Contemporary Clinical Trials 62 (2017) 140–145

Contents lists available at ScienceDirect

Contemporary Clinical Trials

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

On study designs and hypotheses for clinical trials with predictive biomarkers

Weichung J. Shih[a,b,*], Yong Lin[a,b]

[a] Department of Biostatistics, Rutgers School of Public Health, Rutgers University, The State University of New Jersey, 683 Hoes Lane West, Piscataway, NJ 08854, USA
[b] Division of Biometrics, Rutgers Cancer Institute of New Jersey, Rutgers University, The State University of New Jersey, 683 Hoes Lane West, Piscataway, NJ 08854, USA

ARTICLE INFO

Keywords:
Precision medicine
Stratification
Randomization
Hypothesis
Efficiency

ABSTRACT

Recent interest in conducting clinical trials with predictive biomarkers has generated research in comparing relative efficiency of different trial designs. We find these comparisons of efficiency mostly misleading since they are based on different hypotheses. In this paper, we discuss several commonly used trial designs and consider the hypotheses that each design is capable to address. We first consider the ideal situation of no classification errors, then the more realistic situation where marker assay's sensitivity, specificity and the rule of classification are imperfect. We pay special attention to the differences between treatment utility versus absolute treatment effect, and marker by treatment interaction versus marker utility.

1. Introduction

With the rapid development and use of biomarkers, the recent trend of precision medicine in utilizing predictive biomarkers has a great impact on the study design and analysis of modern clinical trials. There have increasingly been discussions and proposals of various trial designs and comparisons among them in the literature. See, for example [1–5]. However, these comparisons of trial designs have been somewhat confusing, since most of them are primarily based on the statistical efficiency without regard to the merits of the design features in terms of their differences in aims and hypotheses. In this article, we review several kinds of designs commonly used in oncology and other clinical trials involving one dichotomized biomarker or a combination of biomarkers, namely, “targeted and untargeted designs”, “stratified designs”, “precision-medicine designs”, and “enrichment designs”. We discuss the salient features of these designs and delineate what clinical hypotheses could be of interest and tested by the respective design. We consider the case of two treatment groups, test (T) versus control/standard of care (C), in the context of mean difference of a continuous endpoint. One can generalize to binary and survival endpoints as well. In the following, let μ_{ij} denote the true mean response of treatment group i (i = T or C) for patients in marker cohort j (j = 0 or 1; j = 0 being marker-negative and j = 1 being marker-positive). For the sake of clarity in presenting the hypotheses that each design is capable to test, we first consider the ideal situation of perfect sensitivity and specificity of a biomarker assay, then followed by the more realistic situation of imperfect sensitivity and specificity.

2. Designs and hypotheses in the absence of classification errors

2.1. Targeted and untargeted designs

In a targeted design, only marker-positive patients, who are predicted to be responsive, are randomized to the treatment groups. In contrast, an untargeted design is an all-comers design, which is the traditional randomized trial without the biomarker information. Simon and Maitournam [1] and Maitournam and Simon [2] reported that targeted designs are more efficient (i.e., requiring fewer subjects) than untargeted or all-comers designs. However, it is obvious that the treatment effect in a targeted design is limited to the marker-positive cohort only, while for the untargeted design, the treatment effect refers to the overall un-selected population. Expressed in notation, the null hypothesis is

$$H_1: \mu_{T\#} = \mu_{C\#}$$

for untargeted designs, where

$$\mu_{T\#} = p\mu_{T1} + (1-p)\mu_{T0} \quad (1)$$

$$\mu_{C\#} = p\mu_{C1} + (1-p)\mu_{C0} \quad (2)$$

and p is the prevalence rate of marker-positives.

For targeted designs, the null hypothesis is

$$H_2: \mu_{T1} = \mu_{C1}$$

Since there is no biomarker information available for patients in the

* Corresponding author at: Department of Biostatistics, Rutgers School of Public Health, Rutgers University, The State University of New Jersey, Room 216, 683 Hoes Lane West, Piscataway, NJ 08854, USA.
E-mail address: shihwj@sph.rutgers.edu (W.J. Shih).

http://dx.doi.org/10.1016/j.cct.2017.08.014
Received 13 February 2017; Received in revised form 13 July 2017; Accepted 18 August 2017
Available online 31 August 2017
1551-7144/

untargeted design, the treatment group means are a mixture of means of the two (un-identified) marker cohorts. H_1 is interpreted as the hypothesis of treatment group effects or treatment utility, as opposed to the absolute treatment effect (see H_4).

The two hypotheses H_1 and H_2 are not compatible unless there is no treatment effect for the marker-negatives (i.e., $\mu_{T0} = \mu_{C0}$). The premise for a targeted design is that the investigator has strong preliminary data (such as retrospective study) indicating that marker-positive patients may benefit from the test treatment, and is interested in focusing on this cohort only. For the untargeted design, the biomarker is not recognized as being predictive, patients are not 'typed' by the biomarker. We caution that statistical efficiency comparison has to be based on testing the same hypothesis.

2.2. Stratified designs

With untargeted designs, as seen above, treatment groups may not be balanced with respect to the biomarker-defined cohorts, despite randomization. A stratified design (Fig. 1), which also includes all eligible patients, but recognizing the predictive value of the biomarker, randomizes treatments within each marker stratum so that treatment groups are more likely balanced with respect to the marker status. Therefore, in addition to the hypotheses H_1 and H_2, a stratified design also allows testing for the treatment effect in the biomarker-negative cohort (H_3), the absolute treatment effect (H_4), the biomarker effect (H_5), as well as the marker by treatment interaction effect (H_6). That is, we may test

H_3: $\mu_{T0} = \mu_{C0}$

H_4: $\mu_{T0} + \mu_{T1} = \mu_{C0} + \mu_{C1}$

H_5: $\mu_{T0} + \mu_{C0} = \mu_{T1} + \mu_{C1}$

H_6: $\mu_{T0} - \mu_{C0} = \mu_{T1} - \mu_{C1}$

Hypothesis H_5 relates to the prognostic effect of the biomarker across treatment groups. Contrasting H_1 with H_4, the latter does not factor in the prevalence, thus is termed absolute (or marginal) treatment effect. The untargeted design can only test H_1 and is not capable to test H_4. It is debatable which hypothesis, H_1 or H_4 is more relevant.

The interaction hypothesis H_6 is of particular interest to Sargent et al. [3] and Mandrekar and Sargent [4–6] so that this design was called "marker-by-treatment interaction design" in these papers. It checks whether and how the biomarker status modifies the treatment effect. If the interaction effect is significant, then a closer examination is needed to find out whether the modification is directional (qualitative interaction) or in magnitude (quantitative interaction). In this case, hypotheses H_2 and H_3 specifically on the marker-positive and -negative cohorts, respectively, will be of interest, while the absolute treatment effect hypothesis H_4 will not be.

However, we think that the composite hypothesis

H_7: H_2 and H_3 followed by the individual component hypotheses H_2 and H_3 are more meaningful (than the interaction hypothesis H_6)

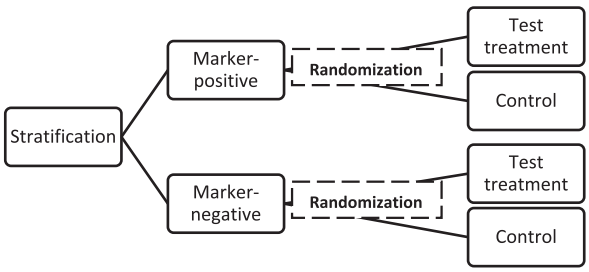

Fig. 1. Stratified design: Treatment randomized within each marker stratum.

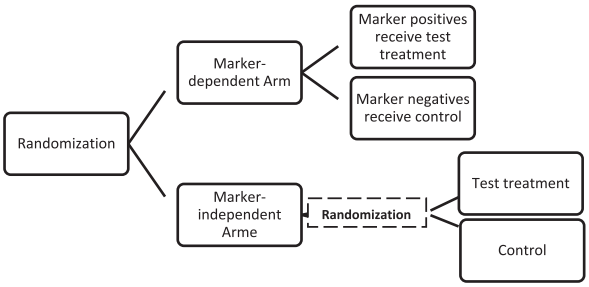

Fig. 2. Precision medicine design.

because they answer the clinical questions more directly regarding the treatment effect in each marker-defined stratum.

2.3. Precision medicine designs

Precision medicine designs randomize patients into two arms (Fig. 2). The first arm is a marker-dependent arm, and the second arm is a marker-independent/guided arm. For the marker-independent arm, patients are further randomized to treatment group T or C without biomarker information. For the marker-dependent arm, marker-positive patients all receive the test treatment T and marker-negative patients all receive the control treatment C (standard of care, SOC). This kind of designs is also called "marker-based strategy design" in Mandrekar and Sargent [4–6] and Young, Laird and Zhou [7].

The marker-independent arm is just like that in untargeted designs; hence the pooled treatment means $\mu_{T\#}$ and $\mu_{C\#}$ are estimated directly in this arm. The marker-dependent arm provides information to estimate μ_{T1} and μ_{C0}. With known or an estimate of the prevalence rate p of the marker-positive patients and by the relationship (1), μ_{T0} will be estimated. Similarly, μ_{C1} can be estimated via the relationship (2).

Having all μ_{ij}, i = T or C, j = 0 or 1 estimable, the hypotheses of marker-specific treatment effects, H_2 and H_3, the treatment and marker (main) effects, H_4 and H_5, the treatment by marker interaction effect, H_6, and the composite H_7 can all be tested.

Therefore, all the hypotheses H_1 to H_7 are testable in both stratified designs and precision medicine designs. Mandrekar and Sargent [4–6] suggested that the stratified design has greater efficiency in terms of overall sample size and events than the precision medicine design. They also suggested that the reason for better efficiency of the stratified design is because the precision medicine design includes patients treated with the same treatment on both marker-independent and marker-dependent arms. They indicated that this overlap increases as the prevalence of marker defined cohorts increases, leading to greater efficiency loss for the precision medicine design.

We agree with the above suggestions but caution the comparison of statistical efficiency they offered. In fact, Mandrekar and Sargent [4] calculated sample sizes required to test the hypotheses H_2 and H_3 for the stratified design, while they calculated the sample size required for comparing the marker-independent arm versus the marker-dependent arm for the precision medicine design. These are not compatible hypotheses for a meaningful comparison of efficiency. Shih and Lin [7] calculated the relative efficiency between the stratified design and the precision medicine design in full details with respect to testing the same hypotheses.

Obviously the stratified design is simpler in structure and all the effects are estimated directly from the marker-treatment components. In contrast, the precision medicine design has a double-randomization structure and the treatment by marker interaction and the marker (main) effects are estimated indirectly, as shown above. Thus, intuitively, the precision medicine design would require a larger overall sample size. However, the precision medicine design has the merit that it provides direct information on clinical utility of the biomarker, which may be provided, but is not so apparent, from the stratified design.