.872	0.741

and as a result increases the overall power of the proposed test.

When designing a trial with a potentially delayed treatment effect one should consider where a change point is likely to occur and what the treatment effect size is likely to be beyond the change point. The total number of events for a trial adopting the proposed two-stage design should provide sufficient events beyond the change point so that a treatment effect can be detected with sufficient power. This requires reasonable estimates of survival functions for both treatment arms based on data from early phase trials.

3. Numerical results

In this section we consider a similar trial design as in Ref. [11] to compare the performance of the proposed approach with that of the standard log-rank test. A randomized trial is designed using the conventional exponential distribution assumption, and 512 events are required to detect an overall hazard ratio of 0.75 between two treatment arms using a log-rank test with a two-sided Type I error rate of 5% and power of 90%. The accrual duration for 680 randomized patients is assumed to be 12 months, and the median survival for the control arm is assumed to be 6 months. The total number of 512 events is considered sufficient to ensure sufficient power for detecting a treatment effect beyond the change point (if one exists) as only 90 events are needed to ensure 90% power for a hazard ratio of 0.5, and a change point is expected to be in the range of 4–8 months.

Table 1 shows that the proposed design adequately controls the Type I error rate when the experimental treatment is not beneficial. Under the alternative of proportional hazards with hazard ratio of 0.75 the power of the proposed test is slightly less than that of the log-rank test. Under the alternative of a delayed treatment effect the proposed approach is markedly more powerful than the log-rank test for various locations of the change point in the hazard ratio function. For each scenario a total of 2000 trials were simulated to compare the performance of the two designs. Intuitively, for the proposed design to significantly outperform the log-rank test the algorithm by He et al. [9] needs to be able to accurately detect the change point with high probability, which requires sufficient number of events to be observed both before and after the change point. If the change point occurs very early in the hazard ratio function the proposed design may not provide a substantial advantage over the standard log-rank test. If the change point occurs late the long time interval with no treatment effect leads to reduced power for both approaches even though the power of the proposed design is relatively higher than that of the log-rank test.

4. Discussions and conclusions

When designing randomized clinical trials for immuno-oncology therapies the standard design based on proportional hazards assumption and log-rank test may not be optimal if there is a reasonable chance of a delayed treatment effect based on the mechanism of action and the available pre-clinical and clinical data to date. A two-stage design and analysis approach is proposed, which first tries to determine if a delayed effect exists. When such an effect is detected testing should be done for the time periods before and after the change point separately. If a delayed effect is not detected the proposed algorithm proceeds with the standard log-rank test. The tradeoff between a small loss in power under the proportional hazards setting and a marked gain when a delayed effect exists may be considered favorable for certain classes of therapies. When considering the two-stage approach extensive simulations should be conducted based on reasonable assumptions on the important trial parameters to determine whether it is indeed favorable over the standard design. The proposed approach may be most valuable when sufficient numbers of events are expected for the time periods before and after the change point. In this case, the standard log-rank test is expected to markedly lose power, and the proposed approach not only has a high likelihood of accurately detecting the change point but also provides sufficient sample size to characterize the treatment effect for both periods. The change point detection algorithm of He et al. [9] is capable of detecting multiple change points in the hazard ratio function, which may be utilized to develop a broad class of designs where treatment effect can be detected for at least one time interval with no harm demonstrated in the other time intervals. Finally, the proposed approach can be easily extended to group sequential trials with an alpha-spending function. In particular, one may spend a small portion (e.g. 20%) of the α to be spent at each interim on identifying a potential change point. Alternatively, since it's unlikely to detect a change point at the early interims given the immaturity of the K–M curves one may start to detect change points at the later interims or only at the final analysis.

References

[1] D.R. Cox, Regression models and life-tables, J. R. Stat. Soc. Ser. B Methodol. 34 (2) (1972) 187–220.
[2] F.S. Hodi, et al., Improved survival with ipilimumab in patients with metastatic melanoma, N. Engl. J. Med. 363 (8) (2010) 711–723.
[3] C. Robert, et al., Ipilimumab plus dacarbazine for previously untreated metastatic melanoma, N. Engl. J. Med. 364 (26) (2011) 2517–2526.
[4] D. Schoenfeld, The asymptotic properties of nonparametric tests for comparing survival distributions, Biometrika 68 (1981) 316–319.
[5] G.D. Fine, Consequences of delayed treatment effects on analysis of time-to-event endpoints, Drug Inf. J. 41 (2007) 535–539.
[6] D.P. Harrington, T.R. Fleming, A class of rank test procedures for censored survival data, Biometrika 69 (1982) 553–566.
[7] D.M. Zucker, E. Lakatos, Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment, Biometrika 77 (4) (1990) 853–864.
[8] T.R. Fleming, D.P. Harrington, Counting Processes and Survival Analysis, John

Wiley & Sons, New York, 1991.
[9] P. He, et al., A sequential testing approach to detecting multiple change points in the proportional hazards model, Stat. Med. 32 (7) (2013) 1239–1245.
[10] N.E. Breslow, Discussion of the paper by D. R. Cox, J. Roy. Stat. Soc. B 34 (1972) 216–217.
[11] T.T. Chen, Statistical issues and challenges in immuno-oncology, J. Immunother. Cancer 1 (2013) 18.