Contemporary Clinical Trials 32 (2011) 755–759

Contents lists available at ScienceDirect

Contemporary Clinical Trials

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

A hybrid approach to predicting events in clinical trials with time-to-event outcomes

Liang Fang, Zheng Su *

Genentech Inc., South San Francisco, CA 94080, USA

ARTICLE INFO

Article history:
Received 22 March 2011
Received in revised form 16 May 2011
Accepted 25 May 2011
Available online 30 May 2011

Keywords:
Clinical trial
Event prediction
Parametric
Non-parametric
Change point

ABSTRACT

In many clinical trials with time-to-event outcomes there are interim analyses planned at pre-specified event counts. It is of great value to predict when the pre-specified event milestones can be reached based on the available data as the timeline for a study is essential to the study sponsors and data monitoring committees for logistic planning purposes. Both parametric and non-parametric approaches exist in the literature for estimating the underlining survival function, based on which the predictions of future event times can be determined. The parametric approaches assume that the survival function is smooth, which is often not the case as the survival function usually has one or multiple change points and the hazard functions can differ significantly before and after a change point. The existing non-parametric method bases predictions on the Kaplan–Meier survival curve appended with a parametric tail to the largest observation, and all of the available data is used to estimate the parametric tail. This approach also requires a smooth survival function in order to achieve an accurate estimate of the tail distribution. In this article, we propose a hybrid parametric, non-parametric approach to predicting events in clinical trials with time-to-event outcomes. The change points in the survival function are first detected, and the survival function before the last change point is estimated non-parametrically and the tail distribution beyond the last change point is estimated parametrically. Numerical results show that the proposed approach provides accurate predictions for future event times and outperforms the existing approaches.

1. Background

In many clinical trials with time-to-event outcomes there are interim analyses planned at pre-specified event counts. It is of great value to predict when the pre-specified event milestones can be reached based on available data as the timeline for a study is essential to the study sponsors and data monitoring committees involved for logistic planning purposes. Both parametric and non-parametric approaches have been proposed in the literature to estimate the underlining survival function for the patient population under study, which is essential to the predictions of arrival times of future events. The parametric approaches are based on either the exponential model for the survival function as proposed by Bagiella and Heitjan [1] or the Weibull distribution as in Ying and Heitjan [2].

The parametric approaches assume that the underlining survival function is smooth and can be accurately estimated by a parametric function with 1 or 2 parameters, which are often not the case as the survival functions usually have one or multiple changes points and the hazard functions can differ significantly before and after a change point. For example, for hospitalized patients with a major cardiovascular adverse event (MACE) the risk of having another MACE is of the greatest during the weeks post the first event while the patient is still in the hospital. The risk will be lower between the release from the hospital and about one year post the event, and it will be the lowest beyond one year of the first event (see e.g. Ref. [3]). As a result, the survival function for patients who just had a MACE can have multiple change points. Another example is given in Goodman et al. [4] where they identified two change points in

* Corresponding author. Tel.: +1 650 467 1190; fax: +1 650 742 8619.
E-mail address: su.zheng@gene.com (Z. Su).

1551-7144/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.cct.2011.05.013

the survival function for prostate cancer mortality for men diagnosed between 1973 and 2002.

Ying et al.[5] proposed a non-parametric method that bases predictions on the Kaplan–Meier [6] survival curve. This non-parametric approach does not assume a smooth survival function but the stepwise estimate for the tail part of the survival function has relatively larger variances and is thus unreliable. In addition, the Kaplan–Meier estimate does not extend beyond the largest observation available. Ying et al. considered appending a parametric tail to the largest observation that declines to zero and they proposed to use all of the available data to estimate the parametric tail. This approach does not address the issue that the tail part of the Kaplan–Meier estimate is unreliable. More importantly, as for the parametric approaches the estimation of the parametric tail is based on all of the available data and as a result will be inaccurate when the survival function is not smooth.

The rest of the paper is structured as follows. In Section 2, we propose a hybrid parametric, non-parametric two-step approach to predicting events in clinical trials with time-to-event outcomes. In the first step, the change point algorithm proposed by Goodman et al.[4] for approximating the survival function with piecewise exponential distributions will be applied to the available data to identify the change points in the underlying survival function. In the second step, the survival function before the final change point will be estimated by the Kaplan–Meier estimate, and the tail of the survival function beyond the final change point will be estimated parametrically and extrapolated via the exponential model. We also generalized the change point detection approach of Goodman et al.[4] to piecewise Weibull distributions, which has the exponential distribution as a special case. This hybrid approach takes advantage of the fact that the non-parametric Kaplan–Meier estimate provides a relatively more accurate estimate for the initial part of the survival function. More importantly, the parametric estimate for the tail distribution is obtained with only the data that falls in the final piece of the survival function. When no change points are detected the proposed algorithm essentially reduces to either the parametric or non-parametric approaches available in the literature and thus the proposed approach can be viewed as a generalization of the existing approaches. Numerical results based on simulated clinical trials are given in Section 3, which show that the proposed approach can accurately predict when the pre-specified event counts can be reached and is thus useful in practice. Some discussions and concluding remarks are given in Section 4.

2. Methods

An accurate estimate for the tail of the survival function is essential to the prediction for the arrival of future events. For example, suppose that at a certain point of a clinical trial a total of N patients have been enrolled and enrollment is closed for the study. Suppose that n events have been observed at this point and there are still M patients in the study who have been followed for the time t_i, $i=1, \ldots, M$ and have not experienced an event yet. Given the survival function $S(t)$ the expected time T that it takes to observe another $m(<M)$ events can be solved from the equation below:

$$\sum_{i=1}^{M} \mathrm{P}(\text{an event between } t_i \text{ and } t_i + T \mid \text{no event up to time } t_i) = \sum_{i=1}^{M} \frac{S(t_i) - S(t_i + T)}{S(t_i)} = m. \quad (1)$$

As a result, in cases where some of the survival functions $S(t_i)$ are relatively small the estimate for T may have a large bias if the tail of the survival function $S(t)$ cannot be accurately estimated. Similarly, if a trial is still open to enrollment and up to K additional patients may be enrolled in the study with expected arrival times $a_1 \le \ldots \le a_K$, the expected time T can be solved from the equation below:

$$\sum_{i=1}^{M} \frac{S(t_i) - S(t_i + T)}{S(t_i)} + \sum_{i=1}^{K(T)} \{1 - S(T - a_i)\} = m. \quad (2)$$

where $K(T)$ denotes the number of a_i's that are no larger than T.

Goodman et al.[4] proposed an approach to detecting multiple change points in survival functions and to approximating a survival function with piecewise exponential distributions. Let $X_1, \ldots, X_n$ denote 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. Only the pairs (T_i, δ_i), $i=1, \ldots, n$ are observed, where $T_i = min(X_i, C_i)$ and $\delta_i = I\{T_i \le C_i\}$. They proposed to use the following change point model to estimate the survival function $S(t)$:

$$\lambda(t) = \begin{cases} \alpha_1 & 0 \le t \le \tau_1 \\ \alpha_2 & \tau_1 \le t \le \tau_2 \\ \vdots & \\ \alpha_{k+1} & t > \tau_k, \end{cases}$$

where $\lambda(t)$ is the hazard function, $0 = \tau_0 < \tau_1 < \ldots < \tau_{k+1} = \infty$ are the change points, k is the number of change points in the model that will be determined in a data-driven way, and α_j is the value of the hazard function between the time points τ_{j-1} and τ_j.

Goodman et al. proposed to obtain the estimates for the parameters via maximizing the following log profile likelihood function:

$$log\, L(\alpha_1, \ldots, \alpha_{k+1}, \tau_1, \ldots, \tau_k) = \sum_{j=1}^{k+1} \left[X\left(\tau_j\right) - X\left(\tau_{j-1}\right)\right] log\alpha_j - \sum_{i=1}^{n} \sum_{j=1}^{k+1} \alpha_j \left(T_i \wedge \tau_j - T_i \wedge \tau_{j-1}\right),$$

where $X(t)$ is the number of events observed up to time t. They showed that the maximum likelihood estimates $\hat{\tau}_j$'s for τ_j's can be obtained via maximizing

$$\sum_{j=1}^{k+1} \left\{X\left(\tau_j\right) - X\left(\tau_{j-1}\right)\right\} log\left(\frac{X\left(\tau_j\right) - X\left(\tau_{j-1}\right)}{\sum_{i=1}^{n} \left(T_i \wedge \tau_j - \tau_{j-1}\right) I\left(T_i > \tau_{j-1}\right)}\right),$$