Contemporary Clinical Trials xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

Contemporary Clinical Trials

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

Sample size and threshold estimation for clinical trials with predictive biomarkers

Howard M. Mackey *, Thomas Bengtsson

Genentech, South San Francisco, CA 94080, USA

ARTICLE INFO

Article history:
Received 3 July 2013
Received in revised form 10 September 2013
Accepted 16 September 2013
Available online xxxx

Keywords:
Personalized medicine
Cut-point
Diagnostic
Predictive
Prognostic
Lung cancer

ABSTRACT

With the increasing availability of newly discovered biomarkers personalized drug development is becoming more commonplace. Unless evidence of the dependence of clinical benefit on biomarker classification is *a priori* unequivocal, personalized drug development needs to jointly investigate treatments and biomarkers in clinical trials. Motivated by the development of contemporary cancer treatments, we propose targeting three main questions sequentially in order to determine (1) whether a drug is efficacious, (2) whether a biomarker can personalize treatment, and (3) how to define personalization. For time-to-event data satisfying the Cox proportional hazards model, we show that (1) and (2) may not directly involve the variance of an interaction term but of a contrast with smaller variance. An asymptotically exact covariance matrix for the parameter vector in the CPH model is derived to construct sample size formulae and an inference approach for thresholds of continuous biomarkers. The covariance matrix also reveals strategies for greater efficiency in trial design, for example, when the biomarker is binary or does not modulate the effect of treatment in the control arm. We motivate our approach by studying the outcome of a contemporary cancer study.

1. Introduction

The goal of personalized drug development is to identify and treat only those patients who are likely to derive clinical benefit from a new medicine and to spare the likely non-responders from the cost, time, and potentially adverse side effects of a new treatment. In addition to testing for efficacy and safety, a drug development plan for a personalized medicine has to reveal a clearly defined patient subgroup for whom the new drug is to be administered (unless this group is *a priori* well-defined). Early clinical studies, where such subgroups are typically identified, may therefore be more resource intensive (*e.g.* require larger sample sizes) than studies of traditional, non-personalized medicines. Here, based on time-to-event efficacy data, we investigate the design and analysis of clinical trials that involve a potentially predictive (and/or prognostic) biomarker quantified by a diagnostic test on a continuous, ordinal, or binary scale.

As noted in Royston and Sauerbrei [17], the potential personalization of a medicine can often be exploited by studying the interaction between biomarker and treatment. For survival data, sample size calculations for studies with a hypothesized interaction between a discrete valued biomarker and the treatment variable are given in Peterson and George [15], Xiang et al. [25], Russek-Cohen and Simon [18], and Schmoor et al. [20]. For continuous valued diagnostic markers with time-to-event data, the STEPP procedure [3,4] and the use of fractional polynomials [16,19] can be successfully applied to characterize and test the biomarker by treatment interaction effect. An adaptive method for estimating the biomarker threshold is described by Jiang et al. [11]. Although the preceding work sheds considerable light on the issue of describing and estimating interaction terms from survival data, it does not directly address the problem of how to design a clinical trial with a potentially significant biomarker by treatment interaction. Here, rather than focusing

* Corresponding author. Tel.: +1 650 467 8964.
E-mail address: mackey.howard@gene.com (H.M. Mackey).

1551-7144/$ – see front matter © 2013 Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.cct.2013.09.005

on the interaction term directly, we pose the problem in terms of contrasts that have a meaningful clinical interpretation, can be targeted in a development plan, and serve as the basis for inference and sample size calculations.

In order to jointly assess drug efficacy and the potential necessity of a biomarker, we propose three stages of inquiry consisting of hypothesis testing and inference that focus on the development of a new, potentially personalized medicine with the goal of approval by a national regulatory authority. Our work assumes that the Cox proportional hazards model (CPH) adequately explains the relationship between treatment, biomarker, and a time-to-event endpoint of interest. Our main contribution includes a sample size formula to be used when designing trials involving predictive biomarkers. This work extends the results of Schoenfeld [21], Peterson and George [15], Hsieh and Lavori [10], and Schmoor et al. [20], to include non-binary biomarker by treatment interactions. Using asymptotic methods, general analytical expressions are derived which can be applied to situations with correlated predictor variables without invocation of variance inflation factors. For continuous biomarkers, we additionally provide a method for approximate inference of the biomarker threshold identifying patients that are likely to derive clinical benefit.

To motivate our methodology, Section 2 presents an exploratory analysis of the biomarker–treatment interaction from a recent phase II oncology clinical trial. Section 3 describes a hypothesis testing framework for clinical studies involving a diagnostic marker, along with inference on the threshold for continuous biomarkers. Sections 4 and 5 provide model assumptions and sample size formulae. Our technical results are illustrated by two examples in Section 6. Our work is concluded in Section 7 with a discussion. Technical results and simulations are provided in the Appendix A.

2. Met expression and its functional relationship to survival in advanced non-small cell lung cancer (NSCLC)

Onartuzumab is a monoclonal antibody targeting the Met pathway where signaling can develop abnormally and cause healthy cells to become cancerous [2,5]. A recent phase II study of onartuzumab for the treatment of advanced stage metastatic NSCLC (OAM4558g) showed that patients with tumors that expressed high levels of the protein Met survived significantly longer when treated with onartuzumab plus erlotinib compared to erlotinib alone, with reported hazard ratios (HR) in the Met biomarker positive and negative groups of .37 and 3.02, respectively [23]. In this trial, tumor tissue samples were collected from each patient and analyzed by immunohistochemistry (IHC) to determine Met protein expression levels. The IHC test gives a score on an ordinal scale between 0 and 3+, indicating low- to high levels of Met proteins on the cell surface. In the efficacy analysis of OAM4558g, patients were considered biomarker positive if their IHC score was either 2+ or 3+, and biomarker negative otherwise [26].

An alternative method used to quantify the IHC assay involved an 'H-Score' which takes values from 0 to 300, indicating very low- to very high expression levels, respectively. To investigate the association between the H-score and survival, we fit (without inclusion of covariates) two separate CPH models to the control and treatment arms, respectively. Based on the output from these models, Fig. 1 shows the Martingale residuals plotted versus the H-score with regression lines estimated using a smoothing spline. The top- and bottom rows of Fig. 1 correspond to the original- and the percentile scales of the H-score, while the left- and right panels are for the control and treatment arms, respectively.

Fig. 1 shows that the risk of death appears to decrease with increasing H-score in the treatment arm with the opposite effect in the control arm, indicative of a strong biomarker–treatment interaction. These results are consistent with those reported in Spigel et al. [23] and Yu et al. [26]. The residual plots of Fig. 1 also suggest a functional form for the H-score when Met level is included as a covariate in the CPH model [24]. Specifically, when expressed on the percentile scale, Met levels appear to approximately linearly modulate the hazard for death in both groups.

Next, we propose an approach for investigating whether an experimental treatment has the potential to be personalized.

3. Personalized versus non-personalized medicines

While a personalized medicine can often be characterized by a biomarker–treatment interaction, not all such interactions indicate the potential for personalization. Indeed, the interaction must help distinguish between subsets of patients that do and that do not derive benefit that is clinically meaningful, not merely statistically significant (Pan and Wolfe 1997).

To illustrate when a biomarker–treatment interaction is indicative of a medicine with the potential for personalization, we consider three types of relationships between the biomarker and clinical outcome. For simplicity, we assume that the relationship between biomarker and clinical outcome is linear and that higher biomarker values are associated with decreased risk. We note that binary, ordinal monotonic, or non-linear monotonic relationships can be similarly described, but eschew from consideration of non-monotonic relationships.

In Fig. 2, the risk of a disease related event is plotted as a function of the biomarker value for three different medicines. The dashed horizontal line separates clinically meaningful effects from effects considered not clinically meaningful. The non-personalized medicine at the bottom of the figure has no dependence on the biomarker, while the line just above represents a 'pseudo-personalized' medicine that modulates the effect of treatment with all patients expected to derive a clinically meaningful benefit. Only the personalized medicine, which crosses the clinically meaningful effect boundary, requires the biomarker to determine patients most appropriate for treatment.

As illustrated, while it is necessary that a personalized medicine has a biomarker–treatment interaction, it is not sufficient. Moreover, while designing (sizing) a study around an expected biomarker by treatment interaction may be a reasonable approach in some situations, not all interactions indicate the potential for personalization. Since it is typically not known in the early stages of development which of the three scenarios depicted in Fig. 2 an investigational drug fits into, the development plan must allow for jointly testing treatment efficacy as well as the potential necessity of a biomarker. Clearly, there are many approaches to testing for

Please cite this article as: Mackey HM, Bengtsson T, Sample size and threshold estimation for clinical trials with predictive biomarkers, Contemp Clin Trials (2013), http://dx.doi.org/10.1016/j.cct.2013.09.005