Journal of Statistical Planning and Inference 140 (2010) 3505–3519

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Compound adaptive GPU design for clinical trials

Ao Yuan *, Paul Bezandry, George Bonney

Howard University, Washington, DC 20059, USA

ARTICLE INFO

Article history:
Received 27 October 2009
Received in revised form
12 March 2010
Accepted 17 May 2010
Available online 27 May 2010

MSC:
primary, 62K05;
secondary, 62H99

Keywords:
Adaptive design
Clinical trials
Compound generalized Pólya urn
Optimal design

ABSTRACT

In clinical trials, several competing treatments are often carried out in the same trial period. The goal is to assess the performances of these different treatments according to some optimality criterion and minimize risks to the patients in the entire process of the study. For this, each coming patient is allocated sequentially to one of the treatments according to a mechanism defined by the optimality criterion. In practice, sometimes different optimality criteria, or the same criterion with different regimes, need to be considered to assess the treatments in the same study, so that each mechanism is also evaluated through the trail study. In this case, the question is how to allocate the treatments to the incoming patients so that the criteria/mechanisms of interest are assessed during the trail process, and the overall performance of the trial is optimized under the combined criteria or regimes. In this paper, we consider this problem by investigating a compound adaptive generalized Pólya urn design. Basic asymptotic properties of this design are also studied.

1. Introduction

1.1. Brief review of clinical trails and the Pólya urn design

In clinical trials, there are often several competing treatments to be assessed during the trial process while patients come sequentially over the long period of study. Based on the then current clinical knowledge of the treatments, each coming patient is allocated to one of them according to some defined mechanisms, so that the overall treatment loss is minimized by the given optimality criteria. However, there are cases in which the treatments needs to be assessed under different criteria and the criteria themselves are to be assessed during the trial process, thus the patients' allocation needs to be determined by the different random mechanisms, in which multiple criteria with different regimes are considered in the same study. In practice, often the same treatments under different mechanisms results differently, thus may require different criterion for each mechanism. As each treatment has its strength and weakness, thus we want to assess the treatments under each of the criterion and allocate the patients so as to minimize the overall risk. The results under each criterion/mechanism will be analyzed separately so that each of them is to be evaluated. As the treatments performances unknown in prior, and are gradually learned through the trial process, to minimize the overall loss of the trial, a compromised criterion between the two is desirable. As the compound criterion depends on the unknown performances of all the treatments, it cannot be determined in prior, instead, it is implemented along the trial process. In this case, the question is how to allocate the coming patients so that the allocation is optimized under the compound criterion. Here, we attempt to explore this problem by using a compound generalized Pólya urn design, and will come back at the problem in

* Corresponding author.
E-mail addresses: ayuan@howard.edu (A. Yuan), pbezandry@Howard.edu (P. Bezandry), ge_bonney@howard.edu (G. Bonney).

0378-3758/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2010.05.022

the Application section. For this, we first review briefly some commonly used methods in clinical trials. Some of these methods apply to either discrete response, continuous response, or both. The generalized Pólya urn (GPU) is a well-known design in clinical trial; there are many extensive results for it. Many of the existing models are applied to the discrete responses, and are not optimally adaptive.

The adaptive design uses accumulating data to update aspects of the study as it continues without undermining the validity and integrity of the trial (Hayre, 1979; Melfi and Page, 1998; Jennison and Turnbull, 2000; Hu and Rosenberger, 2003; Hu and Zhang, 2004, among others). A comprehensive review of works in this filed can be found in Hu and Rosenberger (2006). The purpose of optimal design is to achieve some targeting objective criteria for the allocation proportions (Eisele, 1994; Eisele and Woodroofe, 1995; Rosenberger et al., 2001; Hu et al., 2006; Hu et al., 2009; for example). Recently, Zhang et al. (2006), thereafter ZHC, proposed a sequential estimation-adjusted urn model, Yuan and Chai (2008) independently of ZHC, thereafter YC, studied an adaptive GPU design, which has some similarity to that of ZHC. This design was studied further in Yuan (2008) for the cases of delayed response, staggered/censored entry, heterogeneity and longitudinal/repeated observations. These methods optimize any given target functional of the trial distribution and are applicable to both discrete and continuous responses.

1.2. An exposition of a two-treatment clinical trial

To help understanding of our motivation, consider a simple example. Suppose we have two treatments to investigate in the clinical trial. The success rate of the two treatments are p_1 and p_2 respectively, $N_1(n)$ and $N_2(n)$ are the numbers of patients allocated the each treatment at time n. The commonly used Neyman allocation (Melfi and Page, 1998) is designed to maximize the power of detecting the difference p_1-p_2 of mean performances of the trials, which leads to the allocation proportion $N_1(n)/N_2(n)\to\sqrt{p_1(1-p_1)}/\sqrt{p_2(1-p_2)}$. While the criterion in Rosenberger et al. (2001) is to minimize the expected treatment failure and leads to the proportion $N_1(n)/N_2(n)\to\sqrt{p_1/p_2}$. The two criteria give very different allocation results. When the treatment differences are small, the former will be better in terms of distinguishability, thus smaller sample size will be needed in the study, which is much desired since patients are cost to get in practice. But also it may lead to more life losses which will be of grave consequence as the subjects in study are human beings. The second criterion can lead to less treatment losses when the differences are relatively significant but may not have the desirable power to detect the differences between the treatments. We will see in the application section that, as a result of optimizing the two criteria simultaneously with our method, assume $p_1>p_2$, the compound method will give the allocation proportion $((1-p_1+p_2)/(\sqrt{p_1(1-p_1)}+\sqrt{p_2(1-p_2)}))(\sqrt{p_1(1-p_1)},\sqrt{p_2(1-p_2)})+((p_1-p_2)/(\sqrt{p_1}+\sqrt{p_2}))(\sqrt{p_1},\sqrt{p_2})$. This proportion is a combination/compromise of the two, and is more robust than either one of two criterion used along, in that it keeps much of the power in detecting the difference while maintain less treatment losses.

In this paper, to assess the treatments by different criteria in the same trial and keep the advantages of each trail and avoid their weakness, we propose and study a compound version of the design in YC. The rest part is organized as follows. In Section 2, we describe the existing adaptive sequential GPU design of YC. Next, we present compound versions with and without adaptive features and their basic asymptotic properties in Section 3, and an illustrative application in Section 4. The relevant technical proofs are given in the Appendix.

2. Description of the optimal adaptive GPU design

We first describe the adaptive optimal GPU design (YC). Suppose there are k treatments under study, and a Pólya urn with some initial components, corresponding to records of patient assignments to these treatments $\mathbf{X}_0=(X_{0,1},\ldots,X_{0,k})$. Here, the components of the urn can be discrete, real valued or of mixed type. Let $\mathbf{r}_n$ be the (multiple) response of the n-th patient under one of the treatments and $f(\mathbf{r}_n)$ be the summary score for this response. Without loss of generality we assume $0\le f(\cdot)<\infty$. Let A_i be the event that the study is under treatment i, $\mu_i=E(f(\mathbf{r}_1)|A_i)$ be the expected performance, or success rate, of the i-th treatment, and $\sigma_i^2:=Var(f(\mathbf{r}_1)|A_i)<\infty$ be its variance ($i=1,\ldots,k$). Set $\boldsymbol{\mu}=(\mu_1,\ldots,\mu_k)$ and $\boldsymbol{\sigma}^2=(\sigma_1^2,\ldots,\sigma_k^2)$. For any vector $\mathbf{x}=(x_1,\ldots,x_k)$, denote $|\mathbf{x}|=\sum_{i=1}^k x_i$ and $\{\mathbf{x}\}=\mathrm{diag}(\mathbf{x})$, the diagonal matrix for $\mathbf{x}$. A vector $\mathbf{x}$ is *normalized* if $|\mathbf{x}|=1$. At time n, the urn composition is $\mathbf{X}_n=(X_{n,1},\ldots,X_{n,k})$, the total number of patients assigned to treatment i at time n is $N_i(n)$, and denote $\mathbf{N}(n)=(N_1(n),\ldots,N_k(n))$. In the GPU model, to assign the entering n-th patient to one of the treatments, a random variable is drawn from the multinomial distribution with probabilities $\mathbf{X}_n/|\mathbf{X}_n|$. If it is type i, the patient is assigned to the i-th treatment, a random vector of masses $\boldsymbol{\xi}_i$ is added to the urn compositions, and the response $\mathbf{r}_n$ is used to update the estimate of $\boldsymbol{\mu}$ in the next step. Let $\boldsymbol{\xi}_i=(\xi_{i1},\ldots,\xi_{ik})$ be the increments to the urn composition given the patient is assigned treatment i, $\boldsymbol{\xi}=(\xi_{ij})_{i,j=1,\ldots,k}$ be the matrix representation of the ξ_{ij}'s, and for each n, $\boldsymbol{\xi}_n$ is an i.i.d. version of $\boldsymbol{\xi}$. To simplify the expressions of the asymptotic variances to be derived later, we assume throughout this article that $\boldsymbol{\xi}$ is independent of the response observations. The random vector $\boldsymbol{\xi}_i$ is termed the adding rule, and $\boldsymbol{v}=E(\boldsymbol{\xi})=(v_{ij})$ the design matrix with $v_{ij}=E(\xi_{ij})$ (known). The first eigenvalue λ of the design matrix and its normalized first left (row) eigenvector $\mathbf{v}$ play a key role in the asymptotic properties of the GPU design. Many authors (for instance Athreya and Karlin, 1967; Gouet, 1997; Janson, 2004) studied asymptotic properties of $\mathbf{X}_n$ and $\mathbf{N}(n)$ and proved that

$$(\mathbf{X}_n/|\mathbf{X}_n|,\mathbf{N}(n)/n)\to(\mathbf{v},\mathbf{v})\quad(a.s.)$$

For a comprehensive review in this field, see Rosenberger and Lachin (1993) and other related recent papers.